AN INTERACTING QUANTUM ATOMS APPROACH TO CONSTRUCTING
A CONFORMATIONALLY DEPENDENT BIOMOLECULAR FORCE FIELD
BY GAUSSIAN PROCESS REGRESSION:
POTENTIAL ENERGY SURFACE SAMPLING AND VALIDATION

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

2016

Salvatore Cardamone
School of Chemistry
Contents

Abstract 20

Declaration 22

Copyright Statement 23

Acronyms and Abbreviations 25

Acronyms 25

Acknowledgements 29

1 Introduction 31

2 Exposition 36

2.1 Molecular Mechanics ................................. 36
   2.1.1 Molecular Dynamics ............................. 36
   2.1.2 Classical Force Fields ............................ 41
   2.1.3 \textit{ab initio} Molecular Dynamics ................. 52
2.2 Atomic Partitioning ............................................. 55
  2.2.1 Hirshfeld Partitioning ................................. 56
  2.2.2 Bader’s Atoms in Molecules ....................... 58
  2.2.3 Interacting Quantum Atoms ....................... 64
2.3 Multipole Moment Electrostatics ....................... 70
  2.3.1 Mathematical Details ................................. 70
  2.3.2 Implementation .................................. 81
2.4 Machine Learning ........................................ 97
  2.4.1 Overview ....................................... 97
  2.4.2 Kriging ....................................... 101
  2.4.3 Particle Swarm Optimisation .................... 106
2.5 The Quantum Chemical Topological Force Field .... 108
  2.5.1 Previous Work ................................ 108
  2.5.2 Proposed Implementation ....................... 110

3 Conformational Sampling ................................. 112
  3.1 Introduction ..................................... 112
  3.2 Stationary Point Vibrations ....................... 116
    3.2.1 Kinetic Energy ................................ 116
    3.2.2 Potential Energy .............................. 117
    3.2.3 Equations of Motion ......................... 118
  3.3 Generalised Vibrations ............................. 121
  3.4 TYCHE - Conformational Sampling Software .... 124
5 Subset Selection for Machine Learning

5.1 Introduction .............................................. 204

5.2 Greedy Heuristics ......................................... 206
  5.2.1 Experimental Details ................................. 208
  5.2.2 MaxMin .............................................. 210
  5.2.3 Deletion ............................................ 213
  5.2.4 Alternative Distance Metric ......................... 218

5.3 Sequential Selection ...................................... 225

5.4 Iterative Voronoi Subset Selection ..................... 233

5.5 Domain of Applicability ................................ 239
  5.5.1 Introduction ....................................... 239
  5.5.2 The Convex Hull ................................... 242
  5.5.3 Test Case ......................................... 245
  5.5.4 Largest $d$-Polytope .............................. 247

5.6 Conclusion .............................................. 257

6 Raman Optical Activity ................................... 259

6.1 General Theory .......................................... 261
  6.1.1 Classical Raman Scattering ......................... 261
  6.1.2 The Electromagnetic Field ......................... 264
  6.1.3 Hamiltonian in an Electromagnetic Field .......... 268
  6.1.4 Wavefunction Perturbed by Periodic Field ....... 272
  6.1.5 The Raman Optical Activity Tensors .......... 276
List of Tables

2.1 The number of geometric minima predicted by a variety of force fields for serine and cysteine. Also given are the number of minima that the various force fields predicted but that were not represented in the set of minima generated by ab initio calculations, and the mean average deviations (MAD) for the molecular energies at each geometry. .............................................. 95

2.2 Free energies of hydration for a variety of molecules predicted by use of AMOEBA and TIP3P-like water potentials. Corresponding experimental values are also given. Energies are in kcal mol$^{-1}$. . . . 97

3.1 Memory requirements for TYCHE. .............................................. 140

3.2 Conformational sampling of water at a number of values of $n_{\text{period}}$. . 150

3.3 Conformational sampling of NMA at a number of values of $n_{\text{period}}$. . 150

3.4 Conformational sampling of histidine at a number of values of $n_{\text{period}}$. 151

3.5 Conformational sampling of water using the Jacobian and Hessian from a number of different levels of theory. .............................................. 152

3.6 Conformational sampling of NMA using the Jacobian and Hessian from a number of different levels of theory. .............................................. 152

3.7 Conformational sampling of histidine using the Jacobian and Hessian from a number of different levels of theory. .............................................. 152
3.8 Sampling ranges of four prominent bond stretching degrees of freedom in zwitterionic alanine as the sampling temperature is increased. The \textit{ab initio} Hessian and Jacobian are taken from gas phase frequency calculations. In the bottom row, we include the ranges of these degrees of freedom as obtained from a 300 K \textit{in vacuo} MD trajectory. In red, we have refined these sampling ranges down so that they represent the sampling ranges taken by over 98\% of conformers over the course of the MD trajectory.  

3.9 Sampling ranges of four prominent valence angle bending degrees of freedom in zwitterionic alanine as the sampling temperature is increased. The \textit{ab initio} Hessian and Jacobian are taken from gas phase frequency calculations. In the bottom row, we include the ranges of these degrees of freedom as obtained from a 300 K \textit{in vacuo} MD trajectory. In red, we have refined these sampling ranges down so that they represent the sampling ranges taken by over 98\% of conformers over the course of the MD trajectory.

3.10 Sampling ranges of four prominent bond stretching degrees of freedom in zwitterionic alanine as the sampling temperature is increased. The \textit{ab initio} Hessian and Jacobian are taken from implicitly solvated (CPCM) frequency calculations. In the bottom row, we include the ranges of these degrees of freedom as obtained from a 300 K TIP3P-solvated MD trajectory. In red, we have refined these sampling ranges down so that they represent the sampling ranges taken by over 98\% of conformers over the course of the MD trajectory.

3.11 Sampling ranges of four prominent valence angle bending degrees of freedom in zwitterionic alanine as the sampling temperature is increased. The \textit{ab initio} Hessian and Jacobian are taken from implicitly solvated (CPCM) frequency calculations. In the bottom row, we include the ranges of these degrees of freedom as obtained from a 300 K TIP3P-solvated MD trajectory. In red, we have refined these sampling ranges down so that they represent the sampling ranges taken by over 98\% of conformers over the course of the MD trajectory.

3.12 Sampling ranges of four prominent bond stretching degrees of freedom in zwitterionic alanine, as obtained through \textsc{tyche} with a single seeding geometry (\textsc{tyche} (Single)), with five seeding geometries (\textsc{tyche} (Multi)) and through a MD trajectory.
3.13 Sampling ranges of four prominent valence angle bending degrees of freedom in zwitterionic alanine, as obtained through TYCHE with a single seeding geometry (TYCHE (Single)), with five seeding geometries (TYCHE (Multi)) and through a MD trajectory. 169

4.1 Mean errors associated with the CDEs given in Figure 4.3. 193

4.2 Mean errors corresponding to the CDEs depicted in Figure 4.6. Note the ‘bunching’ of prediction error when 60 energetic minima or higher are used as seeds for the conformational sampling. 201

5.1 Maximum prediction errors for each small molecule kriging model as the training set size is increased. Maximum prediction errors for the randomly constructed training sets are presented in the final row. 212

5.2 Comparison of the greedy MaxMin and greedy Deletion algorithms on a candidate set in $\mathbb{R}^2$, with the aim of constructing a subset, where $|S| = 4$. The first row, corresponding to the steps taken by the MaxMin algorithm, adds two points (red) at each iteration to the subset, whose distance is the largest in the entire candidate set. The second row corresponds to the steps taken by the Deletion algorithm. Here, the two points in the candidate set which are closest together are selected (green). Their nearest neighbours (orange) are then evaluated, and the point with the closest nearest neighbour is removed from the candidate set. 214

5.3 Maximum prediction errors for each small molecule kriging model as the training set size is increased. Maximum prediction errors for the randomly constructed training sets are presented in the final row. 217

5.4 MaxMin-constructed training set maximum prediction errors for each small molecule kriging model as the training set size is increased. Maximum prediction errors for the randomly constructed training sets are presented in the final row. 222

5.5 Deletion-constructed training set maximum prediction errors for each small molecule kriging model as the training set size is increased. Maximum prediction errors for the randomly constructed training sets are presented in the final row. 223
5.6 Average errors associated with the SS1, SS10 and SS1040 kriging models relative to the random benchmark. ............................................. 231

5.7 Maximum errors associated with the SS1, SS10 and SS1040 kriging models relative to the random benchmark. ............................................. 231

5.8 Standard deviation or errors associated with the SS1, SS10 and SS1040 kriging models relative to the random benchmark. ........... 232

6.1 Conformer Boltzmann weights exceeding 1% for the 35 microsolvated histidine systems using the system-weighting and solute-weighting schemes. ......................................................... 308
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Simulation techniques available to the computational chemist.</td>
<td>32</td>
</tr>
<tr>
<td>2.1</td>
<td>A contour plot of the electron density of in molecular plane of furan superimposed onto a representative collection of gradient paths. Atoms are represented by black circles, where the gradient paths terminate. Interatomic surfaces are highlighted as solid curves, and contain bond critical points (black squares). A ring critical point (triangle) is also shown in the centre of the furan.</td>
<td>62</td>
</tr>
<tr>
<td>2.2</td>
<td>Vectors ( \mathbf{r} ) and ( \mathbf{r}' ) in a Cartesian axis system, along with their colatitudeal and azimuthal degrees of freedom, ( (\Theta, \Phi) ) and ( (\theta, \phi) ), respectively.</td>
<td>72</td>
</tr>
<tr>
<td>2.3</td>
<td>Vector quantities required for an expansion, in multipole moments, of the interaction between the charge distributions centred at ( \mathbf{R}_A ) and ( \mathbf{R}_B ).</td>
<td>74</td>
</tr>
<tr>
<td>2.4</td>
<td>The overlap monopole (upper graph) and dipole (lower graph) moments at various overlap centres along the hydrogen fluoride internuclear axis. The nuclear charges have been added to the atomic centres.</td>
<td>80</td>
</tr>
<tr>
<td>3.1</td>
<td>Distribution of molecular energies obtained through TYCHE by use of the energy allocation scheme of (3.4.12). The lowest energy point corresponds to the \textit{ab initio} minimum.</td>
<td>130</td>
</tr>
<tr>
<td>3.2</td>
<td>Distribution of molecular energies obtained through TYCHE by use of the energy allocation scheme of (3.4.13). The lowest energy point corresponds to the \textit{ab initio} minimum.</td>
<td>131</td>
</tr>
</tbody>
</table>
3.3 Distribution of molecular energies obtained through TYCHE by use of the energy allocation scheme of (3.4.14). The lowest energy point corresponds to the \textit{ab initio} minimum. \\

3.4 One-dimensional analytical PES (blue) and the TYCHE-generated PES (red), recreated through the seeding geometries A-G. Stochastic dynamics are performed upon this tessellation of LWs in an attempt to generate a sample set that would be representative of continuous dynamics on the \textit{ab initio} PES. \\

3.5 The conventionally invoked valence coordinates. (a) corresponds to bond stretching, (b) corresponds to valence angle bending, and (c) corresponds to dihedral torsion. The major components that will be required in the following discussion are shown. \\

3.6 TYCHE output outlining the normal mode information of water. Normal modes 1, 2 and 3 correspond to the valence angle bending, symmetric stretching and asymmetric stretching modes, respectively. The RIC contributions to each normal mode are given at the bottom of the figure. \\

3.7 Range of CH bond stretch degrees of freedom in methane (green), ethane (red), neopentane (yellow) and cyclopentane (blue) as the temperature is increased. The black lines correspond to the range of sampling undertaken in Maple et al. \\

3.8 Range of HCH valence angle degrees of freedom in methane (green), ethane (red), neopentane (yellow) and cyclopentane (blue) as the temperature is increased. The black lines correspond to the range of sampling undertaken in Maple et al. \\

3.9 Range of CN bond stretch degrees of freedom in formamide (green), N-methylacetamide (red), methylazocyclopropanone (yellow) and N,N-dimethylformamide (blue) as the temperature is increased. The black lines correspond to the range of sampling undertaken in Maple et al. \\

3.10 Range of OCN valence angle degrees of freedom in formamide (green), N-methylacetamide (red), methylazocyclopropanone (yellow) and N,N-dimethylformamide (blue) as the temperature is increased. The black lines correspond to the range of sampling undertaken in Maple et al.
3.11 Alanine zwitterion atomic labelling. ........................................... 161

3.12 Histograms for the sampling of the alanine sidechain dihedral angle by use of the various TYCHE sampling schemes (by seeding with energetic minima (red), MD snapshot conformations (green) and the selective mode distortion (green)), and the MD sampling (blue). 171

3.13 Helix stretching and bending modes of motion for the alanine-isoleucine helix. ................................................................. 174

3.14 Helix breathing motion recovered from selectively evolving those modes comprising dihedral (red), valence angle (green) and bond stretching (blue) degrees of freedom. ............................................ 175

3.15 Helix bending motion recovered from selectively evolving those modes comprising dihedral (red), valence angle (green) and bond stretching (blue) degrees of freedom. ............................................ 176

4.1 The Atomic Local Frame for the carbon atom of methanol. From the global coordinate system, \( x, y, z \), one is able to define an ALF, the basis vectors of which are denoted \( x_{ALF}, y_{ALF}, z_{ALF} \), from the vectors \( R^{\Omega_0}, R^{\Omega_x}, R^{\Omega_{xy}} \). Any atom not defining the ALF, e.g. \( R^n \), is defined in the ALF with the spherical polar coordinates \( r^n, \theta^n, \phi^n \). 184

4.2 Erythrose and Threose in the open chain and furanose forms. .... 190

4.3 CDEs corresponding to all erythrose (top) and threose (bottom) systems studied. The open chain forms are systematically better predicted than the corresponding furanose forms. ......................... 194

4.4 CDEs for erythrose open chain at various training set sizes. Note the progression of the CDEs towards the lower prediction errors as the training set size increases. However, owing to the logarithmic abscissa, this does not correspond to a uniform enhancement of a kriging model given a consistently larger training set size. ............. 196
4.5 Mean prediction errors associated with CDEs for erythrose open chain as the training set size is increased. With the old kriging engine (left), a distinct plateau formed after roughly 1200 training examples, corresponding to no further improvement in the kriging model despite additional training points. However, the new kriging engine (right) appears to avoid premature plateauing, with additional kriging model improvement at higher training set sizes. Regression fits of the FEREBUS errors against training set size, of functional forms $A + B/n$ (blue) and $C + D/\sqrt{n}$ (black), where $n$ is the training set size, are also given.

4.6 S-curves depicting the power of a kriging model as more energetic minima are utilised for conformational sampling. As the number of minima utilised increases, the CDEs tend towards higher prediction errors. The kriging models which underlie these CDEs have a fixed training set size of 700.

5.1 Average molecular energy error, with accompanying standard deviations, in the testing set conformations given a training set constructed by the MaxMin algorithm from the respective candidate sets at a number of training set sizes. In each pane, the dark black line corresponds to the average molecular energy error in the testing set given a training set of size 1000, generated randomly from the candidate set.

5.2 Average molecular energy error, with accompanying standard deviations, in the testing set conformations given a training set constructed by the Deletion algorithm from the respective candidate sets at a number of training set sizes. In each pane, the dark black line corresponds to the average molecular energy error in the testing set given a training set of size 1000, generated randomly from the candidate set.

5.3 Average molecular energy error, with accompanying standard deviations, in the testing set conformations given a training set constructed by the MaxMin algorithm utilising the kriging covariance as a distance metric. In each pane, the dark black line corresponds to the average molecular energy error in the testing set given a training set of size 1000, generated randomly from the candidate set.
5.4 Average molecular energy error, with accompanying standard deviations, in the testing set conformations given a training set constructed by the Deletion algorithm utilising the kriging covariance as a distance metric. In each pane, the dark black line corresponds to the average molecular energy error in the testing set given a training set of size 1000, generated randomly from the candidate set. 221

5.5 Cumulative distribution functions of errors associated with the sequential selection kriging models. SS1 (green), SS10 (blue) and SS1040 (yellow) are compared to a randomly constructed training set benchmark (red). Whilst the SS1 model performs similarly to the random benchmark, the SS10 and SS1040 both considerably outperform the random benchmark. 230

5.6 Voronoi tessellation of 12 points in $\mathbb{R}^2$, denoted by black points. The Voronoi cell corresponding to each point is coloured. Voronoi cells are partitioned from one another by lines. Where three or more Voronoi cells meet, a vertex is formed. Image generated by the interactive Voronoi diagram generator, located at http://www.alexbeutel.com/webgl/voronoi.html. 235

5.7 Cumulative distribution function of errors for a training set constructed by the Iterative Voronoi Subset Selection (blue) and a randomly constructed training set (red). 237

5.8 The Ackley function in two dimensions. The function is projected down onto the $xy$-plane as a colour-contour plot. 245

5.9 Absolute error (left) and MSE (right) of predictions made from within the convex hull (yellow) and outside of the convex hull (blue) of the training set. 246

5.10 Progression of average error ratio (left abscissa) and $\Delta$ (right abscissa) with increasing acceptance percentage for the water monomer ($\mathbb{R}^3$). 251

5.11 Definition of three vertices used in the construction of $A_\gamma$ in $\mathbb{R}^3$. 253

5.12 Progression of average error ratio (left abscissa) and $\Delta$ (right abscissa) with increasing acceptance percentage for the methanol ($\mathbb{R}^{12}$). 254
5.13 Progression of average error ratio (left abscissa) and $\Delta$ (right abscissa) with increasing acceptance percentage for the NMA ($R^{30}$).

5.14 Progression of average error ratio (left abscissa) and $\Delta$ (right abscissa) with increasing acceptance percentage for the water decamer ($R^{84}$).

6.1 Zwitterionic histidine (His$^0[N_{\pi}H]$) where the imidazole nitrogen atoms have been labelled as indicated in the text. The two sidechain dihedral angles, $\chi_1, \chi_2$ are also labelled.

6.2 Zwitterionic histidine protonation states and tautomers.

6.3 Fractional contribution of the various histidine charge states as a function of pH. Blue, green, red and yellow curves correspond to the fractional contributions of the His$^{2+}$, His$^{1+}$, His$^0$ and His$^{1-}$ states, respectively. We have limited the pH range between 0 and 12, and so omit the fractional contribution of the His$^{2-}$ state.

6.4 Sidechain dihedral angles ($\chi_1, \chi_2$) taken by conformers within the three histidine ensembles obtained by a Kabsch filtering of the MD trajectories; His$^{1+}$ (blue circles), His$^0[N_{\pi}H]$ (green diamonds) and His$^0[N_{\tau}H]$ (red triangles). Grey hatched regions correspond to those conformers used in the work of Deplazes et al.

6.5 Raman spectra of the different Boltzmann weighting schemes. The top pane shows the Raman spectrum resulting from the system-weighting and the middle pane from the solute-weighting. The bottom pane is the experimental Raman spectrum of zwitterionic histidine.

6.6 ROA spectra of the different Boltzmann weighting schemes. The top pane shows the ROA spectrum resulting from system-weighting and the middle pane from the solute-weighting. The bottom pane is the experimental ROA spectrum of zwitterionic histidine.
6.7 Raman spectra of zwitterionic histidine computed from a number of levels of geometrical optimisation. The top pane contains the Raman spectrum computed directly from unoptimised MD snapshots, while the middle two contain Raman spectra computed from conformers which have been optimised with the “Loose” and “Regular” convergence criteria. The bottom panel contains the experimental Raman spectrum of histidine.

6.8 ROA spectra of zwitterionic histidine computed from a number of levels of geometrical optimisation. The top pane contains the ROA spectrum computed directly from unoptimised MD snapshots, while the middle two contain Raman spectra computed from conformers which have been optimised with the “Loose” and “Regular” convergence criteria. The bottom panel contains the experimental ROA spectrum of histidine.

6.9 Raman spectra resulting from the optimisation methodologies we have outlined. Enumeration corresponds to the labels we have used in the main text.

6.10 ROA spectra resulting from the optimisation methodologies we have outlined. Enumeration corresponds to the labels we have used in the main text.

6.11 Raman spectra of microsolvated histidines. From the top pane to the bottom pane, we show histidine microsolvated by 5, 10, 15 and 20 explicit water molecules. The final pane contains the experimental histidine Raman spectrum.

6.12 ROA spectra of microsolvated histidines. From the top pane to the bottom pane, we show histidine microsolvated by 5, 10, 15 and 20 explicit water molecules. The final pane contains the experimental histidine ROA spectrum.

6.13 Raman spectra of microsolvated cationic histidine. From the top pane to the bottom pane, we show cationic histidine microsolvated by 5, 10, 15 and 20 explicit water molecules. The final pane contains the experimental histidine Raman spectrum.
6.14 ROA spectra of microsolvated cationic histidine. From the top pane to the bottom pane, we show cationic histidine microsolvated by 5, 10, 15 and 20 explicit water molecules. The final pane contains the experimental histidine ROA spectrum.
Abstract

The energetics of chemical systems are quantum mechanical in origin and dependent upon the internal molecular conformational degrees of freedom. “Classical force field” strategies are inadequate approximations to these energetics owing to a plethora of simplifications—both conceptual and mathematical. These simplifications have been employed to make the in silico modelling of molecular systems computationally tractable, but are also subject to both qualitative and quantitative errors. In spite of these shortcomings, classical force fields have become entrenched as a cornerstone of computational chemistry.

The Quantum Chemical Topological Force Field (QCTFF) has been a central research theme within our group for a number of years, and has been designed to ameliorate the shortcomings of classical force fields. Within its framework, one can undertake a full spatial decomposition of a chemical system into a set of finite atoms. Atomic properties are subsequently obtained by a rigorous quantum mechanical treatment of the resultant atomic domains through the theory of Interacting Quantum Atoms (IQA). Conformational dependence is accounted for in the QCTFF by use of Gaussian Process Regression, a machine learning technique. In so doing, one constructs an analytical function to provide a mapping from a molecular conformation to a set of atomic energetic quantities. One can subsequently conduct dynamics with these energetic quantities.

The notion of “conformational sampling” is shown to be of key importance to the proper construction of the QCTFF. Conformational sampling is a key theme in this work, and a subject that we will expatiate. We suggest a novel conformational sampling scheme, and attempt a number of conformer subset selection strategies to construct optimal machine learning models. The QCTFF is then applied to carbohydrates for the first time, and shown to produce results well within the commonly invoked threshold of “chemical accuracy” - $O(\beta^{-1})$, where $\beta$ is the thermodynamic beta. Finally, we present a number of methodological developments to aid in both the accuracy and tractability of predicting ab initio vibrational spectroscopies.
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Several publications from the work outlined in this thesis are included in the addendum. They are listed here for convenience:


Those publications in which I am the first author have been used to construct this thesis. Only those portions that I have written have been reproduced.

This thesis does not exceed 80,000 words.

Salvatore Cardamone,
November 22, 2016.
Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property
and/or Reproductions described in it may take place is available in the University IP Policy, in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations and in The University’s Policy on Presentation of Theses.
**Acronyms and Abbreviations**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADMP</td>
<td>Atom Centred Density Matrix Propagation</td>
</tr>
<tr>
<td>AIM</td>
<td>Atoms In Molecules</td>
</tr>
<tr>
<td>ALF</td>
<td>Atomic Local Frame</td>
</tr>
<tr>
<td>AMBER</td>
<td>Assisted Model Building with Energy Refinement</td>
</tr>
<tr>
<td>AMOEBA</td>
<td>Atomic Multipole Optimized Energetics for Biomolecular Applications</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>aug-cc-pVDZ</td>
<td>Augmented Correlation Consistent Polarisable Valence Double Zeta</td>
</tr>
<tr>
<td>B3LYP</td>
<td>Becke 3-Lee Yang Parr</td>
</tr>
<tr>
<td>CCSD(T)</td>
<td>Coupled Cluster Singles and Doubles (with Perturbative Triples)</td>
</tr>
<tr>
<td>CCT</td>
<td>Cartesian Coordinate Tensor Transfer</td>
</tr>
<tr>
<td>CDE</td>
<td>Cumulative Distribution of Errors</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>CHARMM</td>
<td>Chemistry at Harvard Macromolecular Mechanics</td>
</tr>
<tr>
<td>CHELPG</td>
<td>Charges from Electrostatic Potentials using a Grid-Based Method</td>
</tr>
<tr>
<td>COSMO</td>
<td>Conductor-Like Screening Model</td>
</tr>
<tr>
<td>CSD</td>
<td>Cambridge Structural Database</td>
</tr>
<tr>
<td>DoA</td>
<td>Domain of Applicability</td>
</tr>
<tr>
<td>DFT</td>
<td>Density Functional Theory</td>
</tr>
<tr>
<td>DMA</td>
<td>Distributed Multipole Analysis</td>
</tr>
<tr>
<td>GROMACS</td>
<td>Groningen Machine for Chemical Simulations</td>
</tr>
<tr>
<td>GPR</td>
<td>Gaussian Process Regression</td>
</tr>
<tr>
<td>HF</td>
<td>Hartree Fock</td>
</tr>
<tr>
<td>IAS</td>
<td>Interatomic Surface</td>
</tr>
<tr>
<td>IQA</td>
<td>Interacting Quantum Atoms</td>
</tr>
<tr>
<td>IVSS</td>
<td>Iterative Voronoi Subset Selection</td>
</tr>
<tr>
<td>LdP</td>
<td>Largest $d$–Polytope</td>
</tr>
<tr>
<td>LW</td>
<td>Local Well</td>
</tr>
<tr>
<td>M06</td>
<td>Minnesota Functional 06</td>
</tr>
<tr>
<td>MD</td>
<td>Molecular Dynamics</td>
</tr>
<tr>
<td>MEP</td>
<td>Molecular Electrostatic Potential</td>
</tr>
<tr>
<td>Term</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>MP(n)</td>
<td>Møller-Plesset (n)th order Perturbation Theory</td>
</tr>
<tr>
<td>mRMSD</td>
<td>Minimised Root Mean Squared Deviation</td>
</tr>
<tr>
<td>NMA</td>
<td>N-methyl Acetamide</td>
</tr>
<tr>
<td>OPLS-AA</td>
<td>Optimized Potentials for Liquid Simulations All Atoms</td>
</tr>
<tr>
<td>p-RDM</td>
<td>(p)th Order Reduced Density Matrix</td>
</tr>
<tr>
<td>PCM</td>
<td>Polarisable Continuum Model</td>
</tr>
<tr>
<td>PIMD</td>
<td>Path Integral Molecular Dynamics</td>
</tr>
<tr>
<td>PES</td>
<td>Potential Energy Surface</td>
</tr>
<tr>
<td>PSO</td>
<td>Particle Swarm Optimisation</td>
</tr>
<tr>
<td>QCT</td>
<td>Quantum Chemical Topology</td>
</tr>
<tr>
<td>QCTFF</td>
<td>Quantum Chemical Topological Force Field</td>
</tr>
<tr>
<td>RAH-LdP</td>
<td>Restricted Affine Hull of the Largest (d)–Polytope</td>
</tr>
<tr>
<td>RDF</td>
<td>Radial Distribution Function</td>
</tr>
<tr>
<td>RESPA</td>
<td>Reversible Reference System Propagator Algorithm</td>
</tr>
<tr>
<td>RIC</td>
<td>Redundant Internal Coordinate</td>
</tr>
<tr>
<td>ROA</td>
<td>Raman Optical Activity</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root Mean Squared Deviation</td>
</tr>
<tr>
<td>SS</td>
<td>Sequential Selection</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular Value Decomposition</td>
</tr>
<tr>
<td>TAE</td>
<td>Transferable Atom Equivalents</td>
</tr>
<tr>
<td>TDSCF</td>
<td>Time-Dependent Self-Consistent Field</td>
</tr>
<tr>
<td>TDSE</td>
<td>Time-Dependent Schrödinger Equation</td>
</tr>
</tbody>
</table>
Acknowledgements

From commencing my university studies in mechanical engineering, quickly switching to biochemistry, undertaking a PhD in theoretical chemistry, and now transitioning to physics for a postdoctoral position, my academic trajectory has been rather unconventional.

I can probably say, with some degree of certainty, that if it weren’t for my parents’ continual support throughout my periods of indecision, I would have been in no position to submit a thesis. I can’t possibly express the full extent of my gratitude in prose. The greatest accolade I can strive for is to consider myself half as accomplished as they are. They have made this work possible, and so it is dedicated to them in its entirety.

I am indebted to Professor P L A Popelier for his guidance and direction over the past four years. I can appreciate the risk associated with taking me on as a PhD candidate given my previous propensity for switching disciplines. I can only hope that my contributions to his group’s work go some way to repaying his efforts.

I would like to thank Dr E W Blanch, and all the members of the Popelier group, past and present, with whom I have worked over the course of my research. Thanks are also extended to all administrative and technical staff at The University of Manchester with whom I have liaised.

Finally, I would like to thank the BBSRC for a fully-funded DTP studentship.
“... Chemistry is that branch of natural philosophy in which the greatest improvements have been and may be made: it is on that account that I have made it my peculiar study; but at the same time I have not neglected the other branches of science. A man would make but a very sorry chemist if he attended to that department of human knowledge alone. If your wish is to become really a man of science, and not merely a petty experimentalist, I should advise you to apply to every branch of natural philosophy, including mathematics.”

FRANKENSTEIN

Mary Shelley
Chapter 1

Introduction

There are a number of commonly echoed sentiments that enunciate the difficulty in applying quantum mechanics to chemical systems. The statements by Paul Dirac\cite{1} and Walter Kohn\cite{2} are perhaps two of the most overused. To keep this work as uncontrived as possible, the reader is spared from having to read them once more. One can distil their observations down to the fact that the quantum mechanical state vector, the wavefunction, is a complex, high-dimensional object, and its representation requires exponential complexity. Indeed, it has even been suggested that the Schrödinger equation cannot be solved accurately when the number of particles is greater than ten, by virtue of this “curse of dimensionality”\cite{3}.

However, by the above reasoning, one is immediately met with a contradiction, in that there exist a plethora of methods for calculating with, or manipulating wavefunctions for systems of many particles. It can be argued that the complexity alluded to by Dirac and Kohn is illusory. Consider two completely different macroscopic wavefunctions of a system, i.e. they occupy completely different regions of the system’s Hilbert space. Both wavefunctions satisfy the Schrödinger
equation for the system. Then, it is formally valid to write a superposition of these two wavefunctions that satisfies the Schrödinger equation. Of course, this line of reasoning is equivalent to the famous Schrödinger cat gedanken experiment.

However, for the vast majority of physically realisable (low energy) states of the system, nature is not required to explore the entire Hilbert space. As such, superpositions of completely different wavefunctions are irrelevant to an analysis of the underlying physics of the system. Put more concisely, there are no Schrödinger cats in quantum chemistry[4]. One is then able to confine their analysis to small regions of the quantum Hilbert space of the system, and consequently reduce the complexity of the mathematics alluded to by Dirac and Kohn. In so doing, a number of computationally feasible schemes arise for undertaking quantum mechanical calculations on chemical systems.

Figure 1.1: Simulation techniques available to the computational chemist.

Broadly speaking, the computationally tractable methodologies for analysing a
system can be categorised based on their accuracy and computational feasibility, as depicted in Figure 1.1. Those techniques towards the left-hand side of the diagram are typically cheap to evaluate, but also very approximate in nature. Moving further to the right, the techniques become more accurate, but also more expensive to evaluate. Those techniques at the far right of the diagram are only viable options for systems containing a very small number of atoms. The computational chemist is then required to select an appropriate technique based on the desired accuracy of a calculation and the computational facilities that are available to them.

Biochemical systems are very difficult to treat with the lower accuracy methodologies of Figure 1.1. Biochemical systems are subject to a wealth of chemical interactions: hydrogen bonding, halogen bonding, entropic effects, interactions with heavy elements requiring a relativistic treatment, reactions, etc. In addition, biochemical systems can rarely be treated as small, closed thermodynamic systems; the chemical environment typically plays an important role in dictating the chemistry of these systems. As such, the system size required for a meaningful simulation is large. Both the complexity and size of the systems render the majority of the techniques depicted in Figure 1.1 unideal strategies for dealing with biochemical systems.

The aim of this work is to develop a methodology that sits somewhere in the middle of the categorisation of Figure 1.1. Ideally, we would like to combine the speed of pair potential techniques with the accuracy of density functional theory and wavefunction-based techniques. Density functional theory (DFT) is of such importance to computational and theoretical chemistry, that its formulation is given in Appendix A. The methodology outlined will be referred to herein as the Quantum Chemical Topological Force Field (QCTFF). We make no attempt to clarify the jargon that has just been invoked; its meaning will be revealed over the
course of the next chapter.

We immediately make a clarification on the nomenclature used throughout this work. Bader\textsuperscript{[5]} has made a useful distinction between the notion of “configuration” and “geometry”. A configuration of a molecular system corresponds to the molecular graph of the system, i.e. the bonding pattern of the system. In contrast, a molecular geometry denotes the positions of the constituent atoms in some basis, Cartesian or other. Whilst the molecular geometry is subject to change with respect to an \textit{infinitesimal} displacement in atomic positions, the molecular configuration is invariant.

Rather than use the term “geometry”, we choose to use the term “conformation”. We find this term less ambiguous, since geometry can be used for any number of mathematical quantities, whereas the notion of conformation is more restrictive. The idea of a conformation and the sampling of conformations is a predominant theme in this work. We will use $r \in \mathbb{R}^3$, to denote Cartesian three-vectors. $x \in \mathbb{R}^a$, with $a > 3$, will be used for higher-dimensional Cartesian vectors. $q \in \mathbb{R}^a$ will be used for a generalised vector, whose basis vectors are not necessarily orthogonal. We attempt to adhere to this convention throughout.

In Chapter 2, we outline the general principles that will be implemented in the construction of the QCTFF. The mathematical details and derivations are self-contained, and are necessarily rigorous.

In Chapter 3, we outline the importance of conformational sampling of a molecular species for the QCTFF. Further, we introduce a novel conformational sampling methodology that allows the user to conduct a conformational sampling of the physically relevant regions of molecular conformational space.
Chapter 4 is a general proof-of-concept of the QCTFF methodology that has been developed in our lab over the past decade. For the first time, the QCTFF is applied to carbohydrates, and the resulting errors are shown to be well within the limits of “chemical accuracy”.

In Chapter 5, we enumerate a number of problems that have been encountered with the construction of the QCTFF. A variety of remedies are attempted, both from techniques that are available in the literature and a number of novel strategies that have been developed over the past few years.

Finally, Chapter 6 discusses the seemingly unrelated topic of spectroscopic prediction, and introduces a number of novel formalisms for improving the accuracy of these methods. It is hoped that spectroscopic prediction will be a key tool for validating the QCTFF in the near future.
Chapter 2

Exposition

2.1 Molecular Mechanics

2.1.1 Molecular Dynamics

Our starting point is the time-dependent Schrödinger equation (TDSE) for a molecular system comprising $N_e$ electrons and $N_N$ nuclei,

$$i\hbar \frac{\partial}{\partial t} \Psi(r, R; t) = \mathcal{H} \Psi(r, R; t),$$

(2.1.1)

where $\Psi$ is the wavefunction of the molecular system, a function of the $3N_e$—dimensional vector, $r$, containing electronic positions, the $3N_N$—dimensional vector, $R$, containing nuclear positions, and depends parametrically on time$^1$. The Hamiltonian

---

$^1$To be more precise, $r$ and $R$ are functions of time, and therefore correspond to electronic and nuclear trajectories, respectively. The wavefunction is then $\Psi(r(t), R(t))$. 

36
2.1. MOLECULAR MECHANICS

operator, $\mathcal{H}$, has standard form,

$$
\mathcal{H} = -\sum_i \frac{\hbar^2}{2M_i} \nabla_i^2 - \sum_i \frac{\hbar^2}{2m_i} \nabla_i^2 + \sum_{i<j} \frac{q_i q_j}{|r_i - r_j|} - \sum_{I,i} \frac{q_i q_I}{|R_i - r_i|} + \sum_{I<j} \frac{q_I q_J}{|R_J - R_J|},
$$

(2.1.2)

where we use capital Roman letters for indices over nuclear degrees of freedom, and lower case Roman letters for indices over electronic degrees of freedom. $q$ and $m$ are the charge and mass of the particles, respectively, to which the subscripted indices correspond. $\nabla^2$ is the Laplacian operator, and is used in the conventional sense of denoting the divergence of the gradient of a scalar field, i.e. $\nabla^2 = \nabla \cdot \nabla$.

These operators are referred to as; the nuclear kinetic energy, the electronic kinetic energy, the electronic repulsion, the electronic-nuclear attraction and the nuclear repulsion operators, respectively. Rewriting (2.1.2) in a more concise form for the sake of later brevity,

$$
\mathcal{H} = -\sum_i \frac{\hbar^2}{2M_i} \nabla_i^2 - \sum_i \frac{\hbar^2}{2m_i} \nabla_i^2 + \mathcal{V}_{ne}(r, R) = -\sum_i \frac{\hbar^2}{2M_i} \nabla_i^2 + \mathcal{H}_e(r, R),
$$

(2.1.3)

where we have introduced two operators $\mathcal{V}_{ne}$ and $\mathcal{H}_e$,

$$
\mathcal{V}_{ne} = \sum_{i<j} \frac{q_i q_j}{|r_i - r_j|} - \sum_{I,i} \frac{q_i q_I}{|R_i - r_i|} + \sum_{I<j} \frac{q_I q_J}{|R_J - R_J|}
$$

(2.1.4)

$$
\mathcal{H}_e = -\sum_i \frac{\hbar^2}{2m_i} \nabla_i^2 + \mathcal{V}_{ne},
$$

(2.1.5)

the electron-nuclear potential and electronic Hamiltonian operators, respectively.

By invoking the Born-Oppenheimer approximation, we invoke the product ansatz for the molecular wavefunction[6], whereby the nuclear and electronic wavefunctions are entirely separable,

$$
\Psi(r, R; t) = \psi(r; t) \chi(R; t).
$$

(2.1.6)
Substituting (2.1.3) and (2.1.6) into (2.1.1),
\[ i\hbar \frac{\partial}{\partial t} \psi(r; t) \chi(R; t) = \mathcal{H} \psi(r; t) \chi(R; t) \]
\[ = - \sum_I \psi(r; t) \frac{\hbar^2}{2M_I} \nabla_I^2 \chi(R; t) - \sum_i \chi(R; t) \frac{\hbar^2}{2m_i} \nabla_i^2 \psi(r; t) \] (2.1.7)
\[ + \mathcal{V}_{ne}(r, R) \psi(r; t) \chi(R; t). \]
Note that we have eliminated those terms containing \( \nabla_I^2 \psi(r; t) \) and \( \nabla_i^2 \chi(R; t) \) since the nuclear degrees of freedom are independent of the electronic degrees of freedom, and vice versa. Multiplying (2.1.7) through by \( \langle \chi | \) and \( \langle \psi | \), we obtain the following relations,
\[ i\hbar \frac{\partial}{\partial t} \psi(r; t) = - \sum_i \frac{\hbar^2}{2m_i} \nabla_i^2 \psi(r; t) + \langle \chi | \mathcal{V}_{ne}(r, R) | \chi \rangle \psi(r; t), \] (2.1.8)
\[ i\hbar \frac{\partial}{\partial t} \chi(R; t) = - \sum_I \frac{\hbar^2}{2M_I} \nabla_I^2 \chi(R; t) + \langle \psi | \mathcal{H}_e(r, R) | \psi \rangle \chi(R; t), \] (2.1.9)
respectively. Iteratively solving these two equations to self-consistency yields solution to (2.1.1), and is referred to as the time-dependent self-consistent field (TD-SCF) method[7]. The first equation, (2.1.8), corresponds to evolving the electronic wavefunction in the presence of a potential resulting from the nuclear degrees of freedom, and the second, (2.1.9), to evolving the nuclear wavefunction in the potential exerted by the electronic degrees of freedom.

The Born-Oppenheimer approximation allows us to assume that the electronic degrees of freedom adjust instantaneously to any displacement in the nuclear degrees of freedom. This assumption comes about because the nuclei are more than three orders of magnitude heavier than electrons, and so adjust more slowly to external fields than their much lighter counterparts. Considering the manifestation of this in the TDSE, the nuclear wavefunction can be represented as a series of delta functions, while the electronic wavefunction is much more diffuse. We can model
this by treating the nuclei as classical particles[8], a common approach to which is writing

\[
\chi(\mathbf{R}; t) \approx A(\mathbf{R}; t) \exp \left[ \frac{iS(\mathbf{R}; t)}{\hbar} \right],
\]

where \( A(\mathbf{R}; t), S(\mathbf{R}; t) \) are real functions, and correspond to an amplitude and phase factor, respectively. Inserting the nuclear wavefunction (2.1.10) into (2.1.9), we obtain

\[
i\hbar \frac{\partial}{\partial t} A(\mathbf{R}; t) \exp \left[ \frac{iS(\mathbf{R}; t)}{\hbar} \right] = \left( \langle \psi | \mathcal{H}_e | \psi \rangle - \sum_I \frac{\hbar^2}{2m_I} \nabla^2_I \right) A(\mathbf{R}; t) \exp \left[ \frac{iS(\mathbf{R}; t)}{\hbar} \right].
\]

After expansion, this expression can be separated into its real and imaginary components,

\[
\frac{\partial S}{\partial t} + \sum_I \frac{1}{2M_I} (\nabla_I S)^2 + \langle \psi | \mathcal{H}_e | \psi \rangle = \hbar^2 \sum_I \frac{1}{2M_I} \nabla^2_I \frac{A^2}{A} \quad (2.1.11)
\]

\[
\frac{\partial A^2}{\partial t} + \sum_I \nabla_I \left( A^2 \frac{\nabla_I S}{m} \right) = 0. \quad (2.1.12)
\]

The first of these, (2.1.11), can be written in a form that is isomorphic to the classical Hamilton-Jacobi equation in the classical limit, i.e. \( \hbar \to 0 \), leading to

\[
\frac{\partial S}{\partial t} + \sum_I \frac{1}{2M_I} (\nabla_I S)^2 + \langle \psi | \mathcal{H}_e | \psi \rangle = 0. \quad (2.1.13)
\]

This is the equation we will be primarily concerned with, but it should be noted that (2.1.12) resembles a continuity equation, since \( \langle \chi | \chi \rangle = A^2 \) can be interpreted as a nuclear density, as we can verify from (2.1.10). (2.1.11) and (2.1.12) together form the equations of “quantum hydrodynamics”\(^2\). Noting that the second and

\(^2\)As a matter of historical note, the equations of quantum hydrodynamics were derived by Erwin Madelung, after whom they are sometimes named, shortly after Schrödinger’s seminal publication. They also act as a springboard into Bohmian mechanics from the Schrödinger equation.
third terms of (2.1.13) resemble a kinetic and potential energy, we can combine them into a Hamiltonian, where the momentum conjugate to each $R_I$ is $P_I = \nabla_I S$,

$$\frac{\partial S}{\partial t} + \mathcal{H}(R, \nabla_I S) = 0.$$  (2.1.14)

Invoking the Newtonian equations of motion, $\dot{P}_I = -\nabla_I V_e(R_I)$, where $V_e(R_I)$ is the potential to which the nuclei are subjected, $\langle \psi | \mathcal{H}_e | \psi \rangle$,

$$\frac{dP_I}{dt} = -\nabla_I \langle \psi | \mathcal{H}_e | \psi \rangle$$

$$M_I \ddot{R}_I = -\nabla_I \langle \psi | \mathcal{H}_e | \psi \rangle = -\nabla_I V_e(R_I).$$  (2.1.15)

Taking the expectation value of $\mathcal{H}_e$ over all electronic degrees of freedom makes $V_e(R_I)$ a function of the nuclear degrees of freedom. The coupled electronic degrees of freedom are evolved in time by invoking (2.1.8), which in the classical limit, where the nuclei are described by delta functions, becomes

$$i\hbar \frac{\partial}{\partial t} \psi(r; t) = -\sum_i \frac{\hbar^2}{2m_i} \nabla^2_i \psi(r; t) + V_{ne}(r; t)$$

$$= \mathcal{H}_e(r; R) \psi(r; R).$$  (2.1.16)

Here, both $\mathcal{H}_e(r; R)$ and $\psi(r; R)$ depend parametrically on the nuclear positions. (2.1.15) and (2.1.16) constitute the TDSCF equations in a form that is amenable to computation. The potential energy function, $V_e(R_I)$, is often referred to as a potential energy surface (PES), and describes a mapping from the conformational manifold of the nuclear coordinates to the molecular energy.

Behler[9] has given a useful semantic characterisation of the two methods one can implement in approximating the PES. The “physical” approach consists of employing a number of functions that intuitively recreate the behaviour one would expect from a potential. Classical force fields, as we discuss in Section 2.1.2, are one such physical approach. Embedded atom type (EAM) potentials can be used
2.1. MOLECULAR MECHANICS


The “non-physical” approach entails fitting a number of functionally flexible, but physically irrelevant, functions to ab initio data. Splines are the natural choice for such a task, and have been employed for a number of low-dimensional PESs[13], but their implementation in high-dimensional spaces renders them of little use for the majority of molecular systems[14]. A number of other techniques exist for this purpose, such as Sheppard interpolation[15], Gaussian approximation potentials[16], machine learning[17] and genetic programming[18]. We briefly review a few of these methods in Section 2.4.

A third category, which Behler does not allude to, is the use of quantum mechanical methods to construct the PES “on-the-fly”, and perform dynamics directly on the ab initio PES. Naturally, this method is more costly than both the physical and non-physical approaches discussed above, but has proven to be an invaluable tool in the study of molecular dynamics[19, 20]. We review a few of these methods in Section 2.1.3.

2.1.2 Classical Force Fields

A commonly invoked approximation to the molecular PES is the use of a number of simple empirical potentials that are (typically) parameterised from quantum mechanical calculations. Broadly speaking, the PES is split up into two general contributions; the bonded and non-bonded energies, $E_b$ and $E_{nb}$, respectively. $E_{nb}$ is further split into three contributions; the electrostatic $E_{elec}$, van der Waals $E_{vdW}$ and correlation $E_{corr}$ energies.
Given that the PES arises from a quantum mechanical treatment of the molecular system in question, we can immediately see that this splitting is artificial, since such a distinction is not made in the quantum mechanical treatment. However, the distinction between bonded and non-bonded terms allows for an entirely classical modelling of the molecular system, with physically intuitive origins, as we now demonstrate.

**Bonded Terms**

The bonded contributions to the PES approximately model the energetics arising from atomic valency. Bonded atoms are treated as hard, impenetrable spheres connected by flexible linkages. The energetics of this system can then be modelled classically by expressing the bonded energetics as a power series in the displacements of the system form an minimum energy conformation. Modelling is typically undertaken in an internal coordinate basis, the mathematics of which we rigorously outline in Section 3.5. For the moment, we qualitatively outline the main features of the internal coordinates.

The bonded terms are formed from \( n \)-body interactions, where \( n = 2, 3, 4 \), subject to there being some bonding network linking the atoms involved in the \( n \)-body interaction\(^3\). The energy associated with each interaction is then denoted \( E_{1-n}(r_1, \ldots, r_n) \), where \( r \) is an atomic position vector.

Given two atoms with position vectors \( r_i \) and \( r_j \), we seek a functional form to the interatomic potential, \( E_{1-2}(r_i, r_j) \). The distance\(^4\) between two such atoms is

\[^3\text{Defining this bonding network as “covalent” is not necessarily valid here, since hydrogen bonds and ionic bonds can also be subject to the treatment outlined.}\]

\[^4\text{For the remainder of this work, we will denote all normed quantities as } |\mathbf{a}|_p = (a_1^p + \ldots + a_d^p)^{1/p}, \text{ where } d \text{ is the dimensionality of the vector } \mathbf{a} \text{ and } p \text{ corresponds to the } L^p \text{ space upon which we are invoking the norm. For the case where } p = 2, \text{ we obtain the standard Euclidean}\]
denoted \( r_{ij} = ||r_i - r_j|| \). The interatomic potential arising between these two bonded atoms is conventionally modelled as a power series, truncated at finite order,

\[
E_{1-2}(r_i, r_j) = E(r_0) + k_2^r (r_{ij} - r_0)^2 + k_3^r (r_{ij} - r_0)^3 + ... ,
\]

where \( E(r_0) \) is the value of the interatomic potential at the minimum energy interatomic distance, \( r_0 \), and \( k_2^r, k_3^r, ... \) are suitably chosen bond “force constants”, which act as expansion coefficients to the quadratic, cubic, etc., terms of the \( E_{1-2}(r_i, r_j) \) expansion. Truncation of the expansion at the quadratic term yields a harmonic potential for interatomic interactions, while higher order terms offer anharmonic corrections to the harmonic function.

A number of difficulties arise with the above model. For instance, it has been found that the force constants are actually dependent upon the interatomic distance[21]. Aside from the dynamical nature of the force constants, one can approximate their values by a number of techniques. The most popular technique appears to be approximating the harmonic force constants from the quantum mechanical Hessian of the system (see for example [22]). On occasion, the resultant force constants are empirically modified to agree with spectroscopic data and band assignments[23]. Determining the order to which the power series is to be truncated is not entirely clear; a number of popular force fields simply truncate the power series at the quadratic term, whilst more accurate, unfavoured force fields continue to the quartic term[24]. As a final note, we wish to point out that a number of alternative functional forms are available for the modelling of \( E_{1-2}(r_i, r_j) \), which do not use the power series given in (2.1.17), such as the shifted finitely-extensible non-linear elastic (FENE) potential[25] and the AMOeba bond potential[26].

\footnote{This will be the default norm invoked over the course of this work, so we drop the subscript and simply refer to this quantity as ||a|| for clarity. This is in contrast to the magnitude of a scalar quantity, \( a \), which we denote by |\( a \)|.}
Consider now that the atom at $r_i$ is involved in 1-2 interactions with atoms at $r_j$ and $r_k$. Then, the angle subtended by $r_j, r_k$ from $r_i$, $\theta_{ijk}$, can also be used in a power series expansion for the angular energy,

$$E_{1-3}(r_i, r_j, r_k) = E(\theta_0) + k_2^\theta(\theta_{ijk} - \theta_0)^2 + k_3^\theta(\theta_{ijk} - \theta_0)^3 + \ldots,$$

(2.1.18)

where $E(\theta_0)$ is the value of the angular potential and the minimum energy angle, $\theta_0$, and the force constants $k_2^\theta, k_3^\theta, \ldots$ are expansion coefficients. Similar to the $E_{1-2}(r_i, r_j)$ potentials, the vast majority of force fields use a truncated power series at the quadratic term, whilst some more specialist force fields use terms up to the quartic. A number of more complex forms do exist, such as that outlined in [27], where multiplicative polynomials are used to make the potential resemble more of a damped harmonic oscillator. The three-body potential developed for flexible water molecules by Kumagai et al.[28] is an attempt to recreate the vibrational modes of motion for the molecule, and uses a number of additional functions to achieve this. However, much like the case for $E_{1-2}(r_i, r_j)$, such expressions are only implemented in highly specialised small molecule force fields, and do not enjoy mainstream implementation.

The final bonded term we consider is the four body potential, where $r_i$ is linked to $r_k$, $r_j$ is linked to $r_l$, and $r_i$ is linked to $r_j$. We are able to define two planes; the first contains the position vectors $r_i, r_j, r_k$, while the second contains the position vectors $r_i, r_j, r_l$. The angle between these two planes when viewed down the vector $r_i - r_j$, $\tau_{ijkl}$, is then expanded in a power series to approximate the torsional potential,

$$E_{1-4}(r_i, r_j, r_k, r_l) = E(\tau_0) + k_2^\tau(\tau_{ijkl} - \tau_0)^2 + k_3^\tau(\tau_{ijkl} - \tau_0)^3 + \ldots,$$

(2.1.19)

where $E(\tau_0)$ is the vale of the torsional potential at the minimum potential value, $\tau_0$, and the force constants $k_2^\tau, k_3^\tau, \ldots$ are expansion coefficients. One typically
finds, however, that low order power series are inadequate approximations to the torsional potential. The torsional potential is found to be periodic, with energetic barriers easily surmountable over the course of a conventional MD simulation. As such, one is usually required to employ a truncated Fourier series to account for the periodicity\cite{29},

$$E_{1-4}(r_i, r_j, r_k, r_l) = \sum_n \frac{k_n^\tau}{2} \left(1 + \cos(n\tau_{ijkl} + \phi_n)\right), \quad (2.1.20)$$

where the $k_n^\tau$ are the torsional force constants, $n$ is the order to which the Fourier series is truncated, and $\phi_n$ are phase factors. Few alternative functional forms for $E_{1-4}(r_i, r_j, r_k, r_l)$ exist; the truncated Fourier series is an excellent function for modelling the energetics of torsional degrees of freedom. However, it is the coupling between torsional degrees of freedom and the electrostatics of a system\cite{30} that can lead to problems in parameterising $E_{1-4}(r_i, r_j, r_k, r_l)$. However, decoupling the two is not feasible in the majority of classical force fields\cite{31}, and so is not discussed further here.

**Electrostatic Energy**

Consider the Coulombic interaction between two charged entities in a vacuum, placed at positions $r_i$ and $r_j$. The force, $F_{ij}$, between the two is inversely proportional to the square of the distance separating them, $r_{ij}$, and is given by

$$F_{ij} = \frac{q_i q_j}{4\pi \epsilon_0 r_{ij}^2} \frac{(r_i - r_j)}{r_{ij}}, \quad (2.1.21)$$

where $q_i$ and $q_j$ are the charges of the respective particles and $\epsilon_0$ is the permittivity of free space. Introducing a polar solvent between the two particles leads to the assimilation of flux lines between the two particles by the solvent, reducing the
strength of interaction. This effect can be modelled in an average way by multi-
plying the permittivity of free space, $\epsilon_0$, by an empirical factor, $\epsilon$, the permittivity
of the medium which separates the particles, such that Coulomb’s law is modified
to read

$$F_{ij} = \frac{q_i q_j}{4 \pi \epsilon_0 \epsilon r_{ij}^2} \left( \frac{r_i - r_j}{r_{ij}} \right).$$  \hspace{1cm} (2.1.22)

Comparison between (2.1.21) and (2.1.22) shows that $r_{ij}^2$ has been scaled to $r_{ij}^2 \epsilon$, essentially increasing the effective distance over which the force acts between the two particles. Considering the permittivity of water relative to vacuum is 80, the distance between the two particles is effectively increased by a factor of $\sqrt{80} \approx 9$ when the particles are solvated in water.

The implementation of so-called “implicit solvation” models is, however, not quite as trivial as one might envisage. For a molecular system, there is a surrounding cavity which cannot be penetrated by other molecules owing to Pauli repulsion. The surface of this cavity is referred to as the “solvent-accessible surface area”, and is typically modelled as a superposition of atom-centred spheres of radius equal to the atomic van der Waals radius. Hence, the dielectric permittivity is only considered to be greater than unity outside of the cavity[32].

A plethora of such implicit solvation models exist, perhaps the most popular being the “Polarisable Continuum Model” (PCM)[33]. Consider a system (and associated cavity) in a dielectric medium of a given permittivity. The electrostatics\textsuperscript{5} associated with the system polarises, and subsequently induces charges, in the surrounding dielectric medium. As a result of these induced charges in the

\footnotetext[5]{We leave the term “electrostatics” purposely vague- the subtleties will be discussed in Section 2.3.}
2.1. MOLECULAR MECHANICS

medium, the electrostatics of the system are polarised. Obviously then, this sequence of mutual polarisation between system and dielectric requires iteration to self-consistency, leading to the “Self-Consistent Reaction Field” (SCRF) model.

The energetics associated with this process acts as a perturbation to the isolated system energetics. So, for instance, while a charge-separated species (such as a zwitterion, a molecular system we work with in great detail in Chapter 6) is not stable in the gas phase, the energetics associated with the SCRF stabilises, and permits the existence of, the charge-separated species in a high dielectric solvent (such as water). The “Conductor-Like Screening Model” (COSMO)[34] is another popular implicit solvation model, except here the cavity is embedded in an infinite-dielectric medium (i.e. a perfect conductor). In so doing, there is no iteration to self-consistency since the charges induced in the dielectric are immediately dissipated. Neglecting the iteration to self-consistency leads to a significant increase in computational speed.

The prevalent means for modelling molecular electrostatics is by the use of point charges placed at atomic centres[35, 36]. The simplicity of such a model allows for quick computation of electrostatic interactions. By use of Coulomb’s law, the total electrostatic energy of a molecular system is given by

$$E_{elec} = \frac{1}{2} \sum_{i \neq j}^{N} \frac{q_i q_j}{4\pi \epsilon_0 \epsilon_{ij}} ,$$

(2.1.23)

where $N$ is the number of atoms, and the factor of a half corrects for double-counting. However, this method is not as simplistic as first appears. An atom in a molecule rarely has an integer charge; electrons are free to distribute themselves throughout the molecule, subject to the electronegativity equalisation principle[37]. As such, we are required to use the set of partial charges, $\{q_N\}$, and fit them so
as reproduce the true electrostatic potential. We discuss such a method, which utilises the \textit{ab initio} molecular electrostatic potential (MEP), $\Phi(r)$, defined by\[38\]

$$
\Phi(r) = \sum_{i=1}^{N} \frac{Z_i}{||R_i - r||} - \int dr' \frac{\psi^*(r') \psi(r')}{||r' - r||},
$$

(2.1.24)

where the first term denotes the potential arising from the $i^{th}$ nucleus ($Z_i$ is the corresponding atomic number while $R_i$ is its position), and the second term is the potential resulting from the electronic distribution of the system. Partial charges are then assigned to each of the $N$ atoms by choosing them such that they reproduce the MEP. It may be readily seen that doing so is equivalent to minimising a sum of residuals error function, $\mathcal{E}[39]$, 

$$
\mathcal{E} = \sum_{n=1}^{n_{tot}} \left( \sum_{i=1}^{N} \frac{q_i}{||R_i - r_n||} - \Phi(r_n) \right)^2,
$$

(2.1.25)

where $q_i$ is the partial charge associated with the $i^{th}$ atom and $n_{tot}$ is the number of external points at which the MEP is sampled.

The simplicity of such a model is, of course, appealing. However, the appeal of computationally efficient models should wane with an increase in computational power. Partial charges have been used in simulation since the advent of \textit{in silico} chemical modelling, and so one would now assume them to be redundant. Several key problems arise with the usage of partial charges. Analysis of the above function should quickly reveal that this model is only applicable to a static molecule in a single conformation. If a molecule adopts even a slightly different conformation, a new set of partial charges is required to account for electronic redistribution throughout the molecule, resulting in an entirely different MEP.
It seems apparent that as \( n_{\text{tot}} \to \infty \) in (2.1.25), the set of partial charges should become as optimal as possible in reconstructing the MEP. However, it has been shown\[40\] that for large values of \( N \), no finite value of \( n_{\text{tot}} \) can produce a unique set of \( \{q_N\} \) which give an optimal solution to the least squares problem. Indeed, there are solutions for which certain charges within the set \( \{q_N\} \) may be set to zero, needless to say a nonsense situation. Groups in the past have attempted to validate such solutions by the proposition of ‘buried charges’, where certain atoms in a molecule wouldn’t contribute to the MEP as they were positioned too deeply within the molecule, resulting in their individual potentials being sequestered by other atomic potentials\[41\]. However, such a theory is speculative, and is perhaps attributable to the fitting used in deriving the partial charges\[42\].

A final point that requires discussion is that \( E_{\text{elec}} \) in the form employed in (2.1.23) is not amenable to computation. The inverse distance dependence means that \( E_{\text{elec}} \) decays very slowly, so a great deal of interactions require evaluation if one is to capture all meaningful contributions to the electrostatic energy. One can in principle employ a “cutoff distance”, where \( E_{\text{elec}} \) is evaluated if \( r_{ij} < r_{\text{cutoff}} \). However, the use of a cutoff distance is inadvisable, since major artifacts are found to arise over the course of a simulation\[43, 44\].

The matter is complicated further when periodic boundary conditions are employed, where (2.1.23) must be modified to sum over periodic images of the simulation cell,

\[
E_{\text{elec}} = \frac{1}{8\pi\epsilon_0} \sum_n' \sum_{i,j}^N q_i q_j r_{ij} \left( \frac{1}{\|Ln\|} \right),
\]

(2.1.26)

where the sum over \( n \) is a sum over all periodic box replica lattice vectors, \( L \) is the length of the box and the primed summation implies we omit the \( i = j \) term when \( n = 0 \). This sum is conditionally convergent, and diverges when the system is not electrically neutral\[45\].
The Ewald summation[46], and its various incarnations, have been employed to split the above Coulombic interaction into a sum of a short-ranged real space term and a long-ranged reciprocal space term, both of which are absolutely convergent. We present the result without derivation,

\[ E_{\text{elec}} = \frac{1}{8\pi\epsilon_0} \sum_{i,j} q_i q_j \left[ \sum_{n} \text{erfc}(\alpha ||r_{ij} + n||) + \frac{1}{\pi L^3} \sum_{k \neq 0} \frac{4\pi^2}{||k||^2} \exp \left( -\frac{\pi^2 ||k||^2}{\alpha^2} \right) \cos(k \cdot r_{ij}) \right], \]

(2.1.27)

where \( \alpha \) is a damping factor, controlling the convergence of the Ewald summation and \( k \) is a reciprocal vector, equal to \( 2\pi n/L^2 \)[47]. Much literature is available on the Ewald summation and its computational complexity, and the reader is directed to several excellent reviews for a more thorough discussion[48, 49].

van der Waals and Correlation Energies

The van der Waals and correlation energies are best explained by first introducing a popular potential for their modelling. The Lennard-Jones function[50], \( V^{LJ} \), is one such potential, and is given by

\[ E_{\text{vdW}}(r_i, r_j) + E_{\text{corr}}(r_i, r_j) = V^{LJ}_{ij} = 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right], \]

(2.1.28)

where \( r_{ij} \) is the distance between the atoms at \( r_i \) and \( r_j \), \( \sigma_{ij} \) is the distance at which the inter-particle interaction is equal to zero, and \( \epsilon_{ij} \) is the depth of the well at the lowest energy point on \( V^{LJ}_{ij} \). It is worth briefly mentioning that both the parameters \( \sigma \) and \( \epsilon \) are defined for atoms, and mixing rules must be invoked to obtain the quantities \( \sigma_{ij} \) and \( \epsilon_{ij} \). These mixing rules are entirely empirical, and are not based upon a physical theory. A number exist (see [51] for an enumeration), but we do not detail them further here.
The $r_{ij}^{-12}$ term is repulsive, and dominates the short-range portion of $V_{ij}^{LJ}$, ensuring that particles never come too close to one another. The theoretical basis for this term originates from the quantum mechanical exchange interaction (Pauli repulsion), in that fermions (the electrons) cannot occupy the same quantum states. While one is unable to derive a $r_{ij}^{-12}$ functional form from theory, it is suitable for preventing particles from approaching one another, and is also computationally ideal since one can compute it from squaring the $r_{ij}^{-6}$ term, which we now describe.

In contrast, the $r_{ij}^{-6}$ term is attractive everywhere. Unlike the repulsion term, this “dispersion” term is theoretically-grounded, and derives from the van der Waals interactions between two particles. Atoms and molecules possess dipole moments that are either permanent or induced. The permanent dipole moments are features of a higher order treatment of the multipole moments of a system, as we describe in Section 2.3. The induced dipole moments are a result of the molecular electrostatics being conformationally-dependent; as atoms reconfigure themselves, electronic distributions, and consequentially multipole moments, change. However, no classical force field currently allows for a conformationally-dependent treatment of electrostatics. As such, the interactions involving induced dipole moments are introduced as “bolt-on” terms.

The van der Waals interactions comprise three distinct components: (1) KEESOM INTERACTIONS, Permanent dipole - Permanent dipole (“Orientation”); (2) DEBYE INTERACTIONS, Permanent dipole - Induced dipole (“Induction”), and; (3) LONDON INTERACTIONS, Induced dipole - Induced dipole (“Dispersion”). One can show that each of these terms combined leads to an $r_{ij}^{-6}$ law for the potential energy, which we are unable to derive in completion here since it involves a lengthy derivation. The reader is, however, directed to the excellent review of van der Waals interactions for a complete mathematical derivation[52].
Alternative functional forms are available beyond the Lennard-Jones potential. The majority fall into one of two categories; the first category are the Lennard-Jones-like potentials, using two independent inverse distance terms to describe the repulsion and dispersion interactions. Examples include the shifted force n-m potential[53], the shifted Weeks-Chandler-Anderson potential[54] or the AMOEBA 14-7 pair potential[26]. The second category employs exponential functions, such as the Morse, Buckingham and Born-Huggins-Meyer potentials[55].

2.1.3 *ab initio* Molecular Dynamics

Let us consider the computation of $\langle \psi | H_e | \psi \rangle$ by expanding the electronic wavefunction into a suitably chosen basis set. For our purposes, we will find it convenient to use the eigenfunctions of the electronic Hamiltonian, $\{ \phi_k(r; R) \}$, or “adiabatic basis”,
\[
\psi(r; t) = \sum_{k=0}^{\infty} c_k(t) \phi_k(r; R),
\]
where $c_k(t)$ is a time-dependent complex coefficient, and the constraint $\sum_k |c_k(t)|^2 = 1$ is required to ensure the wavefunction is normalised. We also note that the adiabatic basis functions are orthogonal, i.e. $\langle \phi_i | \phi_j \rangle = \delta_{ij}$, the Dirac delta function. Invoking the TDSE of (2.1.1),
\[
i\hbar \frac{\partial}{\partial t} \psi(r; t) = H_e \psi(r; t)
\]
\[
i\hbar \sum_{k=0}^{\infty} \frac{\partial}{\partial t} c_k(t) \phi_k(r; R) = \sum_{k=0}^{\infty} H_e c_k(t) \phi_k(r; R)
\]
\[
i\hbar \sum_{k=0}^{\infty} \left[ \phi_k(r; R) \dot{c}_k(t) + c_k(t) \dot{\phi}_k(r; R) \right] = \sum_{k=0}^{\infty} c_k(t) E_k \phi_k(r; R),
\]
where $E_k$ is the energy associated with eigenfunction $\phi_k(r; R)$, and we have used the dot notation in the conventional sense of denoting a (partial) temporal deriva-
2.1. MOLECULAR MECHANICS

tive. Multiplying through by \( \langle \phi_m | \) ,

\[
\begin{align*}
\sum_{k=0}^{\infty} \left[ \langle \phi_m | \phi_k \rangle \dot{c}_k(t) + c_k(t) \langle \phi_m | \dot{\phi}_k \rangle \right] &= \sum_{k=0}^{\infty} c_k(t) E_k \langle \phi_m | \phi_k \rangle \\
i \hbar \sum_{k=0}^{\infty} c_k(t) \langle \phi_m | \dot{\phi}_k \rangle &= c_m(t) E_m.
\end{align*}
\] (2.1.31)

Writing the above expression as the time evolution of \( c_m(t) \), we have

\[
i \hbar \dot{c}_m(t) = c_m(t) E_m - i \hbar \sum_{k=0}^{\infty} c_k(t) \langle \phi_m | \dot{\phi}_k \rangle.
\] (2.1.32)

We can use the chain rule to rewrite the time derivative of the electronic eigenfunctions, where

\[
\frac{\partial}{\partial t} \phi_k = \nabla_I \phi_k \cdot \frac{\partial R_I}{\partial t} = \dot{R}_I \cdot \nabla_I \phi_k,
\]

and we rewrite (2.1.32) as

\[
i \hbar \dot{c}_m(t) = c_m(t) E_m - i \hbar \sum_{k=0}^{\infty} c_k(t) \sum_I \dot{R}_I \cdot \langle \phi_m | \nabla_I \phi_k \rangle.
\] (2.1.33)

The coupled equations of motion using the above results can then be formally given in terms of the expansion coefficients

\[
M_I \dot{R}_I = -\sum_{k=0}^{\infty} |c_k(t)|^2 \nabla_I E_k - \sum_{k,m} c^*_k(t) c_m(t) (E_k - E_m) \langle \phi_k | \nabla_I \phi_m \rangle,
\] (2.1.34)

\[
i \hbar \dot{c}_m(t) = c_m(t) E_m - i \hbar \sum_{k=0}^{\infty} c_k(t) \sum_I \dot{R}_I \cdot \langle \phi_m | \nabla_I \phi_k \rangle,
\] (2.1.35)

where the first equation can be trivially obtained through expanding \( \langle \psi_k | H_e | \psi_m \rangle \).

This formalism is referred to as Ehrenfest Molecular Dynamics, and includes non-adiabatic transitions between electronic eigenfunctions. When the electronic wavefunction exists in a superposition of quantum states, i.e., more than one coefficient is non-zero, the PES exists as a linear combination of contributions from the relevant eigenfunctions, and the nuclei are said to evolve classically upon the “mean-field” PES.
One can immediately appreciate the computational effort required to use Ehrenfest MD. Since both the electronic wavefunction and nuclei require numerical integration, the user is limited to a dynamical timestep which allows appropriate integration of the electronic degrees of freedom. Since the electrons move far quicker than nuclei, this timestep is orders of magnitude smaller than the conventional timestep one is permitted to use in classical MD, and the process is highly time-consuming for chemically relevant timescales[8]. Naturally, one also requires a finite basis to replace the infinite summations over basis states to make the problem computationally tractable.

One can also conceive of a contrasting form of dynamics, where a dynamical timestep suitable for the nuclear degrees of freedom can be used at the expense of solving the time-independent Schrödinger equation at each step. This approach is referred to as Born-Oppenheimer MD. However, a slight caveat is also introduced, in that dynamics can only be performed on the PES corresponding to a single basis state at any one time, i.e. there is no mean-field variant of Born-Oppenheimer MD where basis states are allowed to interfere. Of course, one can perform dynamics on the PES arising from an excited state, but non-adiabatic transitions cannot be modelled. We constrain our discussion to the ground state eigenfunction, $\phi_0$.

A number of quantum chemical software packages possess some form of Born-Oppenheimer MD functionality, typically in the form of Atom Centred Density Matrix Propagation (ADMP)[59, 60], which has equivalent functionality to Born-Oppenheimer MD, but with a great reduction in computational cost. The coupled equations of motion for Born-Oppenheimer MD are

\begin{align}
M_I \ddot{\mathbf{R}}_I &= -\nabla I \min_{\phi_0} \left\{ \langle \phi_0 | \mathcal{H}_e | \phi_0 \rangle \right\} \\
\mathcal{H}_e \phi_0 &= E_0 \phi_0
\end{align}

(2.1.36) (2.1.37)
At every dynamical step, a minimisation of the expectation of the electronic Hamiltonian with respect to the basis state $\phi_0$ is required. The variation of $\phi_0$ such that $\langle \phi_0 | H_e | \phi_0 \rangle$ is minimised can be solved by the conventional quantum chemical techniques, such as the Hartree-Fock or density functional theories\cite{61}.

It is impossible to systematically outline every type of \textit{ab initio} molecular dynamics that is available. However, we make special mention of the path integral MD (PIMD)\cite{62} and Car-Parrinello\cite{63} approaches. PIMD employs the fact that quantum mechanical path integrals of a system are isomorphic to a ring polymer, whose dynamics are amenable to computational treatment. The Car-Parrinello is arguably the most successful form of \textit{ab initio} MD, and builds upon the Born-Oppenheimer treatment discussed above. However, as opposed to continually solving the electronic Schrödinger equation and updating the nuclear positions in the resultant field, the Car-Parrinello scheme includes electronic dynamics as fictitious degrees of freedom in an extended Lagrangian.

### 2.2 Atomic Partitioning

The finite nature of matter has been a topic of scrutiny since ancient times. The musings of Democritus, Empedocles and Anaxagoras, while fascinating, are superfluous to our current discussion. The inherent wave-nature of electrons prevents their localisation to points in space. Instead, an electron is described by a wave-function, $\psi(r)$, where $r \in \mathbb{R}^3$. The Born rules allows one to characterise the expectation value of an observable, represented by an operator, $\mathcal{O}$, as

$$
\langle \mathcal{O} \rangle = \int_{\mathbb{R}^3} d\mathbf{r} \psi^*(\mathbf{r}) \mathcal{O} \psi(\mathbf{r}) .
$$

(2.2.1)
The integral over all space is problematic when one wishes to assign an observable to an atom within a molecule. Rather, it would be preferable to integrate over some atomic domain, and rigorously demonstrate that the resultant expectation of that observable corresponds to an atomic property.

Several methods have been developed which allow for the rigorous partitioning of an “atom in a molecule”. Such partitioning schemes allow for the electron density to be compartmentalised into finite domains. These domains can subsequently be integrated over to yield electronic properties associated with an atom in a molecule. For example, integration of the electron density, $\rho(r)$, over an atomic domain, $\Omega$, characterises the electronic population associated with that atom,

$$ q(\Omega) = \int_{\Omega} d\mathbf{r} \rho(\mathbf{r}). $$

(2.2.2)

Note that if one wished to compute the atomic charge associated with the atom, the nuclear charge would be subtracted from the above equation. In the following, we give a brief outline of a number of popular methods for characterising an atom in a molecule.

### 2.2.1 Hirshfeld Partitioning

We can model the electron density of a molecule comprising $N$ atoms as a linear combination of $N$ spherically symmetric nuclear-centred atomic electron densities,

$$ \rho^{pr}(\mathbf{r}) = \sum_{i=1}^{N} \rho^{at}_i(\mathbf{r}), $$

(2.2.3)

where $\rho^{at}_i(\mathbf{r})$ are the individual atomic electron density functions, and we have introduced the notion of a “promolecule” density, $\rho^{pr}(\mathbf{r})$[64]. A scalar field for
2.2. ATOMIC PARTITIONING

each atom, termed the sharing function, is introduced,

\[ w_i(r) = \frac{\rho_{i}^{at}(r)}{\rho_{pro}(r)}, \quad (2.2.4) \]

denoting the proportion of the electron density at \( r \) belonging to the \( i^{th} \) atom. Naturally the sum of weighting functions for a given \( r \) is normalised, i.e. \( \sum_i w_i(r) = 1 \).

Now, if one possesses the true\(^6\) electron density of the molecular system, \( \rho(r) \), atomic electron densities can be constructed by use of the sharing function,

\[ \rho_i(r) = w_i(r)\rho(r), \quad (2.2.5) \]

where \( \rho_i(r) \) denotes the “true” electron density function associated with the \( i^{th} \) atom. Atomic properties, such as the charge associated with an atom, are then defined by

\[ q_i = -\int dr \rho_i(r). \quad (2.2.6) \]

We can immediately take issue with the Hirshfeld partitioning scheme by noting that the partitioning does not define finite atomic domains- the atomic domains are said to be overlapping. Atomic electron densities are delocalised throughout all space, so we conclude that atomic properties are dictated by the system in which the atom is situated. This conclusion is in contrast to the notion of a functional group, where certain atomic properties are conserved regardless of the system[65].

We take the example of lithium fluoride and lithium hydride. Contour diagrams of the molecular electron density reveal the similarities in the electron density of lithium. In spite of this, Hirshfeld partitioning predicts atomic charges of +0.59 and +0.41 for the lithium atom in LiF and LiH, respectively[66].

Independent of the criticism we have just outlined, Parr and Nalewajski[67] have used information theory to show that the Hirshfeld partitioning corresponds to a

---

\(^6\) We use the word “true” loosely. By true electron density, we mean an electron density as obtained from experimental or \textit{ab initio} data.
maximal conservation of information by isolated atoms. Later work has shown
this result to be completely dependent upon the entropy function used to measure
the information content of a system, and the work of Parr and Nalewajski to be
valid only for logarithmic entropy measures[68].

Owing to its simplicity, the Hirshfeld partitioning scheme has been used extensively
for deriving a number of atomic properties. For example, the Fukui function[69],
a quantity arising in conceptual DFT to probe the reactivity of a system when
electrons are either added or removed, have been localised to atoms in a molecule.
The computed atomic Fukui functions correlated well with reaction site selectivities
as found from experimental studies[70].

Whilst initially formulated to yield atomic properties, Hirshfeld partitioning is
easily extensible to molecular properties. Water molecule polarisabilities from
clusters[71] and the partitioning of crystalline electron densities into molecular
electron densities[72] have been shown to be amenable to such a treatment.

2.2.2 Bader’s Atoms in Molecules

The development of molecular orbital theory largely caused a decline in chemical
understanding of the theoretical description of molecular systems. The fact that
each electron occupies a molecular orbital dispersed across the spatial extent of
the molecule gave rise to a valid query: why do functional groups impart some
property, such as reactivity, to the molecule, when the distribution of electrons
throughout the system is essentially no different to the “un-functional”7 molecule?
Surely there must be some localisation of electrons to the functional group, which

---

7We use “un-functional” as a loose term, denoting the molecular species lacking the functional
group
permits subsequent functionality. If this is not the case, then even the most fundamental chemical concepts, such as those of nucleophiles and electrophiles, have no theoretical grounding. These terms are used to denote a property of an atom in a molecule, which is not recovered by molecular orbital theory. For example, if one assesses butanol by means of molecular orbital theory, the electrons ‘belonging’ to the hydroxyl group are dispersed throughout the entirety of the molecule. If this is truly the case, then it becomes particularly problematic when one attempts to explain why the presence of the functional group imparts reactivity (an electronic phenomenon), when the electrons are not localised.

Bader\cite{5} has addressed this problem by suggesting that atoms are open systems within a molecule, partitioned from one another by boundary surfaces that are recovered by the Laplacian of the molecular electron density. In this way, one is able to unambiguously define an “atom in a molecule”.

The gradient of a scalar field is a vector field, the vectors of which are directed along the path of greatest increase in the scalar field. As such, the vectors which define the field $\nabla \rho(r)$ point towards the greatest increase in the scalar field $\rho(r)$. By simple vector calculus, one is able to show that the vectors $\nabla \rho(r)$ orthogonally intersect constant electron density isosurfaces\cite{73}.

Points in the vector field satisfying $\nabla \rho(r) = 0$ are termed critical points, and represent a maximum, minimum or point of inflexion within the scalar field $\rho(r)$. The identity of each critical point is revealed by assessing the curvature of $\rho(r)$ at each point, i.e. calculation of the Hessian of $\rho(r)$, $H(\rho)$.

Since $H(\rho)$ is real and symmetric (Hermitian), it may be diagonalised by its eigenvectors, the principal axes of curvature. The resultant eigenvalues correspond to the values of the curvature along each of the principal axes. The nature of the
critical point in question is then given by two easily evaluated parameters; the rank \((\omega)\) and signature \((\sigma)\) of the critical point, where the former is defined as the number of non-zero eigenvalues of \(\rho(r)\), and the latter being the sum of the signs of the eigenvalues. As such, each critical point is fully characterised by \((\omega, \sigma)\), all variations being summarised below\[74\].

<table>
<thead>
<tr>
<th>((\omega, \sigma))</th>
<th>Description</th>
<th>Corresponds to</th>
</tr>
</thead>
<tbody>
<tr>
<td>((3, -3))</td>
<td>Maximum in all three principal axes of curvature.</td>
<td>Maximum in (\rho(r)). Found primarily at nuclear positions (nuclear attractors).</td>
</tr>
<tr>
<td>((3, -1))</td>
<td>Maximum in two principal axes but minimum along third axis.</td>
<td>Found at inter-nuclear positions between nuclei (bond critical points).</td>
</tr>
<tr>
<td>((3, +1))</td>
<td>Minimum in two principal axes but a maximum along the third.</td>
<td>Found within the centre of ring structures (ring critical point).</td>
</tr>
<tr>
<td>((3, +3))</td>
<td>Minimum in all three principal axes of curvature.</td>
<td>Found within the centre of cage structures (cage critical point).</td>
</tr>
</tbody>
</table>

A fundamental result in the topology of, of great importance in the following, is the partitioning of a molecular system into topological atoms. A key feature necessary to achieve this result is the gradient path. An easy way to grasp what this is to think of a succession of very short gradient vectors, one after the other and constantly changing direction. In the limit of infinitesimally short gradient vectors, one obtains a smooth and (in general) curved path, which is the gradient path. A gradient path always originates at a critical point and terminates at another critical point. Bundles of gradient paths form a topological object depending
2.2. ATOMIC PARTITIONING

on the signature of the critical points that the object connects. All possibilities have been exhaustively discussed before\cite{75} but three ubiquitous possibilities are specified as follows:

1. The topological atom is a bundle of gradient paths originating at infinity and terminating at the nucleus.

2. The bond path (or more generally atomic interaction line) is the set of two gradient paths, each originating at a bond critical point and terminating at a different nucleus.

3. The interatomic surface (IAS), which is a bundle of gradient paths originating at infinity and terminating at a bond critical point.

For IASs only, the following condition is met

\[
\nabla \rho(r) \cdot \mathbf{n}(r) = 0,
\]

where \(\mathbf{n}(r)\) is defined as the vector normal to the IAS. By finding all surfaces which obey this condition, the molecule is completely partitioned into distinct atomic volumes, or “atomic basins”, \(\Omega_i\), where the subscript denotes the atomic basin associated with the \(i^{th}\) atom in a molecule. All key topological features in \(\nabla \rho(r)\) are summarised in Figure 2.1.

Integration over these atomic basins allows atomic properties \(P_j(\Omega_i)\) to be defined and calculated. The universal formula from which all atomic properties can be calculated is
Figure 2.1: A contour plot of the electron density of in molecular plane of furan superimposed onto a representative collection of gradient paths. Atoms are represented by black circles, where the gradient paths terminate. Interatomic surfaces are highlighted as solid curves, and contain bond critical points (black squares). A ring critical point (triangle) is also shown in the centre of the furan.

\[
P_f(\Omega_i) = \int_{\Omega_i} d\tau f(\mathbf{r}),
\]

where integration with respect to \(d\tau\) denotes a triple integration over all three Cartesian coordinates, confined to the atomic volume \(\Omega_i\), and \(f(\mathbf{r})\) denotes a property density. For example, if \(f(\mathbf{r})\) equals the electron density \(\rho(\mathbf{r})\) then the corre-
2.2. ATOMIC PARTITIONING

The corresponding atomic property is the electronic population of the topological atom,

\[ N(\Omega_i) = \int_{\Omega_i} \rho(r) d\tau. \]  (2.2.9)

If \( f(r) = 1 \) then we obtain the atomic volume,

\[ v(\Omega_i) = \int_{\Omega_i} d\tau, \]  (2.2.10)

and when \( f(r) = \rho(r)R_{\ell m}(r) \) the topological atom’s multipole moments[76], where \( R_{\ell m}(r) \) is a spherical tensor[77] of rank \( \ell \) and \( m \). Others have shown[78, 79] the better agreement with reference electrostatic potentials of topological multipole moments compared to CHELPG[80] charges. A further advantage of AIM is that the finite size and non-overlapping nature of the topological atoms avoids the penetration effect, which may otherwise appear in the calculation of intermolecular interaction energies.

We wish to clarify the nomenclature that we will adopt throughout this work. Bader has been insistent that the topological atoms as found through an AIM partitioning are open quantum mechanical systems, i.e. quantum atoms. In other words, an AIM partitioning defines an unequivocal quantum atom, i.e. \( \Omega_A \) is a topological atom if and only if \( \Omega_A \) is a quantum atom.

In computing the kinetic energy of an atom, one requires the integral of a kinetic energy density over the corresponding atomic domain. It can be shown that the addition of an arbitrary constant to the kinetic energy density recovers a unique kinetic energy[81] upon integration over the domain of a quantum atom[82]. However, one can show that there exist other spatial domains, that are not topological atoms, which also produce unique kinetic energies when the kinetic energy density

\[ K(\Omega_i) = \int_{\Omega_i} \frac{1}{2} \left( \frac{\partial \psi}{\partial r} \frac{\partial \psi}{\partial r} \right) d\tau, \]  (2.2.11)
is integrated over the domain[83]. Therefore, if $\Omega_A$ is a topological atom, it is a quantum atom, but the reverse does not hold.

We are unconcerned with this one-way implication; the fact that topological atoms are not the only atoms with a unique kinetic energy is unimportant. We are purely interested in the fact that topological atoms partition a molecular system into well-defined atomic domains, over which one can integrate and obtain atomic expectation values[84]. As such, we decouple our usage of the theory of AIM from the claim that AIM provides the atoms of chemistry by referring to the partitioning as Quantum Chemical Topology (QCT).

### 2.2.3 Interacting Quantum Atoms

Energetic partitioning is a commonly-invoked methodology in quantum chemistry. Being able to obtain individual energetic components can prove to be highly informative, and can be used to correlate chemical intuition with quantum mechanical processes. Two popular energetic partitioning schemes are attributable to Morokuma[85], and Ziegler and Rauk[86]. We concern ourselves with an energy partitioning scheme, “Interacting Quantum Atoms”, which utilises the atomic partitioning outlined in 2.2.2, which is gaining in popularity[87].

Strictly speaking, we are interested in partitioning the Hamiltonian of a molecular system, $H$, into atomic contributions,

$$H = \sum_A \left[ T^A + V^{AA} + \sum_{B \neq A} V^{AB} \right], \quad (2.2.11)$$

where $T^A$ is an atomic kinetic energy, $V^{AA}$ is an intra-atomic atomic potential energy and $V^{AB}$ is an inter-atomic potential energy. The potential energy terms can be further decomposed by qualitative inspection. $V^{AA}$ is composed of electronic
2.2. ATOMIC PARTITIONING

repulsion and electron-nuclear attraction operators, \( V^{AA}_{ee} \) and \( V^{AA}_{en} \), respectively, where the electrons and nuclei “belong to” atom A,

\[
V^{AA} = V^{AA}_{ee} + V^{AA}_{en} \quad \text{(2.2.12)}
\]

\( V^{AB} \) similarly comprises the electronic repulsion and electron nuclear attraction energies, in addition to a nuclear-nuclear repulsion energies, \( V^{AB}_{nn} \), but interactions are between particles in different atoms, i.e.

\[
V^{AB} = V^{AB}_{ee} + V^{AB}_{en} + V^{AB}_{ne} + V^{AB}_{nn} \quad \text{(2.2.13)}
\]

allowing us to rewrite (2.2.11)

\[
H = \sum_A \left[ T_A + V^{AA}_{ee} + V^{AA}_{en} + \sum_{B \neq A} \left( V^{AB}_{ee} + V^{AB}_{en} + V^{AB}_{ne} + V^{AB}_{nn} \right) \right]. \quad \text{(2.2.14)}
\]

Note that the nuclear-nuclear repulsion energy can by immediately given by the classical electrostatic energy between point charges

\[
V^{AB}_{nn} = \frac{Z_A Z_B}{r_{AB}}, \quad \text{(2.2.15)}
\]

where \( Z_A, Z_B \) are the nuclear charges associated with atoms A and B, and \( r_{AB} \) is the interatomic distance. A note is in place on the nomenclature we have adopted when denoting the potential energy terms. The subscript and superscript are in correspondence with one another, so that \( V^{AB}_{en} \) denotes the potential energy operator for an electron in atom A interacting with the nucleus of atom B. For this reason, we see that \( V^{AB}_{en} \neq V^{AB}_{ne} \), but \( V^{AB}_{en} = V^{BA}_{ne} \).

We now detail the methods required to accomplish the above partitioning for the electronic terms.
Density Matrices

For a system containing $N$ electrons, we introduce the concept of the "$p^{th}$ order reduced density matrix" (p-RDM)\(^8\), $\Gamma_p(r_1,...,r_p;r'_1,...,r'_p)$,

$$\Gamma_p(r_1,...,r_p;r'_1,...,r'_p) = \left( \begin{array}{c} N \\ p \end{array} \right) \int dr_{p+1}... \int dr_N \Psi^*(r'_1,...,r'_p,r_{p+1},...,r_N)\Psi(r_1,...,r_p,r_{p+1},...,r_N).$$  \hspace{1cm} (2.2.16)

To familiarise ourselves with the p-RDM, we take the special case of the 1-RDM, $\Gamma_1(r_1;r'_1)$,

$$\Gamma_1(r_1;r'_1) = N \int dr_2... \int dr_N \Psi^*(r'_1,...,r_N)\Psi(r_1,...,r_N).$$ \hspace{1cm} (2.2.17)

Integration over the position vector of every electron apart from one removes the dependence of the 1-RDM on the electrons that have been “integrated out”, and we are left with the coherence between a single electron positioned at $r$ and $r'$.

Take the one-electron operator, $\hat{O}_1 = \sum_{i=1}^N \hat{o}_i(r_i)$, the expectation of which is given by

$$\langle \hat{O}_1 \rangle = \sum_{i=1}^N \int dr_1... \int dr_N \Psi^*(r_1,...,r_N)\hat{o}_i\Psi(r_1,...,r_N).$$ \hspace{1cm} (2.2.18)

Since fermions are indistinguishable, it can be shown that each term in the summation is equivalent, allowing us to write

$$\langle \hat{O}_1 \rangle = N \int dr_1... \int dr_N \Psi^*(r_1,...,r_N)\hat{o}_1\Psi(r_1,...,r_N),$$ \hspace{1cm} (2.2.19)

where the choice of $\hat{o}_1$ is arbitrary. We can reformulate this in terms of the 1-RDM by

$$\langle \hat{O}_1 \rangle = \int dr_1 \left[ \hat{o}_1 \Gamma_1(r_1;r'_1) \right],$$ \hspace{1cm} (2.2.20)

\(^8\)Formally, we only concern ourselves with the spatial p-RDM, where we make no reference to the spin degree of freedom of each electron. However, one can easily extend the following to account for spin by simply appending the spin to degree of freedom to each electronic state vector, $x = r \wedge \sigma$. 
where we see that the primed coordinate prevents the operator $\hat{o}_1$ from operating on the complex conjugate part of the 1-RDM. Once the operator acts upon $\Psi(r_1, ..., r_N)$, we allow $r_1 = r'_1$, and we recover (2.2.18). Thus, the primed notation tacitly implies that $r_1$ and $r'_1$ are distinct until the operator has acted on the wavefunction, at which point they become equivalent and integration can be performed.

We can extend this approach for two-electron operators, $\hat{O}_2$, where we introduce the 2-RDM

$$\Gamma_2(r_1, r_2; r'_1, r'_2) = \frac{N(N-1)}{2} \int dr_3 ... \int dr_N \Psi^*(r'_1, r'_2, ..., r_N)\Psi(r_1, r_2, ..., r_N).$$

(2.2.21)

Then,

$$\langle \hat{O}_2 \rangle = \int dr_1 \int dr_2 \left[ \hat{o}_{12} \Gamma_2(r_1, r_2; r'_1, r'_2) \right],$$

(2.2.22)

where $\hat{o}_{12}$ is a generic two-electron operator. We have adopted the convention of Löwdin[88], where a factor of a half is included in the second order density matrix to account for only the unique pairs of electrons. We note that this is in keeping with the combinatorial term in (2.2.16). McWeeny[89], for a not immediately obvious reason, omits this factor, which is seemingly erroneous, but is followed by Blanco et al.[90] in the original work on IQA. We will follow Löwdin’s original definition and maintain consistency with (2.2.16).

**One-Electron Potential**

We introduce a Heaviside function, $\Theta_\Omega(r)$, which is defined as being equal to unity if $r$ lies within the atomic basin $\Omega$, and is equal to zero otherwise. Since a QCT partitioning is spatially exhaustive, the 1-RDM can be partitioned into atomic
contributions by use of this Heaviside function,

\[ \Gamma_1(r_1; r'_1) = \sum_A \Gamma_A^1(r_1; r'_1) = \sum_A \Gamma_A^1(r_1; r'_1) \Theta(r'_1). \] (2.2.23)

We observe that such a partitioning is also valid for single-electron operator quantities, since

\[ \langle \mathcal{O}_1 \rangle = \int dr_1 \left[ \hat{o}_1 \Gamma_1(r_1; r'_1) \right] = \sum_A \int dr_1 \left[ \hat{o}_1 \Gamma_A^1(r_1; r'_1) \right] = \sum_A \int d\Omega_A \hat{o}_1 \Gamma_A^1(r_1; r'_1) = \sum_A \langle \mathcal{O}_A^1 \rangle, \] (2.2.24)

where \( \langle \mathcal{O}_A^1 \rangle \) is the expectation of the one-electron operator \( \mathcal{O}_1 \) associated with the atom \( A \).

The kinetic energy can be written

\[ T^A = -\frac{\hbar^2}{2m} \int_{\Omega_A} dr_1 \nabla_1^2 \Gamma_1(r_1; r'_1), \] (2.2.25)

where \( m \) is the mass of an electron and \( \nabla_1^2 \) is the Laplacian with respect to \( r_1 \).

The electron-nuclear attraction can be written

\[ V_{en}^{AB} = -\int_{\Omega_A} dr_1 \frac{\rho(r_1)Z_B}{r_{en}}, \] (2.2.26)

where the 1-RDM is reduced to the electron density, \( \rho(r_1) \) since \( \mathcal{V}_{en}^{AB} \) does not alter the wavefunction it operates on. By permuting \( A \) and \( B \), we obtain an equivalent expression for \( V_{ne}^{AB} \). By taking the case where \( A = B \), we also recover the expression for \( V_{en}^{AA} \).
2.2. ATOMIC PARTITIONING

Two-Electron Potential

The operator $\gamma^{AB}_{ee}$ requires the 2-RDM, which can be split into three distinct components:

$$\Gamma_2(r_1, r_2; r'_1, r'_2) = \rho(r_1)\rho(r_2) - \Gamma_1(r_1; r_2)\Gamma_1(r_2; r_1) + \Gamma^{corr}_2(r_1, r_2; r'_1, r'_2), \quad (2.2.27)$$

where the first term refers to the uncorrelated Coulomb density, the second term to the quantum mechanical electron exchange, and the third term to the electron correlation. Then,

$$V^{AB}_{ee} = \int_{\Omega_A} dr_1 \int_{\Omega_B} dr_2 \frac{\Gamma_2(r_1, r_2; r'_1, r'_2)}{r_{12}}$$

$$= \int_{\Omega_A} dr_1 \int_{\Omega_B} dr_2 \frac{\rho(r_1)\rho(r_2)}{r_{12}} - \int_{\Omega_A} dr_1 \int_{\Omega_B} dr_2 \frac{\Gamma_1(r_1; r_2)\Gamma_1(r_2; r_1)}{r_{12}}$$

$$+ \int_{\Omega_A} dr_1 \int_{\Omega_B} dr_2 \frac{\Gamma^{corr}_2(r_1, r_2; r'_1, r'_2)}{r_{12}}$$

$$= V^{AB}_{ee, coul} + V^{AB}_{ee, exch} + V^{AB}_{ee, corr} \quad (2.2.28)$$

where $V^{AB}_{ee, coul}$ is the Coulomb energy (which we deal with in Section 2.3.2), $V^{AB}_{ee, exch}$ is the Fock-Dirac exchange energy, and $V^{AB}_{ee, corr}$ is the electron correlation energy. Note that by taking the case where $A = B$, we obtain the expression for $V^{AA}_{ee}$. We have then accounted for all of the terms in (2.2.14), and illustrated a complete energy partitioning through IQA.


CHAPTER 2. EXPOSITION

2.3 Multipole Moment Electrostatics

2.3.1 Mathematical Details

The Multipole Moments

The Poisson equation in three spatial dimensions is given by

$$\nabla^2 \phi(r) = -\frac{\rho(r)}{\epsilon_0},$$

(2.3.1)

where $\phi(r)$ is the electric potential at $r$, $\epsilon_0$ is the permittivity of free space and $\rho(r)$ is a charge distribution. We assume the system is in some arbitrary coordinate frame. The Green’s function, $G(r, r')$ for the Poisson equation is defined as

$$\nabla^2 G(r, r') = \delta(r - r'),$$

(2.3.2)

where the Dirac delta function, $\delta(r - r')$ has been introduced. By analysis, and a little foresight, we see that $r'$ corresponds to a position at which the electric potential diverges, i.e. a charge, as can be verified from Coulomb’s law. By definition, $\phi(r)$ is then the convolution of the Green’s function and the inhomogeneous term for the Poisson equation, i.e.

$$\phi(r) = -G(r, r') * \frac{\rho(r)}{\epsilon_0} = -\frac{1}{\epsilon_0} \int_{\mathbb{R}^3} G(r, r') \rho(r') dr'. $$

(2.3.3)

The Green’s function which solves (2.3.2) is the Newton kernel,

$$G(r, r') = -\frac{1}{4\pi ||r - r'||},$$

(2.3.4)

leading us to the solution of (2.3.3)

$$\phi(r) = \frac{1}{4\pi \epsilon_0} \int_{\mathbb{R}^3} \frac{\rho(r')}{||r - r'||} dr'. $$

(2.3.5)
We rewrite the factor of $||r - r'||^{-1}$ by use of a well-known formula from the theory of spherical harmonics\[91\]

$$
||r - r'||^{-1} = (r^2 + (r')^2 - 2rr' \cos \chi)^{-1/2}
$$

$$
= \frac{1}{r} \left( 1 + \left( \frac{r'}{r} \right)^2 - 2 \left( \frac{r'}{r} \right) \cos \chi \right)^{-1/2}
$$

$$
= \frac{1}{r} \sum_{\ell=0}^{\infty} \left( \frac{r'}{r} \right)^{\ell} P_\ell(\cos \chi) = \sum_{\ell=0}^{\infty} \left( \frac{r'}{r^{\ell+1}} \right)^{\ell} P_\ell(\cos \chi),
$$

(2.3.6)

where we have used $r = ||r||$, $r' = ||r'||$, $\chi$ is the angle between $r$ and $r'$, and $P_\ell(r) = \frac{\ell!}{2^\ell \ell!} \frac{d^\ell}{dr^\ell} (r^2 - 1)^\ell$ is the $\ell$th order Legendre polynomial. Now, consider the spherical angles $\{\Theta, \Phi\}$ and $\{\theta, \phi\}$ that $r$ and $r'$ make with some fixed Cartesian frame of reference. The addition theorem of spherical harmonics expresses the Legendre polynomial in terms of these spherical angles as

$$
P_\ell(\cos \chi) = \sum_{m=-\ell}^{\ell} \frac{(\ell - |m|)!}{(\ell + |m|)!} P^{|m|}_\ell(\cos \Theta) P^{|m|}_\ell(\cos \theta) e^{im(\Phi - \phi)},
$$

(2.3.7)

Further, introducing the spherical harmonics of rank $\ell, m$,

$$
Y_{\ell m}(\theta, \phi) = (-1)^m \sqrt{\frac{2\ell + 1}{4\pi} \frac{(\ell - |m|)!}{(\ell + |m|)!} \ell m^{|m|} P^{|m|}_\ell(\cos \theta) e^{im\phi}} \quad (m \geq 0)
$$

$$
Y_{\ell,-|m|}(\theta, \phi) = (-1)^{\ell-m} Y^*_{{\ell}|m|}(\theta, \phi) \quad (m < 0),
$$

(2.3.8)

(2.3.6) can be rewritten to read

$$
||r - r'||^{-1} = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \left( \frac{r'}{r^{\ell+1}} \right)^{\ell} \sqrt{\frac{4\pi}{2\ell + 1}} Y_{\ell,-|m|}(\Theta, \Phi) Y_{\ell,m}(\theta, \phi),
$$

(2.3.9)

allowing us to express the electric potential of (2.3.5) as

$$
\phi(r) = \frac{1}{4\pi\epsilon_0} \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \int_{\mathbb{R}^3} dr' \rho(r') \left( \frac{r'}{r^{\ell+1}} \right)^{\ell} \sqrt{\frac{4\pi}{2\ell + 1}} Y_{\ell,-|m|}(\Theta, \Phi) Y_{\ell,m}(\theta, \phi)
$$

$$
= \frac{1}{4\pi\epsilon_0} \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \frac{Q_{\ell,m}}{r^{\ell+1}} \sqrt{\frac{4\pi}{2\ell + 1}} Y_{\ell,-|m|}(\Theta, \Phi),
$$

(2.3.10)
Figure 2.2: Vectors \( \mathbf{r} \) and \( \mathbf{r}' \) in a Cartesian axis system, along with their colatitudinal and azimuthal degrees of freedom, \((\Theta, \Phi)\) and \((\theta, \phi)\), respectively.

In this final expression, we have introduced the multipole moment of rank \((\ell, m)\),

\[
Q_{\ell,m} = \int_{\mathbb{R}^3} d\mathbf{r}' \rho(\mathbf{r}') (\mathbf{r}')^\ell \sqrt{\frac{4\pi}{2\ell + 1}} Y_{\ell,m}(\theta, \phi). \tag{2.3.11}
\]

When \( \ell = 0 \) (and \( m = 0 \) by extension), it is easily shown, with the knowledge that \( Y_{0,0}(\theta, \phi) = \frac{1}{2\sqrt{\pi}} \), that the electric potential becomes

\[
\phi(\mathbf{r}) = \frac{1}{4\pi \epsilon_0 r} \int_{\mathbb{R}^3} d\mathbf{r}' \rho(\mathbf{r}') , \tag{2.3.12}
\]

which is Coulomb’s Law, describing a spherically symmetric potential about \( \mathbf{r} \).

Also, with the three spherical harmonics

\[
Y_{1,0}(\theta, \phi) = \frac{1}{2} \sqrt{\frac{3}{\pi}} \cos \theta \\
Y_{1,1}(\theta, \phi) = -\frac{1}{2} \sqrt{\frac{3}{2\pi}} e^{i\phi} \sin \theta \\
Y_{1,-1}(\theta, \phi) = \frac{1}{2} \sqrt{\frac{3}{2\pi}} e^{-i\phi} \sin \theta \tag{2.3.13}
\]
we find that the multipole moments of rank \( \ell = 1 \) take the form

\[
Q_{1,0} = \int_{\mathbb{R}^3} d\mathbf{r}' \rho(\mathbf{r}') r' \cos \theta
\]

\[
= \mu_z \quad (2.3.14)
\]

\[
Q_{1,1} = -\frac{1}{\sqrt{2}} \int_{\mathbb{R}^3} d\mathbf{r}' \rho(\mathbf{r}') r' \left( \sin \theta \cos \phi + i \sin \phi \sin \theta \right)
\]

\[
= -\frac{1}{\sqrt{2}} \left( \mu_y + i \mu_x \right) \quad (2.3.15)
\]

\[
Q_{1,-1} = \frac{1}{\sqrt{2}} \int_{\mathbb{R}^3} d\mathbf{r}' \rho(\mathbf{r}') r' \left( \sin \theta \cos \phi - i \sin \phi \sin \theta \right)
\]

\[
= \frac{1}{\sqrt{2}} \left( \mu_y - i \mu_x \right) \quad (2.3.16)
\]

where we have introduced components of the dipole moment vector, \( \mathbf{\mu} \),

\[
\mathbf{\mu}_\alpha = \left( \frac{\mathbf{r}'}{r'} \cdot \hat{\alpha} \right) \int_{\mathbb{R}^3} d\mathbf{r}' \rho(\mathbf{r}') r' \quad (\alpha = x, y, z) . \quad (2.3.17)
\]

In the above, \( \hat{\alpha} \) is a unit vector along the Cartesian axis along which \( \mu_\alpha \) is oriented.

Higher order multipole moments are obtained in a similar way.

**Interaction**

Consider the Coulomb interaction between two charge distributions, which requires the computation of an expensive six-dimensional integral

\[
E_{AB} = \sum_{A \neq B} \int_{\mathbb{R}^3} d\mathbf{r}_A \int_{\mathbb{R}^3} d\mathbf{r}_B \frac{\rho(\mathbf{r}_A) \rho(\mathbf{r}_B)}{r_{AB}} . \quad (2.3.18)
\]

The following derivation is taken from the extensive work of Stone[92] on the spherical tensor formulation for multipole moment interactions. The factor of \( 1/r_{AB} \)
Figure 2.3: Vector quantities required for an expansion, in multipole moments, of the interaction between the charge distributions centred at \( R_A \) and \( R_B \).

can be transformed into an expansion in Legendre polynomials, as undertaken in (2.3.6) by recognising that

\[
\begin{align*}
  r_{AB} &= \left\| (R_B + r_B) - (R_A + r_A) \right\| = \left\| R_{AB} - (r_A - r_B) \right\|,
  \\
  \frac{1}{\left\| R_{AB} - (r_A - r_B) \right\|} &= \sum_{\ell_A, \ell_B=0}^{\infty} \sum_{m_A=-\ell_A}^{\ell_A} \sum_{m_B=-\ell_B}^{\ell_B} T_{\ell_A, \ell_B, m_A, m_B} R_{\ell_B, m_B}(r_B) R_{\ell_A, m_A}(r_A).
\end{align*}
\]

We introduce the regular and irregular spherical harmonics,

\[
\begin{align*}
  R_{\ell, m}(r) &= r^\ell Y_{\ell, m}(\theta, \phi) \\
  I_{\ell, m}(r) &= \frac{1}{r^{\ell+1}} Y_{\ell, m}(\theta, \phi)
\end{align*}
\]

respectively, as well as the geometric “interaction tensor”, \( T_{\ell_A, \ell_B, m_A, m_B} \), which is an expression for the mutual orientation of the local axis systems of \( A \) and \( B \), in
addition to their distance from their origins, $R_{AB}$, and takes the form

$$T_{\ell_A,\ell_B,m_A,m_B} = (-1)^{\ell_A} \sqrt{\frac{(2\ell_A + 2\ell_B + 1)!}{(2\ell_A)! (2\ell_B)!}} \times \left[ \ell_A \quad \ell_B \quad \ell_A + \ell_B \atop m_A \quad m_B \quad -(m_A + m_B) \right] I_{\ell_A + \ell_B, -(m_A + m_B)}(R_{AB}).$$

(2.3.22)

The bracketed term in the above expression is a Wigner-3$j$ symbol[93]. The interaction tensor for higher order angular momenta can be formulated recursively, which simplifies any resultant computation[94]. The total Coulomb interaction between two charge distributions is therefore given in its entirety by

$$E_{AB} = \sum_{A \neq B} \sum_{\ell_A,\ell_B=0}^{\infty} \sum_{m_A=-\ell_A}^{\ell_A} \sum_{m_B=-\ell_B}^{\ell_B} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} d\mathbf{r}_A \, d\mathbf{r}_B \, T_{\ell_A,\ell_B,m_A,m_B} \rho(\mathbf{r}_B) R_{\ell_B,m_B}(\mathbf{r}_B) \rho(\mathbf{r}_A) R_{\ell_A,m_A}(\mathbf{r}_A).$$

(2.3.23)

Recalling the expression for multipole moments given in (2.3.11), we can rewrite this to be slightly less unwieldy

$$E_{AB} = \sum_{A \neq B} \sum_{\ell_A,\ell_B=0}^{\infty} \sum_{m_A=-\ell_A}^{\ell_A} \sum_{m_B=-\ell_B}^{\ell_B} T_{\ell_A,\ell_B,m_A,m_B} Q_{\ell_B,m_B}(\mathbf{r}_B) Q_{\ell_A,m_A}(\mathbf{r}_A).$$

(2.3.24)

Two methodological concerns are, however, apparent: (1) infinite spatial integrations, as appear in the definition of multipole moments, (2.3.11), are not feasible, and; (2) neither are infinite summations over angular momenta, as appear in the definition of the Coulomb interaction of (2.3.24). To remedy the first issue of infinite spatial integrations, some form of partitioning scheme is required to localise multipole moment expansions over a finite domain. Real space partitioning is desirable, and a Hirshfeld partitioning has been shown to yield stable multipole moment expansions[95]. In our work, we perform integrations over atomic basins, such that multipole moments are defined by[74]

$$Q_{\ell,m}^A = \int_{\Omega_A} d\mathbf{r}_A \rho(\mathbf{r}_A) r_\ell^A \sqrt{\frac{4\pi}{2\ell + 1}} Y_{\ell,m}(\theta, \phi),$$

(2.3.25)
where $Q^A_{\ell,m}$ denotes a multipole moment associated with the atomic basin, $\Omega_A$. Whilst having to explicitly compute the surfaces of the atomic basin is a computationally intensive process, we have adopted it for the reasons outlined in Section 2.2.2.

The second issue of infinite expansions over angular momenta is easily solved by approximating $E_{AB}$ with a truncated expansion to finite $\ell_A, \ell_B$. Past work has shown that, by introducing the interaction rank, $\mathcal{L} = \ell_A + \ell_B + 1$, $E_{AB}$ converges for values of $\mathcal{L} = 5[96, 97]$, i.e. up to quadrupole - quadrupole, and all rank-equivalent combinations, such as octupole - dipole and hexadecupole - monopole interactions.

A final point that requires clarification is that (2.3.24) involves an irregular spherical harmonic that is a function of $R_{AB}$, and regular spherical harmonics that are functions of $r_A$ and $r_B$. One can identify that the interaction diverges unless $||r_A - r_B|| < ||R_{AB}||$. This inequality is used in the evaluation of interactions between multipole moments; the nuclear separation must be greater than the distance over which multipole moment interactions are evaluated. We can quickly satisfy this condition by only evaluating 1-4 and higher electrostatic interactions, although this does not guarantee convergence of $E_{AB}$. In the case of divergence, one can shift the multipole moment centres to non-nuclear positions in an attempt to enforce convergence[96], and one is in fact able to obtain a value for the “optimal” shift to this end algorithmically[98]. The electrostatic interactions involving 1-2 and 1-3 interactions must be treated with some separate short-range electrostatic term, accounted for by an IQA treatment, as explained in Section 2.5.2. In classical force fields, however, these effects will be incorporated into the “bonded” terms, as has been described in Section 2.1.2.
Distributed Multipole Analysis

While not strictly relevant to the work undertaken in this thesis, Distributed Multipole Analysis (DMA)\[92\] has been the major impetus towards a rigorous treatment of molecular electrostatics. As such, we find that some discussion of DMA is not only interesting, but also warranted. In conventional ab initio treatments, one-electron wavefunctions (molecular orbitals) are expanded in a basis of atomic orbitals, centred on atomic nuclei\(^9\), \(\chi_A(\mathbf{r} - \mathbf{R}_A)\),

\[
\psi(\mathbf{r}) = \sum_A c_A \chi_A(\mathbf{r} - \mathbf{R}_A),
\]

(2.3.26)

where the summation over \(A\) denotes a summation over atomic centres and \(c_A\) is an expansion coefficient. The atomic orbitals are commonly written as Gaussian functions

\[
\chi_A(\mathbf{r} - \mathbf{R}_A) = R_{\ell,m}(\mathbf{r} - \mathbf{R}_A) \exp \left[ -\zeta ||\mathbf{r} - \mathbf{R}_A||^2 \right],
\]

(2.3.27)

where \(\zeta\) is a parameterised exponent. The electron density is subsequently expressed as

\[
\rho(\mathbf{r}) = \sum_{A,B} P_{AB} \chi_A(\mathbf{r} - \mathbf{R}_A) \chi_B(\mathbf{r} - \mathbf{R}_B),
\]

(2.3.28)

where \(P_{AB}\) is an element of the density matrix. The above expansion for the electron density is then composed of products of Gaussians,

\[
\chi_A(\mathbf{r} - \mathbf{R}_A) \chi_B(\mathbf{r} - \mathbf{R}_B) = R_{\ell,m}(\mathbf{r} - \mathbf{R}_A) \exp \left[ -\alpha ||\mathbf{r} - \mathbf{R}_A||^2 \right] \\
\times R_{\ell',m'}(\mathbf{r} - \mathbf{R}_B) \exp \left[ -\beta ||\mathbf{r} - \mathbf{R}_B||^2 \right],
\]

(2.3.29)

where \(\alpha\) and \(\beta\) are the exponents for the respective Gaussians. By virtue of the Gaussian product theorem of Boys\[100\], we are able to write the product of

---

\(^9\)There is no technical constraint requiring that the basis functions are centred on atomic nuclei; one could use, for instance, floating spherical Gaussian orbitals\[99\]. However, we proceed by use of atom-centred basis functions, as the consequent nomenclature is then consistent with what has been introduced in the previous sections.
Gaussians into a single Gaussian centred at a point, $P$, along the line $R_{AB}$

$$\exp \left[ -\alpha \|r - R_A\|^2 \right] \exp \left[ -\beta \|r - R_B\|^2 \right] =$$

$$\exp \left[ -\frac{\alpha \beta}{\alpha + \beta} \|R_A - R_B\|^2 \right] \exp \left[ -(\alpha + \beta) \|r - P\|^2 \right],$$

(2.3.30)

where $P = (\alpha R_A + \beta R_B)/(\alpha + \beta)$, and is referred to as the “overlap centre”[101].

The regular spherical harmonic terms appearing in (2.3.29) must also be moved to $P$. The movement of $R_{\ell,m}(r - R_A)$, for instance, is accomplished by use of a linear combination of regular spherical harmonics up to rank $\ell$,

$$R_{\ell,m}(r - R_A) = \sum_{k=0}^{\ell} \sum_{q=-k}^{k} \left( \frac{\ell + m}{k + q} \cdot \frac{\ell - m}{k - q} \right)^{1/2} R_{k,q}(r - P) R_{\ell-k,m-q}(P - R_A).$$

(2.3.31)

The same scheme is valid for $R_{\ell',m'}(r - R_B)$, yielding a linear combination of spherical harmonics up to rank $\ell'$. In addition, the product of spherical harmonics of rank $\ell$ and $\ell'$ is a linear combination of spherical harmonics, ranging from ranks $|\ell - \ell'|$ to $\ell + \ell'$, (2.3.29). For instance, if $\ell = \ell' = 1$, (2.3.29) would contain a product of expansions containing monopole and dipole moments. The product of monopole moment terms would lead to a monopole moment term. The product of monopole and dipole moment terms would lead to dipole ($\ell - \ell' = 1$ and $\ell + \ell' = 1$) moment terms only. Finally, the product of dipole moment terms would lead to monopole ($\ell - \ell' = 0$), dipole and quadrulpole ($\ell + \ell' = 2$) moment terms. One can then appreciate that (2.3.28), when fully expanded, becomes a rather lengthy linear combination of terms of the form $R_{k,q}(r - P) \exp \left[ -\zeta \|r - P\|^2 \right].$

Now, if we are to evaluate the “overlap multipole moments” associated with the term $R_{k,q}(r - P) \exp \left[ -\zeta \|r - P\|^2 \right]$, we invoke (2.3.11),

$$Q_{\ell,m} = \int_{\mathbb{R}^3} dr R_{k,q}(r - P) \exp \left[ -\zeta \|r - P\|^2 \right] \|r - P\|^{\ell} \sqrt{\frac{4\pi}{2\ell + 1}} Y_{\ell,m}(r - P),$$

(2.3.32)
which is zero unless \((\ell, m) = (k, q)\), since the spherical harmonics are mutually orthogonal. Then, (2.3.29) comprises overlap multipole moments of ranks ranging from 0 to \(\ell + \ell'\). To clarify this, consider the two Gaussians \(\chi_A\) and \(\chi_B\) of (2.3.29) have angular momenta \(\ell = \ell' = 0\). Then, the electron density corresponding to this product of Gaussians is represented by a monopole moment at the overlap centre. If \(\ell = 0\) and \(\ell' = 1\), then a monopole and dipole moment are required at the overlap centre.

For every Gaussian product in the expansion (2.3.28), the overlap centre is typically different to other Gaussian products in the expansion. These different overlap centres arise because \(P\) is defined from the Gaussian exponents, which are typically distinct pairs for each Gaussian product. Therefore, a number of overlap multipole moments are defined at various overlap centres along \(R_{AB}\). The situation is best illustrated by the diagram given by Stone in the original DMA publication. We reproduce this diagram in Figure 2.4, where the values of the overlap monopole and dipole moments for hydrogen fluoride (with a \([6s5p3d/4s3p]\) basis set) are plotted along the internuclear axis at the corresponding overlap centres.

Having overlap multipole moments scattered between nuclei makes the above methodology computationally cumbersome. One can, however, shift the overlap multipole moments to one of the two atomic centres. This shifting is undertaken by a formula similar to that presented in (2.3.31). Denote the distance between the overlap centre and an atomic centre by \(s\). Then, shifting an overlap multipole moment, \(Q_{k,m}\) from \(P\) to the atomic centre requires that the atomic multipole moment be written as a linear combination of multipole moments, of ranks \(\ell \geq k\). Each of these atomic multipole moments is also multiplied by a term proportional to \(s^{\ell-k}\). As such, if \(s\) is small, the higher order multipole moments are contribute negligibly to the molecular electrostatics, and the expansion can be truncated at
CHAPTER 2. EXPOSITION

Figure 2.4: The overlap monopole (upper graph) and dipole (lower graph) moments at various overlap centres along the hydrogen fluoride internuclear axis. The nuclear charges have been added to the atomic centres.

low order. However, this multiplicative term of $s^{ℓ−k}$ also constrains the magnitude of $s$, since the requirement of including higher order atomic multipole moments becomes burdensome.

A further issue is in the selection of the atomic centre to which the overlap multipole moments are shifted; since the overlap centres are distributed between two atomic centres, the choice is somewhat arbitrary. Perhaps the simplest solution is to shift the overlap multipole moments to the closest atomic centre, thus minimizing the shifting distance and limiting the problems discussed in the preceding paragraph[102]. A second solution is to require that each atomic centre receives exactly half of the overlap multipole moments, referred to as the Cumulative Atomic Multipole Moment (CAMM) method. Another popular methodology has proposed the shifting of overlap multipole moments to internuclear centroids[103].
It does not require much consideration to see that the DMA scheme is highly dependent upon the basis set used. Changing basis functions requires that the overlap centres be redefined. Increasing the accuracy of the basis set introduces a greater number of overlap multipole moments of higher angular momentum. As such, the resultant atomic multipole moments fluctuate wildly with respect to a change in basis set. Additionally, DMA struggles with diffuse basis functions; the exponents of these basis functions are very small, and so the overlap centre migrates further from the atomic centre. The instability of the atomic multipole moments has been identified as arising from the partitioning of the density in the Hilbert space of basis functions. Stone has remedied this shortcoming[104] by the introduction of a numerical quadrature scheme, which essentially transfers the partitioning from Hilbert space into real space within which the electron density is defined.

2.3.2 Implementation

Transferability

The idea of an atom type is inextricably linked with that of transferability. Whilst complex definitions of an atom type have been proposed, this area remains a source of debate and competing methods[105]. It is, however, widely regarded as a necessary measure to define electrostatic properties as pre-defined parameters for large scale molecular simulation to be truly viable.

The generation of a transferable set of multipole moments is a far more delicate operation than trying to find a corresponding set of partial charges. Whilst a monopole moment is relatively transferable, higher order multipole moments are
less so due to their increasing directional dependencies. The latter make it more
difficult to obtain a generic set of higher order multipole moments for a given atom
type.

Many molecular and group properties are tractable when attempting to demon-
strate transferability. One finds that experimental heats of formation, for example,
may be reproduced for a generic hydrocarbon \(\text{CH}_3(\text{CH}_2)_x\text{CH}_3\), by fitting to a linear
relationship, \(\Delta H_f = 2A + xB\). Here, \(A\) and \(B\) represent the respective energies of
methyl and methylene groups. Indeed, this function is equally applicable to SCF
single-point energies for equivalent systems, such that \(E = 2E(\text{CH}_3) + xE(\text{CH}_2)\) is
accurate to approximately 0.06 kcal mol\(^{-1}\)[106]. Based on this additivity of single
point energies, the concept is easily extended to imply the additivity of electron
correlation energies. Because the correlation energy is a functional of a group’s
electron density, it implies that electronic properties must additionally follow this
transferability scheme[107].

Armed with this, the demonstration that multipole moments possess[108] some
amenability to atom typing should follow. In one case study[109], a set of small
molecules composed of the functional groups present in proteins underwent DMA
at HF/3-21G level of theory, and the multipole moments of each atom were as-
sessed. Atom typing by atomic number or hybridisation state was seen to be
ineffective, but atom typing by bonding to specific functional groups proved to be
more successful. Two transferable schemes were developed: ATOM and PEPTIDE.
The former utilised the average multipole moments for specific atom types gen-
erated from the data set mentioned previously. The PEPTIDE model features a
single multipole moment expansion centre for each distinct amino acid. As such,
the local environment for each of these expansions centres is conserved for a given
amino acid. The usage of ATOM resulted in substantial deviations from the ab
initio electrostatic potential while the peptide model gave far more satisfactory results. Extending from this, a grossly distorted cyclic undecapeptide (a derivative of the immunosuppressive cyclosporine) was analysed by the above two models. The authors compared the electrostatic potentials generated by these models with one generated from DMA. Again, the peptide model exhibited lower average errors than atom.

Many years later, Mooij et al.[110] focused on the generation of an intermolecular potential function implemented in dimers and trimers of methanol. Using a fitted electrostatic term in the intermolecular potential resulted in relatively favourable results: 0.2 kcal mol$^{-1}$ and 1.6 kcal mol$^{-1}$ deviations in the dimer and trimer energies, respectively, from counterpoise-corrected MP2/6-311+G(2d,2p) calculations. This is still more impressive than similar studies on other less complex systems that have attempted to parameterise point-charge electrostatics[111]. Mooij et al.[110] also worked on methanol - water and methane - dimethylether complexes. Each of these systems was assigned a set of atom-centred multipole moments. Equally impressive results were obtained, with all interaction energies replicated to within 0.2 kcal mol$^{-1}$ of the corresponding ab initio calculations. As such, it was concluded that atom-centred multipole moment expansions are indeed transferable between the same molecules in differing environments.

There are several ways of allocating molecular electronic charge to atoms (e.g. DMA[102] or QCT partitioning[112]). Considering our group’s research interests, we focus here on QCT-based techniques. Working with the molecular energy, Bader and Beddall[82] demonstrated that:

1. The total energy of a molecule is given by a sum over the constituent atomic energies.
2. If the distribution of charge for an atom is identical in two different systems, then the atom will contribute identical amounts to the total energy in both systems.

Although these conclusions are given in terms of energy, they hold for any property density of an electronic distribution over an atomic basin. In light of this fact, it was shown by Laidig\cite{113} that multipole moments, under certain constraints, adhere to the above conclusions, and so exhibit transferability.

The property of transferable multipole moments was successfully adopted by Breneman and co-workers\cite{114}, in a method denoted Transferable Atom Equivalents (TAEs)\cite{115}. Primarily, a library was generated consisting of atom-based electron density fragments generated from a QCT decomposition of a set of molecules. DMA was subsequently performed on each of these fragments. These TAEs may then be geometrically transformed into a novel system for which the electrostatic potential is required. The fact that atomic property densities are additive in QCT implies that this recombination of TAEs is sufficient to reproduce an electrostatic potential of the system to a quasi-\textit{ab initio} level of theory. It should, however, be noted that the transferability of atomic basins is approximate, and so this method will necessarily carry a small error.

The efficacy of this methodology was subsequently demonstrated through three “peptide-capped” molecules: alanine, diglycine and triglycine. The analytical electrostatic potentials were computed on 0.002au isodensity surfaces. Equivalent electrostatic potentials were also generated from TAE-reconstructed systems and Gasteiger point-charges for the extended (open) and \(\alpha\)-conformations. The TAE multipole analysis (TAE-MA) reproduced the features of the electrostatic potential generated at \textit{ab initio} level much better than the Gasteiger point-charges
did. A point-charge electrostatic model is unable to accurately predict extremes in electronic features.

In later work carried out in this group, all 20 naturally occurring amino acids and their constituent molecular fragments were rigorously assessed using QCT[116]. A set of 760 distinct topological atoms were generated and cluster analysis identified a set of 42 atom types in total (21 for carbon, 7 for hydrogen, 6 for oxygen, 2 for nitrogen and 6 for sulphur). The trivially assigned atom types implemented in AMBER were either too fine-grained (e.g. too many atom types for N) or too coarse-grained (e.g. carbon atom types not diverse enough). Later, an extensive investigation[117] was carried out for atom typing by atomic electrostatic potential rather than atomic multipole moments as in the previous study. A retinal-lysine system was considered, a prominent feature in the mechanism of bacteriorhodopsin. This study focused on the aldehyde and terminal amino groups of retinal and lysine, respectively. The electrostatic potentials generated by these groups occurring in the full system were compared with those of smaller derivatives of the system. The electrostatic potential of lysine surrounding the terminal amino group was relatively conserved for all derivatives in which two (methylenic) carbon atoms were maintained along the amino acid sidechain. However, the aldehyde group of retinal was more responsive to more distant environmental effects.

This work has recently been further developed, where the concept of a “horizon sphere” is proposed[118]. This sphere contains all the atoms that a given atom, at the sphere’s centre, “sees” in terms of their polarisation of the electron density on the central atom. An α-helical segment of the protein crambin was studied. The electrostatic energy was probed by consideration of the multipole moment expansion (up to rank $L = 4$) centred at a $C_\alpha$. A new set of multipole moments for $C_\alpha$ was calculated for each structure dictated by the growing horizon sphere.
The interaction energy between C$_\alpha$ and a set of probe atoms was evaluated, leading to the conclusion that formal convergence of this interaction energy is attained at a horizon sphere radius of roughly 12Å. More work is underway to scrutinise the validity and generality of this conclusion.

Crystallographers who strive for the generation of transferable atomic electron densities[119], find qualms with the reconstruction of molecular electron densities from these QCT-derived atomic densities. This is due to the mismatch in interatomic surface topologies between transferred atoms. As such, they believe that it becomes very difficult to generate a continuous electron density from these atomic fragments. Work has therefore been directed towards the generation of pseudoatom databanks that may be utilised to reconstruct experimental electron densities from previously elucidated structures. From this approach follows a natural output in the form of atomic multipole moments. It is important to point out that the aforementioned mismatch can be countered by accepting that the interaction energy between atoms is what ultimately matters, not the perfect construction of gapless sequences of topological atoms. With this premise in mind we have shown that the machine learning method kriging captures[120, 121], within reasonable energy error bars, the way a QCT atom changes its shape in response to a change in the positions of the surrounding atoms.

Jelsch et al.[122], for example, demonstrated the capacity of transferring experimental density parameters for small peptides, based upon the Hansen-Coppens formalism[123], and subsequently built a databank of pseudoatoms. The refinement of high resolution X-ray crystallographic data by referral to this databank has been demonstrated[124]. A more computationally-orientated route has been developed in parallel to the one above[119], whereby the experimental density parameters for a set of pseudoatoms were derived from *ab initio* electron densities.
of tripeptides. This method showed an enhanced amenability to transferability compared to its experimentally-derived counterpart.

More recently this pseudoatom database has been built upon[125]. Atom types were defined by grouping atoms with the same connectivity and bonding partners, while the atom type properties were defined by averaging over all constituent “training set” pseudoatoms. Single point calculations were initially carried out on a test set of amino acid derivatives at B3LYP/6-31G** level. The geometry of each species was taken directly from the Cambridge Structural Database(CSD)[126]. Multipole moment expansions for non-hydrogen atoms were truncated at ranks $L = 4$ and $L = 2$ for the hydrogens. These multipole moments were subsequently averaged and standard deviations defined for the dataset. In terms of performance, the databank model appears to give a slightly more pronounced electrostatic potential surrounding oxygen atoms in carboxylate and hydroxyl groups of serine, leucine and glutamine, compared to the more extensive ab initio calculations. Further work showed that the databank does relatively well in the prediction of most atomic multipole moments. Exceptions take the form of higher order multipole moments, most particularly for oxygens and nitrogens. Finally, we mention that somewhat poorer results are obtained when considering total intermolecular electrostatic energies in dimers. The errors are of the same magnitude as those obtained from AMBER99, CHARMM27 and MM3. The authors ascribe these results to the implementation of a Buckingham-type approximation, whereby non-overlapping electron densities are assumed. This results in the underestimation of short-range interactions. The authors report much-reduced discrepancies in these energies by use of their own refined method, which accounts for this discrepancy[127].

However, in spite of the issues of the reconstruction of crystal structures by use
of QCT, the technique remains amenable to the elucidation of electrostatic properties. Woińska and Dominiak[128] have given a thorough elaboration on the transferability of atomic multipole moments based on various density partitions, most notably directly comparing the Hansen-Coppens formalism to both QCT and Hirshfeld partitioning. In their study, multipole moments (up to $L = 4$) were assigned to each atom in a set of biomolecular constructs, ranging from single amino acids to tripeptides. Atom types were subsequently defined from this molecule set based on criteria resembling those used by a similar study[129]. By averaging the multipole moments for given atom types in differing chemical environments, a standard deviation from this average value was obtained. A lower standard deviation is indicative of a high degree of transferability, and vice versa. A QCT analysis of \textit{ab initio} wavefunctions results in highly non-transferable lower-order multipole moments ($L = 0, 1, 2$). Secondly, for the higher-order Hansen-Coppens multipole moments ($L = 3, 4$) are particularly unstable. The atom types found to be non-transferable from the QCT analysis are generally carbons connected to two electronegative atoms (oxygen or nitrogen), or members of aromatic systems. The decline in the level of transferability for higher-order multipole moments for the Hansen-Coppens method is relatively widespread throughout atom types, most prominently carbon and nitrogen. It is, however, strange to note that for both of these points, the poor level of transferability for QCT and Hansen-Coppens pseudoatoms is constrained to specific atom types; for the rest, these techniques generally give rise to the most transferable multipole moments.

Whilst the lower-order QCT multipole moments are largely non-transferable, they tend to be far more stable when derived from crystallographic data. In spite of this, they are still the least transferable multipole moments in the set, with both Hirshfeld partitioning and the Hansen-Coppens pseudoatom formalism yielding somewhat more stable multipole moments. The authors conclude that the most
transferable multipole moments result from Hirshfeld partitioning. QCT discretely partitions electron density into distinct basins and so is vulnerable to numerical issues when undertaking mathematical operations such as integration over the basin. The Hansen-Coppens formalism, on the other hand, suffers from problems with localisation: distant electron density may be incorrectly assigned to a given nucleus. However, an exhaustive study of standard deviations from average multipole moments for given partitioning methods does little to confirm the dominance of one scheme over another. Transferability matters little if the multipole moments in question are incorrectly defined; their subsequent variances over a dataset are inconsequential. In fact, the difficulty in assigning transferable multipole moments to given atom types may equally be indicative of poor atom type definition, or the sheer inability to define a transferable atom type in terms of multipole moments with any great stability. We make a final note in that the atom types defined in this work have been tailored for pseudoatom usage[130, 131], and so may not be useable with a discrete partitioning scheme (QCT), relative to the so-called ‘fuzzy’ decompositions (Hirshfeld and Hansen-Coppens). In fact, this is concomitant with an analysis of dimer energies obtained from these three techniques[132]. When one uses multipole moments obtained by a QCT decomposition as opposed to a Hirshfeld partitioning, the electrostatic energy obtained more closely resembles that obtained from a Morokuma-Ziegler energy decomposition scheme, by as little as 10%.

Simulation

A recent tour de force regarding the feasibility of biomolecular simulation have seen the computation of time trajectories of systems such as the ribosome[133]. However, this study implemented techniques not optimised for the output of par-
particularly accurate results, using a highly parallelisable CHARM++ interface in conjunction with the AMBER force field and the NAMD molecular dynamics package[134], i.e. a partial charge approximation.

Biomolecular simulation requires the implementation of periodic boundary conditions to accurately model the environment in which a system resides. Moreover, the electrostatic energy of a system is slowly convergent. Many solutions to this problem have been proposed over the years[135, 48]. It should be noted that the interaction involving ‘higher order’ multipole moments (\( \mathcal{L} \geq 1 \)) is more short-range than that between monopole moments. As such, the problem of slowly convergent long-range interactions is shared by both isotropic point-charges and multipolar electrostatics because the latter encompass point-charges.

We wish to raise the issue of the conformational dependence of electronic properties. In reality, this problem is not unique to higher order multipole moments. Conventional force fields, which employ partial charges, choose to reside in a pseudo-reality of an invariable electrostatic representation, and so rarely encounter conformational dependencies. It is rather more difficult to simply ignore the obvious reality of molecules as flexible entities when using multipole moments, and has been emphasised in analyses using both DMA and CAMM algorithms[136, 137]. Use of the electrostatic properties for one conformation correctly reproduces its corresponding electrostatic potential. However, use of this parameterisation in an alternative conformation results in highly unfavourable energies. It should be noted that this insufficiency is equally prominent when using a conserved set of partial charges between conformers. As such, since the issue of flexible molecules is a computational complexity that pervades all electrostatic approximations, it would be unfair to regard this as an additional burden specifically for multipole moments. Instead, it is a hurdle that both techniques are required to overcome in
2.3. **MULTIPOLE MOMENT ELECTROSTATICS**

enhancing the accuracy of simulation.

Evaluating multipole moments as a function of a conformational parameter is appealing. For example, the multipole moments for both atoms in carbon monoxide may be described analytically as a function of the interatomic distance in the molecule[138]. However, scaling this idea up to systems with far more conformational degrees of freedom, such as an amino acid, is an appreciably more difficult task.

An “analytical compromise” has been proposed in the past[30], whereby the multipole moments of an atom in ethanol, glycine and acrolein are represented by a Fourier series truncated at third order, whose free variables correspond to the dihedral angles of the molecule. Instead of analytical methods, machine learning methods can be used to interpolate between a set of multipole moments defined for different molecular conformations[121]. As such, one may then predict the multipole moments of an arbitrary conformation that is not present within the initial training set, which corresponds to a true external validation.

Alternative methods have been developed but the literature on these techniques appears to be relatively sparse. For example, it has been proposed[139] that one may average the atomic multipole moments over all conformers that are sampled during a simulation. This has been done for alanine and glycine by shifting the higher order ($\mathcal{L} \geq 1$) atomic multipole moment expansions to a smaller number of expansion sites distributed throughout the molecule. An additional method, previously tested for energy minimisations of crystal structures, revolves around periodically recalculating the atomic multipole moments for the molecule[140]. This proved to give substantially better results than the implementation of then-current methodologies, particularly for systems whose structures are dictated by strong hydrogen bonding.
Forces (and torques) must be calculated for the molecular translational and rotational degrees of freedom to be sampled during the course of a simulation. These may be formulated directly by first and second derivatives of interactions energies, by use of translational and rotational differential operators. If one considers a molecule as a rigid body, the individual atomic multipole moments of the molecule are invariant relative to their stationary local axis systems. As such, the derivative of the interaction energy between two molecular species is satisfied by the derivative of the interaction tensor only. This has been demonstrated in the spherical tensor formalisms\cite{141, 142} and its application (see e.g. \cite{143}). A simulation package that allows for this rigid-body approximation in conjunction with multipolar electrostatics exists\cite{144} and is called DL\_MULTI.

The invariance of atomic multipole moments with respect to a local axis system no longer holds when abandoning the rigid-body approximation in favour of a realistic flexible-body protocol. As such, differentiation of the interaction energy function requires the derivatives of multipole moments in addition to the interaction tensor. Whilst this requires\cite{145} a more involved series of calculations, it is still an attainable requirement. Somewhat more problematic is the fact that the local axis systems in which the atomic multipole moments are referenced evolve over the course of a simulation. Due to the flexible nature of the molecule, neighbouring atomic positions that make up a local axis system change with respect to time. This results in the subsequent net rotation of the atomic local frames. In the context of QCT and the machine learning method kriging, analytical forces can be calculated for “flexible, multipolar atoms”, although this is not trivial\cite{146}.

Literature quotations of CPU time differences between a molecular dynamics simulation using point-charges versus multipole moments (for a given number of nanoseconds, of a given biomolecule with a given number of water molecules sur-
2.3. MULTIPOLE MOMENT ELECTROSTATICS

rounding it), are virtually non-existent. However, Sagui et al.[147] reported a representative ratio of 8.5 in favour of point-charge electrostatics (as implemented in AMBER 7) when most of the calculation is moved to the reciprocal space (via the PME method) with multipolar interactions up to hexadecapole-hexadecapole being included. The only way a point-charge model can ever match the accuracy of a nucleus-centred multipolar model is via the introduction of extra off-nuclear point charges. What is rarely stated is that these additional charges create an enormous computational overhead in a typical system of tens of thousands of atoms because charge-charge interactions are longer range \((1/r\)-dependence\) than any interaction between multipole moments.

AMOEBA

Arguably one of the most successful next-generation force fields is AMOEBA (Atomic Multipole Optimised Energetics for Biomolecular Applications)[26]. AMOEBA has been proven effective in a variety of biomolecular simulations, ranging from solvated ion systems[148, 149, 150] to organic molecules[151] and peptides[152, 153, 154]. The electrostatic energy component of the force field is broken down into two terms. The first term is concerned with permanent atomic multipole moments (expansions truncated at \(L = 2\)), whilst the second term corresponds to induced multipole moments as a result of polarisation effects. The permanent atomic multipole moments are generated by DMA of a set of small molecules such that atom types may be defined. When implemented during a simulation, these atomic multipole moments may be rotated into various fixed local axis systems within the molecule. AMOEBA proves to be competitive, even with ab initio level calculations. All levels of theory tested perform in a uniform manner: the average \(\Delta E\) values across all conformations at the MP2/TQ, \(\omega\)B97/LP, B3LYP/Q and
AMOEBA levels of theory are 3.73, 3.15, 3.64 and 3.30 kcal mol$^{-1}$, respectively. These are impressive values, but one must remain aware that they correspond to total energies as opposed to those arising specifically from the electrostatic component of the force field.

A true demonstration of the benefits corresponding to multipole moments arises from a direct comparison between AMOEBA and the various force fields that employ point-charges. Kaminsky and Jensen\cite{155}, for example, sampled the number of energetic minima of glycine, alanine, serine and cysteine one recovers at MP2 level. The number of minima and their relative energies were subsequently compared to those recovered by use of AMOEBA and four other point-charge force fields. The results for serine and cysteine are outlined in Table 2.1, where the \textit{ab initio} data suggests 39 and 47 minima, respectively. We see that AMOEBA consistently outperforms the large majority of point-charge force fields in terms of the mean absolute deviation (MAD) of energies relative to the MP2 results. AMOEBA additionally outperforms all other force fields in terms of the number of the minima it predicts for each amino acid. Note that the latter result gives rise to an artificially large MAD value relative to the other force fields. Considering the aforementioned more favourable MAD corresponding to AMOEBA, this only emphasises the predictive capacity of AMOEBA.
2.3. MULTIPOLE MOMENT ELECTROSTATICS

<table>
<thead>
<tr>
<th></th>
<th>Serine</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MM3</td>
<td>OPLS_2005</td>
<td>AMBER99</td>
<td>CHARMM27</td>
<td>AMOEBA</td>
</tr>
<tr>
<td>Number of Correct Minima</td>
<td>15</td>
<td>23</td>
<td>21</td>
<td>19</td>
<td>34</td>
</tr>
<tr>
<td>Number of Erroneous Minima</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Percentage Erroneous</td>
<td>20.0</td>
<td>13.0</td>
<td>33.3</td>
<td>10.5</td>
<td>20.6</td>
</tr>
<tr>
<td>MAD (kJ mol(^{-1}))</td>
<td>14.0</td>
<td>4.1</td>
<td>10.9</td>
<td>11.1</td>
<td>4.2</td>
</tr>
</tbody>
</table>

|                    | Cysteine              |                       |                       |                     |                     |
|                    | MM3                   | OPLS\_2005            | AMBER99               | CHARMM27            | AMOEBA              |
| Number of Correct Minima | 21                    | 25                    | 29                    | 21                  | 44                  |
| Number of Erroneous Minima | 3                    | 3                     | 5                     | 5                   | 1                   |
| Percentage Erroneous        | 14.3                  | 12.0                  | 17.2                  | 23.8                | 2.3                 |
| MAD (kJ mol\(^{-1}\))     | 13.9                  | 4.6                   | 5.4                   | 6.6                 | 3.1                 |

Table 2.1: The number of geometric minima predicted by a variety of force fields for serine and cysteine. Also given are the number of minima that the various force fields predicted but that were not represented in the set of minima generated by ab initio calculations, and the mean average deviations (MAD) for the molecular energies at each geometry.

Another study that has directly compared AMOEBA to a variety of other conventional force fields (AMBER, MM2, MM3, MMFF and OPLS) is that of Rasmussen et al.[156], where relative conformational energies were approximated. A set of minima were generated for several molecules, each with intermediary electrostatic properties ranging from entirely non-polar to zwitterionic. The ability of the various force fields to predict relative energies of the minima was probed, in addition to three separate AMOEBA parameterisation schemes, differing in atoms typing or level of theory. All force fields performed extremely well for the non-polar molecules, largely due to the minor electrostatic contribution to the conformational energy of non-polar molecules. As such, the level at which electrostatics were calculated is essentially irrelevant. However, as the molecular species become more polar in nature, the point-charge force fields begin to display their erroneous nature relative to the AMOEBA parameterisations, which demonstrate...
a more uniform predictive capacity. The zwitterionic species were modelled well by several of the point-charge force fields. This can be explained by the fact that full charges are properly represented by a spherical electrostatic potential as the charge is highly localised. Thus, point-charge implementations of electrostatics can model such a case with relative ease.

The accurate reproduction of the properties of water has long plagued simulation. Being able to account for explicit binding of water molecules, in addition to accurate modelling of bulk properties such as the dielectric constant are necessary features if one wishes to accurately simulate solvated systems. AMOEBA attempts to account for the lack of a universally acceptable water model by specifically parameterising the water molecule[157, 158]. Much like the generic AMOEBA force field, atomic multipole moment expansions up to $\mathcal{L} = 2$ are generated using DMA. Polarisation is accounted for via induced atomic dipoles, and van der Waals interactions are modelled by a buffered 14-7 LJ potential. To the credit of the developers, this model is continually improved upon and reparameterised. Most recently[159], atomic multipole moments were generated for a water model at MP2 level with various basis sets in order to probe the reproduction of hydration free energies for a set of small molecules. Whilst the aug-cc-pVTZ basis set was found to give the best results, 6-311++G(2d,2p) gave a comparable accuracy at a much lower computational cost, and so is recommended for larger simulations.

A direct comparison between an AMOEBA water model parameterised at MP2/6-311++G(2d,2p) level and a widely used point-charge water model reveals the benefits of atomic multipole moment-parameterised electrostatics. A popular choice for explicit solvation is the TIP3P model, which assigns a single point-charge to each atomic centre and implements a 12-6 LJ function. Since the LJ functions differ between the two models, a “TIP3P-like” model was generated, which used the
AMOEBA water model, but removed all multipole moments (static and induced), replacing them with point-charges. TIP3P and TIP3P-like models were shown to be equivalent by comparison of RDFs and bulk simulation properties. Deviation of the computed hydration free energies from experimental benchmarks for a set of small molecules are given in Table 2.2 for both AMOEBA and TIP3P-like models. AMOEBA outperforms the TIP3P-like model quite spectacularly, where the RMSD of the AMOEBA values is roughly three times smaller than the RMSD of the TIP3P-like values.

<table>
<thead>
<tr>
<th></th>
<th>Ethylbenzene</th>
<th>p-cresol</th>
<th>Isopropanol</th>
<th>Imidazole</th>
<th>Acetic Acid</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMOEBA</td>
<td>-0.73</td>
<td>-7.27</td>
<td>-5.58</td>
<td>-10.11</td>
<td>-5.69</td>
<td>0.68</td>
</tr>
<tr>
<td>TIP3P-like</td>
<td>-0.89</td>
<td>-10.72</td>
<td>-5.29</td>
<td>-10.87</td>
<td>-7.46</td>
<td>1.96</td>
</tr>
<tr>
<td>Experiment</td>
<td>-0.70</td>
<td>-6.10</td>
<td>-4.70</td>
<td>-9.60</td>
<td>-6.70</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2.2: Free energies of hydration for a variety of molecules predicted by use of AMOEBA and TIP3P-like water potentials. Corresponding experimental values are also given. Energies are in kcal mol\(^{-1}\).

### 2.4 Machine Learning

#### 2.4.1 Overview

We have explored the means one may employ to obtain a number of \textit{ab initio} atomic properties that can be used for force field construction. An area that is entirely overlooked by the classical force field community is that these quantities are not static with respect to the internal degrees of freedom of a system\textsuperscript{10}. The

\textsuperscript{10}Indeed, these quantities are not static with respect to the external degrees of freedom if placed in some anisotropic medium, but a discussion on external degrees of freedom is not undertaken.
internal degrees of freedom of a molecular system are in a constant state of motion owing to inter- and intramolecular forces. When a molecule reconfigures itself, the associated molecular electron density, and consequently the atomic partitioning, instantaneously adjust. Since atomic properties are obtained from integrals over atomic domains, atomic properties must similarly alter with respect to a shift in molecular conformation.

The question then is how one can obtain a conformationally-dependent force field, without having to resort to obtaining analytical PESs (which limit the system size to a handful of atoms, e.g. [160]). Perhaps the simplest way to do so is to construct a database containing conformations and associated atomic/ molecular properties. Of course, if one were to construct an entire biomolecular force field in this way, the memory required to store these databases for a given accuracy would be overwhelming. However, such databases have been constructed for small molecular systems, and used for geometry optimisation[161] and structure refinement[162] with some degree of success.

An alternative method would be to use some analytical function of a number of molecular degrees of freedom, along which an atomic/ molecular property evolves. Fluctuating multipole models ascribe a multipole moment, or set of multipole moments, to an atom, and allow these to vary in response to some external field[139]. This external field is an average description of perturbations resulting from the atomic environment, and so is computationally efficient. The fluctuating charge model[163] is one such popular model, where each atom is assigned a fluctuating monopole moment. Further work has been done to ascribe higher order fluctuating multipole moments to each atom[164, 165], but this is not a widely-used methodology. The issue with these models is the idealised notion of a molecular

here for the sake of simplicity.
perturbation being well-modelled by a single external field vector. In addition, the fluctuating multipole moments are typically constrained to a harmonic evolution, whereas in reality multipole moments possess a far more complex conformational dependence[166].

An attractive solution is some form of interpolative scheme, whereby atomic properties for given molecular conformations are explicitly computed. Such points are referred to as “training points”. Collectively, the set of training points is referred to as a “training set”. An interpolative scheme can then be employed to predict atomic properties at “untrained” points on the molecular potential energy surface.

Machine learning is a field that is suited to such a task. Machine learning is a broad field of mathematics, dealing with the construction of algorithms that can “learn” mappings from a given input to a response by use of a training set. Machine learning has gained some popularity in being used to predict atomic/ molecular energies for given conformations. We make a brief review of the more popular implementations, but are necessarily brief, and omit a number of methodologies that are deserving of attention if space wasn’t a constraining factor (see, for example, [167]).

Perhaps the most popular machine learning techniques are the Artificial Neural Networks (ANNs), a set of massively parallel interconnected nodes that relate some input vector to a response function[168], the inspiration of which is taken from the functioning of biological neuronal circuits. A number of ANNs have been developed over the years[169, 170]. To date, only feed-forward neural networks, also referred to as multilayer perceptrons, have been used in the construction of a force field at the time of writing[9]. ANNs have been applied to a wide range of problems, such as the calculation of energies of systems in the solid state[171, 172], surface adsorption, water[173] and interatomic potentials[174].
Past work within our group\cite{175, 176} has employed ANNs for the prediction of atomic multipole moments. Some success was attained, but ANNs notoriously deteriorate in quality in high-dimensional spaces\cite{177}. This is problematic since the number of conformational degrees of freedom required for the accurate modelling of conformational dependence exceeds the dimensionality for which ANNs are viable. In addition, ANNs suffer from the phenomenon of “overtraining”, where the number of free parameters used in the construction of the ANN makes its predictive capacity deteriorate\cite{178}.

Gaussian Process Regression\cite{179} is a second popular machine learning technique. In this framework, it is assumed that the best estimate for an atomic property in a given molecular conformation can be expanded in Gaussian basis functions\cite{180}. Gaussian Approximation Potentials\cite{16, 181} have been used to this end. Complex descriptors, such as symmetry functions\cite{182, 183}, are employed in the construction of a conformational basis set that is invariant with respect to global translations, rotations, permutations of atoms, and a number of other occurrences, which render these potentials cumbersome\cite{184}.

In the remainder of this section, we describe an alternative Gaussian Process Regression (GPR) machine learning technique, referred to as kriging. Kriging finds its origins in the field of geostatistics\cite{185}, where mineral concentrations were predicted at arbitrary sites. Kriging has gone on to be used in a number of other disciplines, such as real estate appraisal\cite{186}, the design of integrated circuits\cite{187} and environmental science\cite{188}. We have found kriging to be an ideal choice for the construction of a conformationally dependent force field since it suffers from neither overtraining or deterioration in high-dimensional spaces, both of which plague the implementation of ANNs.
2.4. MACHINE LEARNING

2.4.2 Kriging

In the following, we adopt the convention of enumerating elements of a set with a superscript, and components of a vector or matrix with a subscript. We define a set of \( n \) training points, \( \mathbf{X} := \{ \mathbf{x}^1, \ldots, \mathbf{x}^n \} \), where each training point \( \mathbf{x} \) is a \( d \)-dimensional vector, \( \begin{bmatrix} x_1 & \ldots & x_d \end{bmatrix}^\top \). For each training point, we can evaluate a response function, \( y = f(\mathbf{x}) \), and collect these responses into an \( n \)-dimensional vector, \( \mathbf{y} = \begin{bmatrix} y^1 & \ldots & y^n \end{bmatrix}^\top \) for convenience. Kriging is a machine learning method that allows one to predict the value of the response at untrained points, \( f_{\text{pred}}(\mathbf{x}') \), where \( \mathbf{x}' \notin \mathbf{X} \), as a function of the elements of \( \mathbf{X} \).

Kriging is strictly interpolative, which means that the prediction of the response function at trained points, i.e. \( f_{\text{pred}}(\mathbf{x}), \mathbf{x} \in \mathbf{X} \), is exact. At untrained points, it can be shown that kriging is the best linear unbiased predictor with respect to the minimisation of the sum of squared residuals[189]. The term unbiased means that the expectation value of predictions, \( \langle f_{\text{pred}}(\mathbf{x}') \rangle_{\mathbf{x}' \notin \mathbf{X}} \), is equal to the expectation value of the response function, i.e. the first moment, or mean, of the distribution \( f(\mathbf{x}) \). This definition of an unbiased predictor is used as a constraint placed on the kriging estimator, and features as a Lagrange multiplier in the general derivation for the kriging estimator[189].

The key observation of kriging is that the responses, (i.e. the response function evaluations), at two distinct points, \( f(\mathbf{x}^i) \) and \( f(\mathbf{x}^j) \), are similar if \( \mathbf{x}^i \) and \( \mathbf{x}^j \) are close together in space, if the response function is sufficiently well-conditioned. We can summarise the correlation between training points in an \( n \times n \) symmetric, positive semidefinite matrix, \( \mathbf{R} \), whose \( i^\text{th} \) element corresponds to the correlation between points \( \mathbf{x}^i \) and \( \mathbf{x}^j \). The correlation function, or kernel, we use for this work
is a power correlation function[179], of the form

$$R_{ij} = \exp \left[ -\sum_{h=1}^{d} \theta_h |x_i^h - x_j^h|^{p_h} \right], \quad (2.4.1)$$

where \( h \) is an index running over the \( d \) degrees of freedom of points \( x^i \) and \( x^j \), and \( \theta_h \) and \( p_h \) are the so-called hyperparameters associated with these degrees of freedom, and can be accumulated into the \( d \)-dimensional vectors \( \theta \) and \( p \), respectively.

Rasmussen and Williams[179] have defined the \( \theta_h \) hyperparameter as being equal to \( 1/2r_h \), \( r_h \) being the “characteristic distance” of that particular feature. In this way, \( r_h \) is a measure of the extent of the correlation among different points “along” that feature. Rasmussen and Williams showed the hyperparameters to be bounded as follows

$$0 \leq \theta_h \quad \forall h = 1, \ldots, d$$
$$0 \leq p_k \leq 2 \quad \forall h = 1, \ldots, d \quad (2.4.2)$$

Analysis of (2.4.1) shows that, in the case where \( x^i \) and \( x^j \) coincide, \( R_{ij} = 1 \), whilst in the limit where the two points are infinitely removed from one another, \( R_{ij} \to 0 \). In addition, the hyperparameters \( \{ \theta, p \} \) possess physical significance. When \( \theta_h \) is small, the influence of the \( h^{th} \) degree of freedom is enhanced, leading to an increase in correlation between two points. Similarly, if \( \theta_h \) is large, the influence of the \( h^{th} \) degree of freedom is diminished, leading to a decrease in correlation between the two points. As such, we can conclude that \( \theta_h \) is a measure for the “importance” of a degree of freedom on the correlation between points. The interpretation of the components of \( p \) is a little more abstract, but are typically referred to as the “smoothness” of the process. We see that when \( p_h = 2 \), the correlation between points takes a Gaussian form, so that as two points become more distant, their correlation dies off as a Gaussian. If \( p_h = 1 \), the correlation between two points
dies off as a simple exponential. Owing to the boundary conditions in (2.4.2), \( p_h \) can take on real values, the physical significance of which is not entirely clear.

To obtain optimal predictions by means of kriging, we require that the log-likelihood function, \( L(\theta, p, \sigma^2, \mu) \) is maximised[190]. The log-likelihood function has the form

\[
L(\theta, p, \sigma^2, \mu) = -\frac{n}{2} \ln(\sigma^2) - \frac{1}{2} \ln(|R|) - \frac{(y - 1\mu)^\top R^{-1}(y - 1\mu)}{2\sigma^2},
\]

where \( \mathbf{1} \) is a column vector of ones, \( \mu \) and \( \sigma^2 \) the first and second moments, respectively, of the response function and \( |R| \) the conventional notation for a matrix determinant. We can estimate \( \mu \) and \( \sigma^2 \) from the available data by setting the derivative of (2.4.3) with respect to \( \mu \) and \( \sigma^2 \) equal to zero. Analytical forms for the optimal values of \( \mu \) and \( \sigma^2 \), which we denote \( \hat{\mu} \) and \( \hat{\sigma}^2 \), respectively, and are found to be[191]

\[
\hat{\mu} = \frac{\mathbf{1}^\top R^{-1}y}{\mathbf{1}^\top R^{-1}\mathbf{1}} \quad \hat{\sigma}^2 = \frac{(y - 1\mu)^\top R^{-1}(y - 1\mu)}{n}.
\]

Substitution of (2.4.4) into (2.4.3) yields the concentrated log-likelihood function, \( \hat{L}(\theta, p) \), where explicit dependence on \( \hat{\mu} \) and \( \hat{\sigma}^2 \) is implied through dependence on \( \theta \) and \( p \). The concentrated log-likelihood function has the form

\[
\hat{L}(\theta, p) = -\frac{n}{2} \ln(\hat{\sigma}^2) - \frac{1}{2} \ln(|R|).
\]

We then see that maximisation of the concentrated log-likelihood function requires maximisation with respect to \( \{\theta, p\} \). A number of strategies are available for this purpose, both analytical and numerical, yielding optimal estimations for \( \hat{\theta} \) and \( \hat{p} \). One such optimisation strategy is discussed in Section 2.4.3. We do, however, point out that any optimisation procedure requires iterative calculations of \( |R| \), which is not trivial computationally when \( n \) is large. As such, maximisation algorithms requiring as few iterations as possible will naturally be desirable in order to circumvent the large computational costs associated with evaluating the matrix determinant.
Now, in order that predictions can be made at some arbitrary untrained point, \( x^* \), we compute an \( n \)-dimensional vector, \( r = \begin{bmatrix} R_{1*} & \ldots & R_{n*} \end{bmatrix}^\top \), the components of which, \( R_{i*} \), correspond to the correlation between \( x^* \) and \( x_i \), as given by (2.4.1).

Defining an augmented \((n + 1) \times (n + 1)\) correlation matrix, \( \tilde{R} \), as

\[
\tilde{R} = \begin{bmatrix} R & r \\ r^\top & 1 \end{bmatrix},
\]

we can rewrite the log-likelihood function, (2.4.3), along with our optimal \( \hat{\mu} \) and \( \hat{\sigma}^2 \), (2.4.4), as

\[
L(\theta, p, \hat{\sigma}^2, \hat{\mu}) = -\frac{n}{2} \log(\hat{\sigma}^2) - \frac{1}{2} \log(|\tilde{R}|) - \frac{(\tilde{y} - 1\hat{\mu})^\top \tilde{R}^{-1}(\tilde{y} - 1\hat{\mu})}{2\hat{\sigma}^2},
\]

where we have introduced the augmented response vector, \( \tilde{y} = \begin{bmatrix} y^* & \ldots & y^n \end{bmatrix}^\top \),

where \( y^* \) is the response at the point \( x^* \), i.e. the value we wish to predict. Only the final term in (2.4.7) is dependent upon our augmentation, and so writing this out explicitly

\[
\frac{(\tilde{y} - 1\hat{\mu})^\top \tilde{R}^{-1}(\tilde{y} - 1\hat{\mu})}{2\hat{\sigma}^2} = \begin{bmatrix} y - 1\hat{\mu} \\ y^* - \hat{\mu} \end{bmatrix}^\top \begin{bmatrix} R & r \\ r^\top & 1 \end{bmatrix}^{-1} \begin{bmatrix} y - 1\hat{\mu} \\ y^* - \hat{\mu} \end{bmatrix}.
\]

\( \tilde{R}^{-1} \) can be given explicitly by use of the partitioned inverse formula from Thiele[192], which subsequently allows us to write the augmented log-likelihood function as

\[
\tilde{L}(\theta, p, \hat{\sigma}^2, \hat{\mu}) = -\frac{n}{2} \log(\hat{\sigma}^2) - \frac{1}{2} \log(|R|) - \frac{(y^* - \hat{\mu})^2}{2\hat{\sigma}^2(1 - r^\top R^{-1}r)} + \frac{r^\top R^{-1}(y - 1\hat{\mu})}{\hat{\sigma}^2(1 - r^\top R^{-1}r)} (y^* - \hat{\mu}).
\]

Finding the derivative of this quantity with respect to \( y^* \) and equating the derivative to zero results in the maximisation of the augmented log-likelihood function,

\[
-\frac{(y^* - \hat{\mu})}{2\hat{\sigma}^2(1 - r^\top R^{-1}r)} + \frac{r^\top R^{-1}(y - 1\hat{\mu})}{\hat{\sigma}^2(1 - r^\top R^{-1}r)} = 0,
\]
and solving for \( y^* \) yields
\[
\hat{y}^* = \hat{\mu} + r^\top R^{-1} (y - \hat{1}\hat{\mu}),
\]
(2.4.11)
the **kriging estimator**. Naturally, (2.4.11) can be used for any arbitrary point \( x^* \)[193, 194]. We note that making a prediction requires the evaluation of \( R^{-1} \), which can be a computationally demanding process if \( n \) is large.

We conclude by introducing two informative metrics to assess the quality of a kriging model. The first is the condition number[195] of the correlation matrix, \( \kappa(R) \), given by
\[
\kappa(R) = \begin{cases}
||R|| \quad & \text{if } |R| \neq 0 \\
\infty & \text{if } |R| = 0
\end{cases},
\]
(2.4.12)
where \( ||R|| \) corresponds to the operator norm of \( R \). If \( \kappa(R) = 1 \), then the approximation to a solution set introduces no larger errors than those present within the input set. An ill-conditioned matrix, i.e. one whose condition number is large, typically leads to numerical instabilities, and is naturally problematic. It has been found that a small spacing between training points leads to an increase in the condition number of \( R \), which results in a deterioration in the predictive capacity of the machine learning model[196]. We address this issue in significant detail in Chapter 5.

The second useful metric is the mean squared error (MSE) at a prediction point, \( s^2(x^*) \), given by
\[
s^2(x^*) = \hat{\sigma}^2 \left[ 1 - r^\top R^{-1} r + \frac{(1 - 1^\top R^{-1} r)^2}{1^\top R^{-1} 1} \right].
\]
(2.4.13)
Since kriging is interpolative, \( s^2(x^*) \) is equal to zero if \( x^* \) is a training point. When \( x^* \) is not a training point, the mean squared error reflects the level of uncertainty in the prediction of \( y^* \), i.e. the confidence we have in our prediction. The mean
CHAPTER 2. EXPOSITION

squared error is brought down by adding more training points in the vicinity of a prediction point, and so can be used to ascertain whether a greater training density is required in a given region.

2.4.3 Particle Swarm Optimisation

Particle Swarm Optimisation (PSO) is an optimisation strategy which draws from the dynamics of swarms in nature[197, 198]. Each member of the swarm is aware of its current position and is free to move in any way it pleases, but also appears to be aware of some global swarm properties. This latter knowledge gives the impression that the swarm moves as a single entity, in spite of the members of the swarm being free to move independently of the other members. PSO utilises these swarm dynamics as a means to locate the global optimum of a function; particle positions take the form of the degrees of freedom of the function, sampling the function at these points. The swarm then acts collectively to converge upon the global optimum of the function, by moving towards those positions where the function appears to be optimal.

For the sake of simplicity, we will combine the vectors $\theta$ and $p$ into a single $2d$-dimensional vector, $z$, which denotes a point in $\{\theta, p\}$-space. With PSO, we consider an ensemble of $N_{\text{swarm}}$ particles, each one characterised by a position vector $z_i$. Each particle is free to move in discrete time, $t$, subject to

$$z_i(t + 1) = z_i(t) + v_i(t + 1), \quad (2.4.14)$$

where $v_i(t + 1)$ is the “velocity” of the $i^{th}$ particle. The velocity has functional
form\textsuperscript{11}

\begin{equation}
\mathbf{v}_i(t + 1) = \omega \mathbf{v}_i(t) \\
+ c_1 \mathbf{U}_1(0, 1) \circ (\mathbf{b}_i(t) - \mathbf{z}_i(t)) \\
+ c_2 \mathbf{U}_2(0, 1) \circ (\mathbf{g}_i(t) - \mathbf{z}_i(t)),
\end{equation}

and we discuss each term individually. \(\omega\) is referred to as the “inertia” of the particle (conventionally set to 0.729), and dictates the level to which the velocity at time \(t\) contributes to the velocity at time \(t + 1\). \(c_1\) is the “cognitive learning factor” (set to 1.494 in our work), \(\mathbf{U}_1(0, 1)\) is a vector of uniform random variates, and \(\mathbf{b}_i(t)\) is the “private guide”, corresponding to the best position the \(i^{th}\) particle has visited along its trajectory, allowing us to interpret this term as a driving force towards this previous best position. \(c_2\) is the “social learning factor” (again set to 1.494), \(\mathbf{U}_2(0, 1)\) is a vector of uniform random variates, and \(\mathbf{g}_i(t)\) is the “global guide”, corresponding to the best position over all particles in the swarm along all \(N_{\text{swarm}}\) trajectories, this term therefore acting as a driving force towards that position.

The dynamics outlined above are iterated for each particle in the swarm. Because of the social learning factor, the particle trajectories are drawn to the same area. The private guide allows for a simultaneous exploration of the particle’s local environment. As time progresses, the particles converge upon the same point. The optimisation is considered complete when all particles occupy the same point, to within some numerical tolerance.

A number of modifications to the PSO algorithm exist; adaptive PSO\textsuperscript{[199]}, the swarm and queen algorithm\textsuperscript{[200]}, hybrid PSO\textsuperscript{[201]} (where particle multiplication is permitted), etc. However, we have found the conventional PSO algorithm to suffi-

\textsuperscript{11}In this equation, we use \(\mathbf{x} \circ \mathbf{y}\) to denote the Hadamard product between two vectors, and yields the vector \(\mathbf{x} \circ \mathbf{y} = \begin{bmatrix} x_1 y_1 & x_2 y_2 & \ldots & x_d y_d \end{bmatrix}^\top\).
ciently optimise $\theta$ and $p$. This has been verified within our group by comparison to a number of alternative optimisation strategies, such as differential evolution[202] and analytical optimisation[203].

2.5 The Quantum Chemical Topological Force Field

2.5.1 Previous Work

The first work towards the construction of QCTFF verified that the multipole moments, as computed through a QCT partitioning, reproduced the electrostatics of a system over the course of a MD simulation. Studies on both hydrogen fluoride[204] and water[205] clusters were undertaken to this end. With the water clusters, it was found that the density and potential energy were roughly 0.1% and 3%, respectively, off their experimental values, when multipole moment electrostatics were used. The spatial distribution functions were also found to be in good agreement with experiment, with correlation coefficients above 0.97.

In constructing a conformationally-dependent force field, our group began by using ANNs. The first work on this matter was that of Houlding et al.[176], where multipole moments, up to the hexadecapole moments, were predicted for the hydrogen fluoride dimer at a number of MD snapshot geometries. This work was followed by that of Handley et al.[175], where the same approach was taken for water clusters of increasing sizes (up to the hexamer). In the final such study, Darley et al.[206] applied this methodology to N-methyl acetamide and glycine, demonstrating that the approach was valid for biomolecular systems.
For reasons detailed in Section 2.4, the group subsequently transitioned from the use of ANNs to kriging, demonstrating that the errors associated with kriging were lower relative to those of ANNs for water[207], ethanol[208] and alanine[120]. Hydrated sodium ions[209] have also been treated in the same manner. The electrostatic energy of the system over a number of conformations was predicted to an accuracy of 4 kJ mol\(^{-1}\).

The first MD simulations carried out using the kriging models studied aqueous imidazole[210]. The solution density, diffusion coefficients, radial and spatial distribution functions were used to evaluate performance relative to AMBER. The solution density was well-recovered by the QCTFF implementation up to high concentrations, in contrast to AMBER, where the solution density was consistently underestimated relative to experiment. A similar result was found with the diffusion coefficients. The spatial and radial distributions also suggested a completely different ratio of imidazole stacked to chain conformers relative to the AMBER implementation. This work was later built upon in a study on liquid water over a range of pressures and temperatures[211]. Bulk properties were found to be well-recovered relative to point charge water potentials.

More recently, work has focused on reducing the errors attributable to our machine learning models by the refinement of in-house software[121]. As the systems studied become more complex, the workflow associated with the generation of kriging models requires streamlining so that the resultant kriging models are both highly accurate and small (vis-à-vis memory requirements) as possible[212, 213]. In addition, work has also shifted from the generation of kriging models for electrostatics to kriging models for the IQA components outlined in Section 2.2.3[214, 215]. In this way, we are able to forsake the empirical potentials employed in classical force fields for a fully quantum mechanical treatment of molecular energetics.
Work has also been undertaken on a number of conceptual issues that require resolution before the QCTFF becomes a commercially viable product[216]. For instance, the construction of a force field requires some concept of atom typing. In defining an atom type, one is able to use the same parameter set (or kriging models in our case) across a range of atoms within a system. In so doing, one is able to circumvent parameterising new models for every atom within a system, which is simply not feasible. The idea of an atom type is inextricably linked to that of transferability, which we have discussed in Section 2.3.2. A number of studies have been undertaken within our group to this end in an effort to determine the atom types required for the QCTFF to be generally applicable[217, 218, 219].

2.5.2 Proposed Implementation

Our aim is to construct a conformationally-dependent force field, whose energetic terms derive from an IQA treatment of a molecular system. The conformational dependence is accounted for by the construction of a kriging model, that can predict an IQA energy component for untrained molecular conformations. This project has been named the “Quantum Chemical Topological Force Field” in the literature.

Through previous work in our group, we have found that constructing individual kriging models for each of the components outlined in (2.2.14) yields undesirably large prediction errors. Instead, we have found that the use of the four following terms is ideal,

\[ E_{IQA}^A = E_{\text{self}}^A + \sum_B V_{xc}^{AB} + \sum_{B=4} V_{SRC}^{AB} + \sum_{B=5} V_{LRC}^{AB}. \]  

(2.5.1)

\( E_{\text{self}}^A \) is the “self-energy” of atom \( A \), comprising the kinetic, electron-nuclear at-
traction and electron-electron repulsion energies,

\[ E_{self}^A = T_A^A + V_{en}^{AA} + V_{ee}^{AA} \]
\[ = T_A^A + V_{en}^{AA} + V_{ee,coul}^{AA} + V_{ee,exch}^{AA} + V_{ee,corr}^{AA} \].

(2.5.2)

\[ V_{xc}^{AB} \] is the inter-atomic exchange-correlation energy, and is given by

\[ V_{xc}^{AB} = V_{ee,exch}^{AB} + V_{ee,corr}^{AB} \].

(2.5.3)

The terms \( V_{SRC}^{AB} \) and \( V_{LRC}^{AB} \) correspond to the short- and long-ranged classical energies, respectively. The long-ranged classical energy takes the form

\[ V_{LRC}^{AB} = V_{ee,coul}^{AB} + V_{en}^{AB} + V_{ne}^{AB} + V_{nn}^{AB} \quad B \in 1-5 \text{ and higher}, \]

(2.5.4)

which can be expanded in multipole moments. As such, we require the constraint that the interaction between \( A \) and \( B \) be 1-5 and higher, otherwise the multipole moment interaction will diverge, as discussed in Section 2.3.1. We are then required to express \( V_{SRC}^{AB} \) explicitly as

\[ V_{SRC}^{AB} = V_{ee,coul}^{AB} + V_{en}^{AB} + V_{ne}^{AB} + V_{nn}^{AB} \quad B \in 1-4 \text{ and lower} \].

(2.5.5)

Through all of these terms, we find that we need three separate kriging models for \( E_{self}^A, \sum_B V_{xc}^{AB}, \sum_{B_1-4} V_{SRC}^{AB} \), and a further twenty five separate kriging models for multipole moments (up to the hexadecapole moment) of \( \sum_{B_1-5} V_{LRC}^{AB} \). Therefore, we require twenty eight individual kriging models for every atom in the system.
Chapter 3

Conformational Sampling

3.1 Introduction

Our aim is the construction of a kriging model that can be used over the course of a MD trajectory to make reliable predictions. Accordingly, we require training points that span a domain of conformational space which is available to the system under the simulation conditions. Concurrently, some regions of the aforementioned conformational domain are difficult to model. These regions require a greater sampling density to capture the undulant features of the PES. We are therefore required to develop a conformational sampling methodology that satisfies this twofold precondition.

The conformational sampling of small molecular species has historically been undertaken by some form of systematic variation of a small number of molecular degrees of freedom, for instance, torsional angles[220, 221, 222]. These systematic techniques were later adapted to vary a small number of degrees of free-
3.1. INTRODUCTION

don, whether these be internal[223, 224, 225] or Cartesian[226, 227] coordinates, stochastically, by typically invoking some form of Metropolis Monte Carlo scheme[228]. However, for any molecular species containing more than 12 rotatable bonds, sampling a subset of the molecular degrees of freedom, systematically or stochastically, is not a feasible strategy for the generation of a physically realistic conformational ensemble.

Another strategy involves invoking some form of classical MD[229, 230], and selecting relevant conformers based on either;

1. The lowest energy conformers as determined by some classical force field;
2. The output of a conformer at regular intervals along a MD trajectory;
3. Maximising the conformational diversity of the conformers by analysis of some group of degrees of freedom.

A number of statistical tools can be used to demonstrate the inadequacy of MD as a means for conformational sampling. Essential dynamics[231], a form of principal component analysis, is one such prevalent tool. Past work has shown that the conformational flexibility of a protein is attributable to a small number of low frequency, delocalised modes of vibration[232]. Indeed, there is a large literature on using frequency analysis to decompose a MD trajectory into a small number of vibrational degrees of freedom, see for example [233, 234, 235, 236]. However, the finite time scale of a MD simulation does not allow for a proper sampling of these modes, as verified from an essential dynamics trajectory analysis[237].

Large scale conformational changes occur over relatively long timescales. A complete exploration of conformational space with MD, therefore, takes a significant amount of time and computational resources. A number of developments
have, however, made large scale conformational sampling feasible. The first of these is the RESPA (Reversible rEference System Propagator Algorithm), an integration scheme which follows from a Liouville operator treatment of Newtonian dynamics[238]. The RESPA allows for a number of dynamical timesteps to be used over the course of a MD trajectory; the slower degrees of freedom are integrated with longer timesteps than their higher frequency counterparts. RESPA then reduces the computational complexity of the problem, and subsequently allows for longer MD trajectories to be computed. It has even been shown that certain biomolecular systems are amenable to outer timesteps of the order of 100fs[239].

If the PES features a number of low energy regions separated from one another by high energy transition barriers, then even the RESPA is rarely adequate for a complete sampling of conformational space. The timescales required for a trajectory to surmount these high energy transition barriers are typically of the order of milliseconds, and transitions are referred to as “rare events”[240]. Metadynamics[241] and high temperature MD[242] have been used to accelerate the transitioning to higher energy regions of the PES, thereby accelerating conformational sampling. However, results have shown the technique to be susceptible to error owing to the somewhat arbitrary choice of enhancement parameters, such as the choice of the high temperature for transition[243]. It has also been shown that the conformational regions explored using high temperature MD do not overlap with some systematic conformational search methodologies[244].

An interesting methodology, which is somewhat similar to the conformational sampling strategy that we will outline in this chapter, has proposed a “mode-following” protocol[245, 246]. With this, an arbitrary initial molecular conformation is discretely evolved along one or a subset of its lowest frequency normal modes. Note that the molecular Hessian is not updated along the trajectory. Periodically, the
It is not the sampling strategies that we are adverse to. Indeed, there exist elegant algorithms, such as the Wang-Landau algorithm\cite{247} or the nested sampling algorithm\cite{248} for conformational sampling\cite{249, 250, 251}. However, each method we have outlined utilises a classical force field or some other empirical potential. It would be unsightly to initiate the construction of our force field with the very methodology that we are trying to supersede. Aside from this conceptual concern, conventional force fields are notorious for significant disagreement in virtually every molecular property\cite{252, 151}. The choice of force field would then impact on the training of our own force field. Typifying this criticism, a recent article has shown the disparity between the popular classical force fields in predicting the conformational preferences of a number of small proteins\cite{253}.

In the following, we present a method with which rigorous conformational sampling can be undertaken for the purpose of constructing kriging training sets. We approximate the \textit{ab initio} PES by tessellating with a set of second order Taylor expansions at a number of distinct seeding geometries. We refer to these second order Taylor expansions as “Local Wells” (LWs) since they are quadratic in the molecular energy. Stochastically hopping between these LWs, followed by vibration within the LW, allows for dynamics to be undertaken on this approximate “tessellated” PES.
3.2 Stationary Point Vibrations

3.2.1 Kinetic Energy

For a molecular system with \( N \) atoms, we define a molecular Cartesian state vector,
\[
x = \begin{bmatrix} x_1 & x_2 & \ldots & x_{3N} \end{bmatrix}^\top,
\]
and introduce a difference coordinate, \( \Delta x \), relative to some arbitrary configuration,
\[
x^* = \begin{bmatrix} x_1^* & x_2^* & \ldots & x_{3N}^* \end{bmatrix}^\top,
\]
such that
\[
\Delta x = x - x^* = \begin{bmatrix} x_1 - x_1^* & x_2 - x_2^* & \ldots & x_{3N} - x_{3N}^* \end{bmatrix}^\top.
\]

In the following, we will use Greek indices to enumerate atoms, and Latin indices to enumerate degrees of freedom. Constraining \( x^* \) to be invariant, we can use \( \Delta x \) as a conventional state vector in the following derivation. The classical expression for the kinetic energy of a system is then given by
\[
T(\Delta \dot{x}_1, \ldots, \Delta \dot{x}_{3N}) = \frac{1}{2} \sum_{i=1}^{3N} m_i \left( \frac{d}{dt} \Delta x_i \right)^2 = \frac{1}{2} \sum_{\alpha=1}^{N} \frac{m_\alpha}{2} \Delta x_\alpha \cdot \Delta \dot{x}_\alpha,
\]
where we have given an example of the formulation in terms of individual degrees of freedom and in terms of atomic position vectors, \( \Delta x_\alpha \). It will be convenient to introduce a set of mass-weighted Cartesian coordinates, \( q_i = \Delta x_i \sqrt{m_i} \), so that
\[
(3.2.2) \quad T(\dot{q}_1, \ldots, \dot{q}_{3N}) = \frac{1}{2} \sum_{i=1}^{3N} \left( \frac{d}{dt} q_i \right)^2 = \frac{1}{2} \sum_{i=1}^{3N} \dot{q}_i^2.
\]
3.2. STATIONARY POINT VIBRATIONS

3.2.2 Potential Energy

The potential energy corresponding to a state, \( V(\mathbf{x}) = V(x_1, ..., x_{3N}) \) is given by a Taylor series about the previously introduced state \( \mathbf{x}^* \), leading to

\[
2V(x_1, ..., x_{3N}) = 2V_0(x_1^*, ..., x_{3N}^*) + 2 \sum_{i=1}^{3N} (x_i - x_i^*) \frac{\partial V}{\partial x_i} \bigg|_{\mathbf{x}^*} + \sum_{i,j=1}^{3N} (x_i - x_i^*) (x_j - x_j^*) \frac{\partial^2 V}{\partial x_i \partial x_j} \bigg|_{\mathbf{x}^*} + ... .
\]

(3.2.4)

The derivative factors are constant owing to the evaluation condition. Introducing some shorthand notation for the derivatives\(^1\),

\[
\frac{\partial V}{\partial x_i} \bigg|_{\mathbf{x}^*} = J_i, \quad \frac{\partial^2 V}{\partial x_i \partial x_j} \bigg|_{\mathbf{x}^*} = H_{ij},
\]

(3.2.5)

such that

\[
2V(x_1, ..., x_{3N}) = 2V_0(x_1^*, ..., x_{3N}^*) + 2 \sum_{i=1}^{3N} (x_i - x_i^*) J_i + \sum_{i,j=1}^{3N} (x_i - x_i^*) (x_j - x_j^*) H_{ij} + ... .
\]

(3.2.6)

The first order and second order spatial derivatives of the potential energy correspond to elements of the Jacobian and Hessian, respectively. By choosing \( \mathbf{x}^* \) such that it represents a stationary point on the potential energy surface, we are free to set \( V(\mathbf{x}^*) = 0 \). The term involving the Jacobian goes to zero at this point, by definition. Omitting all terms higher than second order, we obtain a harmonic approximation to the potential energy

\[
2V(x_1, ..., x_{3N}) \approx \sum_{i,j=1}^{3N} (x_i - x_i^*) (x_j - x_j^*) H_{ij}.
\]

(3.2.7)

\(^1\)The Jacobian, \( J \), is defined as the matrix of all first-order partial derivatives of a function, \( f : \mathbb{R}^a \rightarrow \mathbb{R}^b \), with respect to those degrees of freedom over which \( f \) is defined. Taking the case of \( a = 1 \), we see that \( J \) takes the form of \( \left[ \frac{\partial f}{\partial x_1} ... \frac{\partial f}{\partial x_a} \right]^T \), which is the form used here. Of course, Jacobian is equivalent to the gradient of a scalar field, \( \nabla f \).
It is useful to express the potential energy in the same coordinates as those used for the kinetic energy. This can be achieved by using the definition of our mass-weighted Cartesian coordinates, leading to the relation

$$\frac{\partial}{\partial q_i} = \frac{1}{\sqrt{m_i}} \frac{\partial}{\partial (x_i - x_i^*)} = \frac{1}{\sqrt{m_i}} \frac{\partial}{\partial x_i}.$$ 

Modifying (3.2.7) to account for the above transformation, we obtain

$$2V(x_1, \ldots, x_{3N}) = \sum_{i,j=1}^{3N} \Delta x_i \Delta x_j \frac{\partial^2 V}{\partial x_i \partial x_j} \bigg|_{x^*} = \sum_{i,j=1}^{3N} \frac{q_i q_j}{\sqrt{m_i m_j}} \frac{\partial^2 V}{\partial q_i \partial q_j} \bigg|_{q^*}$$

$$= \sum_{i,j=1}^{3N} H_{ij} q_i q_j = 2V(q_1, \ldots, q_{3N}),$$

where the Hessian is now expressed in mass-weighted Cartesian form, i.e.

$$H_{ij} = \frac{1}{\sqrt{m_i m_j}} \frac{\partial^2 V}{\partial x_i \partial x_j} \bigg|_{x^*} = \frac{\partial^2 V}{\partial q_i \partial q_j} \bigg|_{q^*}.$$ 

### 3.2.3 Equations of Motion

Given our expressions for the kinetic and potential energy, (3.2.2) and (3.2.8), respectively, we can evaluate the Euler-Lagrange equations of motion

$$\frac{d}{dt} \frac{\partial T}{\partial \dot{q}_k} + \frac{\partial V}{\partial q_k} = 0 \quad \forall k = 1, \ldots 3N$$

leading to

$$\frac{d}{dt} \left( \frac{1}{2} \sum_{i=1}^{3N} \dot{q}_i^2 \right) + \frac{\partial}{\partial q_k} \left( \frac{1}{2} \sum_{i,j=1}^{3N} H_{ij} q_i q_j \right)$$

$$= \frac{d}{dt} \left( \frac{1}{2} \sum_{i=1}^{3N} \dot{q}_i^2 \right) + \frac{1}{2} \sum_{i,j=1}^{3N} \frac{\partial}{\partial q_k} H_{ij} q_i q_j$$

$$= \frac{d}{dt} \left( \sum_{i=1}^{3N} \delta_{ik} \dot{q}_i \right) + \frac{1}{2} \sum_{i,j=1}^{3N} H_{ij} \left( \dot{q}_i \frac{\partial q_j}{\partial q_k} + \dot{q}_j \frac{\partial q_i}{\partial q_k} \right)$$

$$= \frac{d}{dt} \left( \sum_{i=1}^{3N} \delta_{ik} \dot{q}_i \right) + \frac{1}{2} \sum_{i,j=1}^{3N} H_{ij} \left( \dot{q}_i \frac{\partial q_j}{\partial q_k} + \dot{q}_j \frac{\partial q_i}{\partial q_k} \right)$$
3.2. STATIONARY POINT VIBRATIONS

\[ \frac{d}{dt} \ddot{q}_k + \frac{1}{2} \sum_{i,j=1}^{3N} H_{ij} (q_i \delta_{jk} + q_j \delta_{ik}) = 0, \]  
\[ \frac{d^2}{dt^2} q_k + \frac{1}{2} \left( \sum_{i=1}^{3N} H_{ik} q_i + \sum_{j=1}^{3N} H_{kj} q_j \right) = 0, \]  
\[ \frac{d^2}{dt^2} q_k + \sum_{i=1}^{3N} H_{ik} q_i = 0, \]  
(3.2.10)

where we have invoked the symmetric nature of the Hessian \((H_{ij} = H_{ji})\) and the fact that the last two sums are identical since \(i, j\) are dummy indices.

We are now required to solve a second order homogeneous differential equation, the solution of which is a simple superposition of sinusoids of angular frequency \(\omega\) and amplitudes \(A_k, B_k\) for the \(k^{th}\) equation of motion

\[ q_k(t) = A_k \cos(\omega t) + B_k \sin(\omega t). \]

We choose to use the more compact notation of a single sinusoid with a phase factor, \(\phi\),

\[ q_k(t) = A_k \cos(\omega t + \phi). \]  
(3.2.11)

Substituting (3.2.11) into (3.2.10), we obtain

\[ \frac{d^2}{dt^2} A_k \cos(\omega t + \phi) + \sum_{i=1}^{3N} H_{ik} A_i \cos(\omega t + \phi) = 0 \]
\[ -\omega^2 A_k \cos(\omega t + \phi) + \sum_{i=1}^{3N} H_{ik} A_i \cos(\omega t + \phi) = 0. \]  
(3.2.12)

The next step involves the cancellation of the factor \(\cos(\omega t + \phi)\) in each term. This cancellation places a constraint on the form of (3.2.11): if \(\omega t + \phi = (2n + 1)\pi/2\), where \(n \in \mathbb{N}\), cancellation is not permitted since the cosine of this quantity is zero. However, prior to cancellation, we recover \(q_k(t) = 0\) from (3.2.12), and so the constraint is irrelevant. Continuing with non-zero \(\cos(\omega t + \phi)\), we obtain

\[ -\omega^2 A_k + \sum_{i=1}^{3N} H_{ik} A_i = 0 \quad \therefore \quad \sum_{i=1}^{3N} A_i (H_{ik} - \omega^2 \delta_{ik}) = 0. \]  
(3.2.13)
CHAPTER 3. CONFORMATIONAL SAMPLING

This equation constitutes an eigensystem, for which there exist $3N$ values of $\omega$ yielding non-trivial solutions for the $q_k(t)$, i.e. where $A_k \neq 0$. The conventional way to solve the eigensystem is by formation of the secular determinant,

$$
\begin{vmatrix}
H_{11} - \omega^2 & \cdots & H_{1,3N} \\
\vdots & \ddots & \vdots \\
H_{3N,1} & \cdots & H_{3N,3N} - \omega^2
\end{vmatrix} = 0. \quad (3.2.14)
$$

Note that if the Hessian is in diagonal form, the determinant takes the form $
\prod_{i=1}^{3N}(H_{ii} - \omega^2)$. Equality to zero follows from at least one term being equal to zero in the expansion, which is satisfied only if $\omega^2$ is equal to one of the diagonal elements of the Hessian. Therefore, if the Hessian is in diagonal form, the angular frequencies satisfying (3.2.13) are given by the square roots of the diagonal elements of the Hessian.

It is worth pointing out that if one of the diagonal elements of the diagonalised Hessian is less than zero, then the corresponding $\omega$ is imaginary. The resultant equation of motion, (3.2.11) then becomes

$$
q_k(t) = A_k \cos(i\omega t + \phi)
= A_k \cos(i\omega t) \cos(\phi) - A_k \sin(i\omega t) \sin(\phi)
= A_k \cosh(\omega t) \cos(\phi) - A_k \sinh(\omega t) \sin(\phi), \quad (3.2.15)
$$

which diverges after a certain amount of time. However, our timestep selection outlined in Section 3.4.2 guarantees that the equation of motion is never evolved for long enough to diverge. Imaginary frequencies are encountered at non-stationary point conformations on the PES. When we generalise our method to accommodate non-stationary point conformations, this point is worth bearing in mind.

Consider now a molecule under the influence of no external fields or field derivatives. If we rigidly shift the molecule in space, we find that the internal state of
the molecule is not altered. There are then three basis vectors with no associated equations of motion. This behaviour is manifest in three of the angular frequencies being equal to zero, i.e. by use of (3.2.11), \( q_k(t) = 0 \) for these three degrees of freedom. A further three degrees of freedom can be argued to be similarly time-invariant. Consider rigid rotations about the three Euler angles of the system. As with the global translational degrees of freedom, rigid rotations do not affect the internal state of the molecule, and so an appropriate coordinate system yields another three angular frequencies equal to zero.

The choice of coordinate system we have alluded to which renders six angular frequencies equal to zero are referred to as the *normal coordinates*. The most computationally amenable way to transform the Hessian into a normal coordinate basis is by constructing an appropriate transformation matrix, \( D : \mathbb{R}^{3N} \to \mathbb{R}^{3N} \). Note that the mapping to normal coordinates conserves the dimensionality of the space, an orthogonal direct sum of external and internal vector spaces, i.e. \( \mathbb{R}^{3N} = \mathbb{R}^6_{\text{ext}} \oplus \mathbb{R}^{3N-6}_{\text{int}} \). It is the elements of \( \mathbb{R}^{3N-6}_{\text{int}} \) that represent the internal degrees of freedom of the system. The elements of \( \mathbb{R}^6_{\text{ext}} \) correspond to global translational and rotational degrees of freedom, and there is no point to sampling within this space since all conformations are equivalent.

### 3.3 Generalised Vibrations

It is clear from the previous section that setting the spatial first derivative of the potential to zero simplifies the resultant Euler-Lagrange equations of motion. We now assume that spatial first derivative does not vanish, and proceed with a general derivation for the equations of motion of a system that is not situated at a stationary point on the potential energy surface.
Rewriting (3.2.7) to include the Jacobian term,

\[
2V(x_1, \ldots, x_{3N}) \approx 2 \sum_{i=1}^{3N} (x_i - x_i^*) J_i + \sum_{i,j=1}^{3N} (x_i - x_i^*) (x_j - x_j^*) H_{ij},
\]

(3.3.1)

where \(x^*\) now corresponds to an arbitrary position vector. Transformation into the mass-weighted Cartesian form is accomplished as before by a simple modification

\[
2V(q_1, \ldots, q_{3N}) \approx 2 \sum_{i=1}^{3N} \frac{J_i}{\sqrt{m_i}} q_i + \sum_{i,j=1}^{3N} \frac{q_i q_j}{\sqrt{m_i m_j}} H_{ij}
\]

\[
\approx 2 \sum_{i=1}^{3N} J_i q_i + \sum_{i,j=1}^{3N} q_i q_j H_{ij},
\]

(3.3.2)

where the Jacobian and Hessian have been re-expressed in terms of the mass-weighted Cartesians,

\[
J_i = \frac{1}{\sqrt{m_i}} \frac{\partial V}{\partial x_i} \bigg|_{x^*} = \frac{\partial V}{\partial q_i} \bigg|_{q^*},
\]

(3.3.3)

\[
H_{ij} = \frac{1}{\sqrt{m_i m_j}} \frac{\partial^2 V}{\partial x_i \partial x_j} \bigg|_{x^*} = \frac{\partial^2 V}{\partial q_i \partial q_j} \bigg|_{q^*}.
\]

(3.3.4)

From (3.2.9), the only additional calculation we require is the spatial derivative of the term involving the Jacobian in (3.3.2). As such,

\[
\frac{\partial}{\partial q_k} \sum_{i=1}^{3N} J_i q_i = \sum_{i=1}^{3N} J_i \frac{\partial q_i}{\partial q_k} + \sum_{i=1}^{3N} q_i \frac{\partial J_i}{\partial q_k} = \sum_{i=1}^{3N} J_i \delta_{ik} = J_k,
\]

(3.3.5)

where we have introduced the Kronecker delta, \(\delta_{ik}\), such that \(\delta_{ik} = 1\) if \(i = k\), and \(\delta_{ik} = 0\) otherwise. The term involving the partial derivative of the Jacobian is equal to zero owing to the evaluation constraint we have placed on the Jacobian. The partial derivative of the potential with respect to a spatial degree of freedom
3.3. GENERALISED VIBRATIONS

is then

$$\frac{\partial V}{\partial q_k} = J_k + \sum_{i=1}^{3N} H_{ik} q_i.$$  (3.3.6)

The expanded Euler-Lagrange equation of (3.2.10) now reads

$$\frac{d^2}{dt^2} q_k + \sum_{i=1}^{3N} H_{ik} q_i = -J_k,$$  (3.3.7)

which is an inhomogeneous second order differential equation. Physical insight into this equation can be garnered by observing that $-J_k$ is a force, and so (3.3.7) is the equation of motion governing a forced harmonic oscillator, where the driving force is a constant.

Solution to an inhomogeneous differential equation is found by solving the underlying homogeneous differential equations, as we have done in (3.2.11), and appending a “particular solution”, which is dependent on the inhomogeneous term, $\xi(J_k)$. The form of $\xi(J_k)$ is found by substituting it for the independent variable in our original homogeneous differential equation, i.e.

$$\frac{d^2}{dt^2} \xi(J_k) + \sum_{i=1}^{3N} H_{ik} \xi(J_k) = -J_k$$

$$\sum_{i=1}^{3N} H_{ik} \xi(J_k) = -J_k$$

$$\xi(J_k) = -\frac{J_k}{\sum_{i=1}^{3N} H_{ik}},$$  (3.3.8)

where the temporal derivative of $\xi(J_k)$ is zero since the Jacobian possesses no time dependence. Then, the general equation of motion for our forced harmonic oscillator is

$$q_k(t) = A_k \cos(\omega t + \phi) - \frac{J_k}{\sum_{i=1}^{3N} H_{ik}}.$$  (3.3.9)

We can validate this solution by dimensional analysis, where the particular solution has dimensions of force over force constant, i.e. distance, as would be expected.
Also, we note that at an energetic minimum, $J_k = 0$ and we recover the stationary point solution of (3.2.11).

## 3.4 Tyche - Conformational Sampling Software

In this section, we present a number of additional concepts and their computational implementations. These will then come together to form Tyche, a piece of software that we have developed over the past few years. *Tyche* is the major conformational sampling methodology used within our group for the construction of kriging training sets.

### 3.4.1 Transformation to Normal Coordinates

By finding the six basis vectors which span $\mathbb{R}^6_{\text{ext}}$, and exploiting the fact that the normal coordinates form a mutually orthogonal basis, the basis vectors spanning $\mathbb{R}^{3N-6}_{\text{int}}$ can be found by a Gram-Schmidt orthogonalisation relative to the six external basis vectors.

The basis vectors of $\mathbb{R}^6_{\text{ext}}$ are found by invoking the Sayvetz [254] (or Eckart[255]) conditions. These correspond to defining a reference frame for the system, whose origin is located at the molecular centre of mass, and whose axes are aligned with the principal axes of inertia of the system. As such, this reference frame translates and rotates with the system, such that for an observer in this reference frame, the molecule would appear stationary.

First, the molecule is translated to a frame of reference whose origin is at the molecular centre of mass. In terms of $\mathcal{D}$, the Sayvetz conditions are contained
within the first six columns. We specify the \( i \)th column of \( D \) with \( d_i \). The first three columns correspond to the translational Sayvetz conditions,

\[
\begin{align*}
  d_1 & = \begin{bmatrix} \sqrt{m_1} & 0 & 0 & \sqrt{m_2} & 0 & 0 \cdots \sqrt{m_N} & 0 & 0 \end{bmatrix}^T, \\
  d_2 & = \begin{bmatrix} 0 & \sqrt{m_1} & 0 & 0 & \sqrt{m_2} & 0 \cdots 0 & \sqrt{m_N} \end{bmatrix}^T, \\
  d_3 & = \begin{bmatrix} 0 & 0 & \sqrt{m_1} & 0 & 0 & \sqrt{m_2} \cdots 0 & 0 \sqrt{m_N} \end{bmatrix}^T.
\end{align*}
\]

These are clearly mutually orthogonal, but we also see that the scalar product with a generalised coordinate yields

\[
d_k \cdot q = \sum_{i=1}^{3N} m_i x_i = 0, \quad \forall k = 1, 2, 3.
\]

Equality to zero follows from the frame of reference having its origin coincide with the molecular centre of mass.

Defining \( d_4, d_5, d_6 \) is significantly more involved. We introduce some novel notation for this purpose; an atomic position vector is defined as

\[
x_\alpha = \begin{bmatrix} x_\alpha & y_\alpha & z_\alpha \end{bmatrix}^T,
\]

where \( x_\alpha, y_\alpha, z_\alpha \) are the components of the atomic position vector along the three Cartesian axis. Then, defining the moment of inertia tensor,

\[
I = \begin{bmatrix}
  \sum_{\alpha=1}^{N} m_\alpha (y_\alpha^2 + z_\alpha^2) & -\sum_{\alpha=1}^{N} m_\alpha x_\alpha y_\alpha & -\sum_{\alpha=1}^{N} m_\alpha x_\alpha z_\alpha \\
  -\sum_{\alpha=1}^{N} m_\alpha y_\alpha x_\alpha & \sum_{\alpha=1}^{N} m_\alpha (x_\alpha^2 + z_\alpha^2) & -\sum_{\alpha=1}^{N} m_\alpha y_\alpha z_\alpha \\
  -\sum_{\alpha=1}^{N} m_\alpha z_\alpha x_\alpha & -\sum_{\alpha=1}^{N} m_\alpha z_\alpha y_\alpha & \sum_{\alpha=1}^{N} m_\alpha (x_\alpha^2 + y_\alpha^2)
\end{bmatrix}.
\]

The moment of inertia tensor can be diagonalised, \( P^T I P = I' \), where \( P \) contains the eigenvectors, or principal moments, of \( I \) and \( I' \) is a diagonal matrix containing the associated eigenvalues. Then, the elements of \( d_4, d_5, d_6 \) are of the form

\[
\begin{align*}
d_{4,\alpha} & = \left( x_\alpha \cdot P_2^T \right) P_3 - \left( x_\alpha \cdot P_3^T \right) P_2 \sqrt{m_\alpha}, \\
d_{5,\alpha} & = \left( x_\alpha \cdot P_3^T \right) P_1 - \left( x_\alpha \cdot P_1^T \right) P_3 \sqrt{m_\alpha}.
\end{align*}
\]
where the numerical index of $\mathbf{P}$ denotes a column, and $\mathbf{d}_{k,\alpha}$ corresponds to the triplet of entries in $\mathbf{d}_k$ related to the three degrees of freedom of the $\alpha$th atom[256]. These rotational Sayvetz conditions correspond to cross products of the principal moments and atomic position vectors. However, for the sake of brevity, we simply refer the reader to the previously referenced works for a more in-depth derivation[254, 255].

We now have the six external basis vectors which span $\mathbb{R}^6_{\text{ext}}$, and one can verify that they form a mutually orthogonal set. The $3N - 6$ basis vectors spanning $\mathbb{R}^{3N-6}_{\text{int}}$ can then be found by a Gram-Schmidt orthogonalization, which we now outline. We define the projection operator as a mapping from a vector, $\mathbf{v}$, onto a vector $\mathbf{u}$,

$$\text{proj}_\mathbf{u}(\mathbf{v}) = \frac{\mathbf{v} \cdot \mathbf{u}}{\mathbf{u} \cdot \mathbf{u}} \mathbf{u}. \quad (3.4.6)$$

Denote the set of vectors defined by the columns of $\mathbf{D}$, $\{\mathbf{d}_1, \ldots, \mathbf{d}_{3N}\}$. The first six such vectors, $\mathbf{d}_1, \ldots, \mathbf{d}_6$, are those which span $\mathbb{R}^6_{\text{ext}}$. We generate $\{\mathbf{d}_7, \ldots, \mathbf{d}_{3N}\}$ randomly, and “project out” the components of these vectors that are not orthogonal to the six external basis vectors $\{\mathbf{d}_1, \ldots, \mathbf{d}_6\}$. Invoking the projection operator, we iterate the following

$$\mathbf{d}_i = \mathbf{d}_i - \sum_{j=1}^{i-1} \text{proj}_{\mathbf{d}_j}(\mathbf{d}_i) \quad \forall i = 7, 3N, \quad (3.4.7)$$

until $\mathbf{d}_i \cdot \mathbf{d}_j = \delta_{ij}, \forall i, j = 1, \ldots, 3N$, where $\delta_{ij}$ is the Kronecker delta. Note that we only Gram-Schmidt orthogonalise $\{\mathbf{d}_7, \ldots, \mathbf{d}_{3N}\}$ since $\{\mathbf{d}_1, \ldots, \mathbf{d}_6\}$ are already mutually orthogonal. Once all columns of $\mathbf{D}$ are mutually orthogonal, we normalise each so that the transformation described by $\mathbf{D}$ is norm-conserving.

Introducing the $3N \times 3N$ diagonal matrix, $\mathbf{M}$, whose elements are given by $M_{ii} =
$1/\sqrt{m_i}$, the $3N \times 3N$ matrix $\mathcal{E}$ comprising the eigenvectors of the mass-weighted Hessian, and the $3N \times 3N$ diagonal matrix $\Omega$, containing elements $\Omega_{ii} = \omega_i^2$. We subsequently subject the Hessian to the transformation

$$\mathcal{E}^\top D^\top M H M D \mathcal{E} = \Omega. \quad (3.4.8)$$

The first six diagonal entries of $\Omega$, $\omega_1^2, \ldots, \omega_6^2$ are equal to zero, and the associated eigenvectors correspond to the global translational and rotational degrees of freedom of the system. The remaining $3N - 6$ entries of $\Omega$ correspond to the frequencies of the internal degrees of freedom of the system.

A useful quantity that we will make use of in the next section is the reduced mass of a normal mode, which we conveniently group together into the $3N$ vector $\mu$, the individual elements of which are given by

$$\mu_k = \left[ (MdE)_k \cdot (MdE)_k \right]^{-1}, \quad (3.4.9)$$

where $(MdE)_k$ denotes the $k^{th}$ column of the matrix product $MdE$. Note then by the equations of simple harmonic motion that the force constant of the $k^{th}$ harmonic oscillator is given by $\omega_k^2 \mu_k$. Indeed, the quantity $MdE$ will be of great importance when we come to discuss redundant internal coordinates in Section 3.5. $MdE$ is a $3N \times 3N$ matrix, the columns of which correspond to individual normal modes, and the rows contain the Cartesian displacements associated with each normal mode.

### 3.4.2 Dynamics

We have previously given the equations of motion for each normal coordinate in (3.3.9), which is restated here in numerical form,

$$q_k(t + \Delta t_k) = q_k(t) + A_k \cos(\omega \Delta t_k + \phi) - \frac{J_k}{\sum_{i=1}^{3N} \overline{H_{ik}}}. \quad (3.4.10)$$
There are three free parameters for each equation of motion; the amplitude, $A_k$, the integration timestep $\Delta t_k$, and the phase, $\phi$. The amplitude can be discerned from the conventional equations of simple harmonic motion, where

$$A_k = \sqrt{\frac{2E_k}{\omega_k^2 \mu_k}}. \quad (3.4.11)$$

$E_k$ is the energy we have made available to the $k^{th}$ normal mode, $\omega_k$ its angular frequency and $\mu_k$ its reduced mass, the formulations for which we have given in Section 3.4.1.

### Energy Allocation

The total thermal energy that is available to a single degree of freedom is given by a standard equipartition of energy, $k_B T/2$. An immediate solution is to simply set each $E_k = k_B T/2$. However, we find this choice to be unnecessarily restrictive. The equipartition value $k_B T/2$ is the expectation of the energy in each degree of freedom. In reality, the energy available to each degree of freedom is a dynamic quantity which fluctuates about the equipartition value as energy is redistributed through the degrees of freedom.

We are able to sample the microcanonical ensemble by working with the total thermal energy available to a system, $(3N - 6)k_B T/2$. By stochastically distributing this energy through all degrees of freedom, we have

$$E_k = (3N - 6)\frac{k_B T}{2} U(0, 1), \quad (3.4.12)$$

where $U(0, 1)$ is a random number selected from the uniform distribution on $[0, 1]$. One can subsequently rescale all energies so that $\sum_k E_k = (3N - 6)k_B T/2$. Therefore the total energy within the system remains constant throughout the sampling.

\footnote{While an integration timestep is typically the same across all degrees of freedom, we have suggestively allowed the integration timestep to vary across the degrees of freedom.}
Such a choice grants a great deal of conformational freedom for the subsequent sampling. For example, there is the possibility that a single degree of freedom possesses the vast majority of the total thermal energy. The resultant equations of motion consequently permit the exploration of the high energy portions of the potential well along a single degree of freedom.

However, one also needs to safeguard against physically unrealistic conformations, where chemical bonds have actually been broken. Such conformations are of no use over the course of an MD simulation, and are more frequently encountered when exploring the high energy regions of conformational space. The aforementioned safeguarding can be undertaken by requiring all generated conformers to be constrained within some conformational domain. For instance, one can constrain a subset of valence coordinates (e.g. all bond lengths and valence angles) to lie within some range of their equilibrium values, e.g. each bond length, \( b \), lies within the range \( c^{-1}b_0 \leq b \leq cb_0 \), where \( c \) is a suitably chosen “stretching” parameter and \( b_0 \) is the equilibrium bond length. Indeed, GUIs which display chemical structures use such a criterion in ascertaining whether a bond should be drawn between two atoms. \( b_0 \) there is the sum of the van der Waals radii of two atoms, and \( c \) is conventionally set to be 1.2.

The distribution of molecular energies as obtained through allocating the energy by (3.4.12) is depicted in Figure 3.1. The sampling well is deep; the range of molecular energies is roughly 120 kJ mol\(^{-1}\), and spans the physically relevant regions of conformational space. However, we note that there is a significant discontinuity between the \textit{ab initio} minimum energy and the \textsc{tyche} sampling energy. This discontinuity is not ideal, but could be reduced by an adequate parameterisation of the stretching parameter discussed above.

We have, however, decided against this methodology. The choice of the stretch-
CHAPTER 3. CONFORMATIONAL SAMPLING

Figure 3.1: Distribution of molecular energies obtained through TYCHE by use of the energy allocation scheme of (3.4.12). The lowest energy point corresponds to the \textit{ab initio} minimum.

The parameter is entirely arbitrary, and renders our conformational sampling more akin to a systematic sampling within some hypercube domain on the valence coordinates. To circumvent the necessity of parameterising the conformational domain upon which we sample, while maximising the amount of conformational space available for sampling, we can instead sample within the canonical ensemble by choosing

\[ E_k = \frac{k_B T}{2} N(1,1), \]  

where \( N(1,1) \) is a random number selected from a Gaussian distribution with expectation value and standard deviation of unity. No rescaling to enforce \( \sum_k E_k = (3N - 6)k_B T/2 \) is taken with this approach, enhancing the amount of conformational space available to the system. \( N(1,1) \) limits the accessibility to the physically unrealistic high energy portions of conformational space.
The distribution of molecular energies as obtained through allocating the energy by (3.4.13) is depicted in Figure 3.2. We see that the range of energies, roughly 50 kJ mol\(^{-1}\), has been significantly reduced relative to Figure 3.1, restricting our sampling to the lower energy regions of conformational space. We have, however, reduced the discontinuity between the \textit{ab initio} minimum energy and the TYCHE sampling energy by a great deal.

Figure 3.2: Distribution of molecular energies obtained through TYCHE by use of the energy allocation scheme of (3.4.13). The lowest energy point corresponds to the \textit{ab initio} minimum.

This discontinuity remains undesirable, even in Figure 3.2. Ideally, we would like the sampling to be continuous across the molecular energies. To this end, we alter (3.4.13) to read

\[ E_k = \frac{k_B T}{2} \mathcal{U}(0, 1) \mathcal{N}(1, 1). \]  

(3.4.14)

We have added a uniform random variate to modify the temperature, allowing it to take on values lower than that specified by \( T \). In this way, we essentially force
sampling of the lower energy regions of the PES ($U(0, 1) \rightarrow 0$), while retaining the ability to sample the higher energy regions of the PES ($U(0, 1) \rightarrow 1$).

Figure 3.3: Distribution of molecular energies obtained through TYCHE by use of the energy allocation scheme of (3.4.14). The lowest energy point corresponds to the ab initio minimum.

From Figure 3.3, we see that adopting the energy allocation scheme of (3.4.14), we are able to essentially remove the discontinuity in energy between the seeding geometry and the sampling conformations, while still retaining the ability to sample the high energy regions of the well; the range of sampling energies is now roughly 85 kJ mol$^{-1}$. Unfortunately this energy allocation scheme is not equivalent to a particular thermodynamic ensemble, but we find the sampling more in line with our requirements for the construction of a kriging model training set.
Integration Timestep

Our choice for the integration timestep is helped by ascertaining the associated time period of the degree of freedom, $T_k = 2\pi/\omega_k$. For each period of the motion, it seems reasonable to require a number of points, $n_{\text{period}}$, be sampled along the trajectory. Therefore, the integration timestep is given by, $\Delta t_k = T_k/n_{\text{period}}$. Our reason for individually parameterising the integration timestep for each degree of freedom is the range of frequencies spanned by the normal modes (typically greater than an order of magnitude). If we were to set a global integration time step, governed by the highest frequency degree of freedom, then the quantity $q_k(\Delta t) = q_k(t + \Delta t) - q_k(t)$ for a low frequency degree of freedom will be extremely small. By allowing each degree of freedom to have an associated integration timestep governed by its own dynamics, we can ensure an expansive conformational sampling. The reader is directed towards the discussion of the RESPA in Section 3.1, where parallels can be drawn. We note that a single parameter is required to this end, $n_{\text{period}}$, which is user-defined.

Phase

The final free variable in (3.4.10) is the phase of the equation of motion, $\phi$. The meaning associated with this value, as far as we can ascertain, is largely irrelevant. There appears to be no physically intuitive manner to define these phases. The only requirement is that $0 \leq \phi \leq 2\pi$, whose implementation is not strictly necessary since our equations of motion are periodic in $\phi$. As such, each $\phi$ is randomly selected, and is redefined after the output of each conformer.
Sample Output

All free variables have now been given and we are in a position to undertake a conformational sampling. In practice, we do not output each conformer generated for every integration timestep. Conformers from successive integration timesteps are highly correlated with one another, and are typically similar. Ideally, we would like to remove all correlation between outputted conformers. To do so, we evolve the system with an initial parameterisation of the \( \{E_k, \phi\} \). After a complete time period has been completed, the \( \{E_k, \phi\} \) are randomly redefined and the process is iterated. The output of a conformation is governed by, for each time period, selecting a random integer between \([1, n_{\text{period}}]\), and outputting the conformer which has been generated after that number of timesteps. In this way, we avoid the necessity of adding another parameter that governs the frequency with which conformers are output. Since conformers from systems governed by differing \( \{E_k, \phi\} \) are uncorrelated (if the random numbers used to define these values are uncorrelated), the sampled conformers are also uncorrelated.

3.4.3 Markov Chains on Tessellated PESs

Our formulation so far permits a local exploration of the \textit{ab initio} PES (to second order) about a seeding geometry. However, to construct training sets that are of use for MD, we require a more thorough exploration of conformational space. One way to achieve a global exploration of the PES is to use a number of local explorations. In this way, one reconstructs the PES by tessellating with LWs. In the limit of infinite such LWs, one obtains the full \textit{ab initio} PES.

Parallels between the tessellation of the PES by a number of LWs and the con-
struction of a PES from interpolating between points of the PES\cite{15, 257} can be drawn. However, while the interpolation schemes use just the molecular energy (and sometimes first derivatives) for the fit, we use the higher order topological information contained within the Jacobian and Hessian, which is arguably a superior use of \textit{ab initio} data. In addition, the interpolation schemes are typically only viable for small molecular PESs\cite{258, 259}, whereas our method is not particularly bounded by system size.

An issue with the tessellation approach is that a series of LWs cannot be trivially linked into a continuous PES upon which one can conduct classical dynamics. Since each LW is approximated as quadratic in the displacement from the seeding geometry, the sides of the LW cannot be joined in a continuous fashion without interpolating between them in some way. This is perhaps an avenue that could be explored in the future, but is avoided in our current formulation.

Instead, we have decided to forsake the benefits associated with continuous dynamics for a stochastic approach. Indeed, it could be argued that continuous dynamics are of no use for the purposes of conformational sampling. To this end, we have proposed the use of dynamics within a LW, followed by “hopping” to another LW, and iterating this process. The trajectory is then continuous within a well, and discontinuous with respect to hopping events. Naturally, the ideal candidate for the hopping scheme is the Metropolis method, leading to the hopping events taking the form of a Markov chain.

A Markov chain is defined as a stochastic process where the probability of an event (the probability of a state $x_n$ given the past states of the system $x_{n-1}, ..., x_1$, $P(x_n|x_{n-1}, ..., x_1)$) is equal to its probability conditional only on the previous event\cite{260}, i.e.

$$P(x_n|x_{n-1}, ..., x_1) = P(x_n|x_{n-1}).$$

(3.4.15)
In other words, the history of the system is irrelevant in discerning the next state of the system. One can subsequently form the probability of a sequence of states as

\[
P(x_n, ..., x_1) = P(x_n|x_{n-1})P(x_{n-1}|x_{n-2})...P(x_2|x_1).
\]

(3.4.16)

Consider a stochastic matrix of such conditional probabilities, \( \Gamma_{ij} = P(x_j|x_i) \). We require \( \Gamma \) to be row-normalised, i.e. \( \sum_j \Gamma_{ij} = 1 \), which satisfies the requirement that the system can transition from \( x_i \) to another state within the sample space[261]. Naturally, each element of \( 0 \leq \Gamma_{ij} \leq 1 \). There exists a unique left-eigenvector, \( \gamma \), of the stochastic matrix, such that

\[
\gamma^T \cdot \Gamma = \gamma,
\]

(3.4.17)

where \( \gamma \) is a Perron-Frobenius eigenvector, with associated eigenvalue of one. All other eigenvectors of the stochastic matrix possess eigenvalues smaller than the Perron-Frobenius eigenvector. \( \gamma \) then corresponds to an equilibrium state of the system, to which a Markov chain will converge to, given the system is ergodic[262].

We require that any closed trajectory of the form \( P(x_1, x_n, ..., x_1) \) satisfy the Kolmogorov criterion and form a reversible Markov chain, whereby

\[
P(x_1|x_n)P(x_n|x_{n-1})...P(x_2|x_1) = P(x_1|x_2)...P(x_{n-1}|x_n)P(x_n|x_1).
\]

This condition can be enforced by the detailed balance condition\(^3\),

\[
\Gamma_{ij}P(x_i) = \Gamma_{ji}P(x_j).
\]

(3.4.18)

Expanding out the elements of the stochastic matrix and rearranging, we obtain

\[
\frac{P(x_j|x_i)}{P(x_i|x_j)} = \frac{P(x_j)}{P(x_i)}.
\]

(3.4.19)

\(^3\)In fact, the detailed balance condition is excessively strict for our purposes; one can construct Markov processes with Frobenius-Perron eigenvectors that do not satisfy detailed balance. However, detailed balance is a condition that is readily implementable.
We write the elements of the stochastic matrix as a product of two separate terms, \( P(x_j|x_i) = g(x_j|x_i)A(x_j|x_i) \). The proposal distribution, \( g(x_j|x_i) \), and the acceptance distribution, \( A(x_j|x_i) \), correspond to the probability of proposing and accepting the state \( x_j \) from the state \( x_i \), respectively. Rewriting (3.4.19),

\[
\frac{A(x_j|x_i)}{A(x_i|x_j)} = \frac{P(x_j)g(x_i|x_j)}{P(x_i)g(x_j|x_i)}.
\]  

(3.4.20)

Our final step requires that we choose an acceptance probability that satisfies the above condition. We require that the acceptance probability be constrained from above by unity, and so introducing the Metropolis choice,

\[
A(x_j|x_i) = \min \left[ 1, \frac{P(x_j)g(x_i|x_j)}{P(x_i)g(x_j|x_i)} \right].
\]  

(3.4.21)

The Metropolis algorithm is presented in Algorithm 1. For our purposes, we have no condition to place on the proposal distribution, so for simplicity we set it to be the uniform distribution. The acceptance distribution takes the form the the standard Boltzmann weighting scheme, where \( A(x_j|x_i) = \exp[-\Delta E(x_j|x_i)/k_BT_{met}] \), where \( \Delta E(x_j|x_i) = E(x_j) - E(x_i) \), the energy difference between the seeding geometries. We have also introduced \( T_{met} \), the Metropolis temperature, which is distinct from the temperature with which we perform the dynamics. This distinction allows for a more flexible hopping scheme, where the user can artificially increase or decrease the transition probability if more or less (respectively) sampling is required.

3.4.4 Algorithm

Given \( n_{seed} \) seeding geometries, an ab initio calculation is performed for each, and the energy, Jacobian and Hessian are found for each. Our aim is then the generation of a conformational sampling set, \( S \), comprising \( n_{sample} \) distinct molecular conformations.
Select a random starting state, $x_1$
$x_i = x_1 \ ; \ x_j = 0$
while $n_{\text{markov}} < n$ do
  while $x_j = 0$ do
    Propose a new state, $x'$, in accordance with $g(x'|x_i)$
    Accept a transition to $x'$ with probability $A(x'|x_i)$ given by (3.4.21)
    Transition to $x'$ and add to Markov chain, $x_j = x'$
  end
  $x_i = x_j \ ; \ x_j = 0$
  $n_{\text{markov}} = n_{\text{markov}} + 1$
end

Algorithm 1: The Metropolis Algorithm

For each seeding geometry, the transformation matrix $D$ is computed and used to transform the Jacobian and Hessian into a normal coordinate basis. The Hessian is diagonalised, and the Cartesian state vector $x$, angular frequencies $\omega$, reduced masses $\mu$, inhomogeneous terms $\xi$ and matrix $MDE$ are stored.

The sampling itself is conducted by the formation of a Markov chain on the seeding geometries. At each step of the Markov chain, the seeding geometry undergoes vibrational analysis subject to the methodology outlined in Section 3.4.2. Samples are added to $S$ until it contains $n_{\text{sample}}$ conformations, at which point TYPCH terminates.

In Figure 3.4, we give a simple one-dimensional example of the TYPCH methodology. Seven seeding geometries are used, and labelled A-G. About each of these seeds, we approximate the PES as a LW, upon which dynamics can be conducted. It is worth noting that point F constitutes a non-stationary point seed; the dynamics about this point are governed by the motion of a driven harmonic oscillator, driving the minimum of the LW towards the local minimum on the PES.

Of course, we appreciate that the tessellation of LWs becomes a poorer approximation to the $ab$ initio PES in higher dimensions, and as the density of seeding
3.4. TYCHE - CONFORMATIONAL SAMPLING SOFTWARE

Figure 3.4: One-dimensional analytical PES (blue) and the TYCHE-generated PES (red), recreated through the seeding geometries A-G. Stochastic dynamics are performed upon this tessellation of LWs in an attempt to generate a sample set that would be representative of continuous dynamics on the *ab initio* PES.

geometry decreases. In reality, the more seeding geometries included, the better the representation of the PES and the more realistic the dynamics.

For the sake of argument, we commence our sampling within the LW of C. Dynamics are performed within this well and a sample is output. Another seeding point is then selected, B, and an attempt to transition to this LW is attempted. Notice that the transition from C to B is unfavourable since B is of a higher energy than C. The transition to C is subsequently rejected, and another seeding geometry is selected, G. The transition from C to G is favourable since G is lower in energy than C, and so the transition occurs, and dynamics are conducted within the LW of G. This process is iterated until the requested number of samples have been output.

In Table 3.1, we present the memory requirements of running the TYCHE algorithm.
Assume we have a system comprising 100 atoms. Given a modest 2 GB of memory, we are then able to reconstruct the PES with almost 22000 seeding geometries, which is certainly a significant number. Of course, each one of these geometries would need to undergo an \textit{ab initio} calculation in order that the Jacobian and Hessian could be computed for each, which would be prohibitively expensive for 22000 conformers comprising 100 atoms. However, the \textsc{tyche} methodology does not appear to be limited by memory requirements. The algorithm also runs extremely quickly, and so computational time is not prohibitive either.

### 3.5 Redundant Internal Coordinates

#### 3.5.1 Overview

The redundant internal coordinates (RICs) of a molecular system constitute a basis within which the molecular geometry can be full specified, but the number of basis functions exceeds the $3N - 6$ internal degrees of freedom of the system. Historically, RICs have been a useful choice for geometry optimisations of molecular systems. It has been found that expressing the Hessian in a RIC basis minimises the couplings between the degrees of freedom, i.e. the off-diagonal elements of the Hessian.
are minimised. Herein, we will use \( N \) to represent the number of non-redundant degrees of freedom, \( M \) the number of RICs, where \( M = N + R \), \( R \) being the number of redundancies in the RIC system in question[263].

The choice of coordinates with which a molecular conformation is expressed can have profound consequences on a number of computational methodologies. For example, it can be shown that a rectilinear basis, e.g. Cartesians, is not suitable for expressing large displacements involving curvilinear degrees of freedom, such as torsional motions[264]. Attempting to express a displacement in a rectilinear basis can lead to artificial lengthening of a number of other curvilinear degrees of freedom, such as bond stretches[265].

Transforming from an \( N \)-dimensional rectilinear coordinate system, \( \mathbf{x} \), to an \( M \)-dimensional curvilinear coordinate system, \( \mathbf{q} \), requires a Taylor expansion of the curvilinear coordinates in terms of the rectilinear coordinates. Or, for the \( i \)th curvilinear coordinate, \( q_i(x_1, \ldots, x_N) \),

\[
q_i - q_i^0 = A_i + \sum_{j=1}^{N} \frac{\partial q_i}{\partial x_j}(x_j - x_j^0) + \frac{1}{2} \sum_{j,k=1}^{N} \frac{\partial^2 q_i}{\partial x_j \partial x_k}(x_j - x_j^0)(x_k - x_k^0) + \ldots \quad (3.5.1)
\]

where the derivatives are constrained by the evaluation condition \( q = q_0 \). Introducing the difference coordinates, \( \Delta \mathbf{x} = \mathbf{x} - \mathbf{x}^0 \) and \( \Delta \mathbf{q} = \mathbf{q} - \mathbf{q}^0 \), and the partial derivatives

\[
B_{ij} = \left. \frac{\partial q_i}{\partial x_j} \right|_{q=q_0} \quad C_{ijk} = \left. \frac{\partial^2 q_i}{\partial x_j \partial x_k} \right|_{q=q_0}
\]

we can drop the arbitrary shift \( A_i \) and rewrite the above as

\[
\Delta q_i = \sum_{j=1}^{N} B_{ij} \Delta x_j + \frac{1}{2} \sum_{j,k=1}^{N} C_{ijk} \Delta x_j \Delta x_k + \ldots \quad (3.5.2)
\]

\[
\Delta \mathbf{q} = \mathbf{B} \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^\top \mathbf{C} \Delta \mathbf{x} + \ldots
\]

If we consider only infinitesimal displacements in the difference coordinates, then displacements in \( \mathbf{q} \) will become first order with respect to displacements in \( \mathbf{x} \). Thus,
we are able to neglect all second order and higher terms in the above expansion, and subsequently

\[
\delta q_i \approx \sum_{j=1}^{N} B_{ij} \delta x_j
\]

\[
\delta q \approx B \cdot \delta x
\]

(3.5.3)

For infinitesimal displacements, the transformation from a Cartesian to a redundant internal coordinate basis is given by the “Wilson B-matrix”, \( B : \mathbb{R}^N \to \mathbb{R}^M \), whose elements are the derivatives of the redundant internal coordinate with respect to the Cartesian degrees of freedom.

Writing the potential energy of a molecular system as a Taylor series in the RICs,

\[
V(\Delta q) = V_0 + g_q^\top \Delta q + \frac{1}{2} \Delta q^\top H_q \Delta q,
\]

(3.5.4)

where \( g_q \) and \( H_q \) are the gradient and Hessian of the potential energy in the RIC basis, respectively. Substituting the power series in \( \Delta x \) for \( \Delta q \), and equating equal powers in the displacement with a Taylor series of the potential energy in the Cartesian coordinates, we obtain the following transformations\[266, 263\]

\[
g_x = B^\top g_q \quad H_x = B^\top H_q B + g_q^\top C.
\]

(3.5.5)

It is worth pointing out that since the Hessian transforms linearly in the gradient, it is, by definition, a covariant tensor. Then, \( H_q \) requires modification with Christoffel symbols for it to take the form of a tensor comprising second derivatives with respect to the RICs\[267\].

### 3.5.2 Valence Coordinates

The most intuitive set of RICs are perhaps those defined by the connectivity of the molecule. The valence coordinates are such a set, comprising bond stretch, valence
angle and dihedral torsional degrees of freedom, and are depicted in Figure 3.5. We describe each of these in turn and their respective derivatives with respect to the Cartesian degrees of freedom.

In evaluating the derivatives of the valence coordinates, one can adopt either of two approaches. The first is purely analytical, with a great deal of cumbersome algebraic manipulation[268], whereas the second uses the fact that the gradient vector of a scalar field is directed towards the maximum increase in the scalar field, which can be ascertained by inspection[269]. We adopt the former approach since we deal only with simple valence coordinates.

**Bond Stretching**

Given two atomic position vectors, \( \mathbf{x}_\alpha \) and \( \mathbf{x}_\beta \), we define the bond vector between the two as \( \mathbf{x}_{\alpha\beta} = \mathbf{x}_\beta - \mathbf{x}_\alpha \) (i.e. the vector directed from \( \alpha \) to \( \beta \)), and the bond length \( r_{\alpha\beta} = ||\mathbf{x}_{\alpha\beta}|| \). For this case, the RIC is the bond length, i.e. \( q_{\alpha\beta}^{\text{bond}} = r_{\alpha\beta} \).

The derivative of \( q_{\alpha\beta}^{\text{bond}} \) with respect to the Cartesian degrees of freedom of atoms \( \gamma \neq \alpha, \beta \) are obviously zero, since the displacement of these atoms does not effect \( q_{\alpha\beta}^{\text{bond}} \).

The derivatives of \( q_{\alpha\beta}^{\text{bond}} \) with respect to the components of \( \mathbf{x}_\alpha \) are given by the components of the unit vector that maximise \( q_{\alpha\beta}^{\text{bond}} \) when displacing atom \( \alpha \) along the unit vector. By inspection, we verify that the unit vector is simply \( -\mathbf{x}_{\alpha\beta}/r_{\alpha\beta} \). Similarly, the derivative of the components of \( \mathbf{x}_{\alpha\beta} \) with respect to the elements of \( \mathbf{x}_\beta \) are given by the unit vector that maximises \( r_{\alpha\beta} \), which is \( \mathbf{x}_{\alpha\beta}/r_{\alpha\beta} \). Therefore,
Figure 3.5: The conventionally invoked valence coordinates. (a) corresponds to bond stretching, (b) corresponds to valence angle bending, and (c) corresponds to dihedral torsion. The major components that will be required in the following discussion are shown.

The derivatives of the bond stretching RICs are given by

\[
\frac{\partial q_{\text{bond}}}{\partial x_i} = \begin{cases} 
-\frac{(x_{\alpha\beta})_i}{r_{\alpha\beta}} & (x_i \in x_\alpha) \\
\frac{(x_{\alpha\beta})_i}{r_{\alpha\beta}} & (x_i \in x_\beta) \\
0 & \text{otherwise}
\end{cases},
\]

where \((x_{\alpha\beta})_i\) denotes the \(i^{th}\) component of the vector.
3.5. REDUNDANT INTERNAL COORDINATES

Valence Angles

Take three atomic position vectors, \( \mathbf{x}_\alpha, \mathbf{x}_\beta \) and \( \mathbf{x}_\gamma \). This system comprises two bond vectors, \( \mathbf{x}_{\gamma\alpha} \) and \( \mathbf{x}_{\gamma\beta} \), directed from the apex atom \( \gamma \) to the two terminal atoms \( \alpha, \beta \), respectively. As before, the bond lengths of these bond vectors are given by \( r_{\gamma\alpha} = ||\mathbf{x}_{\gamma\alpha}|| \) and \( r_{\gamma\beta} = ||\mathbf{x}_{\gamma\beta}|| \). The valence angle RIC, \( \alpha_{\beta, \gamma} \), is the angle subtended by \( \mathbf{x}_{\gamma\alpha} \) and \( \mathbf{x}_{\gamma\beta} \) from \( \mathbf{x}_\gamma \), and is defined as

\[
\alpha_{\beta, \gamma} = \arccos \left( \frac{\mathbf{x}_{\gamma\alpha} \cdot \mathbf{x}_{\gamma\beta}}{r_{\gamma\alpha} r_{\gamma\beta}} \right) \tag{3.5.7}
\]

The derivatives of \( \alpha_{\beta, \gamma} \) with respect to the components of one of the terminal atoms, say \( \mathbf{x}_\alpha \) for argument, are given by the components of the unit vector that maximise \( \alpha_{\beta, \gamma} \) when displacing \( \alpha \) along the unit vector. By inspection, we see that this unit vector is perpendicular to \( \mathbf{x}_{\gamma\alpha} \) in the plane of the page, and can be expressed as the double cross product \( \mathbf{x}_{\gamma\alpha} \times \mathbf{x}_{\gamma\beta} \times \mathbf{x}_{\gamma\alpha} \). Then, normalising this quantity,

\[
\frac{\partial \alpha_{\beta, \gamma}}{\partial x_i} = \begin{cases} 
\mathbf{x}_{\gamma\alpha} \times \mathbf{x}_{\gamma\beta} \times \mathbf{x}_{\gamma\alpha} / r_{\gamma\alpha} & (x_i \in \mathbf{x}_\alpha) \\
\mathbf{x}_{\gamma\beta} \times \mathbf{x}_{\gamma\alpha} \times \mathbf{x}_{\gamma\beta} / r_{\gamma\beta} & (x_i \in \mathbf{x}_\beta)
\end{cases} \tag{3.5.8}
\]

where the expression for \( x_i \in \mathbf{x}_\beta \) follows by symmetry. A rigid translation of the molecule does nothing to alter the valence angle. Suppose we shift the apex atom \( \gamma \) an infinitesimal amount in one direction, and subsequently shift the entire molecule by an opposite amount; the change in valence angle can then be expressed in terms of the sum of contributions from the shift in the terminal atoms with reversed sign, i.e.

\[
\frac{\partial \alpha_{\beta, \gamma}}{\partial x_i} = -\frac{x_{\alpha\gamma} \times x_{\beta\gamma} \times x_{\alpha\gamma}}{r_{\alpha\gamma}} - \frac{x_{\beta\gamma} \times x_{\alpha\gamma} \times x_{\beta\gamma}}{r_{\beta\gamma}} \quad (x_i \in \mathbf{x}_\gamma) \tag{3.5.9}
\]
Torsion and Additional Valence Coordinates

Take four atomic position vectors, \( \mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_\gamma \) and \( \mathbf{x}_\delta \). The system comprises three bond vectors, \( \mathbf{x}_{\alpha\beta}, \mathbf{x}_{\gamma\delta} \) and \( \mathbf{x}_{\beta\gamma} \), and two valence angles, \( q_{\alpha\gamma,\beta}^{\text{angle}} \) and \( q_{\delta\beta,\gamma}^{\text{angle}} \). The torsional angle, \( q_{\alpha\beta\gamma\delta}^{\text{torsion}} \), is defined as the angle between the planes spanned by \( \{ \mathbf{x}_{\alpha\beta}, \mathbf{x}_{\beta\gamma} \} \) and \( \{ \mathbf{x}_{\gamma\delta}, \mathbf{x}_{\beta\gamma} \} \), as observed along the \( \mathbf{x}_{\beta\gamma} \) bond vector. Analytically, we can express \( q_{\alpha\beta\gamma\delta}^{\text{torsion}} \) as

\[
q_{\alpha\beta\gamma\delta}^{\text{torsion}} = \arccos \left( \frac{(\mathbf{x}_{\alpha\beta} \times \mathbf{x}_{\beta\gamma}) \cdot (\mathbf{x}_{\beta\gamma} \times \mathbf{x}_{\gamma\delta})}{r_{\alpha\beta}r_{\gamma\delta}^2 \sin(q_{\alpha\gamma,\beta}^{\text{angle}}) \sin(q_{\delta\beta,\gamma}^{\text{angle}})} \right).
\]

(3.5.10)

It is hoped that the author will be forgiven for not presenting a formal derivation of the derivatives of this quantity with respect to the Cartesian degrees of freedom of the system. The derivatives can be obtained by the same methodology employed in the previous sections, and so are simply presented

\[
\frac{\partial q_{\alpha\beta\gamma\delta}^{\text{torsion}}}{\partial x_i} = \begin{cases} 
-\frac{\mathbf{x}_{\alpha\beta} \times \mathbf{x}_{\beta\gamma}}{r_{\alpha\beta}^2 r_{\beta\gamma}^2 \sin^2(q_{\alpha\gamma,\beta}^{\text{angle}})} & (x_i \in \mathbf{x}_\alpha) \\
\frac{\mathbf{x}_{\gamma\delta} \times \mathbf{x}_{\beta\gamma}}{r_{\beta\gamma}^2 r_{\alpha\beta}^2 \sin^2(q_{\alpha\gamma,\beta}^{\text{angle}})} - \frac{\mathbf{x}_{\beta\gamma} \times \mathbf{x}_{\gamma\delta}}{r_{\beta\gamma}^2 r_{\gamma\delta}^2 \tan(q_{\delta\beta,\gamma}^{\text{angle}}) \sin(q_{\delta\beta,\gamma}^{\text{angle}})} & (x_i \in \mathbf{x}_\beta) \\
P_{\alpha\delta}P_{\beta\gamma}(x_i \in \mathbf{x}_\beta) & (x_i \in \mathbf{x}_\gamma) \\
P_{\alpha\delta}P_{\beta\gamma}(x_i \in \mathbf{x}_\alpha) & (x_i \in \mathbf{x}_\delta) 
\end{cases}
\]

(3.5.11)

where we have introduced the permutation operator, \( P_{ab} \), which permutes the indices of the function it operates on. In the above case, the permutation operator acts on either the expression for \( (x_i \in \mathbf{x}_\beta) \) or \( (x_i \in \mathbf{x}_\alpha) \), the explicit forms of which have been given.

A number of additional valence coordinates can be defined, such as the four-body angle between a bond and a plane (out-of-plane RIC)[270]. Since, by definition, there is no upper bound on the number of RICs employed to define a molecular system, one can introduce as many valence coordinates as desired. However, we note
3.5. REDUNDANT INTERNAL COORDINATES

that for geometry optimisation, simply adding more RICs results in more coupling between the degrees of freedom, and so the Hessian loses its sparse characteristic.

3.5.3 Implementation

The normal modes of motion with which we evolve the system are inconvenient if we wish to direct our sampling to physically meaningful regions of conformational space. For example, if we are able to identify that we require a rigorous sampling of a certain internal coordinate, we are unable to predominantly sample that degree of freedom in a normal mode basis.

However, if we are able to decompose the normal modes of a system into a valence coordinate basis, we would be able to select those normal modes that are dominated by certain physically meaningful motions, and direct our sampling by selectively evolving those normal modes.

Decomposing the normal modes of a system into contributions from valence coordinates is quite a simple process. Recall in Section 3.4.1 we introduced the quantity \( \mathbf{MDE} \) as the Cartesian displacements associated with each normal mode. Since a Cartesian state vector transforms to a RIC basis through (3.5.3), we simply require evaluation of the matrix multiplication \( \mathbf{BMDE} \) to transform the normal modes into a valence coordinate basis.

Demonstrating the results of this computation, we recall that the three normal modes of water are the “symmetric stretch”, “asymmetric stretch” and “valence angle bending” modes of motion. The first two involve only bond stretching degrees of freedom, while the third involves predominantly the angle bending degree of freedom, along with a small amount of bond stretching. Figure 3.6 shows the
output of TYCHE concerning the three normal modes of water.

Figure 3.6: TYCHE output outlining the normal mode information of water. Normal modes 1, 2 and 3 correspond to the valence angle bending, symmetric stretching and asymmetric stretching modes, respectively. The RIC contributions to each normal mode are given at the bottom of the figure.

<table>
<thead>
<tr>
<th>Normal Mode</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1622.4800</td>
<td>3803.8164</td>
<td>3938.8590</td>
</tr>
<tr>
<td>Reduced Mass</td>
<td>1.0823</td>
<td>1.0455</td>
<td>1.812</td>
</tr>
<tr>
<td>Force Constant</td>
<td>1.6777</td>
<td>8.9071</td>
<td>9.8733</td>
</tr>
<tr>
<td>Inhomogeneity</td>
<td>0.800005</td>
<td>-0.000014</td>
<td>0.000018</td>
</tr>
</tbody>
</table>

Normal Coordinates

<table>
<thead>
<tr>
<th>Atom Number</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>O 1</td>
<td>-0.00</td>
<td>0.00</td>
<td>0.07</td>
<td>-0.00</td>
<td>0.00</td>
<td>-0.06</td>
<td>0.00</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>H 2</td>
<td>0.00</td>
<td>0.43</td>
<td>-0.56</td>
<td>0.00</td>
<td>0.58</td>
<td>0.48</td>
<td>-0.00</td>
<td>-0.56</td>
<td>-0.43</td>
</tr>
<tr>
<td>H 3</td>
<td>0.00</td>
<td>-0.43</td>
<td>-0.56</td>
<td>-0.00</td>
<td>-0.58</td>
<td>0.48</td>
<td>-0.00</td>
<td>-0.56</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Redundant Internal Coordinates

<table>
<thead>
<tr>
<th>Number/ Type</th>
<th>Contribution</th>
<th>Contribution</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Stretch)</td>
<td>8.8922</td>
<td>0.4973</td>
<td>6.9600</td>
</tr>
<tr>
<td>2 (Stretch)</td>
<td>8.8922</td>
<td>0.4973</td>
<td>6.9600</td>
</tr>
<tr>
<td>3 (Angle)</td>
<td>8.8922</td>
<td>0.4973</td>
<td>6.9600</td>
</tr>
</tbody>
</table>

The section labelled “Normal Coordinates” reveals the Cartesian displacements associated with each normal mode, as computed through MDE. The section labelled “Redundant Internal Coordinates” denotes the valence coordinate contributions to each normal mode. For the low frequency mode, we find that 89% of the motion corresponds to the valence angle bending, while the remaining 11% results from bond stretching. For the two higher frequency modes, we find that the overwhelming majority of motion derives from bond stretching.

This methodology is applicable to any system, allowing us to propose a simple means by which we can selectively sample regions of valence coordinate space. The TYCHE input has two keywords: “WhichModes” and “MaxModes”. “WhichModes” can take three separate values; “Bonds”, “Angles” and “Dihedrals”. When one of these options is selected, TYCHE will take the “MaxModes” normal modes with the largest average contribution from the valence coordinates specified by
“WhichModes”. For the sake of demonstration, if the user set “MaxModes = 10” and “WhichModes = Dihedrals”, TYPHE will sample the 10 normal modes with the highest contribution from dihedral valence coordinates.

One can envisage a number of ways in which TYPHE can sample using this information. For instance, it could bias the total energy into a subset of the normal modes, while still putting some energy into the remaining normal modes. Currently, TYPHE will only vibrate “MaxModes” normal modes. We demonstrate the power of this methodology in Section 3.6.3.

3.6 Results

3.6.1 Validation and Benchmarking

In the following, we present a systematic parameterisation of those free variables we have alluded to in previous sections. We have attempted to makes this section as terse and condensed as possible for the sake of brevity, and so necessarily omit the results of auxiliary experiments, leaving only those results of significance.

Parameterisation of $n_{period}$

Our first benchmarking study assesses the effects of choice of $n_{period}$. Our aim is to obtain as much conformational sampling as possible, and so we have chosen to determine the optimum choice of $n_{period}$ to be that which maximises the degree of conformational sampling. In Tables 3.2, 3.3 and 3.4, we present the range and standard deviations of a prominent degree of freedom for water (valence
angle), NMA (peptide dihedral) and zwitterionic histidine (sidechain dihedral), where the chosen degrees of freedom are specified in the respective parentheses. The CPU time required to obtain conformational ensembles comprising 4000 conformers is also presented. The Jacobian and Hessian have been calculated at the B3LYP/aug-cc-pVDZ level of theory. We have used a temperature of 1750 K for the sampling, which seems excessive, but is in fact arbitrary for our purposes, and simply amplifies the results obtained at lower temperatures.

<table>
<thead>
<tr>
<th>( n_{\text{period}} )</th>
<th>( \text{Range (degrees)} )</th>
<th>( \text{Standard Deviation (degrees)} )</th>
<th>( \text{CPU Time (seconds)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 1 \times 10^1 )</td>
<td>34.3</td>
<td>5.3</td>
<td>0.116</td>
</tr>
<tr>
<td>( 1 \times 10^2 )</td>
<td>32.9</td>
<td>5.3</td>
<td>0.204</td>
</tr>
<tr>
<td>( 1 \times 10^3 )</td>
<td>33.3</td>
<td>5.3</td>
<td>1.243</td>
</tr>
<tr>
<td>( 1 \times 10^4 )</td>
<td>33.3</td>
<td>5.3</td>
<td>11.505</td>
</tr>
<tr>
<td>( 1 \times 10^5 )</td>
<td>33.3</td>
<td>5.3</td>
<td>113.318</td>
</tr>
</tbody>
</table>

Table 3.2: Conformational sampling of water at a number of values of \( n_{\text{period}} \).

<table>
<thead>
<tr>
<th>( n_{\text{period}} )</th>
<th>( \text{Range (degrees)} )</th>
<th>( \text{Standard Deviation (degrees)} )</th>
<th>( \text{CPU Time (seconds)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 1 \times 10^1 )</td>
<td>108.6</td>
<td>19.3</td>
<td>0.807</td>
</tr>
<tr>
<td>( 1 \times 10^2 )</td>
<td>107.5</td>
<td>19.3</td>
<td>1.899</td>
</tr>
<tr>
<td>( 1 \times 10^3 )</td>
<td>106.9</td>
<td>19.3</td>
<td>12.469</td>
</tr>
<tr>
<td>( 1 \times 10^4 )</td>
<td>106.9</td>
<td>19.3</td>
<td>114.530</td>
</tr>
<tr>
<td>( 1 \times 10^5 )</td>
<td>106.9</td>
<td>19.3</td>
<td>1152.896</td>
</tr>
</tbody>
</table>

Table 3.3: Conformational sampling of NMA at a number of values of \( n_{\text{period}} \).
3.6. RESULTS

<table>
<thead>
<tr>
<th>$n_{\text{period}}$</th>
<th>Range (degrees)</th>
<th>Standard Deviation (degrees)</th>
<th>CPU Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1 \times 10^1$</td>
<td>56.8</td>
<td>5.6</td>
<td>3.416</td>
</tr>
<tr>
<td>$1 \times 10^2$</td>
<td>52.0</td>
<td>5.6</td>
<td>6.081</td>
</tr>
<tr>
<td>$1 \times 10^3$</td>
<td>52.1</td>
<td>5.6</td>
<td>33.423</td>
</tr>
<tr>
<td>$1 \times 10^4$</td>
<td>52.1</td>
<td>5.6</td>
<td>298.152</td>
</tr>
<tr>
<td>$1 \times 10^5$</td>
<td>52.2</td>
<td>5.6</td>
<td>2867.963</td>
</tr>
</tbody>
</table>

Table 3.4: Conformational sampling of histidine at a number of values of $n_{\text{period}}$.

Our results are conclusive, in that the value of $n_{\text{period}}$ is largely irrelevant, in both its magnitude and the system in which it is used. In fact, the range of conformational sampling appears to decrease slightly as $n_{\text{period}}$ increases. This is a particularly useful trend, since the time required to generate the conformational ensemble increases linearly in $n_{\text{period}}$, and so becomes prohibitive for large systems. Therefore, it seems advisable to set $n_{\text{period}}$ to some suitably low value. For the remainder of our work, we will use $n_{\text{period}} = 10$.

**Level of Theory**

Next, we investigate the effects of the level of theory with which the Jacobian and Hessian are computed. Once again, our criterion for selecting the optimal level of theory is the degree of conformational sampling we are able to obtain with each level of theory. This study could be all the more exhaustive by investigating level of theory/ basis set combinations, but we feel this to be superfluous. Our systems of choice are as above (water, NMA and zwitterionic histidine), where the same degrees of freedom are evaluated. We have used the aug-cc-pVDZ basis set for each test case, and once again a temperature of 1750 K.
Computing the Jacobian and Hessian at the Hartree Fock level of theory seems to quite severely limit the amount of conformational space available to the systems. In water, only the lowest frequency mode is able to sample the valence angle degree of freedom.
of freedom, so by looking at the force constant of this mode across the levels of theory, we are able to see the reason why Hartree Fock is so restrictive. The force constants associated with the valence angle mode are 2.10, 1.70, 1.68 and 1.68 mDyne Å$^{-1}$ for the HF, B3LYP, M06 and MP2 levels of theory, respectively. This higher force constant prohibits sampling of the valence angle mode given the same amount of energy, and so conformational flexibility is reduced.

It seems as if B3LYP allows for the greatest conformational flexibility across the levels of theory investigated, certainly with regards to the NMA peptide dihedral. Both M06 and MP2 perform relatively well, but appear to be inconsistent in the amount of conformational flexibility that they allow. As such, we recommend the use of the B3LYP level of theory if the aim is to obtain as diverse a conformational sampling as possible.

**Temperature**

Our final parameterisation involves ascertaining the temperature at which we can construct a conformational ensemble spanning the relevant regions of conformational space. An approximation as to what constitutes the relevant regions of conformational space is obtained by benchmarking against the conformational ensembles used in the construction of alternative force fields. A particularly useful set of papers to this effect present a number of useful statistics regarding the conformational ensembles used to parameterise a generic force field for use with alkanes[271], organic compounds[272], peptides[273], carbohydrates[274] and ionic species[275]. The conformational ensembles in these studies have been constructed by a systematic sampling along each normal mode of motion for given molecules in their global minimum energy conformations.
In the following discussion, we make reference to the “temperature”, which can
be ambiguous since its invocation can be used in the context of an MD trajectory
or the temperature used in tyche as defined in Section 3.4.2. When discussing
the temperature at which an MD trajectory has been computed, we explicitly
write “MD temperature”, or a phrase to that effect. If the word temperature is
used without any further clarification, the sampling temperature used for tyche
is implied.

We immediately reject the notion that using a temperature equal to that at which
the desired MD trajectory is to be computed is adequate for the construction of
a valid conformational ensemble. Our approximation of the LW necessitates a
revised attitude to our use of temperature. The amount of energy\(^4\) required to dis-
tort a system from the seeding geometry increases quadratically in the distortion.
Escape from the LW is therefore never permitted, whilst in reality, after a short
displacement from the seeding geometry, the potential energy need not increase
monotonically. Unless the seeding geometry lies within a particularly deep well
on the true PES, it is highly likely that a true local exploration of conformational
space about the seeding geometry is equivalent to a high temperature exploration
of the LW.

In the following, we benchmark against three separate sets of data. The first two
sets of data derive from the aforementioned references involving the parameteri-
sation of a generic force field for use with alkanes\(^271\) and peptides\(^273\). Both of
these papers use seventeen distinct molecules for sampling, but we have chosen four
from each, representing a cross-section of the types of molecules studied. For the
alkanes, we have selected methane, ethane, neopentane and cyclopentane, while

\(^4\)The notion of “energy” and “temperature” are interchangeable since they are linearly propor-
tional. Our discussion is invariant with regards to the proportionality constant linking these
two quantities.
for the peptides we have selected formamide, N-methylacetamide, methylazacyclopropanone and N,N-dimethyl formamide The third set of data is a MD trajectory for zwitterionic alanine that we have computed using the GROMACS MD package in conjunction with the AMBER99SB force field. The MD trajectory has been computed at 300 K in the NPT ensemble with a Parrinello-Rahman barostat. We are brief on these details since they are identical to the MD trajectories that are computed in Section 6.4.1, where the reader is directed if they wish for a thorough description.

We have generated conformational ensembles using TYCHE at a number of temperatures. For each computation, we have used $n_{\text{period}} = 10$ and computed the Hessian and Jacobian at the B3LYP/aug-cc-pVDZ level of theory. We use only the global minimum as a seeding geometry for the molecules enumerated above. For this reason, we only benchmark against bond stretching and valence angle bending degrees of freedom since the torsional degrees of freedom are only properly sampled by seeding with a number of distinct molecular geometries (which we tackle in Section 3.6.2). We have used 500 conformers in the construction of each conformational ensemble obtained through TYCHE.

The resultant conformational ensemble computed using TYCHE is compared to that obtained from the previously discussed sources. We have decided to use the range of sampling of a number of selected degrees of freedom for benchmarking. For example, suppose a system possesses a number of similar bonds (e.g. CH, NH, etc.). The range of this degree of freedom is then defined by the maximum and minimum sampled values of that bond length across all similar bonds within the system. It is worth noting that our goal is not to faithfully recreate the ranges of sampling in references [271, 273] and the zwitterionic alanine MD trajectory. Instead, we see these as minimum requirements for sampling, and indicative of the
relevant ranges of values taken by those degrees of freedom.

For the alkanes, the degrees of freedom against which we benchmark are the CH bond stretching and HCH valence angle bending degrees of freedom. Alkane sampling is summarised in Figures 3.7 and 3.8. These diagrams show the range of sampling as obtained through TycHE of the CH bond stretch and HCH valence angle bending degrees of freedom, respectively, for the four alkanes we have mentioned. As the temperature is increased, the range of sampling increases, as expected. Also shown in these figures are the maximum and minimum values of the corresponding degrees of freedom taken by over 99% of conformers in reference [271]. We see that for the CH bond stretch, we sample roughly the same range of bond lengths as the referenced work at a temperature of 900 K. Similarly, for the HCH valence angle bend, we sample roughly the same range of valence angles as the reference work at a temperature of 900 K.

Moving onto the peptides, the degrees of freedom against which we benchmark are the CN bond stretching and OCN valence angle bending degrees of freedom. Peptide sampling is outlined in Figures 3.9 and 3.10. These diagrams show the range of sampling as obtained through TycHE of the CN bond stretch and OCN valence angle bending degrees of freedom, respectively, for the four peptides we have mentioned.
Figure 3.7: Range of CH bond stretch degrees of freedom in methane (green), ethane (red), neopentane (yellow) and cyclopentane (blue) as the temperature is increased. The black lines correspond to the range of sampling undertaken in Maple et al.

Figure 3.8: Range of HCH valence angle degrees of freedom in methane (green), ethane (red), neopentane (yellow) and cyclopentane (blue) as the temperature is increased. The black lines correspond to the range of sampling undertaken in Maple et al.
Figure 3.9: Range of CN bond stretch degrees of freedom in formamide (green), N-methylacetamide (red), methylazocyclopropanone (yellow) and N,N-dimethylformamide (blue) as the temperature is increased. The black lines correspond to the range of sampling undertaken in Maple et al.

Figure 3.10: Range of OCN valence angle degrees of freedom in formamide (green), N-methylacetamide (red), methylazocyclopropanone (yellow) and N,N-dimethylformamide (blue) as the temperature is increased. The black lines correspond to the range of sampling undertaken in Maple et al.
3.6. RESULTS

From Figure 3.9, it is evident that TYCHE vastly oversamples the higher end of the CN bond length range relative to reference [273], particularly the methylazocyclopropanone species. Contrastingly, the lower end of the CN bond length range is not sampled at all through TYCHE. A similar result is seen in Figure 3.10, where we only really sample the full range as obtained in reference [273] by 1800 K. Even at 1800 K, half of the range is sampled by a single molecular species, methylazocyclopropanone, and so our sampling can hardly be considered successful.

We note that the sum of the covalent radii of carbon and nitrogen is roughly 1.4 Å. However, the range of sampling claimed in reference [273] for the CN bond length is 1.22 – 1.51 Å. To claim that the range of sampling of this degree of freedom is bounded from below by a value that is roughly 20% smaller than the sum of the atomic covalent radii seems spurious. Indeed, from the MD trajectory of zwitterionic alanine presented later in this section, the sampling range of the CN bond length is roughly 1.35 – 1.60 Å, almost perfectly in keeping with the results that we have obtained in Figure 3.9 at a temperature of roughly 600 K. Similarly, the equilibrium peptide bond angle OCN is roughly 120° is the majority of systems, whereas the range of sampling claimed in reference [273] for this degree of freedom is 114 – 146°. Once more, claiming that the range of sampling of this degree of freedom is bounded from above by a value greater than 20% of the equilibrium bond angle seems excessive.

We stop short of claiming the data given in [273] is incorrect. Analysis of Figure 3.10 sheds a little more light on the discrepancies we have encountered. For example, two groups of peptide species sample completely opposite ends of the sampling range; formamide, N-methylacetamide and N,N-dimethylformamide sample the low end of the range, while methylazocyclopropanone samples the other, with
virtually no overlap between the two groups. This result can be immediately accounted for since the nitrogen in methylazocyclopropanone is situated in a three-membered ring, and thus its motion is severely constrained relative to the other molecules. Indeed, the three-membered ring sterically clashes with the ketone portion of the molecule when the OCN valence angle is $120^\circ$, and so the equilibrium OCN valence angle is more obtuse.$^5$

In [273], a sample set of seventeen distinct molecules are used. We therefore presume that given the full sampling of each of these molecules, one could sample the full range of the CN bond stretching and OCN valence angle bending degrees of freedom reported. However, since we only consider a small subset of the molecules used, it stands to reason that we are not able to sample the full range reported. Unfortunately, without an analysis of the seventeen molecules used in [273], it is difficult to be conclusive about the temperature that should be used for sampling with TYCHE.

$^5$One could question why this discrepancy is not evident with the alkanes we have analysed. In response, we indicate that alkanes are typically very similar, and are not amenable to adopting conformations the disrupt the tetrahedral bonding pattern of carbon.
Finally, we use an \textit{in vacuo} MD trajectory of zwitterionic alanine (See Figure 3.11 for an overview of the atomic labelling scheme used herein) to parameterise the temperature used for conformational sampling. We have selected a number of prominent degrees of freedom whose sampling ranges we will benchmark against. The MD trajectory was analysed with the in-house software \textsc{hermes}\[276]. The bond stretching degrees of freedom that we use are N1-H2, C1-C2, N1-C1 and C3-O1. The valence angle bending degrees of freedom that we use are O1-C3-O2, H2-N1-H3, N1-C1-C3 and H5-C2-H6. The sampling ranges of these degrees of freedom are presented in Tables 3.8 and 3.9 as the sampling temperature is increased. Also provided are the ranges taken by these degrees of freedom over the course of the MD trajectory.
### Table 3.8: Sampling ranges of four prominent bond stretching degrees of freedom in zwitterionic alanine as the sampling temperature is increased. The *ab initio* Hessian and Jacobian are taken from gas phase frequency calculations. In the bottom row, we include the ranges of these degrees of freedom as obtained from a 300 K *in vacuo* MD trajectory. In red, we have refined these sampling ranges down so that they represent the sampling ranges taken by over 98% of conformers over the course of the MD trajectory.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>N1-H2</th>
<th>C1-C2</th>
<th>N1-C1</th>
<th>C3-O1</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>[1.08 - 0.98]</td>
<td>[1.59 - 1.48]</td>
<td>[1.56 - 1.47]</td>
<td>[1.29 - 1.24]</td>
</tr>
<tr>
<td>600</td>
<td>[1.10 - 0.97]</td>
<td>[1.62 - 1.47]</td>
<td>[1.58 - 1.46]</td>
<td>[1.30 - 1.23]</td>
</tr>
<tr>
<td>900</td>
<td>[1.13 - 0.95]</td>
<td>[1.64 - 1.45]</td>
<td>[1.60 - 1.44]</td>
<td>[1.31 - 1.22]</td>
</tr>
<tr>
<td>1200</td>
<td>[1.17 - 0.94]</td>
<td>[1.65 - 1.44]</td>
<td>[1.61 - 1.44]</td>
<td>[1.31 - 1.21]</td>
</tr>
<tr>
<td>1500</td>
<td>[1.20 - 0.94]</td>
<td>[1.67 - 1.43]</td>
<td>[1.63 - 1.43]</td>
<td>[1.32 - 1.21]</td>
</tr>
<tr>
<td>1800</td>
<td>[1.24 - 0.93]</td>
<td>[1.68 - 1.42]</td>
<td>[1.64 - 1.42]</td>
<td>[1.32 - 1.20]</td>
</tr>
<tr>
<td>MD</td>
<td>[1.15 - 0.92]</td>
<td>[1.65 - 1.43]</td>
<td>[1.63 - 1.35]</td>
<td>[1.35 - 1.19]</td>
</tr>
<tr>
<td></td>
<td>[1.11 - 0.96]</td>
<td>[1.61 - 1.45]</td>
<td>[1.55 - 1.43]</td>
<td>[1.30 - 1.21]</td>
</tr>
</tbody>
</table>

We see from these tables that we recover the upper bound on the sampling range for the majority of the degrees of freedom by 900-1200 K, but do not recover the lower bound, if at all, even by a temperature of 1800 K. However, we have also presented the ranges of sampling from the MD trajectory by over 98% of the conformational ensemble, and we find these ranges to be much more agreeable with our results. Indeed, the majority of ranges are recovered by a temperature of roughly 900 K, which is entirely in keeping with the results we obtained for the alkanes.

We propose that the reason for the full MD sample spanning a larger range than that obtained by *Tyche* is that the molecule is permitted to undergo large conformational changes over the course of an MD trajectory. As we have mentioned,
3.6. RESULTS

<table>
<thead>
<tr>
<th>Temperature</th>
<th>O1-C3-O2</th>
<th>H2-N1-H3</th>
<th>N1-C1-C3</th>
<th>H5-C2-H6</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>[131.5 – 124.8]</td>
<td>[120.2 – 99.9]</td>
<td>[115.8 – 107.7]</td>
<td>[116.8 – 101.1]</td>
</tr>
<tr>
<td>600</td>
<td>[132.7 – 123.3]</td>
<td>[125.2 – 96.6]</td>
<td>[117.5 – 106.1]</td>
<td>[120.2 – 98.2]</td>
</tr>
<tr>
<td>900</td>
<td>[133.6 – 122.1]</td>
<td>[128.9 – 94.0]</td>
<td>[118.9 – 104.9]</td>
<td>[122.8 – 96.0]</td>
</tr>
<tr>
<td>1200</td>
<td>[134.3 – 121.2]</td>
<td>[132.1 – 91.9]</td>
<td>[120.1 – 103.9]</td>
<td>[124.9 – 94.1]</td>
</tr>
<tr>
<td>1500</td>
<td>[135.0 – 120.3]</td>
<td>[134.8 – 90.0]</td>
<td>[121.1 – 103.0]</td>
<td>[126.7 – 92.4]</td>
</tr>
<tr>
<td>1800</td>
<td>[135.6 – 119.6]</td>
<td>[137.2 – 88.4]</td>
<td>[122.0 – 102.2]</td>
<td>[128.4 – 90.9]</td>
</tr>
<tr>
<td>MD</td>
<td>[134.3 – 115.8]</td>
<td>[128.3 – 90.2]</td>
<td>[117.6 – 97.3]</td>
<td>[126.2 – 92.8]</td>
</tr>
<tr>
<td></td>
<td>[131.0 – 120.6]</td>
<td>[123.2 – 96.0]</td>
<td>[117.3 – 103.4]</td>
<td>[121.3 – 95.4]</td>
</tr>
</tbody>
</table>

Table 3.9: Sampling ranges of four prominent valence angle bending degrees of freedom in zwitterionic alanine as the sampling temperature is increased. The \textit{ab initio} Hessian and Jacobian are taken from gas phase frequency calculations. In the bottom row, we include the ranges of these degrees of freedom as obtained from a 300 K \textit{in vacuo} MD trajectory. In red, we have refined these sampling ranges down so that they represent the sampling ranges taken by over 98% of conformers over the course of the MD trajectory.

TYCHE is far more restrictive when seeded with a single conformer, and so does not permit such large-scale conformational changes. Allowing the system to adopt conformations that differ significantly from the seeding geometry results in a number of atomic interactions not present at the seeding geometry. For instance, a steric clash in one conformation may weaken an interatomic bond in the molecule, and grant the bond a great deal more flexibility (or restrict the bond further), and subsequently lead to an entirely different sampling range. Alternatively, an intramolecular hydrogen bond in one conformation may restrict a number of degrees of freedom relative to the conformer with no intramolecular hydrogen bond.

To address this issue, we require the use of a number of seeding geometries to permit the exploration of a greater extent of conformational space. This will be undertaken in Section 3.6.2. However, 900 K appears to be a sensible choice of temperature to reproduce the vast majority of sampling ranges we have encoun-
tered across a wide variety of molecular systems, and so we advocate its use when sampling with TYCHE.

As a matter of interest, we have also acquired an explicitly solvated MD trajectory of zwitterionic alanine with TIP3P explicit water molecules. The presence of water should dampen down the electrostatic interaction between atoms, and therefore alter the range of vibrations for a number of degrees of freedom. To sample a conformational ensemble that is representative of a solvated system with the TYCHE methodology, we would ideally require an \textit{ab initio} Hessian and Jacobian from the explicitly solvated system. However, this is obviously not feasible for systems involving even a moderate number of explicit water molecules.

As such, we have computed the \textit{ab initio} Hessian and Jacobian for the zwitterionic alanine embedded in a polarisable continuum implicit solvation field ($\epsilon = 80$). We expect this field to dampen the motion of any interactions involving partially charged atoms. As such, we anticipate that the upper bounds to the sampling ranges of the previously sampled degrees of freedom will be shifted higher. The sampling of these degrees of freedom are presented in Tables 3.10 and 3.11.
3.6. **RESULTS**

<table>
<thead>
<tr>
<th>Temperature</th>
<th>N1-H2</th>
<th>C1-C2</th>
<th>N1-C1</th>
<th>C3-O1</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>[1.09 – 0.98]</td>
<td>[1.58 – 1.49]</td>
<td>[1.56 – 1.48]</td>
<td>[1.30 – 1.25]</td>
</tr>
<tr>
<td>600</td>
<td>[1.15 – 0.97]</td>
<td>[1.60 – 1.48]</td>
<td>[1.58 – 1.47]</td>
<td>[1.32 – 1.24]</td>
</tr>
<tr>
<td>900</td>
<td>[1.21 – 0.95]</td>
<td>[1.62 – 1.47]</td>
<td>[1.60 – 1.46]</td>
<td>[1.33 – 1.23]</td>
</tr>
<tr>
<td>1200</td>
<td>[1.27 – 0.94]</td>
<td>[1.64 – 1.46]</td>
<td>[1.61 – 1.45]</td>
<td>[1.34 – 1.22]</td>
</tr>
<tr>
<td>1500</td>
<td>[1.32 – 0.93]</td>
<td>[1.65 – 1.45]</td>
<td>[1.62 – 1.44]</td>
<td>[1.35 – 1.22]</td>
</tr>
<tr>
<td>1800</td>
<td>[1.38 – 0.93]</td>
<td>[1.66 – 1.44]</td>
<td>[1.63 – 1.44]</td>
<td>[1.35 – 1.21]</td>
</tr>
<tr>
<td><strong>MD</strong></td>
<td>[1.13 – 0.97]</td>
<td>[1.68 – 1.40]</td>
<td>[1.63 – 1.36]</td>
<td>[1.33 – 1.19]</td>
</tr>
<tr>
<td></td>
<td><strong>[1.12 – 0.97]</strong></td>
<td><strong>[1.62 – 1.47]</strong></td>
<td><strong>[1.58 – 1.43]</strong></td>
<td><strong>[1.32 – 1.22]</strong></td>
</tr>
</tbody>
</table>

Table 3.10: Sampling ranges of four prominent bond stretching degrees of freedom in zwitterionic alanine as the sampling temperature is increased. The *ab initio* Hessian and Jacobian are taken from implicitly solvated (CPCM) frequency calculations. In the bottom row, we include the ranges of these degrees of freedom as obtained from a 300 K TIP3P-solvated MD trajectory. In red, we have refined these sampling ranges down so that they represent the sampling ranges taken by over 98% of conformers over the course of the MD trajectory.
Table 3.11: Sampling ranges of four prominent valence angle bending degrees of freedom in zwitterionic alanine as the sampling temperature is increased. The ab initio Hessian and Jacobian are taken from implicitly solvated (CPCM) frequency calculations. In the bottom row, we include the ranges of these degrees of freedom as obtained from a 300 K TIP3P-solvated MD trajectory. In red, we have refined these sampling ranges down so that they represent the sampling ranges taken by over 98% of conformers over the course of the MD trajectory.

<table>
<thead>
<tr>
<th>Temperature</th>
<th>O1-C3-O2 (degrees)</th>
<th>H2-N1-H3 (degrees)</th>
<th>N1-C1-C3 (degrees)</th>
<th>H5-C2-H6 (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>[130.9 – 125.9]</td>
<td>[117.2 – 101.2]</td>
<td>[115.2 – 107.1]</td>
<td>[116.4 – 101.0]</td>
</tr>
<tr>
<td>600</td>
<td>[131.8 – 124.8]</td>
<td>[120.9 – 98.6]</td>
<td>[116.7 – 105.4]</td>
<td>[119.5 – 98.0]</td>
</tr>
<tr>
<td>900</td>
<td>[132.5 – 123.9]</td>
<td>[123.8 – 96.7]</td>
<td>[117.8 – 104.0]</td>
<td>[121.7 – 95.8]</td>
</tr>
<tr>
<td>1200</td>
<td>[133.1 – 123.1]</td>
<td>[126.2 – 95.2]</td>
<td>[118.8 – 102.9]</td>
<td>[123.6 – 93.9]</td>
</tr>
<tr>
<td>1500</td>
<td>[133.7 – 122.4]</td>
<td>[128.3 – 93.9]</td>
<td>[119.6 – 101.9]</td>
<td>[125.1 – 92.3]</td>
</tr>
<tr>
<td>1800</td>
<td>[134.1 – 121.8]</td>
<td>[130.1 – 92.8]</td>
<td>[120.4 – 101.0]</td>
<td>[126.5 – 90.8]</td>
</tr>
<tr>
<td>MD</td>
<td>[129.4 – 112.9]</td>
<td>[122.6 – 93.6]</td>
<td>[127.1 – 100.9]</td>
<td>[124.7 – 91.0]</td>
</tr>
<tr>
<td></td>
<td>[127.3 – 114.1]</td>
<td>[122.0 – 97.2]</td>
<td>[120.3 – 105.3]</td>
<td>[120.9 – 95.6]</td>
</tr>
</tbody>
</table>
3.6. RESULTS

For a number of the degrees of freedom, we see that there is typically a poor overlap between the sampling ranges obtained from TYCHE and the MD trajectory until the TYCHE sampling temperature is set excessively high. Once again, this is improved significantly by restricting the MD sampling range to include 98% of conformers from the MD trajectory, at which point the optimal TYCHE sampling temperature appears to be roughly 900 K, in keeping with our previously acquired results. However, this restriction is not as successful as when implemented with the in vacuo conformers. For example, the N1-H2 bond stretching and O1-C3-O2 valence angle bending sampling ranges still do not match up with those obtained through TYCHE.

We note that those degrees of freedom with poor disagreement with the TYCHE conformational ensemble involve the more partially charged atoms, i.e. the zwitterionic components of the molecule. One can assume that our inability to correctly sample these degrees of freedom correctly is resultant from the lack of explicit solvent interactions with the TYCHE sampling. In an explicitly solvated environment, these atoms predominantly form hydrogen bonds with the solvent, which subsequently impacts on the sampling ranges of the internal degrees of freedom involving these atoms. One is then unable to recreate these sampling ranges with TYCHE unless some explicit solvation is included. However, it is certainly encouraging that we can recover the sampling ranges of the other degrees of freedom quite readily, suggesting that TYCHE, parameterised with an implicitly solvated Hessian and Jacobian, is readily adaptable to solvated systems that don’t interact with the solvent as strongly.
3.6.2 Markov Chain Conformational Sampling

Previous work in our group has located the local minima of each of the naturally occurring amino acids \textit{in vacuo}[277]. In the referenced work, the amino acids are “capped” with peptide bonds at both the N- and C-terminii. The capping removes the capacity for the amino acids to take zwitterionic forms. For capped alanine, eleven distinct energetic minima were found. So that we can compare with the MD trajectory obtained for zwitterionic alanine in the previous section, we have removed the capping groups from the capped alanine minima. The zwitterionic conformers were subsequently optimised at the B3LYP/aug-cc-pVDZ level of theory with implicit solvation so that the zwitterion remains stable. Upon optimisation, we have performed single point frequency calculations on the optimised conformers \textit{in vacuo} to obtain the gas phase Hessian and Jacobian.

\textsc{tyche} sampling has been undertaken at a temperature of 900 K, to remain consistent with our findings in Section 3.6.1. To allow for transition between the seeding geometries, we have selected a transition temperature of 300 K, since we see no reason why our LW approximation should require a higher temperature. Out of the original eleven minima, five were found to have Boltzmann weights exceeding $10^{-3}$. The remaining six minima were discarded from our sampling as seeding geometries.

In Tables 3.12 and 3.13, we present the sampling ranges of the bond stretching and valence angle bending degrees of freedom used in Section 3.6.1 as obtained from a single seeding geometry (\textsc{tyche} (Single)), from the five seeding geometries found to have a Boltzmann weight greater than $10^{-3}$ (\textsc{tyche} (Multi)), and from the MD trajectory of zwitterionic alanine. We make no reference to the restricted sampling ranges invoked in 3.6.1 since the multiple seeding geometries grant the
capacity to undertake a full conformational sampling.

<table>
<thead>
<tr>
<th></th>
<th>N1-H2</th>
<th>C1-C2</th>
<th>N1-C1</th>
<th>C3-O1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TYCHE (Single)</td>
<td>[1.13 – 0.95]</td>
<td>[1.64 – 1.45]</td>
<td>[1.60 – 1.44]</td>
<td>[1.31 – 1.22]</td>
</tr>
<tr>
<td>TYCHE (Multi)</td>
<td>[1.19 – 0.91]</td>
<td>[1.63 – 1.45]</td>
<td>[1.59 – 1.37]</td>
<td>[1.32 – 1.19]</td>
</tr>
<tr>
<td>MD</td>
<td>[1.15 – 0.92]</td>
<td>[1.65 – 1.43]</td>
<td>[1.63 – 1.35]</td>
<td>[1.35 – 1.19]</td>
</tr>
</tbody>
</table>

**Table 3.12:** Sampling ranges of four prominent bond stretching degrees of freedom in zwitterionic alanine, as obtained through TYCHE with a single seeding geometry (TYCHE (Single)), with five seeding geometries (TYCHE (Multi)) and through a MD trajectory.

<table>
<thead>
<tr>
<th></th>
<th>O1-C3-O2</th>
<th>H2-N1-H3</th>
<th>N1-C1-C3</th>
<th>H5-C2-H6</th>
</tr>
</thead>
<tbody>
<tr>
<td>TYCHE (Single)</td>
<td>[131.5–124.8]</td>
<td>[120.2–99.9]</td>
<td>[115.8–107.7]</td>
<td>[116.8–101.1]</td>
</tr>
<tr>
<td>TYCHE (Multi)</td>
<td>[133.2–117.4]</td>
<td>[126.0–91.0]</td>
<td>[114.4–97.7]</td>
<td>[123.4–91.8]</td>
</tr>
<tr>
<td>MD</td>
<td>[134.3–115.8]</td>
<td>[128.3–90.2]</td>
<td>[117.6–97.3]</td>
<td>[126.2–92.8]</td>
</tr>
</tbody>
</table>

**Table 3.13:** Sampling ranges of four prominent valence angle bending degrees of freedom in zwitterionic alanine, as obtained through TYCHE with a single seeding geometry (TYCHE (Single)), with five seeding geometries (TYCHE (Multi)) and through a MD trajectory.

From both Tables 3.12 and 3.13, we see that the inclusion of more than a single seeding geometry results in an excellent agreement of sampling ranges with those of the MD trajectory. In contrast, with a single seeding geometry, we found in Section 3.6.1 that the lower bounds of the sampling ranges were poorly reproduced,
and the upper bounds of the sampling ranges were only found to be in agreement with the MD upper bound at very high temperatures.

We omitted reporting the sampling of the dihedral degrees of freedom in Section 3.6.1 since we anticipated a single seeding geometry to be inadequate for sampling these degrees of freedom. Typically, the dihedral degrees of freedom adopt a number of rotamer conformations that are similar in energy but distinct in terms of conformation. A single LW is incapable of capturing such broad sampling ranges.

In the following, we analyse the sampling of the H1-N1-C1-C2 dihedral degree of freedom of zwitterionic alanine. From our analysis, we have found this to be the most widely varying dihedral angle over the course of the MD simulation, and so it ideal for our purposes. The tetrahedral amino group is able to freely rotate in the zwitterion, and so we expect three distinct maxima in the sampling frequency at $-120^\circ, 0^\circ, 120^\circ$. Conformational studies of peptides are conventionally validated with Ramachandran plots, which report on the peptide backbone dihedral angles. A Ramachandran plot, is, however, of little use when working with a single amino acid zwitterion since the terminal groups can freely rotate, and so the majority of the Ramachandran plot is typically sampled.

In Figure 3.12, we present histograms showing the sampling frequency of the possible H1-N1-C1-C2 dihedral angles of zwitterionic alanine. The red and blue traces correspond to the sampling obtained through Tyche and the MD trajectory, respectively. The MD trajectory appears to recover the three rotamers, albeit the peaks are extremely broad, particularly the one centred at $0^\circ$. Indeed, virtually the entire range of values the dihedral degree of freedom can take is sampled, which implies that a small number of seeding geometries will be unable to recover the sampling obtained through MD. When seeding with the gas-phase minima, the three rotamers are sampled, but there is very little sampling obtained about the
3.6. RESULTS

Seeding geometries—roughly a range of $40^\circ$ about each rotamer. It then seems as if seeding with local minima is inadequate for an extensive sampling of conformational space.

Figure 3.12: Histograms for the sampling of the alanine sidechain dihedral angle by use of the various TYCHE sampling schemes (by seeding with energetic minima (red), MD snapshot conformations (green) and the selective mode distortion (green)), and the MD sampling (blue).

We can artificially extend the sampling range with the same seeding geometries by increasing the sampling temperature. However, since we have already found an optimal temperature with which to sample the bond stretching and valence angle bending degrees of freedom, increasing the temperature is not a feasible approach. Increasing the temperature would lead to an oversampling of the bond stretching and valence angle bending degrees of freedom, and likely drive the system into undesirable regions of conformational space by, for example, breaking bonds.

We have instead proposed a selective sampling scheme that prevents the oversampling of the bond stretching and valence angle bending degrees of freedom.
By decomposing each normal mode down into its constituent redundant internal coordinates, we can vibrate those modes that are dominated by dihedral degrees of freedom with a higher temperature than that used for vibrating the modes dominate by bond stretching and valence angle bending. For our experiment, we have chosen to vibrate the ten modes with the highest dihedral contributions with a temperature of 1800 K, while maintaining a temperature of 900 K for the remaining normal modes. The sampling from this scheme is depicted by the yellow histogram in Figure 3.12.

We see that the selective high temperature sampling yields a broader sampling about the rotamer seeds than the single temperature. Indeed, the range of sampling about the rotamer at 0° is increased by roughly 40° relative to the single temperature sampling. However, we are unable to sample continuously across the dihedral range as accomplished by the MD trajectory. Whilst the selective high temperature sampling is certainly a useful tool if we are restricted in the number of seeding geometries available. It is encouraging that we can increase the sampling range so much from a minor adjustment to TYCHE.

The final strategy involves the use of more seeding geometries than can be offered by the local minima of the system. The power of the TYCHE methodology is that the seeding geometries can be obtained by any means, whether it be systematic sampling, MD, or even random conformer generation. We have selected 100 random conformers from the MD trajectory against which we are benchmarking. The Jacobian and Hessian from each of these conformers has been evaluated at the B3LYP/aug-cc-pVDZ level of theory, and those conformers with a Boltzmann weight lower than $10^{-3}$ was discarded. This left twenty seeding geometries with which to run TYCHE with. Naturally, since these conformers are not stationary points on the PES, we require the non-stationary point treatment that has been
3.6. RESULTS

outlined in Section 3.3.

The sampling obtained from this MD-seeding approach is depicted by the green histogram in Figure 3.12. Relative to both the minima-seeding and selective high temperature approaches, we recover a greater range of dihedral sampling with the MD-seeding. However, there are a number of discrepancies between the MD trajectory and MD-seeding sampling regions. For example, the rotamer at 0° is oversampled relative to the MD trajectory sampling. We could perhaps improve the sampling by increasing the transition temperature, subsequently including more seeding geometries into our sampling and covering a broader range of the dihedral angle. In addition, we could invoke the selective high temperature strategy, but these strategies are not investigated further here.

3.6.3 Conformational Sampling of Large Molecular Species

As a final study, we demonstrate the capacity of the TYCHE methodology to be used for large biomolecular systems. The dominant conformational degrees of freedom that govern the motion of large biomolecules have been shown to be contained within the low frequency normal modes\cite{278}. These low frequency normal modes have been strongly linked to enzymatic mechanisms in proteins, such as bovine pancreatic trypsin inhibitor\cite{232}, the gramicidin-A dimer channel\cite{279}, lysozyme\cite{280}, F\textsubscript{1}-ATPase\cite{281} and ras p21\cite{282}. Indeed, vibrations have been proposed to be driving forces for coupling energy release to conformational changes. The energy is thought to be transferred by means of a soliton along α-helical motifs\cite{283, 284}. The soliton is thought to propagate along the chains formed along the helix axis originating from peptide backbone interactions. Excitons have also been proposed to be responsible for the coupling of photoexcitation to
conformational changes[285].

Figure 3.13: Helix stretching and bending modes of motion for the alanine-isoleucine helix.

A simple polypeptide system that exhibits large scale motion is the α-helix motif. The α-helix is a spiral conformation of a peptide chain, in which the carbonyl and amide groups from the peptide backbone interact along the helix axis and stabilise the conformation. The α-helix is one of the more stable secondary structure motifs, and is a highly important feature for a number of protein functionalities[286, 287, 288]. Two large-scale conformational motions have been identified in the past; the helix “breathing” mode[289], in which there is a stretching along the helix axis, and the helix “bending” mode[290], where the helix bends about its midpoint. Our test system will be a decapeptide, in which alanine resides sandwich an isoleucine at the centre of the helix. This system is depicted in Figure 3.13.

To track the helix breathing, we use the distance between the $C_\alpha$ atoms at the N- and C-termini of the helix. For the helix bending, we use the angle subtended by the arc of the helix breathing degree of freedom we have just specified, from the
3.6. RESULTS

$C_\alpha$ of the isoleucine residue. We present the evolution of these degrees of freedom in Figures 3.14 and 3.15, where the twenty modes with the highest bond stretch, valence angle or dihedral angle contributions are evolved.

Figure 3.14: Helix breathing motion recovered from selectively evolving those modes comprising dihedral (red), valence angle (green) and bond stretching (blue) degrees of freedom.

We see that for both the helix breathing and bending, sampling those modes that are dominated by dihedral degrees of freedom yields a far greater range of conformers than sampling those modes that are dominated by valence angle and bond stretch degrees of freedom. This is to be expected, since the dihedral-dominated modes are typically far lower in frequency than the other modes. Our result is then in keeping with studies that have found large scale conformational changes to be dominated by low frequency modes of motion, as we have mentioned. We do note that the range of sampling is not extensive. However, this results from the harmonic approximation we have employed in expanding about the seeding
Figure 3.15: Helix bending motion recovered from selectively evolving those modes comprising dihedral (red), valence angle (green) and bond stretching (blue) degrees of freedom.

conformation; the walls of the well become steep rather quickly, and so vibrations about a given point are limited in the amount of conformational space available. One can in principle undertake a much greater conformational search by using a number of additional seeding geometries.

3.7 Conclusion

We have presented a conformational sampling methodology that is both computationally tractable, and in principle, more representative of the sampling one may obtain from the \textit{ab initio} PES than one constructed by a classical force field.

The \textsc{tyche} methodology is ideal for the construction of conformers with which
3.7. CONCLUSION

A machine learning model can be constructed. It naturally implements a form of importance sampling in the form of transitioning between seeding geometries based on Boltzmann weights. It also provides both large-scale conformational sampling, in addition to finer sampling of the higher frequency degrees of freedom. These properties make it an ideal conformational sampling tool for the construction of machine learning models.

Regarding future work, we suggest that the RIC selective sampling requires further development. For instance, one could define separate temperatures for various subsets of modes dominated by certain valence coordinate contributions, e.g. a high temperature for those modes that are dominated by dihedral degrees of freedom, and a lower temperature for those modes that are dominated by bond stretching degrees of freedom.

It may perhaps be feasible to approximately update the Hessian of the system with respect to a distortion. Hessian updating methods are frequently invoked during \textit{ab initio} geometry optimisations\cite{291}, and for small displacements are relatively accurate. By use of these Hessian-updating methodologies, one could re-evaluate the normal modes of the system once it has been displaced from the seeding geometry, and subsequently vibrate along these new normal modes. However, it is anticipated that such a methodology would significantly increase the computational power required for sampling.

Of course, the sample set as output from TYCHE is not ideal for the immediate construction of a machine learning model. One must tailor the sampling to optimise both the predictive capacity and computational tractability of a resultant kriging model. This requires some form of processing of the sample set produced by TYCHE, which we detail in Chapter 5.
Chapter 4

A Novel Carbohydrate Force Field

4.1 Introduction

The computational analysis of biochemical systems is largely biased towards peptides and proteins. One may therefore be forgiven for assuming that they possess a near monopoly in biochemistry. However, it is only by complexation with additional molecular species that peptides and proteins are able to accomplish their myriad roles within biological systems[292]. For example, eukaryotic proteins are subject to post-translational modification, a process in which various carbohydrate sequences are attached to the protein. Vital chemical entities, such as enzymatic cofactors (e.g. ATP, NADP, etc.) and nucleotides, are entirely dependent upon the existence of products from the pentose phosphate pathway, which is synthesised from carbohydrates. In fact, the phosphate pathway would be unable to run at all if it were not for the energy derived from carbohydrates, which undergo glycolysis.
4.1. **INTRODUCTION**

and are subsequently passed into the tricarboxylic acid cycle.

Many biochemical force fields are parameterised by exhaustively sampling quantities arising from peptide atom types. One cannot simply use protein atom types as a direct substitution for carbohydrate atom types for a number of reasons:

1. Peptides possess features that are generally absent in carbohydrates, most prominently the presence of nitrogen and the ability to form structural motifs. Both of these features are particularly perturbative to electrostatic quantities associated with constituent atoms. For example, nitrogen possesses a significant quadrupole moment, which influences other atomic electrostatic quantities anisotropically, and cannot be captured by the standard point charge approximation. Equivalently, structural motifs such as helices and sheets are stabilised by vast intermolecular bonding networks. This dependence on structural motifs necessarily influences the properties of other atoms by constraining them to states that do not necessarily coincide with those of the unfolded non-native state. Although carbohydrates do form structural motifs, they tend to remain flexible under standard biological conditions, and do not typically assemble into the stable secondary structures, which polypeptides do.

2. Carbohydrates exhibit far more conformational freedom than peptides. Electronic quantities vary as a function of the conformational degrees of freedom of a molecular species[109]. As such, the electronic quantities of a more flexible conformation will vary to a greater extent than those of a less flexible one.

3. Many carbohydrate species exhibit a preference for axial rather than equatorial arrangements of electron-rich substituents on an anomeric carbon. The
CHAPTER 4. A NOVEL CARBOHYDRATE FORCE FIELD

The conformational freedom of carbohydrates renders them somewhat troublesome for experimentalists, as they prove to be highly difficult to characterise by conventional high-resolution structural determination techniques, particularly X-ray crystallography. To be precise, carbohydrates tend to be difficult to crystallise, which is problematic because X-ray diffraction techniques are a valuable source of structural information. As such, the structural characterisation of carbohydrates rests with a few experimental techniques, and subsequent validation by computational means. This required harmony between experiment and computation is vastly important, and has been recently explored, and so proves to be a fruitful avenue for development.
Classical force fields such as OPLS-AA, CHARMM, GROMOS and AMBER appear to have characterised carbohydrates as ‘secondary molecular species’ relative to their peptide counterparts. As such, these parameterisations resemble “bolt-on” components. However, force fields that are specifically tailored for carbohydrates do exist and have proven successful. GLYCAM[300] is perhaps the most prominent of these force fields, and has been ported to AMBER. More recently, the advent of DL FIELD has facilitated the use of GLYCAM parameters within DL POLY 4.0[301].

GLYCAM has undergone extensive validation in an attempt to demonstrate its efficacy. Several studies have focused upon its ability to reproduce conformer populations in explicitly solvated molecular dynamics simulations[302, 303], which is obviously important owing to the massive conformational freedom of carbohydrates. The applicability of GLYCAM to larger, more biologically relevant structures, such as the binding of endotoxin to recognition proteins[304] or the dynamics of lipid bilayers[305], has also been demonstrated.

GLYCAM has attempted to break the paradigm of deriving partial charges based on a single molecular configuration. Instead, it has been developed such that the partial charges are averaged over the course of a molecular dynamics simulation, thus (albeit simplistically) accounting for the dynamic nature of electronic properties[306]. However, it must be emphasised that GLYCAM resides within the partial charge approximation to electrostatics, thus severely limiting its predictive capacity. Sugars are particularly amenable to hydration, yet a partial charge approximation to electrostatics cannot recover the directional preferences of hydrogen bond formation without the addition of extra point charges at non-nuclear positions. The isotropic nature of partial charge electrostatics is readily overcome by use of a multipole moment description of electrostatics, which naturally describes anisotropic electronic features such as lone pairs. The benefits of such a multipole moment description over their partial charge equivalents has
been systematically demonstrated over the past 20 years in many dozens of papers, recently reviewed[307]. These benefits are not necessarily outweighed by the common misconception that multipole moment implementations are computationally expensive relative to their point charge counterparts. The long-range nature of point charge electrostatics, $O(r^{-1})$, relative to higher order multipole moments (dipole-dipole interactions, for example, die off as $O(r^{-3})$), means this is not strictly true. Point charges require a larger interaction cutoff radius relative to higher order multipole moments, and therefore form the bottleneck in electrostatic energy evaluation. Given proper handling (e.g. parallel implementation), the computational overheads associated with multipole moment electrostatics can be managed.

In the remainder of this chapter, we shall demonstrate a novel means for modelling electrostatics by use of a multipole moment expansion centred upon each atomic nucleus. The techniques we present will inherently capture the conformational dependence of these multipole moments.

### 4.2 A Basis Set for Machine Learning

A great deal of work has been conducted within our group on the use of an Atomic Local Frame (ALF) as a basis set for denoting molecular conformations. This basis set is ideal since it spans only the internal degrees of freedom of the molecule, making it superior to a Cartesian basis set. In constructing the ALF, one requires the definition of a local frame within which ones can express the positions of all atoms within a molecular system.

We will use $\mathbf{R}$ to denote atomic position vectors in some global coordinate frame.
4.2. A BASIS SET FOR MACHINE LEARNING

We begin by nominating an origin atom, $\Omega o$, at position $R^{\Omega o}$. Then, the bond vector which links the origin atom to its heaviest neighbour $\Omega x$, $R^{\Omega x} - R^{\Omega o}$, acts as the $x$-axis of the local frame. Finally, the $xy$-plane in the local frame is selected to contain both the first and second ($\Omega xy$) heaviest neighbours of the origin atom. When the origin atom is not of the required valency to possess two neighbours, the Cahn-Ingold-Prelog rules are invoked to select $\Omega xy$. The $z$-axis is then specified by its orthogonality to the $xy$-plane.

With these quantities, we construct a rotation matrix, $C : R^n \rightarrow r_\zeta$, where $R^n$ is an atomic Cartesian position vector in the (arbitrary) global frame of reference, and $r_\zeta$ is the corresponding atomic Cartesian position vector in the ALF. The components of $C$ are constructed by standard geometric functions, requiring three quantities

\[ r^{\Omega x} = ||R^{\Omega x} - R^{\Omega o}|| \]  \hspace{1cm} (4.2.1)
\[ r^{\Omega xy} = ||R^{\Omega xy} - R^{\Omega o}|| \]  \hspace{1cm} (4.2.2)
\[ \chi^{\Omega} = \arccos \left( \frac{R^{\Omega x} \cdot R^{\Omega xy}}{r^{\Omega x} \cdot r^{\Omega xy}} \right) \]  \hspace{1cm} (4.2.3)

where $r^{\Omega x}$ is the distance from $\Omega o$ to $\Omega x$, $r^{\Omega xy}$ is the distance from $\Omega o$ to $\Omega xy$, and $\chi^{\Omega}$ is the angle between the two vectors $(R^{\Omega x} - R^{\Omega o})$ and $(R^{\Omega xy} - R^{\Omega o})$. Therefore, three degrees of freedom are required to construct a local axis system centred on $\Omega o$, the ALF.
Figure 4.1: The Atomic Local Frame for the carbon atom of methanol. From the global coordinate system, $x, y, z$, one is able to define an ALF, the basis vectors of which are denoted $x_{ALF}, y_{ALF}, z_{ALF}$, from the vectors $R^{\Omega_o}, R^{\Omega_x}, R^{\Omega_{xy}}$. Any atom not defining the ALF, e.g. $R^n$, is defined in the ALF with the spherical polar coordinates $r^n, \theta^n, \phi^n$.

For all other atoms in the system, $n \notin \{\Omega_o, \Omega_x, \Omega_{xy}\}$, positions are given by the triplet of spherical polar coordinates, $\{r^n, \theta^n, \phi^n\}$, i.e. the radial distance, colatitudinal and azimuthal angles, respectively. Then, we see that $3N - 6$ degrees of freedom are required to describe a molecule in the ALF. We note that we could have simply used Cartesian position vectors for all such atoms not defining the ALF and obtain a basis of the same dimensionality. However, we find the spherical polar coordinates more physically intuitive. Of course, in doing so, our basis vectors
do not have the same units (the radial distance has units of distance while the colatitudinal and azimuthal angles have units of radians, and are periodic), but we have found that this does not affect our methodology.

4.3 Computational Details

The workflow proposed below essentially takes an ensemble of configurations as input, and outputs kriging models for the variation in the atomic multipole moments as a function of the configuration of the system:

1. The test system is sampled in accordance with the methods outlined in Section 3. The general idea is to sample as much of conformational space as would form an ensemble for the true physical system along to the course of a dynamical trajectory. This subsequently allows for the formation of a kriging model that will be used in a purely interpolative fashion.

2. Single-point calculations are performed on each sample and the resultant ab initio electron density is partitioned by QCT software. The multipole moments of the topological atoms are subsequently obtained. This allows a kriging model to evaluate a functional form corresponding to the evolution of the various multipole moments as a function of conformation.

3. The sample set is split into a non-overlapping training set and test set. The training set is utilised for training of our kriging models, i.e. these are the points that the kriging function must pass through. The test set is not trained for, but is used after the construction of the kriging models to evaluate the errors associated with their predictions.
4. A kriging model is built for each multipole moment of each atom, which allows for the generation of a smooth interpolative function, mapping the evolution of the multipole moment against the molecular conformational degrees of freedom. By use of particle swarm optimisation, we optimise our kriging parameters, to obtain an optimal kriging model.

5. The kriging models are assessed by making them predict the multipole moments for each atom in a system whose configuration has not been used for training of the kriging model. However, we do possess the \textit{ab initio} multipole moments for this configuration. As such, we evaluate the energy associated with all $1\rightarrow n$ (i.e. two nuclei separated by four bonds) and higher ($1\rightarrow n$, $n \geq 5$) order interatomic electrostatic interactions as given by the predicted multipole moments from the kriging models, and the equivalent energy as given by the (exact or original) \textit{ab initio} atomic multipole moments. We subsequently assess the deviation of the kriging predictions from the \textit{ab initio} electrostatic energies.

The choice of $1\rightarrow n$ ($n \geq 5$) interactions over the conventional $1\rightarrow n$ ($n \geq 4$) is discussed in Section 2.5.2, and is used to prevent a divergence in the multipole moment interaction energy.

Note that the electrostatic energy\cite{213} is the final arbiter in the validation of the kriging models, rather than the atomic multipole moments themselves, which are the kriging observations. It is worth recalling that every prediction point from a kriging model possesses an associated variance, i.e. uncertainty in the predicted value. We have in the past investigated the magnitude of these variances, but they are orders of magnitude smaller than the actual predicted value, and so inconsequential. We thus proceed under the assumption that uncertainties in predicted
values are negligible. The molecular electrostatic energy is calculated by a well-
known multipolar expansion[77] involving a multitude of high-rank atomic multi-
pole moments[208]. This expansion is truncated to quadrupole-quadrupole (L = 5)
and rank-equivalent combinations (dipole-octopole and monopole-hexadecapole).

Whilst somewhat indirect, the validation through energy rather than multipole
moment has a twofold purpose. Primarily, the energy is the quantity that will be
used for dynamical simulations, and so is the ultimate descriptor that we wish to
evaluate correctly. Secondly, the alternative would be to assess the predictive ca-
pacity of each individual kriging model. For a system with any sizeable number of
atoms, where each atom has 25 individual multipole moment kriging models, the
data analysis obviously becomes overwhelming. However, this analysis is unnec-
essary owing to the uniqueness of the Taylor expansion from which the multipole
moments arise. Since the electrostatic energy is computed from two such unique
series, then if the electrostatic energy is correctly predicted, the multipole mo-
ments must also be correct by deduction. Note that this consideration is valid for
a single atom-atom interaction.

In order to gauge the models’ validity, we plot a cumulative distribution of errors
(CDE). The CDE plots the absolute deviation of the predicted energy from the
\textit{ab initio} energy, predicted from the \textit{ab initio} multipole moments, after having
evaluated the multipole moment interactions. Put more precisely, the predicted
multipole moments form an energy that is subtracted from an energy obtained
from the \textit{ab initio} moments. Then the absolute value of this difference is taken.
Hence compensation of errors is not allowed because first the difference is taken and
then the absolute value. All errors are subsequently summed to give an “energetic
deviation”. These energetic deviations are plotted against the percentile of test
configurations that fall on or below the given energetic deviation. The CDE is
sigmoidal (as such it is coloquially referred to as an “S-curve” in our group, but we adopt the more meaningful CDE nomenclature throughout), and therefore represents a Gaussian distribution of errors.

Our aim is then twofold: the first is to reduce the upper tail of the sigmoid such that the 100th percentile error is convergent at as low an error as possible. This corresponds to the predictions being uniformly good across the test set with no spurious predicted interactions. Our second aim is to shift the CDE as far down the abscissa as possible, which ensures the average error associated with our predictions is as low as possible. The first goal is achieved by certifying that the training points used for the construction of the kriging models form the boundaries of configurational space with respect to our sample set. This boundary checking guarantees that the kriging model is being asked to interpolate from training data. Boundary checking is discussed in Section 5.5.

The second goal is attained by consistent improvement of the kriging engine, and making sure that the test points are uniformly close to the training data, allowing for efficient interpolation. Note that we do not necessarily choose test points that are close to the training points. In other words, the training and test sets are constructed independently. By ensuring that the training data is uniformly distributed throughout the sampling domain, the average “distance” between an arbitrary test point and a training point will equal that between some other arbitrary test point and another training point. This guarantees no spurious predictions in under-trained regions of conformational space. Of course, it is prudent to invoke some form of importance sampling, which yields a greater sampling density in more “important” regions of conformational space. We discuss the methodologies relating to training set construction in Section 5.

We work on the tetrose diastereomers erythrose and threose, the smallest carbo-
hydrates that adopt open chain and furanose forms. The particular conformations studied are given in Figure 4.2. Energetic minima were provided by Prof Ibon Alkorta, who had previously conducted a PES scan of these species[308], and are reported more thoroughly elsewhere[309]. All geometries were subsequently optimised by the program GAUSSIAN03[310] at the B3LYP/6-311++G(d,p) level of theory. The Hessian of each geometry has been computed, and utilised in the TYCHE procedure described in Section 3, allowing for the output of 2000 geometries for each system. Single point calculations were performed on each sample at the B3LYP/apc-1[311] level of theory. The apc-1[311] basis set (which is a polarization-consistent (pc) double-ζ plus polarization basis set with diffuse functions) was used for the DFT calculations, since this family of basis sets has been specifically optimized for DFT. The resultant wavefunctions are then passed on to the program AIMALL[312], which calculates the atomic multipole moments by application of QCT.

We are only interested in the internal degrees of freedom of the molecular configurations. Hence the atomic multipole moments must be expressed in an atomic local frame (ALF) rather than in the global frame. This procedure makes sure that the kriging focuses on the variation of atomic multipole moments within the molecule. Otherwise, when referring to the global frame (rather than the ALF), the three components of an atomic dipole moment, for example, vary upon rigid rotation of the whole molecule. Training for such a variation is useless. The details of the atomic local frame chosen for our work are outlined in Section 4.2.

Kriging of each multipole moment (up to the hexadecapole moment) was performed for each atom by the in-house program FEREBUS 1.4 and models consisting of $N$ training examples were generated. Atom-atom interaction energies of $1 - 5$ and higher were computed in a test set of 200 arbitrary conformations, using the $ab$
Figure 4.2: Erythrose and Threose in the open chain and furanose forms.
initio multipole moments, and compared to the interaction energies from multipole moments predicted by the kriging models. Errors are given in the form of so-called S-curves, which map the percentile of conformations within the test set, predicted up to a maximum error chosen, which is read off on the abscissa.

Here we give a brief overview of how the force field we propose could be utilised within the context of MD. We limit the discussion to the evaluation of electrostatic interactions, but work is currently being undertaken within our group to establish the framework for an entire force field[313], which deviates considerably from the terms arising in a classical force field. The non-electrostatic terms are also obtained via the QCT partitioning of molecular energy, given in Section 2.2.3. At designated points over the course of a MD simulation, the conformational state of the system is evaluated. At this point, atomic multipole moments, up to the hexadecapole moment, can be extracted from the kriging models, and subsequently utilised for the evaluation of interatomic electrostatic interactions. In this way, we capture the conformational dependence of the atomic multipole moments. Separate kriging models are obtained for the non-electrostatic terms, that is, the intra-atomic energy (both kinetic[212] and potential), the short-range interatomic Coulomb energy not obtained by multipolar expansion, and the interatomic exchange energy.

The issue of coordinate frames needs further clarification because the Cartesian MD frame coordinate system is not the same as the local coordinate frame within which we have evaluated the atomic multipole moments. Prior to invoking a kriging model for the evaluation of multipole moments, the conformational state of the system must be converted from a Cartesian frame of reference to an ALF. From here, atomic multipole moments can be evaluated corresponding to the current state of the system. Previous work has derived the forces arising from the interactions of atomic multipole moments within the Cartesian frame of reference[146],
which requires the partial derivatives of the multipole moments with respect to the ALF degrees of freedom. These terms have an analytical functional form, and so computation of the forces in the Cartesian frame of reference can be performed explicitly. These forces can subsequently be utilised by the standard MD procedure.

An embryonic workflow has been integrated within the MD package DL_POLY 4.0. Currently, an atomic kriging model is loaded into memory and the relevant data stored, followed by removal of the kriging model from memory. Memory requirements are discussed in Appendix B. Of course, memory management is crucial to the speed of the proposed methodology, and so will require a great deal of fine-tuning. However, much speedup can undoubtedly be accomplished by a number of techniques, e.g. caching of regularly used quantities and parallel implementation.

Finally, it is too early to extensively comment on the computational cost of the current approach. It would be naive to directly compare the flop count of the current force field with a traditional one without appreciating that (i) extra non-nuclear point charges are needed to match the accuracy of multipole moments and the former propagate over long range, (ii) multipolar interactions drop off much faster than $1/r$, depending on the rank of the interacting multipole moments, which depends on the interacting elements themselves (see extensive testing in the protein crambin[118]), (iii) the efficiency of the multipolar Ewald summation[314] that is being implemented in DL_POLY 4.0, (iv) the dominance of monopolar interactions at long range (vast majority of interactions) and the (v) outstanding fine-tuning of the kriging models at production mode. The current force field may well be an order of magnitude slower than a traditional force field. This estimate and the fact that the current force field contains electronic information invite one to compare
its performance with on-the-fly \textit{ab initio} calculations instead.

4.4 Results and Discussion

4.4.1 Single Minimum

The lowest energy conformer for each system was chosen as an input structure for sampling. CDEs were subsequently generated for each of these training sets by the methodology outlined in the previous section. The CDEs are given in Figure 4.3, and the accompanying mean errors in Table 4.1.

<table>
<thead>
<tr>
<th></th>
<th>Open Chain</th>
<th>α-furanose</th>
<th>β-furanose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erythrose</td>
<td>0.27</td>
<td>1.32</td>
<td>1.30</td>
</tr>
<tr>
<td>Threose</td>
<td>0.34</td>
<td>0.83</td>
<td>1.47</td>
</tr>
</tbody>
</table>

Table 4.1: Mean errors associated with the CDEs given in Figure 4.3.

The first point to notice is that the open chains for both erythrose and threose are modelled by kriging to a significantly better standard than the furanose forms. We may, however, immediately attribute this to the number of 1-5 and higher interactions occurring in these systems. For both open chains, 25 interactions are required to be evaluated for comparison to the energies produced from the \textit{ab initio} multipole moments. The numbers of interactions requiring evaluation for the furanose forms comes to 39, which is roughly 50\% more than the amount evaluated in the open chain forms. We would subsequently expect a proportional
Figure 4.3: CDEs corresponding to all erythrose (top) and threose (bottom) systems studied. The open chain forms are systematically better predicted than the corresponding furanose forms.
relationship between the number of interactions required for evaluation and the mean error attributed to the kriging model. Whilst we see this to be roughly true when comparing the errors on the threo open and α-furanose forms, the errors appear disproportionately higher for the other systems.

Kriging is an interpolative technique, and so is not suited for extrapolation. However, we point out that the kriging engine is still predictive for extrapolation—in this case, the prediction falls to the mean value of the function. Obviously this is not ideal for highly undulatory functions. However, considering how the atomic multipole moments do not fluctuate over vast ranges, the mean will often represent a respectable prediction to the function value. The kriging model can be refined in an iterative fashion, whereby extrapolation points are added to the training set. This is a commonly used technique in the field of machine learning. Whilst not currently implemented within our methodology, the iterative protocol is a technique which is currently being explored.

Regardless of the problem encountered in the above, we see that it is easily remedied by strategic sampling of conformational space. In fact, these problems are ubiquitous to machine learning techniques, and have been encountered in studies which attempt to implement neural networks to predict a PES[9]. In this field, the problems have been solved to some extent by making the prediction engine issue a warning to the user that a point is being predicted which lies outside of the training set. This proves to be advantageous as the user may then recognise that the point should be included within the training set as it obviously lies within an accessible portion of conformational space. The point can then be included within the training set, and one can generate refined models by undergoing this process iteratively.
4.4.2 Training Set Size Dependency

We start by discussing the effects of increasing the training set size on the prediction error for a kriging model. For this purpose we use the erythrose open chain system owing to its higher conformational flexibility, which we assume amplifies the effects of training set size. Kriging models were generated for this system with training sets ranging from 700 to 1500 sampling points, in increments of 100. The same test set (of 200 points) was reserved for prediction by all models. The CDEs for this are given in Figure 4.4.

As expected, the prediction errors of the CDEs in Figure 4.4 systematically decrease as the training set size is increased. In other words, the CDEs move to lower prediction errors. However, owing to the logarithmic abscissa, this does not correspond to a uniform enhancement of a kriging model given a consistently larger training set size.
the left with increasing training set size, although this is not true for all parts of the CDEs because they clearly intersect in many places. Overall the uniform increments of 100 in training set size are not matched by equal uniform strides of improvement in CDE shape and position. An alternative way to gauge the improvement in prediction with increasing training set size is monitoring the average prediction error for each CDE. This value cannot be read off for a CDE in Figure 4.4 but can be easily calculated.

Figure 4.5 plots the average prediction error for each CDE against increasing training set size: red for “Old FEREBUS” and blue for “New FEREBUS”, a development version of our kriging engine, which differs in a number of ways to the “Old FEREBUS”. We include both “Old FEREBUS” and “New FEREBUS” data to establish whether any functional forms of average prediction error against increasing training set size are conserved with respect to improvements in the engine. The “Old FEREBUS” data show a plateau in the average error (left pane) at a training set size of about 1200, after an initial decrease in this error. This plateau would be rather problematic, as it implies some maximum efficiency of the kriging engine, beyond which there is no reward for an extension of the training set. However, this is not the case for the “New FEREBUS” data (right pane).

Learning theory states that for a machine learning method of this type (kriging), the mean prediction error should decrease asymptotically towards zero, with functional form $A + B/n$ or $C + D/\sqrt{n}$, where $n$ is the training set size, and $A, B, C$ and $D$ are fitted constants. Figure 4.5 plots these asymptotes, where $(A_{old} = 0.105; B_{old} = 348.11)$ and $(C_{old} = -0.293; D_{old} = 22.06)$, as determined by regression analysis against the “Old FEREBUS” data, each with $R^2$ coefficients of 0.93. Similarly, for the “New FEREBUS” data, constants of $(A_{new} = 0.097; B_{new} = 293.38)$ and $(C_{new} = -0.193; D_{new} = 18.64)$ were obtained. These fitted asym-
totes both possess $R^2$ values of 0.98. As such, we conclude that the decay of the mean prediction error of our machine learning method possesses, as yet, inconclusive functional form.

The results in Figure 4.5 are consistent with the behaviour seen in similar interpolation methods[315]: for an infinite training set size, the mean prediction error will asymptote to zero. However, for the methodology to remain computationally feasible, some finite training set size will of course be required. So, for example, we find that for a mean prediction error of 0.3 kJ mol$^{-1}$, the training set would require about 1450 sample points for either functional form taken as the decay of the prediction error, i.e. $A + B/n$ or $C + D/\sqrt{n}$.

A comment on the nature of the average error is in place here. In principle, the prediction error consists of the sum of the estimation error and the approximation error. From learning theory, one expects the estimation error only to go to zero. The bias–variance decomposition of a learning algorithm’s error also contains a quantity called the irreducible error, resulting from noise in the problem itself. This error has been investigated some time ago in the context of tests on kriging of ethanol multipole moments[208] and is caused by the small noise generated by the integration quadrature of the atomic multipole moments. Secondly, any bias caused by an inherent error in the $ab\ initio$ method used, compared to the best method available (e.g. CCSD(T) with a complete basis set), is not relevant in our error considerations. The reason is that we always assess the performance of kriging training against the (inevitably approximate) $ab\ initio$ at hand, which we refer to the source of “the” $ab\ initio$ data.
Figure 4.5: Mean prediction errors associated with CDEs for erythrose open chain as the training set size is increased. With the old kriging engine (left), a distinct plateau formed after roughly 1200 training examples, corresponding to no further improvement in the kriging model despite additional training points. However, the new kriging engine (right) appears to avoid premature plateauing, with additional kriging model improvement at higher training set sizes. Regression fits of the \textsc{ferebus} errors against training set size, of functional forms $A + B/n$ (blue) and $C + D/\sqrt{n}$ (black), where $n$ is the training set size, are also given.
CHAPTER 4. A NOVEL CARBOHYDRATE FORCE FIELD

4.4.3 Multiple Minima

The amount of conformational space available to molecular systems reaches levels which are entirely unfeasible for systematic sampling as the number of atoms increases. As such, it becomes all the more prudent to obtain an efficient sampling scheme for our purposes. As we have mentioned, our sampling methodology is limited to local conformational exploration about some given input geometry, since the PES about that point is approximated as a harmonic well. As such, to thoroughly explore conformational space, our methodology requires the usage of a number of such starting geometries. Then, the molecular PES is approximated by a number of harmonic wells. If the input geometries are sufficiently close to one another, the wells will overlap, and the PES may be explored seamlessly.

For the open chain form of erythrose, 174 energetic minima were found by an exhaustive search of conformational space. Figure 4.6 plots the CDEs obtained for samples which have been generated from different numbers of up to 99 minima. The CDEs display increasingly poor prediction results as the number of starting minima increases. The actual mean errors for these CDEs are summarised in Table 4.2. This trend has a logical interpretation. As the number of seeding structures increases, the sampled conformational space grows in size. Given a fixed kriging model size, the sampling density therefore decreases. The kriging model then deviates from the true analytical function, and the results from predictions deteriorate.

Of course, thorough sampling of conformational space is an issue for parameterising any force field, and by no means one that is resultant from our methodology. We may overcome this issue in two ways. The first is the ongoing improvement of our kriging engine to deal with larger training sets comprising more molecular
4.4. RESULTS AND DISCUSSION

Figure 4.6: S-curves depicting the power of a kriging model as more energetic minima are utilised for conformational sampling. As the number of minima utilised increases, the CDEs tend towards higher prediction errors. The kriging models which underlie these CDEs have a fixed training set size of 700.

<table>
<thead>
<tr>
<th>Number of Minima</th>
<th>1</th>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Prediction Error (kJ mol$^{-1}$)</td>
<td>0.27</td>
<td>0.85</td>
<td>0.9</td>
<td>1.23</td>
<td>1.30</td>
<td>1.37</td>
</tr>
</tbody>
</table>

Table 4.2: Mean errors corresponding to the CDEs depicted in Figure 4.6. Note the ‘bunching’ of prediction error when 60 energetic minima or higher are used as seeds for the conformational sampling.
configurations. The second is by undertaking sampling with only a subset of the energetic minima that are available. This is all the more valid an approach if most of the minima are very high in energy relative to the lowest-lying minima. These regions of conformational space will be accessed very infrequently during the course of a MD simulation, and so may be sampled much more coarsely. This selective sampling is quite readily employed, and has been discussed at length in the literature. For example, Brooks and Karplus[232] found that a comprehensive sampling of conformational space for bovine pancreatic trypsin inhibitor could be achieved by evolving only the lowest frequency normal modes of motion. Needless to say, this is readily accomplished by our sampling methodology.

4.5 Conclusion

We have demonstrated that the atomic multipole moments of a set of carbohydrates are amenable to the machine learning technique kriging. Whilst this has been done in the past for a variety of chemical species including naturally occurring amino acids, this is the first foray into the field of glycobiology.

Kriging is able to capture the conformational dependence of the multipole moments and make predictions, such that the error in the electrostatic energy relative to that derived from \textit{ab initio} data is encouraging, given the popular aim is to obtain errors below 4 kJ mol$^{-1}$. Indeed, the presented methodology is immediately extensible to any term arising in an energetic decomposition of a system. If some quantity is conformationally dependent, then the dependence can be modelled by kriging.

As such, an entire force field can be parameterized by the current methodology, reproducing \textit{ab initio} quantities for use in classical MD. This route is preferable
4.5. **CONCLUSION**

... to the computationally intensive approach of *ab initio* MD.
Chapter 5

Subset Selection for Machine Learning

5.1 Introduction

In Appendix B, we give an estimate for the memory requirements of a biomolecular force field based upon the methodology we have outlined. A number of factors are largely out of our control, such as the number of atom types required. However, the size of the training set used to construct a kriging model is within our control, and can be minimised given a scheme for prudent training set construction.

As we have mentioned in Section 2.4.2, the errors associated with a kriging model diminish as we increase the size of the training set. However, we are limited in our ability to improve the accuracy of a kriging model by systematically increasing the training set size because; (1) the construction of a kriging model is an $O(N^3)$ process, and the memory requirements scale as approximately $O(N^2)$; (2) the
clustering of points in a training set is associated with an ill-conditioning of $\mathbf{R}$, and; (3) undulant regions of the response function require denser sampling to capture complex topological features.

With regards to (1), we obviously require that the training set is as small as possible to make the resultant kriging model viable. Regarding (2), we require that training points are as mutually distant from one another as possible to prevent clustering and subsequent ill-conditioning of $\mathbf{R}$. Finally, concerning (3), some regions of conformational space require a higher density of sampling to capture the undulant topology of the response function.

It is virtually impossible to satisfy each of the three conditions without a detailed knowledge of the response function we are trying to model. Unfortunately, the response function is a PES, and detailed knowledge is simply not available for all but the simplest of molecular systems. We will present a scheme in Section 5.3 that samples the response function at a great deal of points in an attempt to gain some information regarding the topology of the response function. Whilst this methodology possesses a number of caveats, it is the closest we can come to concurrently satisfying all conditions.

The methodologies of Sections 5.2 and 5.4 prioritise conditions (1) and (2) while disregarding (3) entirely for the sake of increased computational tractability. Finally, in Section 5.5, we address a separate issue of ensuring the kriging model is always used in an interpolative fashion, since the extrapolative capacity of a kriging model is severely limited.
CHAPTER 5. SUBSET SELECTION FOR MACHINE LEARNING

5.2 Greedy Heuristics

Perhaps the most straightforward methodology available for optimal subset selection is the greedy algorithm. A greedy algorithm parses a large number of points, the candidate set, $\Omega = \{x_1, ..., x_N\}$, into a subset, $S$. $S$ is constructed iteratively, such that points are transferred from the candidate set to the subset based on some selection function, one at a time, until $|S| = n$, where $n < N$. The premise of a greedy algorithm is that a globally optimal subset can be constructed by iterating a locally optimal selection function.

The operation of a greedy algorithm is best demonstrated in the context of graph theory. Consider a set of connected nodes about which we wish to construct the shortest possible Hamiltonian circuit[316]. The connections between nodes, or edges, possess an associated length, and it is these lengths that characterise the length of the Hamiltonian circuit. If each node is connected to every other node, then every solution is an optimal solution, and so we deal explicitly with the case where each node is only connected to a subset of the remaining nodes. We also assume that each edge links two distinct nodes, i.e. a node cannot be connected to itself. Subject to these conditions, the construction of the shortest possible Hamiltonian circuit is referred to as the travelling salesman problem.

Given an origin node, a greedy algorithm will invoke its associated selection function and find the closest linked node, subject to the constraint that no nodes which have been previously “visited” can be used, a characteristic of a Hamiltonian circuit. The closest node then acts as an origin node and the process can be iterated until we have a closed path about all nodes. It is not too difficult to find systems where the greedy algorithm fails to find an optimal Hamiltonian circuit. However, greedy algorithms are not used for their ability to find the opti-
5.2. **GREEDY HEURISTICS**

Greedy algorithms are simple and algorithmically efficient, making them ideal candidates for approximate optimal subset selection on large, complex datasets, where systematic evaluation of each solution is not feasible.

For the construction of a maximally diverse subset $S$ selected from the candidate set $\Omega$, we invoke a variant of the travelling salesman problem, where instead of invoking a selection function to minimise the distance between nodes, we maximise the distance the distance between nodes. The corresponding nodes are then added to $S$ and removed from $\Omega$. This process is iterated until $|S| = n$.

For our purposes, the candidate set will comprise some large pool of molecular conformations. The subset that is generated by the greedy algorithm is then the training set (nomenclature that we adopt herein), with which a kriging model can be built and validated on an external test set, not contained within the candidate set. The choice of distance metric is arbitrary, but for our purposes we initially choose the Euclidean norm. This choice will be subject to revising in Section 5.2.4.

Regarding our expectations, a greedy algorithm should reduce the clustering of points in the training set. We anticipate the following improvements in our kriging models:

1. **Dispersion** More disperse distribution of training points through the training set, making it less probable that predictions will be made in under-represented areas of conformational space.

2. **Uncluttering** Inhibition of the clustering of training points in the training set, improving the global predictive capacity of the kriging model.

From (1), we expect lower maximum errors attributed to a kriging model since the number of predictions in poorly modelled regions of conformational space is
reduced. However, one could also posit local regions of conformational space to be less well-represented by the training set, thereby increasing the average error associated with a kriging model. From (2), we anticipate the average error associated with a kriging model will decrease since the kriging model will not be overtrained in certain regions of conformational space. Based on this qualitative analysis, we anticipate an overall decrease in the maximum error associated with a kriging model, but reserve judgement on the average error, since (1) and (2) appear to yield opposing effects. Whether the effect of one will outweigh the other is unknown.

We are unsure of the effects of dispersion and uncluttering on the MSE of prediction points. Since the MSE is a measure of whether the training density is appropriate in the vicinity of a prediction point, reducing the sampling density should technically increase the MSE of prediction points. However, since we anticipate that both dispersion and uncluttering will improve our kriging models, it is perhaps reasonable to assume that the MSE will also decrease at prediction points.

5.2.1 Experimental Details

We evaluate the performance of the greedy algorithms on a number of small molecule test systems: water, methanol, N-methyl acetamide (NMA) and glycine. Each system has a candidate set of cardinal \(|\Omega| = 3500\), with the exception of water, where \(|\Omega| = 1000\) owing to its limited number of conformational degrees of freedom. We have constructed a series of training sets from these candidate sets by the greedy algorithms, beginning at a training set size of 100 and incrementing this by 100 up to a training set size of 1000. Each of the training sets is then
used to construct a kriging model, and its performance evaluated on an external testing set. The external testing set for each system comprises 500 distinct conformations not contained within $\Omega$. We benchmark each of the kriging models against a kriging model constructed from 1000 random samples of the candidate set, which has thus far been the conventional methodology adopted by our group in the construction of training sets.

To test the greedy subset selection algorithms, we require some function with which to conduct error analysis. For each conformation within the testing set, we predict the $\sum_B V_{ee,exch}^{AB}$, $\sum_B V_{ee,coul}^{AB}$ and $V_{en}^{AA}$ IQA components of each atom within the molecule. With these components, we can define the total prediction error corresponding to a test conformation, $\mathcal{E}(x)$, by

$$\mathcal{E}(x) = \left| \Delta \sum_B V_{ee,exch}^{AB}(x) \right| + \left| \Delta \sum_B V_{ee,coul}^{AB}(x) \right| + \left| \Delta V_{en}^{AA}(x) \right|,$$

(5.2.1)

where the $\Delta$ implies that we take the difference between the actual and predicted values of the quantity it precedes, i.e. the error associated with that quantity. Taking the absolute value of the error in each IQA component ensures there is no cancellation of error between terms, thus making our error analysis as rigorous as possible.

All samples have been obtained using the TYPHE conformational sampling software, described in Section 3. All \textit{ab initio} calculations were performed with the GAUSSIAN09 software package, using the B3LYP/6-31G(d,p) level of theory. IQA decomposition was carried out with AIMALL and kriging models were obtained through FEREBUS.
5.2.2 MaxMin

The MaxMin algorithm judges an optimal subset to be one that maximises the geometric distance between points. For this purpose, the selection function is a distance metric between points, \(d(x_i, x_j)\), and points are parsed from the candidate set to the subset in a manner which makes the subset as sparse as possible\[317\]. The MaxMin algorithm, then, should avoid the ill-conditioning of the \(R\) matrix as a result of constituent points being too close\[196\]. The steps of the algorithm are outlined in Algorithm 2\[318\].

1. Take \(S = \emptyset\)
2. Take \(x_i\) and \(x_j\), where \(x_i, x_j \in \Omega\) such that \(d(x_i, x_j)\) is maximal
3. Add \(x_i\) and \(x_j\) to \(S\)
while \(|S| < n\) do
   1. Find \(x_k \in \Omega \setminus S\), such that \(\min_{x_i \in S} d(x_i, x_k)\) is maximum among \(\Omega \setminus S\)
   2. Add \(x_k\) to \(S\)
end

**Algorithm 2:** MaxMin Algorithm

In Figure 5.1, we present the average errors associated with the prediction of the testing set for each small molecule as the training set size is increased. For each small molecule studied, we see that as the training set size is increased, the average prediction error converges almost exactly to the average prediction error obtained from the random benchmark sampling. There is, therefore, no discernible benefit associated with constructing a training set with the MaxMin algorithm in terms of average errors. We present the maximum errors associated with each kriging model in Table 5.1.
Figure 5.1: Average molecular energy error, with accompanying standard deviations, in the testing set conformations given a training set constructed by the MaxMin algorithm from the respective candidate sets at a number of training set sizes. In each pane, the dark black line corresponds to the average molecular energy error in the testing set given a training set of size 1000, generated randomly from the candidate set.
<table>
<thead>
<tr>
<th>Training Set Size</th>
<th>Water</th>
<th>Methanol</th>
<th>NMA</th>
<th>Glycine</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>9.563</td>
<td>37.596</td>
<td>50.790</td>
<td>47.116</td>
</tr>
<tr>
<td>200</td>
<td>10.436</td>
<td>12.182</td>
<td>22.858</td>
<td>37.353</td>
</tr>
<tr>
<td>300</td>
<td>5.598</td>
<td>10.951</td>
<td>21.368</td>
<td>36.196</td>
</tr>
<tr>
<td>400</td>
<td>6.258</td>
<td>10.820</td>
<td>20.076</td>
<td>38.252</td>
</tr>
<tr>
<td>500</td>
<td>4.537</td>
<td>13.231</td>
<td>19.582</td>
<td>39.373</td>
</tr>
<tr>
<td>600</td>
<td>5.578</td>
<td>8.508</td>
<td>17.738</td>
<td>41.471</td>
</tr>
<tr>
<td>700</td>
<td>5.666</td>
<td>8.908</td>
<td>16.704</td>
<td>39.480</td>
</tr>
<tr>
<td>800</td>
<td>5.081</td>
<td>7.679</td>
<td>15.728</td>
<td>36.916</td>
</tr>
<tr>
<td>900</td>
<td>4.297</td>
<td>7.558</td>
<td>14.846</td>
<td>33.338</td>
</tr>
<tr>
<td>1000</td>
<td>4.495</td>
<td>8.256</td>
<td>15.796</td>
<td>32.683</td>
</tr>
<tr>
<td>Random</td>
<td>4.488</td>
<td>7.213</td>
<td>14.993</td>
<td>15.061</td>
</tr>
</tbody>
</table>

Table 5.1: Maximum prediction errors for each small molecule kriging model as the training set size is increased. Maximum prediction errors for the randomly constructed training sets are presented in the final row.

While there are two examples where the maximum prediction error of a MaxMin constructed training set is lower than that of the random benchmark (Water and NMA with training set sizes of 900), for the most part we can conclude that the MaxMin algorithm makes no consistent improvement on random subset selection. Indeed, it seems more likely that the MaxMin algorithm constructs training sets that are worse than the random benchmarks based on the maximum errors, particularly glycine where the maximum error is double that of the randomly constructed training set.
5.2. GREEDY HEURISTICS

5.2.3 Deletion

Whilst similar in nature to the MaxMin algorithm, the greedy Deletion algorithm judges an optimal subset to be one which minimises the amount of geometric clustering between points in the subset. To achieve this, the greedy Deletion algorithm iteratively removes points from the candidate set by rejecting a point at each iteration that is closest to any other point in the candidate set, as outlined in Algorithm 3.

1. Take $S = \Omega$
   while $|S| > n$
   
   Take $x_i$ and $x_j$, where $x_i, x_j \in S$ such that $d(x_i, x_j)$ is minimal
   Determine $c_i = \min_{k \in S \setminus \{i, j\}} d(x_i, x_k)$ and $c_j = \min_{k \in S \setminus \{i, j\}} d(x_j, x_k)$
   Remove $i$ from $S$ if $c_i < c_j$, or $j$ from $S$ if $c_j < c_i$
   
   Algorithm 3: Deletion Algorithm

Qualitatively, it seems as if both the MaxMin and Deletion algorithms will yield the same subset by approaching the problem from different ends. Whilst the MaxMin algorithm builds a subset from the “bottom up” (i.e. adding points that are furthest from all other points), the Deletion algorithm builds a subset from the “top down” (i.e. removing points that are closest to all other points). However, the selection function is the same for both, so one may be forgiven for assuming the two algorithms to converge upon the same subset. We can demonstrate a simple example where the two give differing subsets in Table 5.2.
Table 5.2: Comparison of the greedy MaxMin and greedy Deletion algorithms on a candidate set in $\mathbb{R}^2$, with the aim of constructing a subset, where $|S| = 4$. The first row, corresponding to the steps taken by the MaxMin algorithm, adds two points (red) at each iteration to the subset, whose distance is the largest in the entire candidate set. The second row corresponds to the steps taken by the Deletion algorithm. Here, the two points in the candidate set which are closest together are selected (green). Their nearest neighbours (orange) are then evaluated, and the point with the closest nearest neighbour is removed from the candidate set.
5.2. **GREEDY HEURISTICS**

We see that both greedy algorithms add the points A and B to the subset, as would be expected if our aim is to construct a geometrically diverse subset. However, the two greedy algorithms differ when selecting points from the cluster, (CDEFG), in the candidate set. The MaxMin algorithm selects two points which qualitatively seem to be the closest two in the cluster, F and G, whilst the Deletion algorithm selects the two points that appear to be maximally separated, D and F. Of course, this example is cherry-picked to demonstrate the difference between the two greedy algorithms, and we do not mean to imply that the Deletion algorithm gives a more geometrically diverse subset than the MaxMin algorithm. Counter examples can be constructed which favour MaxMin over Deletion.

In Figure 5.2, we present the average error in the total molecular energy for the same 500 external test cases that were used in Section 5.2.2 as we increase of the size of the training set constructed by the greedy Deletion algorithm.

The results in Figure 5.2 are very similar to those presented for the MaxMin algorithm in Figure 5.1. If anything, the greedy Deletion algorithm appears to be slightly outperformed by the greedy MaxMin algorithm owing to the slower rate of convergence to the random benchmark in both the water and methanol systems. We also present the maximum errors associated with each kriging model in Table 5.3.
Figure 5.2: Average molecular energy error, with accompanying standard deviations, in the testing set conformations given a training set constructed by the Deletion algorithm from the respective candidate sets at a number of training set sizes. In each pane, the dark black line corresponds to the average molecular energy error in the testing set given a training set of size 1000, generated randomly from the candidate set.
5.2. *GREEDY HEURISTICS*

<table>
<thead>
<tr>
<th>Training Set Size</th>
<th>Water (kJ mol$^{-1}$)</th>
<th>Methanol (kJ mol$^{-1}$)</th>
<th>NMA (kJ mol$^{-1}$)</th>
<th>Glycine (kJ mol$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>11.716</td>
<td>21.181</td>
<td>54.513</td>
<td>47.916</td>
</tr>
<tr>
<td>200</td>
<td>7.692</td>
<td>17.110</td>
<td>32.058</td>
<td>42.534</td>
</tr>
<tr>
<td>300</td>
<td>7.001</td>
<td>16.342</td>
<td>21.750</td>
<td>39.598</td>
</tr>
<tr>
<td>400</td>
<td>7.676</td>
<td>13.813</td>
<td>22.212</td>
<td>50.454</td>
</tr>
<tr>
<td>500</td>
<td>5.589</td>
<td>15.473</td>
<td>20.631</td>
<td>39.508</td>
</tr>
<tr>
<td>600</td>
<td>5.413</td>
<td>14.992</td>
<td>18.562</td>
<td>38.762</td>
</tr>
<tr>
<td>700</td>
<td>5.372</td>
<td>11.325</td>
<td>17.696</td>
<td>36.581</td>
</tr>
<tr>
<td>800</td>
<td>5.335</td>
<td>12.976</td>
<td>16.319</td>
<td>35.033</td>
</tr>
<tr>
<td>1000</td>
<td>4.486</td>
<td>16.217</td>
<td>15.043</td>
<td>30.707</td>
</tr>
<tr>
<td>Random</td>
<td>4.488</td>
<td>7.213</td>
<td>14.993</td>
<td>15.061</td>
</tr>
</tbody>
</table>

Table 5.3: Maximum prediction errors for each small molecule kriging model as the training set size is increased. Maximum prediction errors for the randomly constructed training sets are presented in the final row.

Relative to the maximum errors in Table 5.3, the Deletion methanol model is particularly poor. The training sets constructed for methanol have maximum errors that are virtually double those of the MaxMin training sets. In saying this, the maximum error attributed to the Deletion glycine model is roughly 2 kJ mol$^{-1}$ lower than the equivalent MaxMin model. The prevailing conclusion is that the maximum errors from both greedy algorithms are significantly larger than the random benchmarks in most cases. The large maximum errors attributed to the greedy algorithms are particularly strange, and we have struggled to explain them. Even given the alterations we will make in the next section to the greedy algorithms, it is quite difficult to reason why, in their current incarnations, they perform so poorly relative to the random methodology.


### 5.2.4 Alternative Distance Metric

Before dismissing the greedy heuristics, we need to remedy a conceptual flaw with the above implementations. In both the MaxMin and Deletion schemes, points are added or removed from the subset based on a geometric criterion, which in effect attempts to maximise the Euclidean distance between all points within the resultant subset.

Kriging makes a prediction of the response function at untrained points based on the value of response function at trained points, weighted by the distance between the trained point and the prediction point. So, training points that are further away from the prediction point contribute less to the prediction of the response function at the prediction point, while training points that are closer contribute more. This predictive methodology is captured by the covariance between points, i.e. the kriging kernel, and is governed by Equation \((2.4.1)\).

Equation \((2.4.1)\) scales the \(L^1\)-norm, or taxicab metric, in the \(k^{th}\) degree of freedom by \(\theta_k\). This scaled \(L^1\)-norm is then exponentiated in the \(k^{th}\) degree of freedom by \(p_k\). The exponential of the negative sum of these “scaled-exponentiated” values is then taken over all degrees of freedom.

It stands to reason that the covariance between two points has some informal role as a distance metric. It resembles an \(L^1\)-norm on a scaled Gaussian space in the limit that \(p_k = 2 \ \forall k = 1, \ldots, d\), a scaled exponential space in the limit that \(p_k = 1 \ \forall k = 1, \ldots, d\), and some scaled intermediary space otherwise. At the very least, without wishing to present a rigorous proof that the covariance and distance between points are correlated, we posit that adopting the Euclidean norm as a selection function is not appropriate.
5.2. **GREEDY HEURISTICS**

An issue immediately arises in adapting the MaxMin and Deletion algorithms for the covariance space. We require *a priori* knowledge of $\{\hat{\theta}, \hat{p}\}$, solutions to the maximum log-likelihood function of Equation (2.4.9), to evaluate the correlation between points. This information is obviously not available until a kriging model has been constructed from a training set. A potential workaround for this would be to approximate $\{\hat{\theta}, \hat{p}\}$ using a small training set. We could subsequently use these approximations to approximate the covariance between points.

We have attempted to ascertain whether such an approximation is viable. We have evaluated the evolution of $\theta$ as the training set size increases for each IQA component, for each atom in the water molecule. We have omitted the evolution of $p$ from our analysis since its value remains virtually constant as the training set size increases. We have qualitatively established that the values of $\theta$ stabilise when the training set size reaches roughly 700. We satisfy ourselves with qualitative convergence since we only mean to approximate the hyperparameters, and so we do not attempt to invoke any numerical measure for convergence. We have verified that this qualitative convergence is exhibited for methanol, but it becomes increasingly difficult to analyse the data in higher-dimensional spaces. As such, we have not undertaken any equivalent analysis for NMA or glycine.

We have revised the greedy algorithms of 5.2.2 and 5.2.3 to utilise the approximate kriging hyperparameters, obtained from kriging models of a randomly generated training set for each small molecule. Each training set comprises 1000 points, which is perhaps excessive given that we have established that convergence occurs at roughly 700 training points. The results for the MaxMin and Deletion algorithms with the adapted distance metric are given in Figures 5.3 and 5.4 and Tables 5.4 and 5.5.
Figure 5.3: Average molecular energy error, with accompanying standard deviations, in the testing set conformations given a training set constructed by the MaxMin algorithm utilising the kriging covariance as a distance metric. In each pane, the dark black line corresponds to the average molecular energy error in the testing set given a training set of size 1000, generated randomly from the candidate set.
Figure 5.4: Average molecular energy error, with accompanying standard deviations, in the testing set conformations given a training set constructed by the Deletion algorithm utilising the kriging covariance as a distance metric. In each pane, the dark black line corresponds to the average molecular energy error in the testing set given a training set of size 1000, generated randomly from the candidate set.
Maximum Prediction Error (kJ mol$^{-1}$)

<table>
<thead>
<tr>
<th>Training Set Size</th>
<th>Water</th>
<th>Methanol</th>
<th>NMA</th>
<th>Glycine</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>9.961</td>
<td>20.895</td>
<td>40.027</td>
<td>38.391</td>
</tr>
<tr>
<td>400</td>
<td>7.904</td>
<td>8.171</td>
<td>21.216</td>
<td>15.543</td>
</tr>
<tr>
<td>500</td>
<td>6.995</td>
<td>7.514</td>
<td>20.131</td>
<td>15.056</td>
</tr>
<tr>
<td>600</td>
<td>4.909</td>
<td>8.182</td>
<td>18.683</td>
<td>14.180</td>
</tr>
<tr>
<td>700</td>
<td>4.919</td>
<td>7.015</td>
<td>16.023</td>
<td>14.450</td>
</tr>
<tr>
<td>800</td>
<td>4.351</td>
<td>6.425</td>
<td>16.208</td>
<td>13.331</td>
</tr>
<tr>
<td>900</td>
<td>4.058</td>
<td>6.733</td>
<td>15.426</td>
<td>14.227</td>
</tr>
<tr>
<td>1000</td>
<td>4.490</td>
<td>8.190</td>
<td>16.396</td>
<td>13.979</td>
</tr>
<tr>
<td>Random</td>
<td>4.488</td>
<td>7.213</td>
<td>14.993</td>
<td>15.061</td>
</tr>
</tbody>
</table>

Table 5.4: MaxMin-constructed training set maximum prediction errors for each small molecule kriging model as the training set size is increased. Maximum prediction errors for the randomly constructed training sets are presented in the final row.
5.2. **GREEDY HEURISTICS**

<table>
<thead>
<tr>
<th>Training Set Size</th>
<th>Water</th>
<th>Methanol</th>
<th>NMA</th>
<th>Glycine</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>13.119</td>
<td>18.764</td>
<td>45.476</td>
<td>25.421</td>
</tr>
<tr>
<td>300</td>
<td>7.256</td>
<td>7.892</td>
<td>21.817</td>
<td>12.393</td>
</tr>
<tr>
<td>400</td>
<td>7.636</td>
<td>8.440</td>
<td>16.468</td>
<td>19.797</td>
</tr>
<tr>
<td>500</td>
<td>5.392</td>
<td>6.744</td>
<td>16.781</td>
<td>13.618</td>
</tr>
<tr>
<td>600</td>
<td>5.730</td>
<td>6.472</td>
<td>16.532</td>
<td>13.306</td>
</tr>
<tr>
<td>700</td>
<td>5.586</td>
<td>6.624</td>
<td>16.566</td>
<td>14.537</td>
</tr>
<tr>
<td>800</td>
<td>5.007</td>
<td>7.427</td>
<td>16.242</td>
<td>13.738</td>
</tr>
<tr>
<td>1000</td>
<td>4.484</td>
<td>7.095</td>
<td>13.061</td>
<td>13.132</td>
</tr>
<tr>
<td>Random</td>
<td>4.488</td>
<td>7.213</td>
<td>14.993</td>
<td>15.061</td>
</tr>
</tbody>
</table>

Table 5.5: Deletion-constructed training set maximum prediction errors for each small molecule kriging model as the training set size is increased. Maximum prediction errors for the randomly constructed training sets are presented in the final row.

Based on the average errors attributed to each of the kriging models obtained from the MaxMin and Deletion algorithms, we can always find training sets that outperform the random benchmark training set. For water, both the MaxMin and Deletion algorithms construct training sets that outperform the random benchmark by a training set size of 800 points. For methanol, the MaxMin algorithm outperforms the random benchmark by roughly 0.3 kJ mol\(^{-1}\) at a training set size of 700, but the training set constructed by the Deletion algorithm fails to improve upon the random training set. The situation is reversed for NMA, where the Deletion algorithm constructs a training set that outperforms the random training set...
by roughly 0.1 kJ mol$^{-1}$ at a training set size of 1000, whereas the MaxMin training sets are outperformed by the random training set. Finally, for glycine, we find that both the MaxMin and Deletion training sets outperform the random training set by roughly 0.1 kJ mol$^{-1}$.

In terms of implementing these strategies, our mix of results proves problematic. Ideally, we would like to implement a subset selection technique that consistently improves upon the randomly constructed training sets. Whilst both of the greedy algorithms can outperform the random subset selection, both can also be outperformed by the random training set selection methodology.

If we are to qualify our subset selection strategy by the maximum error attributed to the model, then we see from Tables 5.4 and 5.5 that the Deletion algorithm systematically outperforms the random training set construction by a significant amount, almost 2 kJ mol$^{-1}$ for NMA and glycine. MaxMin, on the other hand, is not quite as successful, being outperformed in this category for methanol and NMA. As such, we have tentatively concluded that the greedy Deletion algorithm is a preferable candidate for subset selection, since it commonly outperforms the random subset selection in terms of average error, and systematically outperforms the random subset selection in terms of maximum error.

A number of problems can be attributed to the implementation of the greedy Deletion algorithm. We note that both the maximum error and average error of a kriging model do not decrease monotonically with training set size. There seems to be no way in which we can determine, a priori, an ideal compromise between minimising the size of the training set and error associated with the resultant kriging model. For example, the optimal training set in terms of average error occurs at training set sizes of 800, 700, 1000 and 1000 for water, methanol, NMA and glycine, respectively. In contrast, the optimal training set in terms of
maximum prediction error occurs at training set sizes of 900, 600, 1000 and 1000. One could systematically increase the training set size until the errors associated with some testing set begin to increase, but the construction of numerous kriging models is time-consuming.

We imagine that as the size of the candidate set increases, there is more variety with which to construct a training set. Therefore, a larger candidate set should yield better training sets given an appropriate subset selection methodology. However, employing a “top-down” approach to subset selection is problematic for a large candidate set, as we can illustrate through example. Given a candidate set consisting of 10,000 molecular conformations. If we wish to construct a training set comprising 1000 molecular conformations, the Deletion subset selection requires 9000 iterations, whereas MaxMin subset selection requires 1000 iterations. The additional computational effort required for Deletion over MaxMin is obviously amplified as the candidate set increases. One can then appreciate that a “bottom-up” approach to subset selection from a large candidate set is preferable.

5.3 Sequential Selection

The clustering of training points in conformational space negatively impacts on the quality of a kriging model, where the global predictive capacity of the kriging model is known to deteriorate. However, if the topology of the response function surface is particularly undulant in a certain area, more training points will be required in this area to capture its topological features. Therefore, a purely geometric criterion for subset selection, no matter how elaborate, can never be seen as an optimal means for subset selection for kriging. Ideally, one would possess information about the response function so that undulant regions can be represented by a higher training
density. In other words, some variant of importance sampling is required to direct sampling to these difficult-to-model regions of conformational space.

The issue with this approach for our purposes is that information about the response function is only available after the computationally intensive \textit{ab initio} calculations required to obtain the atomic properties being kriged. One method for circumventing this issue is by crudely computing the response function at a small number of points in conformational space, and evaluating which regions appear to be unsatisfactorily represented. For example, if two training points, \( \mathbf{x} \) and \( \mathbf{x}' \), are deemed sufficiently close, i.e. \( d(\mathbf{x}, \mathbf{x}') < d_{\text{max}} \), we can impose some requirement that the response function is “sufficiently continuous”, i.e. \( |f(\mathbf{x}) - f(\mathbf{x}')| < \epsilon \), where \( d_{\text{max}} \) and \( \epsilon \) are suitably chosen parameters. If the two points fail to meet this criterion, then another point is added to the training set between \( \mathbf{x} \) and \( \mathbf{x}' \) in some fashion.

The issue with the above methodology is that subset selection becomes an inherently serial task. A set of \textit{ab initio} calculations are performed, new conformations are determined followed by a new set of \textit{ab initio} calculations. This process is then iterated until the response function between all points is “sufficiently continuous”. There is no scope for a parallel workflow since each set of newly proposed conformations is explicitly dependent upon the previous set of \textit{ab initio} calculations. In addition, the choice of parameters \( d_{\text{max}} \) and \( \epsilon \) is arbitrary, and presumably system-dependent.

As such, we adopt the scheme employed in the work of Rennen[319], that of \textit{Sequential Selection}. With sequential selection, a large candidate set is constructed, with the response function having been evaluated at each point. An initially small training set is constructed from the candidate set, \( |S| = n_0 \). Using \( S \), a kriging

\footnote{The manner in which this small training set is constructed is irrelevant if it sufficiently small.}
model is constructed and the response function at each point in the candidate set but not in the training set, $\Omega \setminus S$, is predicted. Some subset of the worst predicted points in $\Omega \setminus S$ is added to the training set, and the whole process iterated until the training set is of adequate size, or all points are predicted with an error below some tolerance. The latter criterion can be formalised by defining the error of prediction for a conformation, $x_i$, to be given by $\mathcal{E}(x_i) = |f_{\text{pred}}(x_i) - f_{\text{act}}(x_i)| < \mathcal{E}$, where $f_{\text{pred}}(x_i)$ and $f_{\text{act}}(x_i)$ are the predicted and actual values of the response function at $x_i$, respectively, and $\mathcal{E}$ is a pre-defined error tolerance. The sequential selection scheme is illustrated in Algorithm 4.

1. Take $S = \emptyset$
2. Select $n_0$ points from $\Omega$ and add to $S$
   while ($|S| < N$) $\lor$ ($\mathcal{E}(x_i) < \mathcal{E}$ $\forall x_i \in \Omega \setminus S$) do
   Calculate $\mathcal{E}(x_i) = |f_{\text{pred}}(x_i) - f_{\text{act}}(x_i)|$ $\forall x_i \in \Omega \setminus S$
   Add some subset of those points whose $\mathcal{E}(x)$ is highest to $S$
end

**Algorithm 4:** Sequential Selection Algorithm

We have left the definition of “some subset of the worst predicted points” purposefully vague as there are a number of strategies one can adopt in defining this subset. The best possible training set will result when the worst predicted point is added to the training set at each iteration. Following the formalism of Rennen, we call this strategy SS1. However, building up a training set one point at a time, whilst having to recompute a kriging model at each iteration requires a great deal of time, and is naturally not an ideal strategy.

A second approach would be to add the $n$ worst predicted points to the training set at each iteration, which we denote SS$n$. We indicate that SS1 is the limiting case to the SS$n$ methodology. Adopting the SS$n$, with $n > 1$, strategy is not equivalent to an accelerated version of SS1. In adding the $n$ worst predicted points, where $n > 1$, from a kriging model, it is likely that a number of these worst predicted
points reside in a similar, poorly-modelled region of conformational space. As such, the addition of these points to the training set results in a clustering of points in the new training set, leading to over-training in that region. However, SS$n$ is a far more efficient means for constructing a training set than SS1. Following Rennen, we have chosen to evaluate the performance of SS10, where the 10 worst predicted points are added to the training set at each iteration.

The final approach we propose is an attempt to circumvent the potential for over-training as we have indicated is a flaw to the SS$n$ scheme. To do this, we add $n$ points to the training set from uniformly sampling the $m$ worst predicted points, where $m > n$. This scheme is denoted SS$mn$. Again, following Rennen, we have chosen the SS1040 methodology, where of the 40 worst predicted points, we select 10 uniformly (i.e. the 40$^{th}$, 36$^{th}$, 32$^{nd}$, ... worst predicted points) and add them to the training set at each iteration. The SS$mn$ methodology works under the assumption that points with similar errors occupy similar regions of conformational space. By uniformly selecting points from the worst points, one prevents the addition of conformationally similar points to the training set.

The sequential selection methodology is highly labour intensive, and so we restrict ourselves to a single test case; the exchange-correlation energy of the carbon atom in methanol. We have found in past work that the exchange-correlation energy appears to be the most difficult to model by kriging, and so seems an appropriate candidate for evaluating the performance of the sequential selection methodologies outlined above. We use the candidate set comprising 3500 models that was used in Section 5.2, and construct random training sets comprising 595 conformations for SS1, and 550 conformations for both SS10 and SS1040. Each training set is then built up to 600 conformations. The performance of the resulting models is evaluated on the rest of the candidate set not included in the training set at
5.3. **SEQUENTIAL SELECTION**

Each iteration and the appropriate points are added to the training set from the candidate set at each iteration. Additionally, we have constructed an external testing set of 200 conformations to compare the performance of each sequential selection algorithm.

Figure 5.5 gives the cumulative distribution functions for the errors associated with each of the sequential selection methodologies along with a random benchmark. It is immediately evident that both the SS10 and SS1040 models outperform the random benchmark quite significantly considering they differ by only 50 training points. Surprisingly the SS1 model performs rather poorly, but we can attribute this to having only added 5 points to the training set, which we assume to have little impact on the quality of the training set. For the SS1 model to perform better, it would be preferable to begin from an initial training set size of 550 and iterate the algorithm 50 times to generate a training set with 600 points. However, as we have previously mentioned, this is a particularly laborious task.
Figure 5.5: Cumulative distribution functions of errors associated with the sequential selection kriging models. SS1 (green), SS10 (blue) and SS1040 (yellow) are compared to a randomly constructed training set benchmark (red). Whilst the SS1 model performs similarly to the random benchmark, the SS10 and SS1040 both considerably outperform the random benchmark.
We can also plot the average error, maximum error and average standard deviations for each method over the course of iteratively building the training sets. The tabulated data for these quantities are presented in Tables 5.6, 5.7 and 5.8, respectively.

### Average Prediction Error (kJ mol\(^{-1}\))

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Random</th>
<th>SS1</th>
<th>SS10</th>
<th>SS1040</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>0.811</td>
<td>0.774</td>
<td>0.790</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.856</td>
<td>0.856</td>
<td>0.638</td>
</tr>
<tr>
<td>3</td>
<td>0.798</td>
<td>0.852</td>
<td>0.643</td>
<td>0.596</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.646</td>
<td>0.638</td>
<td>0.779</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.854</td>
<td>0.653</td>
<td>0.630</td>
</tr>
</tbody>
</table>

Table 5.6: Average errors associated with the SS1, SS10 and SS1040 kriging models relative to the random benchmark.

### Maximum Prediction Error (kJ mol\(^{-1}\))

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>Random</th>
<th>SS1</th>
<th>SS10</th>
<th>SS1040</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>3.317</td>
<td>3.620</td>
<td>2.964</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>4.548</td>
<td>3.344</td>
<td>2.288</td>
</tr>
<tr>
<td>3</td>
<td>3.418</td>
<td>4.230</td>
<td>2.423</td>
<td>2.082</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>2.191</td>
<td>2.281</td>
<td>2.988</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>3.488</td>
<td>2.649</td>
<td>2.376</td>
</tr>
</tbody>
</table>

Table 5.7: Maximum errors associated with the SS1, SS10 and SS1040 kriging models relative to the random benchmark.
Table 5.8: Standard deviation of errors associated with the SS1, SS10 and SS1040 kriging models relative to the random benchmark.

The data in the tables corroborates well with what we observe in the cumulative distribution of errors in Figure 5.5. Each metric confirms that the SS10 and SS1040 algorithms outperform the random benchmark by a great deal. It is a little difficult to quantify whether SS10 outperforms SS1040, or vice versa, as the data presented appear to fluctuate, with one outperforming the other at one iteration, and then reversing for the next. However, a maximum error almost 1 kJ mol\(^{-1}\) and an average error just over 0.2 kJ mol\(^{-1}\) lower than the random benchmark is vastly impressive given so few iterations of the sequential selection algorithm.

As mentioned at the beginning of this section, the sequential selection subset selection strategies are not ideal for our purposes owing to the need for a large volume of \textit{ab initio} calculations that can be parsed into the resultant training set. It is for this reason that we do not pursue the sequential selection strategies in the following work. The results presented in this section confirm that the sequential selection strategies certainly outperform the random benchmark, but they are not feasible strategies owing to their wastefulness of computationally intensive data, and the requirement that kriging models be built iteratively. However, evaluating the performance of the sequential selection methodologies is of use, since they can be used as benchmark subset construction calculations in future work.
Both the greedy heuristics and the sequential selection algorithms require a candidate set of data points from which to generate a subset. The sequential selection algorithms are particularly expensive as the explicit calculation of the response function is required for each data point in the candidate set. Calculations of the response function are so time consuming that they are not feasible for large systems or large candidate sets.

A second unrelated issue is that the construction of a subset from a discrete candidate set is not optimal. The subset that one is able to construct is entirely dependent upon the quality of the points within the candidate set. For example, if each point in the candidate set is clustered, then we have no means to prevent our subset from being clustered. A preferable scheme would allow for the selection of datapoints from the continuous space that we are attempting to model, without the restrictions imposed by selection from a discrete set of points.

Formally speaking, given some pre-defined domain, we wish to add points to a subset from within the domain, such that the added points are as mutually dissimilar as possible. An immediate issue is in the definition of the domain within which we sample. To sample within a domain, one requires boundary conditions, allowing for the elucidation of whether a point lies within the domain or outside of it. We avoid discussion of these boundary conditions momentarily. This is a topic that merits an entirely separate discussion, which we undertake in Section 5.5. Rather, we proceed under the assumption that we have been given some initial subset \( S := \{ x_1, \ldots, x_s \} \), from which we are able to determine the boundary conditions of the subset, i.e. we have knowledge of the polytope on \( \mathbb{R}^d \) bounding \( S \).
Our formal problem is actually a variant of the \textit{Largest Empty Sphere Problem},

**Definition** (Largest Empty Sphere). \textit{For a set of points, }\(S := \{x_1, \ldots, x_s| x \in \mathbb{R}^d\}\), \textit{and the associated bounding polytope, }\(D\), \textit{find the largest hypersphere that contains no elements of }\(S\), \textit{whose centre is in the polytope}. Or,

\[
\max_{x \in D} \min \{||x - x_i|| \mid i = 1, \ldots, s\}.
\]  

(5.4.1)

A qualitative enunciation of the above definition will prove instructive. We are required to find the radial distance between a query point, \(x\), and the closest point in the subset, \(x_i\). We then find the maximum such distance over every conceivable query point in the domain\[320\]. Whilst a conceptually simple problem, its solution is NP-hard, and brute force approaches (which scale as \(O(|S|^4)\)) are conventionally adopted. However, for high-dimensional spaces, such systematic means for solution are not tractable. Note also that the number of possible query points within the domain is uncountably infinite, and so some scheme is required to discretise the space.

We here find it convenient to introduce the concept of a Voronoi tessellation:

**Definition** (Voronoi Tessellation). \textit{Given a set of points, }\(S := \{x_1, \ldots, x_s| x \in \mathbb{R}^d\}\) \textit{and a distance metric, }\(d(x_i, x_j)\). \textit{The Voronoi cell of }\(x_k \in S\), \(V_k\), \textit{is defined as}

\[
V_k := \{x \in \mathbb{R}^d \mid d(x, x_k) \leq d(x, x_m)\},
\]  

(5.4.2)

where \(x_m \in S \setminus x_k\).

The Voronoi tessellation possesses a relatively straightforward interpretation; given a set of points, the Voronoi cell belonging to a query point is equal to the region of space that is closer to (or equal to) the query point than any other point in the set. The equality condition is of particular importance, since it implies the existence
of points that belong to more than one Voronoi cell. When a point belongs to
two Voronoi cells, it defines a partitioning between the Voronoi cells. When a
point belongs to three or more Voronoi cells, is defines a vertex of the Voronoi
tessellation. These features are demonstrated in Figure 5.6.

Figure 5.6: Voronoi tessellation of 12 points in $\mathbb{R}^2$, denoted by black points. The
Voronoi cell corresponding to each point is coloured. Voronoi cells are partitioned
from one another by lines. Where three or more Voronoi cells meet, a vertex is
formed. Image generated by the interactive Voronoi diagram generator, located at

The notion of the Voronoi tessellation is immediately generalisable to a space
of any dimensionality. In $\mathbb{R}^3$, for example, the boundary between two Voronoi
cells is a plane, the boundary between three Voronoi cells is a line, and a vertex is
defined by the boundary between four or more Voronoi cells. A vertex of a Voronoi
tessellation in $\mathbb{R}^d$ is defined by the boundary between $d + 1$ or more Voronoi cells.

Solution to the largest empty sphere problem and the Voronoi tessellation on $S$
are related. It can be shown that the centre of the largest empty sphere in $S$
coincides with the vertex of the Voronoi tessellation of $S$ that is furthest from all
CHAPTER 5. SUBSET SELECTION FOR MACHINE LEARNING

the elements of $S$. This link allows us to recast (5.4.1) as a maximisation over the discrete vertices of a Voronoi tessellation as opposed to a maximisation over the continuous domain enclosed by the bounding polytope\cite{321}. Solution to the largest empty sphere problem by the Voronoi tessellation scales as $O(|S| \log |S|)$, which is certainly an appealing prospect. However, neither the memory requirements or the scaling with dimensionality of the space in which $S$ is defined are well-documented.

Given this link between the Voronoi tessellation and the largest empty sphere, we propose a novel subset construction algorithm, which we enumerate in Algorithm 5, and refer to as the Iterative Voronoi Subset Selection (IVSS) algorithm.

1. Take $S = \{x_1, ..., x_s\}$
   while $|S| < N$ do
     Construct the Voronoi tessellation of $S$
     Collect the vertices of the Voronoi tessellation, $V := \{v_1, ..., v_n\}$
     Evaluate $\max_{v \in V} \min_{x_i \in S} ||v - x_i||$, where $x_i \in S$
     Add corresponding $v$ to $S$
   end

Algorithm 5: Iterative Voronoi Subset Selection Algorithm

We have used water to establish a proof-of-concept for the IVSS. An initial training set was randomly generated, consisting of 10 points. We have then used the freely available qVORONOI\cite{322} code to Voronoi tessellate the training set. qVORONOI operates by first computing the Delaunay triangulation of the points provided, and subsequently obtains the Voronoi tessellation as the dual space to the Delaunay triangulation\cite{323}. The training set for water was iteratively grown until the training set size had reached 100. Comparison was then conducted against a randomly constructed training set also containing 100 points. An external testing set comprising 300 points was used for validation. We present the cumulative distribution function for the errors in Figure 5.7.

From Figure 5.7, it is evident that the Iterative Voronoi scheme significantly out-
Figure 5.7: Cumulative distribution function of errors for a training set constructed by the Iterative Voronoi Subset Selection (blue) and a randomly constructed training set (red).
performs the randomly constructed training set. The mean error associated with
the random training set is 2.64 kJ mol$^{-1}$, whereas that associated with the IVSS
training set is 2.09 kJ mol$^{-1}$. The maximum error associated with the random
training set is 17.76 kJ mol$^{-1}$, whereas that associated with the IVSS training
set is 12.83 kJ mol$^{-1}$. We note that these values are not quite as good as those
obtained in Section 5.2.4, but we can associate this with an unoptimal selection
of points included in the initial subset. Since the IVSS can only add points from
within the domain bounded by the initial subset, an informed choice of these points
is required. As such, we restrict our critique to direct comparison with the random
subset alone.

We have encountered a number of problems in moving to higher dimensional cases.
The qVORONOI algorithm stores each boundary surface in main memory, and
computes the vertices of the Voronoi tessellation as the intersection of $d$ such
hyperplanes. For methanol, where $d = 12$, we have found that the memory required
to store this amount of data exceeds the amount of memory available on the HPC
facilities that are available to us at the present time. Indeed, one can appreciate
that when the Voronoi tessellation is taking place in a high dimensional space,
with a large number of Voronoi cells, the procedure is simply not tractable.

As such, while we have provided a proof-of-concept, illustrating that the IVSS
scheme significantly outperforms a random benchmark, it is unfortunately not a
viable methodology given the high dimensionality of the spaces that we regularly
work in. It is possible that the dimensionality of the problem could be reduced by
a judicious choice of dominant dimensions for the problem, but this process has
been a contentious issue without our group, and is the subject of ongoing research.
5.5 Domain of Applicability

5.5.1 Introduction

A kriging model is valid upon the domain it has been trained. Since the kriging model has no information concerning the topology of the response function outside of the training domain, it cannot make informed predictions of the response function outside of the training domain. This is not to say that kriging is not predictive in these regions. Analysis of Equation (2.4.11) shows that as a query point gets further from all other training points, the second term of Equation (2.4.11) goes to zero. The prediction of the response function at a query point is then equal to the mean of the response function as computed from the optimal hyperparameters, \( \hat{\theta}, \hat{p} \), Equation (2.4.4). Therefore, a kriging model can predict a point outside of the training set to have a response function value equal to the mean of the response function over the training domain. However, kriging is not meant to be extrapolative, and so if the response function has a complex topology in untrained regions, the mean is not a particularly useful prediction.

Our aim is to ensure that the predictions made by a kriging model are consistently accurate. We can go some way to achieving this aim by specifying the domain over which the kriging model has been trained, and is therefore valid. We term this domain the Domain of Applicability[324], or DoA. To characterise a domain, we require some means to define its boundary surface. We can subsequently elucidate, before making a prediction, whether the kriging model is a valid description of the response function at a query point. If the query point does not lie in the DoA, we can either add it to the training set and retrain the kriging model, or flag the query point as being inappropriately modelled by the kriging model.
A number of strategies exist for defining the DoA of a set of data, which we briefly review.

**Range-Based** [325, 326]

One geometric way of defining the boundary surface would be to consider the conformational space of a molecule in $\mathbb{R}^d$. Along each degree of freedom, we can determine

$$x_k^{\text{max}} = \max_{x_k \in S} x_k, \quad x_k^{\text{min}} = \min_{x_k \in S} x_k,$$

(5.5.1)

the maximum and minimum values that the training set spans. By placing a $(d-1)$-dimensional hyperplane orthogonal to each axis $x_k$ at $x_k = x_k^{\text{max}}$ and $x_k = x_k^{\text{min}}$, we can construct a DoA as everything enclosed by these hyperplanes. So, in $\mathbb{R}^3$, a series of planes on $\mathbb{R}^2$ define the boundaries of the DoA, which is a cube. However, it becomes immediately apparent that such a scheme is bound to fail. Imagine four points make up a training set in $\mathbb{R}^3$, and form the vertices of a tetrahedron. Defining the DoA as a cube is inappropriate since a tetrahedron occupies one fifth of the volume of a bounding cube, and so four fifths of the space contained within the DoA is not represented within the training set. Those untrained regions lying within the DoA are referred to as “empty regions” [327].

**Distance-Based** [328, 329]

Distance-Based methods can be used to evaluate the DoA without explicitly computing the boundary surface of the DoA. The Mahalanobis distance [330] is a measure of the distance between a point (the query point) and a distribution (the training set). Given a test set, one can evaluate the distance between each query point within the test set and the training set, and determine points within the...
5.5. **DOMAIN OF APPLICABILITY**

DoA to be those which are less than a threshold Mahalanobis distance from the training set. The existence of a threshold distance is unwelcome as it renders the methodology dependent on some user-defined parameter, which is difficult to characterise.

A variant, and slightly more robust form, of the conventional Distance-Based method is the $k$–nearest neighbours approach, where the distances between a query point and the $k$ nearest training points is found. If all of these distances fall within some threshold distance, then one can assume the query point to be well-predicted[331]. However, the $k$ nearest neighbours approach still relies upon a user-defined parameter, and so is not ideal.

**Probability Density Distribution[332]**

The final methodology we wish to mention is the Probability Density Distribution Method, where for every test point, the training data is evaluated in the corresponding region and the density of training points evaluated. If the density is lower than some threshold value, then the test point is considered to lie outside of the DoA. Probability Density Distributions are able to identify empty regions of the training space, making them rigorous means for defining the DoA. However, it should be noted that Probability Density Distributions are particularly difficult to implement in high-dimensional spaces[333], and the parameterisation of sufficient sampling density is, like the Mahalanobis distance, undesirable.

Alternative methodologies for the evaluation of a DoA do exist, such as Decision Trees[334] and Stepwise Approaches[335], but have been less well-reported in the literature. The attentive reader will notice that each of the references that has been cited in this section comes from the field of Quantitative Structure-Activity
Relationship (QSAR). Without wishing to delve into the complexities of this field, suffice to say that our problems are somewhat different from the problems tackled in QSAR. The biggest difference is in the dimensionality of the space in which we train our kriging models. Values taken by data points are also frequently discrete. The concept of a DoA for our particular problem is not well-characterised, and so we attempt to pursue an alternative methodology to define the DoA.

5.5.2 The Convex Hull

An appealing method for defining the boundary surface of the DoA is the convex hull of the set of the training set[336]. The convex hull, $C_S$, of a set of points, $S$, is defined as the set of all convex combinations of the constituent points of $S$, where:

**Definition** (Convex Combination). For a set of points, $S := \{x_1, ..., x_n| x \in \mathbb{R}^d\}$, their convex combination is defined as

$$k_1x_1 + k_2x_2 + ... + k_nx_n \quad s.t. \quad \begin{cases} 
\sum_{i=1}^{n} k_i = 1 \\
0 \leq k_i \leq 1 \quad \forall i = 1, ..., n
\end{cases}$$

The set of all convex combinations (i.e. each possible set of $\{k_i\}$ subject to the constraints) of $S$ defines a minimal domain that contains all elements of $S$. The convex hull for $S$ is then given by

$$C_S := \bigcup \left\{ \sum_{x_i \in S} k_i x_i \right\} \quad s.t. \quad \begin{cases} 
\sum_{i=1}^{n} k_i = 1 \\
0 \leq k_i \leq 1 \quad \forall i = 1, ..., n
\end{cases}$$

(5.5.2)

The convex hull defines a $d$–polytope surrounding all points in a set on $\mathbb{R}^d$. To construct a $d$–polytope in $\mathbb{R}^d$, one requires at least $d + 1$ points. So, on $\mathbb{R}$,
two points, \( \{A, B\} \), are required to construct a 1-polytope, i.e. a straight line. The convex hull of \( \{A, B\} \) consists of all points between \( \{A, B\} \), each of which can be written as a linear combination of \( \{A, B\} \), subject to the constraints of a convex combination. On \( \mathbb{R}^2 \), three points, \( \{A, B, C\} \), are required to construct a 2-polytope, i.e. a triangle. The convex hull of \( \{A, B, C\} \) consists of all points within the triangle defining the convex hull, each of which can be written as a linear combination of \( \{A, B, C\} \), subject to the constraints of a convex combination. We can generalise this result to as many dimensions as required.

If some query point, \( y \), lies within the convex hull, \( C_S \), it can be written as a convex combination of all the points contained within \( S \) which satisfies the constraints placed on the \( \{k_i\} \). Or, algebraically,

\[
\begin{bmatrix}
    x_1 & \ldots & x_n \\
    1 & \ldots & 1 \\
\end{bmatrix}
\begin{bmatrix}
    k_1 \\
    \ldots \\
    k_n \\
\end{bmatrix} =
\begin{bmatrix}
    y \\
    1 \\
\end{bmatrix},
\]

(5.5.3)

where we have introduced the vector \( k = \begin{bmatrix} k_1 & \ldots & k_n \end{bmatrix}^\top \). Note that the final equation ensures the normalisation of \( k \). (5.5.3) corresponds to a set of \( d + 1 \) linear equations in \( n \) unknowns. For the typical case where \( n \gg d + 1 \), the system of equations is underdetermined, and a solution vector \( k \) cannot be uniquely defined.

One way to circumvent this system of underdetermined equations would be to explicitly compute \( C_S \). However, with conventional software, this is not possible since the best methods available scale as \( O(nd/2) \), which is problematic for the high-dimensional spaces we consider with a great deal of training points. Indeed, we have found that the popular software package, qHull\cite{322}, is not capable of computing the convex hull of a system in \( \mathbb{R}^{12} \) and above. We can, however, make
explicit use of Carathédory’s Theorem to form a system of \( d + 1 \) equations in \( d + 1 \) unknowns. Carathédory’s Theorem states:

**Theorem** (Carathédory’s Theorem). Given a point \( \mathbf{y} \in \mathbb{R}^d \) that lies within \( C_S \), then,

\[
\exists \mathcal{J} := \left\{ \mathbf{x}_1, \ldots, \mathbf{x}_{d+1} \mid \mathcal{J} \subseteq S \right\},
\]

such that \( \mathbf{y} \) lies within \( C_{\mathcal{J}} \).

In other words, we are able to define a subset, \( \mathcal{J} \subseteq S \), of cardinal \( d + 1 \), such that \( \mathbf{y} \) lies within the convex hull of \( \mathcal{J} \). We can then revise (5.5.3) to

\[
\begin{bmatrix}
\mathbf{x}_1 & \ldots & \mathbf{x}_{d+1} \\
1 & \ldots & 1
\end{bmatrix}
\begin{bmatrix}
k_1 \\
\vdots \\
k_{d+1}
\end{bmatrix}
= 
\begin{bmatrix}
\mathbf{y} \\
1
\end{bmatrix},
\tag{5.5.4}
\]

with \( \mathbf{k} = \begin{bmatrix} k_1 & \ldots & k_{d+1} \end{bmatrix}^\top \). (5.5.4) is now a series of \( d + 1 \) equations in \( d + 1 \) unknowns, and has a unique solution. If \( \mathbf{y} \) lies within \( C_{\mathcal{J}} \), then there exists a vector \( \mathbf{k} \) that satisfies the normality constraint of \( \mathbf{k} \), with no individual \( 0 \leq k_i \leq 1 \).

The issue with the above methodology is that if \( \mathbf{y} \) lies within \( C_S \), then we cannot arbitrarily select \( \mathcal{J} := \{ \mathbf{x}_1, \ldots, \mathbf{x}_{d+1} \} \subseteq S \) and expect \( \mathbf{y} \) to lie within \( C_{\mathcal{J}} \), since \( C_S \) is defined as \( \bigcup_{\mathcal{J} \subseteq S} C_{\mathcal{J}} \). Therefore, to rigorously specify whether \( \mathbf{y} \) lies within \( C_S \), we must systematically assess whether it lies in \( \binom{n}{d+1} \) possible convex hulls. Combinatorially scaling methods are certainly not desirable where \( n \gg d + 1 \), and so we address this issue in Section 5.5.4.
5.5. **DOMAIN OF APPLICABILITY**

5.5.3 **Test Case**

To evaluate whether the convex hull is an appropriate boundary surface for the DoA, we must test whether points predicted outside of the convex hull defined by the training set have higher errors than those predicted from within the convex hull. A convenient low-dimensional test case is the Ackley function,

\[
 f(x) = -a \exp \left( -b \sqrt{\frac{1}{d} \sum_{i=1}^{d} x_i^2} \right) - \exp \left( \frac{1}{d} \sum_{i=1}^{d} \cos(c x_i) \right) + a + \exp(1),
\]  

(5.5.5)

where recommended values for the parameters are \( a = 20, b = 0.2, c = 2\pi \). The Ackley function is a standard problem for global optimisation algorithms, as the number of local minima renders optimisation methods based on spatial derivatives inefficient. The Ackley function is shown in Figure 5.8.
We use the Ackley function on $\mathbb{R}^2$ as a test case on the domain $x_1, x_2 \in [0, 20]$. Our convex hull will be the polygon with vertices $(0, 0)$, $(20, 0)$ and $(10, 20)$. 1000 random points were chosen on $x_1, x_2 \in [0, 20]$, and were assigned to either belonging within the convex hull or lying outside of the convex hull. With those points within the convex hull, we have constructed a training set of 50 points, including the vertices of the convex hull. Then, 400 points from inside of the convex hull and 400 points from outside of the convex hull were predicted and compared to the analytical results of the Ackley function at those points. The average errors and mean squared errors associated with each test point are given in Figure 5.9.

![Figure 5.9: Absolute error (left) and MSE (right) of predictions made from within the convex hull (yellow) and outside of the convex hull (blue) of the training set.](image)

Both the average error and average MSE of predictions of points outside the convex hull are more than three and eight times larger, respectively, than their counterparts that lie within the convex hull. This result is indicative of the fact that the convex hull is a natural means for partitioning an unbounded space into a DoA and untrained region. Naturally though, we need to apply this methodology to
relevant systems that we wish to krig e. We undertake this in the next section.

5.5.4 Largest $d$-Polytope

As we have seen, invoking Caratheodory’s Theorem allows us to define a convex hull, $C_\varepsilon$ on $\mathbb{R}^d$ by use of $d + 1$ points. However, it is the union of all such convex hulls of $d + 1$ points from a training set of $n$ points that gives the convex hull of all $n$ points. Computation of the convex hull in this fashion is not feasible owing to the associated combinatorial scaling of the problem.

We have formulated two distinct strategies for circumventing the combinatorial scaling. The first is to use some small number of convex hulls formed by $d + 1$ points, and take their union as representative of the convex hull of all $n$ points. The second is to find the $d + 1$ points within $S$ that form the largest possible convex hull from all possible \( \binom{n}{d+1} \) combinations.

We have chosen to adopt the second strategy. If we find it to be inadequate, then finding the largest possible convex hull formed by $d + 1$ points is still of use, since this convex hull can then be used in the first strategy in a combination with a number of other convex hulls. Any approximation strategy will perceive some points within the true convex hull to lie outside of the approximate convex hull, but our aim is to construct a “guiding” DoA.

To find the $d + 1$ points that form the largest possible $C_\varepsilon$ unfortunately also requires an explicit computation of $C_S$, a task that we are attempting to circumvent. As such, we introduce a further approximation, where we construct $C_\varepsilon$ from the largest $d$-polytope (LdP) on $S$. To find the LdP, we invoke a greedy heuristic where the selection function is one which maximises the distance from a point to
the polytope that can be constructed from the points currently within $\mathcal{S}$, as we now illustrate.

Given two points in $\mathcal{S}$, we can construct a 1-polytope, $F(x_1, x_2) \equiv F_1$, and evaluate the distance between $F_1$ and each point in $S \setminus \mathcal{S}$, i.e. $d(x_k, F_1)$, $\forall k \in S \setminus \mathcal{S}$. The point that is furthest from $F_1$ is then added to $\mathcal{S}$, and a 2-polytope constructed, $F(x_1, x_2, x_3) \equiv F_2$. This process is iterated until $|\mathcal{S}| = d + 1$. Generalising, for the $i^{th}$ iteration of the greedy algorithm, an $i$-polytope, $F_i$, is constructed from the points in $\mathcal{S}$ and $d(x_k, F_i)$, $\forall k \in S \setminus \mathcal{S}$ is evaluated. The most distant point is then added to $\mathcal{S}$ and the process is iterated.

We give an explicit low-dimensional example of the above process to demonstrate its working. We take a set of points in $\mathbb{R}^3$, meaning that we require $|\mathcal{S}| = d + 1 = 4$. The first two points in $\mathcal{S}$ are selected by choosing the two most distant points in $S$. The next point is found by evaluating the distance between all points in $S \setminus \mathcal{S}$ and the line, $F_1$, formed by the two points in $\mathcal{S}$, the most distant of which is added to $\mathcal{S}$. Now there are three points in $\mathcal{S}$, meaning that $F_2$ is a plane, and so we evaluate the distance between all points in $S \setminus \mathcal{S}$ and the plane $F_2$. The most distant point is added to $\mathcal{S}$, and we finish with the four required points to form $C_\mathcal{S}$.

We see that the process can be extended indefinitely so long as we can generalise the notion of the distance between a polytope of arbitrary dimension and a point. We can gain some insight into how to perform such a task by an analysis of water, a low-dimensional case ($d = 3$). To construct the LdP, we require the distance from a point to a line segment ($d = 1$), and the distance from a point to a plane ($d = 2$). Both of these problems have geometrical solutions, where the normal vector to the $d$-polytope is computed, and the vector from a point on the $d$-polytope to some point is projected onto the normal vector. The distance from the point to the
domain of applicability to the $d$-polytope is then obtained.

When attempting to validate the above methodology on a water model, we found that given a candidate set of 1000 points, the LdP contained roughly 30 of those points. Such a low proportion of points lying within the LdP implies that the LdP is a poor approximation to $C_S$. We have chosen to revise the notion of a convex combination of the points forming the LdP. Recall that a convex combination of points is defined as

$$k_1x_1 + k_2x_2 + ... + k_nx_n \quad s.t. \quad \begin{cases} \sum_{i=1}^{n} k_i = 1 \\ 0 \leq k_i \leq 1 \quad \forall i = 1, ..., n \end{cases}$$

A number of alternative combinations do exist. For example, the conical combination of points is defined as

$$k_1x_1 + k_2x_2 + ... + k_nx_n \quad s.t. \quad \begin{cases} \sum_{i=1}^{n} k_i \in \mathbb{R}_+ \\ k_i \geq 0 \quad \forall i = 1, ..., n \end{cases}$$

whilst the affine combination places no restriction on the $k_i$. It can be shown that the convex combination of points is a subset of the conical combination of points, which in turn is a subset of the affine combination of points. The more strict the restrictions on the $k_i$, the less space covered by the respective combination of points. By varying the restrictions we place on the $k_i$, we are able to increase the space covered by the combination of points we choose to define the LdP. In the absence of a formal name for this methodology, we simply refer to it as the restricted affine combination of points, which in turn define a restricted affine hull, $A_S$, of a set of points, $S$. When dealing with $A_S$, we will use the acronym RAH-LdP, the restricted affine hull of the LdP.

To minimise the number of free parameters that we are required to use for the restricted affine combination of points, we have chosen the constraint $-\Delta \leq k_i \leq$
CHAPTER 5. SUBSET SELECTION FOR MACHINE LEARNING

1 + ∆ ∀i = 1, ..., n as well as the normality condition on the $k_i$. Thus, depending on the magnitude of ∆, $A_\varphi$ is the union of $C_\varphi$ and a “halo region” that extends the domain of $C_\varphi$. Intuitively, this methodology is valid for our purposes, since the points closer to the LdP have a higher probability of lying within the true convex hull of the training data. If we are scrupulous about our choice of ∆, we should obtain reasonable approximations to the true convex hull of the training data.

For our preliminary work, we have chosen ∆ such that a certain percentage of points within the candidate set lies within the RAH-LdP (the “acceptance percentage”). A suitable metric to evaluate the performance of the LdP is an “average error ratio”- the ratio of the average error from points within the LdP to the average error from points outside of the LdP. So, if the RAH-LdP is able to distinguish between well-predicted and poorly-predicted points, the average error ratio should be small. For water, the progression of both the average error ratio and ∆ with respect to an the acceptance percentage are plotted in Figure 5.10.

We see from Figure 5.10 that ∆ is linear in the acceptance percentage. When the acceptance percentage is low, we find that the average error of points within the RAH-LdP is roughly four times lower than the average error of points outside of the hull. As we increase past an acceptance percentage of 50%, the average error of points inside of the RAH-LdP is roughly six times lower than the average error of points outside of the RAH-LdP. Therefore, for the low-dimensional case, we find that the RAH-LdP is a successful approximation to the DoA.

Unfortunately, generalising the notion of the distance between a point and a $d$−polytope is not trivial when $d > 2$. We have attempted to scale-up the idea of projecting the position vector of a point onto the normal vector to the $d$−polytope. Computing the normal vector to polytopes where $d > 2$ is not immediately ob-
5.5. *DOMAIN OF APPLICABILITY*

Figure 5.10: Progression of average error ratio (left abscissa) and Δ (right abscissa) with increasing acceptance percentage for the water monomer ($\mathbb{R}^3$).

vious, e.g. the normal vector to a tetrahedron. However, we can exploit the fact that the scalar product of the normal vector and a vector lying in the $d$–polytope is zero. In 3-dimensional systems, the cross product can be used to this effect, but unfortunately such an approach only ensures orthogonality in $\mathbb{R}^3$ and $\mathbb{R}^7$[337]. As such, we are forced to employ an approximate technique to computing the normal vector in spaces of arbitrary dimensions. We have attempted a Gram-Schmidt orthogonalisation procedure to this end, the algorithm of which has been outlined in Section 3.4.1. The algorithmic workflow that we have proposed for our greedy heuristic is shown in Algorithm 6.

However, the Gram-Schmidt orthogonalisation was unsuccessful at producing a usable RAH-LdP in $\mathbb{R}^4$ and above. The resultant RAH-LdP contained a very small number of points, and was simply not viable as an approximation to the convex hull.
1. Take $S = \emptyset$
2. Find $\max_{x_i, x_j \in S} d(x_i, x_j)$ and add $x_i, x_j$ to $S$

while $|S| < d + 1$ do

\[ i = |S| \]

Form the $i$-polytope, $F_i$, from the points in $S$

Find the normal vector to $F_i$, $n_i$

for $x_k \in S \setminus S$ do

Compute vector, $v_k$, from point on $F_i$ to $x_k$

\[ d(x_k, F_i) = v_k \cdot n_i \]

end

Add $x_k$ to $S$, where $\max_{k \in S \setminus S} d(x_k, F_i)$

end

**Algorithm 6:** The LdP Algorithm

Instead, we have attempted to define the LdP by use of a physically intuitive scheme. Recall that we require $d + 1$ points to define the LdP. $d$ points can be trivially obtained by projecting the position vector of each point onto a single basis vector at a time. The maximum such value is then used as a vertex for the LdP along the corresponding basis vector. A case in $\mathbb{R}^3$ is depicted in Figure 5.11, and the generalisation to higher dimensions is straightforward.

In Figure 5.11, the red points correspond to the candidate set. The three basis vectors of the space are $\hat{x}_1, \hat{x}_2, \hat{x}_3$. We find that the point denoted by position vector $a$, when projected onto $\hat{x}_1$, is the maximum such value out of all the candidate set. We therefore use the blue point, $\text{PROJ}_a \hat{x}_1$, as a vertex of the LdP. Equivalently, we can define $\text{PROJ}_b \hat{x}_2$ and $\text{PROJ}_c \hat{x}_3$ as vertices of the LdP, arising from the projections of $b$ and $c$ onto $\hat{x}_2$ and $\hat{x}_3$, respectively. The blue face then corresponds to a 2-polytope formed from the three vertices. The final point required can then be selected by use of the greedy MaxMin heuristic across the entire candidate set. For the sake of argument, assume the point chosen by the MaxMin greedy heuristic resides close to the origin of the coordinate system. Then, the LdP is a tetrahedron of relatively large size, and it can be appreciated that a great proportion of the candidate set lies within the LdP.
Figure 5.11: Definition of three vertices used in the construction of $A_\gamma$ in $\mathbb{R}^3$. 
It is worth noting that in Figure 5.11, we see that a number of points lie outside of the prospective LdP. While this is demonstrative of the LdP being an unideal representation of the convex hull, it is remedied somewhat by use of the restricted affine hull, and the “halo region” associated with the RAH-LdP that it grants. Of course, a LdP that is closer to the actual convex hull of the points is preferable, but to some extent, it makes no real difference.

In Figures 5.12, 5.13 and 5.14, we give the progression of $\Delta$ and the average error ratio with respect to the acceptance percentage for methanol ($\mathbb{R}^{12}$), NMA ($\mathbb{R}^{30}$) and the water decamer ($\mathbb{R}^{84}$), respectively. Candidate sets of 600 points have been used for each example.

![Figure 5.12: Progression of average error ratio (left abscissa) and $\Delta$ (right abscissa) with increasing acceptance percentage for the methanol ($\mathbb{R}^{12}$).](image)

For each of the graphs, we find that $\Delta$ is once again essentially linear in the acceptance percentage, which is a useful trait since its value can be predicted to give a desired acceptance percentage. For methanol, we find that the RAH-LdP is
Figure 5.13: Progression of average error ratio (left abscissa) and $\Delta$ (right abscissa) with increasing acceptance percentage for the NMA ($\mathbb{R}^{30}$).

Figure 5.14: Progression of average error ratio (left abscissa) and $\Delta$ (right abscissa) with increasing acceptance percentage for the water decamer ($\mathbb{R}^{84}$).
a relatively successful means for approximating the DoA of a kriging model. Being able to successfully approximate the convex hull in $\mathbb{R}^{12}$ is particularly noteworthy, since the convex hull cannot be explicitly computed in $\mathbb{R}^{12}$ by use of QHull. However, it is worth mentioning that the difference in average error between those points outside and inside of the RAH-LdP is nowhere near as impressive as that obtained for water.

We see that as the dimensionality of the space increases, the RAH-LdP is a successively poorer approximation to the DoA. In $\mathbb{R}^{30}$, we find that the average error ratio is only less than one when the acceptance percentage is above 80%, while in $\mathbb{R}^{84}$, the average error ratio is roughly unity across all acceptance percentages. These results are perhaps suggestive of the fact that the RAH-LdP is not an appropriate approximation to the DoA in high-dimensional spaces.

We suggest a few potential remedies to the shortcomings of the RAH-LdP as an approximation to the DoA. Of course, a more thorough description of LdP will offer a better approximate to the convex hull of the candidate set. Since construction of the LdP is a geometric problem, a number of strategies can be employed, and are commonly analytical in nature, making their computational implementation trivial. Secondly, and perhaps most importantly, we have not decoupled the quality of the training set from the complete process of defining the DoA. If the candidate set yields a poor kriging model, then kriging model technically has no DoA, i.e. it is not applicable on any domain of space. As such, any further work on the DoA would require a decoupling of the kriging model quality from the RAH-LdP quality.

Regarding the implementation of the DoA within the kriging methodology, we suggest it as a useful validation tool for a kriging model, in a similar manner to that proposed by Li and coworkers[17]. A kriging model can be used over the
5.6 Conclusion

We have presented a number of methodological developments for the selection of an optimal training set for kriging. The greedy subset selection algorithms have been reformulated to use a novel selection function. This novel selection function results in improved accuracy of the kriging model, relative to one constructed by a random subset selection. However, work is still required to optimise this methodology and yield “ideal” training sets.

A novel subset selection algorithm, the Iterative Voronoi algorithm, has been presented. However, this methodology is only viable for low-dimensional cases, and so cannot be implemented for our purposes. In spite of this, if it were possible to reduce the dimensionality of the systems in some way, it may be feasible to implement the Iterative Voronoi algorithm in the future and establish whether it is a successful subset selection methodology for larger systems.

Finally, we have presented a novel means for defining the Domain of Applicability of a machine learning model. By approximating the “restricted affine hull” from a course of a MD simulation, and those points within and outside of the DoA logged. If a great deal of points lie outside of the DoA, it suggests that the kriging model does not properly account for the physically relevant regions of conformational space, and additional training points are required. Alternatively, if only a small number of conformations lie outside of the DoA, the kriging model is essentially validated as spanning the relevant regions of conformational space. In the latter case, the kriging model is suitable for release to end-product users, i.e. those who anticipate using the kriging models as “black boxes”.
set of points by use of a greedy algorithm to construct the largest $d$-polytope on the points, it has been shown that one can approximately determine whether a point lies within the training set of a machine learning model in low-dimensional spaces. Further work is, however, required, to make the methodology more successful in higher dimensional spaces. The approximation of the DoA of a kriging model is of utility in our work since it allows one to ascertain whether a kriging model is predictive in those relevant regions of conformational space.
Chapter 6

Raman Optical Activity

Raman optical activity (ROA) is a powerful chiroptical technique that can be used to probe a number of molecular properties. The polarisation characteristics of Rayleigh and Raman scattered photons were investigated by Atkins and Barron in 1969, leading to the conclusion that the scattered light should carry a small degree of circular polarisation depending upon the chirality of the scattering body\[338\]. Later work\[339\] by the same group remedied a number of simplifications that had been employed in the previous work, and also introduced an experimentally-detectable quantity,

\[
\Delta = \frac{(I_R - I_L)}{(I_R + I_L)}, \tag{6.0.1}
\]

where \(\Delta\) is referred to as the circular intensity difference (CID), while \(I_R\) and \(I_L\) correspond to the intensities of the right and left-scattered light, respectively.

In the early days of its inception, ROA was used to deduce stereochemical information of molecular species by comparison with band patterns of related molecules\[340\]. Unlike a number of other spectroscopies, ROA intensities in the low wavenumber region of the spectrum are rich in information. These regions correspond to, for
example, torsional degrees of freedom[341] and large scale conformational modes of motion[342]. ROA also possesses the capacity to determine the proportions of enantiomers within a sample, which is of great utility in purification processes[343].

The structural features of biochemical systems make them particularly amenable to ROA spectroscopy. ROA offers a far more robust analysis of biomolecules than more common forms of vibrational spectroscopy because of its sensitivity to chirality. The ROA signal of a system is largely dominated by the more rigid and chiral elements of a system, corresponding to secondary structure motifs in polypeptides. This makes ROA useful for structural characterisation of complex biomolecules, in contrast to Raman spectroscopy, where amino acid sidechain signals typically dominate the spectrum. ROA is by no means constrained to peptide systems, and the ROA signals of other biochemical species, such as DNA bases[344] and oligomeric nucleic acids[345], also prove informative. The utility of ROA in structural studies has been exemplified over the years by a number of studies on the folding[346, 347], assembly[348] and dynamical properties[349] of biochemical systems.

A Note on Optical Activity

For the sake of brevity, we do not concern ourselves with the mathematical complexities of optical activity. However, the results of optical activity experiments can be quite easily correlated with molecular structure, without concerning ourselves with mathematical formalities. We introduce the parity operator, $\Pi : x \rightarrow -x$, where we use $x$ as a generic coordinate vector. We are interested in the operation of the parity operator on the electric field vector, $E$, and molecular conformation vector, $x$. Without too much trouble, we can verify that both $E$ and $x$ change sign under $\Pi$, and by definition are referred to as “polar vectors” (in contrast to
the notion of an “axial vector”).

The central concept that we employ is that parity is conserved for the vast majority of physical systems (the $\beta$–decay of $^{60}$Co is a notable exception[350]). Elaborating upon this, suppose we subject an experiment to the parity operator; both experiments, before and after the parity operator has operated, should yield physically realisable experimental results.

Let us take the experimental setup for an optically active molecule which is irradiated by a linearly polarised light beam\textsuperscript{1}. The interaction of the molecule with the light beam results in a rotation of the light to, for argument’s sake, the right. One can subsequently detect the handedness of the light by means of some experimental apparatus. Under $\Pi$, the molecule is replaced by the corresponding enantiomer, and the scattered (a technical term that we define in the next section) light changes its handedness, in this case to the left. These results are experimentally verifiable, in that enantiomers rotate light in opposite directions[351]. Given that parity is always conserved, we can also conclude that enantiomers rotating light the same way is not a physically realisable situation.

6.1 General Theory

6.1.1 Classical Raman Scattering

Qualitatively, the theory of molecular light scattering dictates that a photon of a given frequency (and hence energy) is absorbed by the molecular system. This ab-

\textsuperscript{1}Since linearly polarised light is composed a superposition of left- and right-circularly polarised light, the parity operator has no effect on the linearly polarised light beam.
sorbtion event promotes the system into a vibrationally excited state, as exhibited by a change in the dipole moment of the system. After a short time, the photon is re-emitted, or scattered, and the molecular system returns to its “ground state”. However, since the vibrational energy levels of the system are quantised, only light of certain frequencies can result in a scattering event. We see, then, that if incident light results in some change in the molecular dipole moment, the light must have promoted the system to a vibrationally excited state, and will subsequently be scattered. This explanation is the (classical) basis for vibrational spectroscopy, whereby one can observe the vibrational energy levels of a molecular system by detecting these scattering events\(^2\).

Consider a periodic electric field as a function of time, \( \mathbf{E}(t) = \mathbf{E}^0 \cos(\omega_L t) \), where \( \mathbf{E}^0 \) is the amplitude of the field and \( \omega_L \) its angular frequency. Interaction of this field with a molecule comprising a set of point charges induces a time-dependent electric dipole moment, \( \mu(t) \),

\[
\mu_\alpha(t) = \alpha_{\alpha\beta}E_\beta(t) = \alpha_{\alpha\beta}E^0_\beta \cos(\omega_L t),
\]

(6.1.1)

where we have invoked the Einstein summation convention, i.e. repeated indices are summed over. \( \alpha_{\alpha\beta} \) denotes an element of the anisotropic electric dipole polarisability tensor\(^3\).

We have discussed at length in Section 3 how the normal modes of motion form a convenient basis within which to represent molecular motion. In this spirit, we consider a minimum energy molecular conformation to be denoted by the \( 3N - 6 \) dimensional vector, \( \mathbf{Q}^0 \). We have also shown that infinitesimal displacements can

\(^2\)If the incident light promotes the system into an *electronically* excited state, then the associated technique is termed absorption spectroscopy. Promotion of a system to an electronically excited state typically requires a great deal more energy than promotion to a vibrationally excited state, and so higher energy light is required for these spectroscopies.

\(^3\)In simpler theories, the electric dipole polarisability is assumed to be isotropic, and so is a scalar.
be approximated harmonically, so that $dQ_k(t) = Q_0^k \cos(\omega_k t)$, where $\omega_k$ is the angular frequency of the $k^{th}$ normal mode.

Without wishing to be mathematically rigorous, we propose that the electric dipole polarisability is conformationally dependent. Recall that the electric dipole moment corresponds to the displacement of electron density from a spherical configuration. If the electric dipole polarisability were not conformationally dependent, the electron density would be deformed linearly in the amount of energy supplied. Intuitively, it seems reasonable that the the amount of energy required to deform the electron density should increase as the electron density is further distorted.

Keeping this argument in mind, we express the electric dipole polarisability tensor as a Taylor series in $Q$,

$$\alpha_{\alpha\beta} = \alpha_{\alpha\beta}^0 + \frac{\partial \alpha_{\alpha\beta}}{\partial Q_k} \bigg|_{Q_0} dQ_k + \frac{\partial \alpha_{\alpha\beta}}{\partial Q_k \partial Q_l} \bigg|_{Q_0} dQ_k dQ_l + \ldots$$

$$(6.1.2)$$

where we have truncated the Taylor series at first order, neglecting all quadratic and higher terms, which we reiterate is only valid for infinitesimal displacements. We have also used the notation $\alpha_{\alpha\beta}^0$ to denote the electric dipole polarisability tensor evaluated at $Q_0^0$. Using this quantity in (6.1.1), we obtain

$$\mu_\alpha(t) \approx \alpha_{\alpha\beta}^0 E_\beta \cos(\omega_L t) + \frac{\partial \alpha_{\alpha\beta}}{\partial Q_k} \bigg|_{Q_0} Q_0^k \cos(\omega_k t) E_\beta.$$

$$(6.1.3)$$

Introducing the harmonic form of the electric field,

$$\mu_\alpha(t) \approx \alpha_{\alpha\beta}^0 E_\beta^0 \cos(\omega_L t) + \frac{\partial \alpha_{\alpha\beta}}{\partial Q_k} \bigg|_{Q_0} Q_0^k \cos(\omega_k t) E_\beta^0 \cos(\omega_L t)$$

$$(6.1.4)$$

By invoking the prosthaphaeresis formula for the product of cosines, we obtain

$$\mu_\alpha(t) \approx \alpha_{\alpha\beta}^0 E_\beta^0 \cos(\omega_L t) + \frac{1}{2} \frac{\partial \alpha_{\alpha\beta}}{\partial Q_k} \bigg|_{Q_0} Q_0^k E_\beta^0 \left[ \cos \left( (\omega_L + \omega_k) t \right) + \cos \left( (\omega_L - \omega_k) t \right) \right]$$

$$(6.1.5)$$
We recall that if an event is associated with a change in the dipole moment, it signifies a promotion of the system to a vibrationally excited state. From the above equation, we see there are three distinct components that contribute to the change in dipole moment. The first term on the right hand side corresponds to the scattering of light with angular frequency $\omega_L$, i.e. the frequency of the scattered light is equal to the frequency of the incident light. This scattering event is referred to as Rayleigh scattering, and corresponds to elastic scattering.

The other two terms correspond to scattering events involving light of frequency $\omega_L + \omega_k$ and $\omega_L - \omega_k$. These are referred to as inelastic, or Raman scattering events. For the case of $\omega_L + \omega_k$, the system begins in a vibrationally excited state, and upon the scattering event returns to the ground state, scattering light of higher energy than the incident light. We refer to this as “Anti-Stokes Raman scattering”. Conversely, for $\omega_L - \omega_k$, the system begins in the ground state, and after scattering returns to a vibrationally excited state, scattering light of lower energy than the incident light, referred to as “Stokes Raman scattering”.

This Raman scattered light is experimentally detectable, and forms the basis of a spectroscopy that is of great use to the chemist. The Stokes scattered light is, under standard conditions (room temperature), orders of magnitude more prominent than the anti-Stokes scattered light. As such, it is the Stokes scattered light that is collected in the formation of a Raman spectrum, although in principle the anti-Stokes scattered light could be used to the same effect.

### 6.1.2 The Electromagnetic Field

A charge density, $\rho(\mathbf{x})$, and its corresponding current density, $\mathbf{J} = \rho \mathbf{v}$, where $\mathbf{v}$ is the velocity of the charge distribution, give rise to an electric and magnetic field,
E and B respectively, in free space. Modified fields, D and H, can be related to their free space counterparts by the relations

\[ D(x, t) = \varepsilon \varepsilon_0 E(x, t) \quad \text{and} \quad H(x, t) = \frac{1}{\mu \mu_0} B(x, t). \] (6.1.6)

The constants \( \varepsilon_0 \) and \( \mu_0 \) are the dielectric constant and magnetic permeability of free space, respectively, whilst \( \varepsilon \) and \( \mu \) are their values in an isotropic medium. These constants become functions, \( \varepsilon(x) \) and \( \mu(x) \), if the medium is anisotropic.

The four Maxwell equations in an infinite homogeneous medium have form

\[
\begin{align*}
\nabla \cdot D &= \rho \quad \implies \quad \nabla \cdot E = \frac{\rho}{\varepsilon_0}, \quad (6.1.7) \\
\nabla \cdot B &= 0 \quad \implies \quad \nabla \cdot H = 0, \quad (6.1.8) \\
\nabla \times E &= -\frac{\partial B}{\partial t} \quad \implies \quad \nabla \times D = -\varepsilon \varepsilon_0 \mu_0 \frac{\partial H}{\partial t}, \quad (6.1.9) \\
\nabla \times H &= J + \frac{\partial D}{\partial t} \quad \implies \quad \nabla \times B = \mu_0 \left( J + \varepsilon \varepsilon_0 \frac{\partial E}{\partial t} \right), \quad (6.1.10)
\end{align*}
\]

each of which is of sufficient importance to merit a name: (6.1.7) is Gauss’ Theorem; (6.1.8) is Gauss’ Law for Magnetostatics; (6.1.9) is Faraday’s Law of Magnetic Induction; and (6.1.10) is Ampere’s Law.

Invoking the vector identity for an arbitrary vector function, \( Z \),

\[ \nabla \times (\nabla \times Z) = \nabla (\nabla \cdot Z) - \nabla^2 Z, \]

we can establish the wave equations for the electric and magnetic fields,

\[ \nabla^2 E = \varepsilon \varepsilon_0 \mu_0 \frac{\partial^2 E}{\partial t^2}, \quad (6.1.11) \]

\[ \text{We derive the electric field wave equation from the Maxwell equations, but the magnetic field wave equation follows in a similar manner. From the vector identity,} \]

\[ \nabla \times (\nabla \times D) = \nabla (\nabla \cdot D) - \nabla^2 D. \]

By (6.1.7) and (6.1.9),

\[ -\nabla \times \left( \varepsilon \varepsilon_0 \mu_0 \frac{\partial H}{\partial t} \right) = \nabla \rho \frac{\varepsilon_0}{\varepsilon} - \nabla^2 D. \]

Since the spatial and temporal differential operators commute, we can combine this property
\[ \nabla^2 \mathbf{B} = \varepsilon \varepsilon_0 \mu_0 \frac{\partial^2 \mathbf{B}}{\partial t^2}. \] (6.1.12)

Comparison with the standard wave equation, with which the reader is assumed to be sufficiently familiar that we can omit reproducing it here, leads to an expression for the velocity of the wave, \( v = 1/\sqrt{\varepsilon_0 \mu_0} \), which equates to the speed of light in a given isotropic medium.

Solutions to the wave equation are the familiar plane waves,

\[ \mathbf{E}(\mathbf{r}, t) = \mathbf{E}^0 e^{i(k \cdot \mathbf{r} - \omega t)}, \] (6.1.13)

where \( \mathbf{E}^0 \) denotes the amplitude of the wave, \( \omega \) is the angular frequency and \( \mathbf{k} \) is termed the wavevector, and is related to the wavelength, \( \lambda \), of light by the expression \( |\mathbf{k}| = 2\pi/\lambda \). The wavevector is directed orthogonally to the wave fronts of the wave (surfaces of constant phase).

The four Maxwell equations can be written as two equations involving the more fundamental scalar and vector potentials, \( \phi(\mathbf{r}, t) \) and \( \mathbf{A}(\mathbf{r}, t) \), respectively. If the divergence of a vector function is equal to zero, as in (6.1.8), the underlying vector function can be written as the curl of the vector potential, i.e.

\[ \mathbf{B} = \nabla \times \mathbf{A}. \] (6.1.14)

Then, (6.1.9) becomes

\[ \nabla \times \mathbf{E} = \nabla \times \frac{\partial \mathbf{A}}{\partial t} : \nabla \times \left( \mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} \right) = 0, \]

with (6.1.10)

\[-\varepsilon_0 \mu_0 \frac{\partial^2 \mathbf{D}}{\partial t^2} = \frac{\nabla \rho}{\varepsilon_0} - \nabla^2 \mathbf{D}.\]

The first term of the right hand side goes to zero owing to \( \nabla \rho = 0 \), since the medium is homogeneous. Substitution of \( \mathbf{D} \) for the corresponding vacuum quantity, \( \mathbf{E} \), leads to

\[-\varepsilon_0 \mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} = -\nabla^2 \mathbf{E},\]

and the wave equation follows.
which implies that $\mathbf{E}$ can be written as

$$\mathbf{E} = -\nabla \phi(r, t) - \frac{\partial \mathbf{A}}{\partial t}, \quad (6.1.15)$$

where we have utilised the fact that the curl of the gradient field, $\nabla \phi(r, t)$, is equal to zero. It should then be noted that the choice of the scalar potential, $\phi(r, t)$, is arbitrary, since it does not contribute to the form of $\mathbf{E}$.

By use of (6.1.15), we find that (6.1.7) becomes

$$\nabla \cdot \mathbf{E} = -\nabla \cdot \left( \nabla \phi(r, t) + \frac{\partial \mathbf{A}}{\partial t} \right) = \frac{\rho}{\varepsilon \varepsilon_0} \nabla^2 \phi(r, t) + \frac{\partial}{\partial t} \nabla \cdot \mathbf{A} = -\frac{\rho}{\varepsilon \varepsilon_0}. \quad (6.1.16)$$

Similarly, (6.1.10) combined with (6.1.14) yields

$$\nabla \times (\nabla \times \mathbf{A}) = \mu \mu_0 \left( \mathbf{J} + \varepsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right).$$

Then, by the vector identity we introduced above and (6.1.15)

$$\nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} = \mu \mu_0 \left( \mathbf{J} - \varepsilon_0 \frac{\partial}{\partial t} \left[ \nabla \phi(r, t) - \frac{\partial \mathbf{A}}{\partial t} \right] \right)$$

$$= \mu \mu_0 \left( \mathbf{J} - \varepsilon_0 \left[ \nabla \frac{\partial \phi}{\partial t} - \frac{\partial^2 \mathbf{A}}{\partial t^2} \right] \right)$$

$$= \mu \mu_0 \mathbf{J} - \mu \mu_0 \varepsilon_0 \left( \nabla \frac{\partial \phi}{\partial t} - \frac{\partial^2 \mathbf{A}}{\partial t^2} \right).$$

Finally, by noting that $\mu \mu_0 \varepsilon_0 = v^{-2}$,

$$\nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} + \frac{1}{v^2} \left( \nabla \frac{\partial \phi}{\partial t} - \frac{\partial^2 \mathbf{A}}{\partial t^2} \right) = \mu \mu_0 \mathbf{J}. \quad (6.1.17)$$

The two equations, (6.1.16) and (6.1.17), completely specify four Maxwell equations, (6.1.7) - (6.1.10).

A peculiarity of the vector and scalar potentials means that the transformations

$$\mathbf{A} = \mathbf{A}_0 - \nabla \Lambda(r, t) \quad (6.1.18)$$

$$\phi = \phi_0 + \frac{\partial}{\partial t} \Lambda(r, t), \quad (6.1.19)$$
where $\Lambda(r, t)$ is an arbitrary function and $A_0, \phi_0$ correspond to the vector and scalar potentials at arbitrary points in space and time, leave the electric and magnetic field vectors unaltered. By (6.1.14)

$$B = \nabla \times A = \nabla \times (A_0 - \nabla \Lambda) = \nabla \times A_0,$$

where we have already seen that the curl of the gradient of a scalar function disappears. Similarly by (6.1.15)

$$E = -\nabla \phi - \frac{\partial A}{\partial t} = -\nabla \left( \phi_0 + \frac{\partial}{\partial t} \Lambda(r, t) \right) - \frac{\partial}{\partial t} (A_0 - \nabla \Lambda(r, t))$$

$$= -\nabla \phi_0 - \frac{\partial}{\partial t} A_0.$$

This property is termed *gauge invariance*, and suggests the existence of free variables which can be fixed by placing restrictions on the vector and scalar potentials. One such choice of restrictions is termed the *Coulomb gauge*, such that $\nabla^2 \Lambda(r, t) = \nabla \cdot A_0$, then by the gauge transformation for the vector potential,

$$\nabla \cdot A = \nabla \cdot A_0 - \nabla \cdot \nabla \Lambda(r, t)$$

$$= \nabla \cdot A_0 - \nabla^2 \Lambda(r, t) = \nabla \cdot A_0 - \nabla \cdot A_0,$$

leaving the Coulomb gauge condition

$$\nabla \cdot A = 0. \quad (6.1.20)$$

### 6.1.3 Hamiltonian in an Electromagnetic Field

The classical Lagrangian for a set of particles in an electromagnetic field is given by

$$\mathcal{L}(r, \dot{r}, t) = J(r, t) - V(r, \dot{r}, t) = \sum_{i=1}^{N} \left[ \frac{1}{2} m_i \dot{r}_i^2 - q_i \Phi(r_i, t) + q_i \dot{r}_i \cdot A(r_i, t) - V(r_i) \right], \quad (6.1.21)$$
where $q_i$ is the charge of the $i^{th}$ particle, $\mathcal{T}$ and $\mathcal{V}$ are the kinetic and potential energies, respectively, $V(\mathbf{r}_i)$ is some external potential and $\mathbf{A}$ and $\Phi$ are the vector and scalar potentials describing the electromagnetic field we have introduced in Section 6.1.2. To obtain the Hamiltonian of the system, we perform the Legendre transformation, which necessitates evaluation of the canonical momentum conjugate to a coordinate $r_\alpha$, $p_\alpha$. This is readily achieved by invoking the Euler-Lagrange equation

$$p_\alpha = \frac{\partial \mathcal{L}}{\partial \dot{r}_\alpha} = \frac{1}{2} \frac{\partial}{\partial \dot{r}_\alpha} m_i \dot{r}_\alpha^2 + \frac{\partial}{\partial \dot{r}_\alpha} q_i \dot{r}_\alpha A_\alpha(\mathbf{r}_i, t)$$

$$= m_i \dot{r}_\alpha - q_i A_\alpha(\mathbf{r}_i, t),$$

(6.1.22)

where the index $\alpha$ denotes a generalised coordinate and the index $i$ denotes the atom to which the generalised coordinate belongs. The Hamiltonian is subsequently defined as

$$\mathcal{H}(\mathbf{r}, \mathbf{p}, t) = \sum_{i=1}^{N} p_i \cdot \dot{r}_i - \mathcal{L}(\mathbf{r}, \dot{\mathbf{r}}, t)$$

$$= \sum_{i=1}^{N} \left[ \frac{1}{2m_i} (p_i - q_i A(\mathbf{r}_i, t))^2 + q_i \Phi(\mathbf{r}_i, t) + V(\mathbf{r}_i) \right].$$

(6.1.23)

Expanding the above quantity, we obtain

$$\mathcal{H}(\mathbf{r}, \mathbf{p}, t) = \sum_{i=1}^{N} \left[ \frac{1}{2m_i} (p_i^2 - q_i p_i \cdot A(\mathbf{r}_i, t) - q_i A(\mathbf{r}_i, t) \cdot p_i + q_i^2 A^2(\mathbf{r}_i, t)) \right.$$

$$+ q_i \phi(\mathbf{r}_i, t) + V(\mathbf{r}_i) \right].$$

(6.1.24)

Since we are attempting to describe a quantum phenomenon, we employ the correspondence theorem, in which the momenta are replaced by their quantum me-
chanical operators, \( p_i = -i\hbar \nabla_i \). As such,

\[
H(r, p, t) = \sum_{i=1}^{N} \left[ -\frac{\nabla_i^2}{2m_i} + \frac{i\hbar q_i}{2m_i} \nabla_i \cdot A(r_i, t) + \frac{i\hbar q_i}{2m_i} A(r_i, t) \cdot \nabla_i + \frac{q_i^2}{2m_i} A^2(r_i, t) + q_i \phi(r_i, t) + V(r_i) \right]
\]

(6.1.25)

By working in the Coulomb gauge defined in (6.1.20), we can set the term containing \( q_i \nabla_i \cdot A(r_i, t) = 0 \). We also omit terms proportional to \( A^2(r_i, t) \) and higher since we assume the external field is of low intensity. Rearranging to separate the Hamiltonian into a sum of unperturbed and perturbed Hamiltonians, and using the classical quantities \( r \) and \( p \) once more,

\[
H(r, p, t) = H_0 + H_{\text{int}}
\]

\[
= \sum_{i=1}^{N} \left[ \frac{1}{2m_i} p_i^2 + V(r_i) \right] + \sum_{i=1}^{N} \left[ q_i \phi(r_i, t) - \frac{1}{2m_i} q_i A(r_i, t) \cdot p_i \right].
\]

(6.1.26)

where \( H_{\text{int}} \) is the first order perturbation to the ground state Hamiltonian, \( H_0 \).

Our next step involves the expression of the perturbing fields, \( E(r, t) \) and \( B(r, t) \) as Taylor series about some origin, \( r_0 \), i.e.,

\[
E_\alpha(r, t) = E_\alpha(r_0, t) + r_\beta \nabla_\beta E_\alpha(r_0, t) + \ldots,
\]

\[
B_\alpha(r, t) = B_\alpha(r_0, t) + r_\beta \nabla_\beta B_\alpha(r_0, t) + \ldots,
\]

(6.1.27)

where we use the Greek indices \( \alpha, \beta, \ldots \) to denote a Cartesian degree of freedom.

We also introduce the Einstein summation convention, where repetition of an index in a term implies summation over all degrees of freedom. Since (6.1.26) is

\footnote{Technically, the positions \( r_i \) should be also exchanged for their quantum mechanical operators, \( \hat{r}_i \). However, this substitution does nothing to change the underlying mathematics, and so is omitted here.}
expressed in terms of the vector and scalar potentials, we require that (6.1.27) also be reformulated in terms of \( \phi(r, t) \) and \( A(r, t) \). Barron and Gray[352] have found that one can choose the scalar and vector potentials to be of the form

\[
\phi(r, t) = \phi(r_0, t) - r_\alpha E_\alpha(r_0, t) - \frac{1}{2} r_\alpha r_\beta \nabla_\alpha E_\beta(r_0, t) + \ldots ,
\]

\[
A(r, t) = \frac{1}{2} \epsilon_{\alpha\beta\gamma} B_\beta(r_0, t) r_\gamma + \frac{1}{3} \epsilon_{\alpha\gamma\delta} r_\beta \nabla_\beta B_\gamma(r_0, t) r_\delta + \ldots ,
\]

where \( \epsilon_{\alpha\beta\gamma} \) is the rank three Levi-Civita symbol, and is equal to 1 if the parity of the permutation of indices is positive, and -1 otherwise. For the following, we omit the zeroth order scalar potential, \( \phi(r_0, t) \), as it can be accounted for later as an additive term. Placing these into the expression for \( H_{\text{int}} \) in (6.1.26), we obtain

\[
H_{\text{int}} = \sum_{i=1}^{N} q_i \left[ -r_{i,\alpha} E_\alpha(r_0, t) - \frac{1}{2} r_{i,\alpha} r_{i,\beta} \nabla_\alpha E_\beta(r_0, t) + \ldots \right] \]

\[- \sum_{i=1}^{N} \frac{q_i P_{i,\alpha}}{2m_i} \left[ \frac{1}{2} \epsilon_{\alpha\beta\gamma} B_\beta(r_0, t) r_{i,\gamma} + \frac{1}{3} \epsilon_{\alpha\gamma\delta} r_{i,\beta} \nabla_\beta B_\gamma(r_0, t) r_{i,\delta} + \ldots \right] \]

\[= -\mu_\alpha E_\alpha(r_0, t) - \Xi_{\alpha\beta} \nabla_\alpha E_\beta(r_0, t) - m_\alpha B_\alpha(r_0, t) + \ldots ,\]

where we have used three multipole moments for shorthand,

\[
\mu_\alpha = \sum_{i=1}^{N} q_i r_{i,\alpha} ,
\]

\[
\Xi_{\alpha\beta} = \frac{1}{2} \sum_{i=1}^{N} q_i r_{i,\alpha} r_{i,\beta} ,
\]

\[
m_\alpha = \sum_{i=1}^{N} \frac{q_i}{2m_i} \epsilon_{\alpha\beta\gamma} r_{i,\beta} P_{i,\alpha} ,
\]

the electric dipole, electric quadrupole and magnetic dipole moments, respectively.

It is worth revising (6.1.31) to account for the fact that the quadrupole moment is traceless,

\[
\Theta_{\alpha\beta} = \frac{1}{2} \sum_{i=1}^{N} q_i \left( 3r_{i,\alpha} r_{i,\beta} - r_i^2 \delta_{\alpha\beta} \right) ,
\]

(6.1.33)
and (6.1.29) becomes
\[
\mathcal{H}_{\text{int}} = -\mu_\alpha E_\alpha(r_0, t) - \frac{1}{3} \Theta_{\alpha\beta} \nabla_\alpha E_\beta(r_0, t) - m_\alpha B_\alpha(r_0, t) + \ldots, \tag{6.1.34}
\]
where the factor of \(\frac{1}{3}\) cancels the factor of 3 in (6.1.33).

### 6.1.4 Wavefunction Perturbed by Periodic Field

We write the time-dependent Schrödinger equation
\[
\left( i\hbar \frac{\partial}{\partial t} - \mathcal{H}_0 \right) |\psi\rangle = \mathcal{H}_{\text{int}} |\psi\rangle, \tag{6.1.35}
\]
where we have already encountered
\[
\mathcal{H}_0 = \sum_{i=1}^{N} \left( \frac{1}{2m_i} p_i^2 + V(r_i) \right) \tag{6.1.36}
\]
\[
\mathcal{H}_{\text{int}} = \left( -\mu_\alpha E_\alpha(r_0, t) - \frac{1}{3} \Theta_{\alpha\beta} \nabla_\alpha E_\beta(r_0, t) - m_\alpha B_\alpha(r_0, t) + \ldots \right) \tag{6.1.37}
\]
Herein, we drop the convention of explicitly specifying the arguments of the functions \(E(r_0, t)\) and \(B(r_0, t)\), instead using the shorthand \((E)_0\) and \((B)_0\), respectively. We note that by setting \(\mathcal{H}_{\text{int}} = 0\), we recover the time-independent Schrödinger equation, the solution of which takes the form
\[
|\psi\rangle = \sum_{n} c_n |\phi_n\rangle \exp(-i\omega_n t) = \sum_{n} c_n |\psi_n\rangle \tag{6.1.38}
\]
where \(|\phi_n\rangle\) denotes the \(n^{th}\) eigenfunction with accompanying eigenvalue of \(\hbar\omega_n\), \(\omega_n\) being the angular frequency. \(c_n\) is an expansion coefficient, where the index \(n\) runs over all eigenfunctions of the system. For the remainder of our discussion, we attempt to make our arguments as general as possible, and so assume that the wavefunction and perturbing fields are complex valued. For our purposes, the
6.1. GENERAL THEORY

perturbing electric and magnetic fields take the form of single-frequency harmonic plane waves, with angular frequency $\omega_L$,

\[
(E_0)_{\beta} = E_0^\beta \exp(i\omega_L t) = E_0^\beta \exp(-i\omega_L t)
\]

(6.1.39)

\[
(\bar{E}_0)_{\beta} = \bar{E}_0^\beta \exp(i\omega_L t)
\]

(6.1.40)

\[
(B_0)_{\beta} = B_0^\beta \exp(-i\omega_L t)
\]

(6.1.41)

\[
(\bar{B}_0)_{\beta} = \bar{B}_0^\beta \exp(i\omega_L t)
\]

(6.1.42)

where $E_0^\beta$ and $B_0^\beta$ are the complex $\beta-$components of the electric and magnetic field amplitudes. We have also used the bar notation, $(\bar{E}_0)_{\beta}$, $(\bar{B}_0)_{\beta}$, to denote the complex conjugate. Since we deal explicitly with rank-1 tensors, we have no need to adopt the conjugate transpose notation that is frequently incurred in the literature. We shall make a great deal of use of the following identities

\[
(E_0)_{\beta} = \frac{1}{2} \left[ (E_0)_{\beta} + (\bar{E}_0)_{\beta} \right] = \frac{1}{2} \left[ E_0^\beta \exp(-i\omega_L t) + \bar{E}_0^\beta \exp(i\omega_L t) \right]
\]

(6.1.43)

\[
(\dot{E}_0)_{\beta} = -\frac{i\omega_L}{2} \left[(E_0)_{\beta} - (\bar{E}_0)_{\beta} \right] = -\frac{i\omega_L}{2} \left[ E_0^\beta \exp(-i\omega_L t) + \bar{E}_0^\beta \exp(i\omega_L t) \right]
\]

(6.1.44)

where we have used the familiar dot notation to denote a temporal derivative. One can form the wavefunction of a system perturbed by external fields as an expansion of each eigenfunction, $|\psi'_n\rangle [353, 354]$,

\[
|\psi'_n\rangle = \left[ |\psi_n\rangle + |\psi_n^{a\beta}\rangle (E_0)_{\beta} + |\psi_n^{b\beta}\rangle (\bar{E}_0)_{\beta} + |\psi_n^{c\beta}\rangle (B_0)_{\beta} + |\psi_n^{d\beta}\rangle (\bar{B}_0)_{\beta} + |\psi_n^{e\gamma}\rangle \nabla_{\beta}(E_0)_{\gamma} + |\psi_n^{f\gamma}\rangle \nabla_{\beta}(\bar{E}_0)_{\gamma} + \ldots \right] \exp(-i\omega_n t)
\]

(6.1.45)

with

\[
|\psi^{x\beta}_n\rangle = \sum_{j \neq n} x_j^\beta |\psi_j\rangle \quad x = a, b, c, d
\]

(6.1.46)

\[
|\psi^{x\gamma}_n\rangle = \sum_{j \neq n} x_j^\gamma |\psi_j\rangle \quad x = e, f
\]

(6.1.47)
|ψ⟩ is an excited state wavefunction and \( x^β_{jn} \) and \( x^{βγ}_{jn} \) are expansion coefficients. We are now in a position to develop solutions to (6.1.35). We deal initially with the temporal derivative term of (6.1.45). Since (6.1.46) and (6.1.47) are time-independent, we need only concern ourselves with temporal derivatives of the fields and exponential terms

\[
iℏ\frac{∂}{∂t} |ψ_n^\prime⟩ = ℏω_n [ |ψ_n⟩ + |ψ^{aβ}_n⟩ (E_β)_0 + |ψ^{bβ}_n⟩ (E_β)_0 + |ψ^{cβ}_n⟩ (B_β)_0 + |ψ^{dβ}_n⟩ (B_β)_0 + |ψ^{eβ}_n⟩ \nabla_β (E_γ)_0 + |ψ^{fβ}_n⟩ \nabla_β (B_β)_0 + \exp(-iω_nt) + ℏω_L [ |ψ^{aβ}_n⟩ (E_β)_0 - |ψ^{bβ}_n⟩ (E_β)_0 + |ψ^{cβ}_n⟩ (B_β)_0 - |ψ^{dβ}_n⟩ (B_β)_0 + |ψ^{eβ}_n⟩ \nabla_β (E_γ)_0 + |ψ^{fβ}_n⟩ \nabla_β (B_β)_0 + \exp(-iω_nt)
\]

(6.1.48)

where we have used (6.1.43) to obtain temporal derivatives of the electric and magnetic fields. Expressing the multipole interaction Hamiltonian term, (6.1.37), by use of (6.1.43), we obtain

\[
\mathcal{H}_{int} = -\frac{1}{2}μ_β \left[ (E_β)_0 + (E_β)_0 \right] - Θ_{βγ} \frac{1}{6} \left[ \nabla_β (E_γ)_0 + \nabla_β (E_γ)_0 \right] - m_β \frac{1}{2} \left[ (B_β)_0 + (B_β)_0 \right] + \ldots
\]

(6.1.49)

We have suggestively changed the indices relative to (6.1.37) for later convenience. Our final step involves expression of (6.1.35) in terms of these new quantities,
yielding the unwieldy expression

\[
\hbar \omega_n \left[ |\psi_n^{\alpha \beta}\rangle + |\psi_n^{\alpha 0\beta}\rangle (E_\beta)_0 + |\psi_n^{0\beta\beta}\rangle (E_\beta)_0 + |\psi_n^{\beta\beta\beta}\rangle (E_\beta)_0 + |\psi_n^{\beta\beta\beta}\rangle (\bar{B}_\beta)_0 + |\psi_n^{\beta\beta\beta}\rangle \nabla_\beta (E_\gamma)_0 + |\psi_n^{\beta\beta\beta}\rangle \nabla_\beta (\bar{E}_\gamma)_0 + \ldots \right] \exp(-i\omega_n t) + \\
\hbar \omega_L \left[ |\psi_n^{\alpha \beta}\rangle (E_\beta)_0 - |\psi_n^{\beta \beta\beta}\rangle (E_\beta)_0 + |\psi_n^{\beta\beta\beta}\rangle (B_\beta)_0 - |\psi_n^{\beta\beta\beta}\rangle (\bar{B}_\beta)_0 + |\psi_n^{\beta\beta\beta}\rangle \nabla_\beta (E_\gamma)_0 - |\psi_n^{\beta\beta\beta}\rangle \nabla_\beta (\bar{E}_\gamma)_0 + \ldots \right] \exp(-i\omega_L t) - \\
\mathcal{H}_0 \left[ |\psi_n\rangle + |\psi_n^{\alpha \beta}\rangle (E_\beta)_0 + |\psi_n^{\alpha \beta\beta}\rangle (E_\beta)_0 + |\psi_n^{\beta\beta\beta\beta\beta}\rangle (E_\beta)_0 + |\psi_n^{\beta\beta\beta}\rangle (B_\beta)_0 + |\psi_n^{\beta\beta\beta}\rangle \nabla_\beta (E_\gamma)_0 + |\psi_n^{\beta\beta\beta}\rangle \nabla_\beta (\bar{E}_\gamma)_0 + \ldots \right] \exp(-i\omega_n t)
\]

\[= \left[ -\mu_\beta \frac{1}{2} (E_\beta)_0 + (\bar{E}_\beta)_0 - \Theta_{\alpha\beta\gamma} \frac{1}{6} \left[ \nabla_\beta (E_\gamma)_0 + \nabla_\beta (\bar{E}_\gamma)_0 \right] - m_\beta \frac{1}{2} (B_\beta)_0 + (\bar{B}_\beta)_0 + \ldots \right] \left[ |\psi_n\rangle + |\psi_n^{\alpha \beta}\rangle (E_\beta)_0 + |\psi_n^{\beta \beta\beta}\rangle (E_\beta)_0 + |\psi_n^{\beta\beta\beta\beta\beta}\rangle (E_\beta)_0 + |\psi_n^{\beta\beta\beta}\rangle (B_\beta)_0 + |\psi_n^{\beta\beta\beta}\rangle \nabla_\beta (E_\gamma)_0 + |\psi_n^{\beta\beta\beta}\rangle \nabla_\beta (\bar{E}_\gamma)_0 + \ldots \right] \exp(-i\omega_n t)\]

\[(6.1.50)\]

Collecting terms involving the same field components, we are able to find analytical expressions for the expansion coefficients of (6.1.45), expressed in (6.1.46) and (6.1.47). So, collecting terms involving \((E_\beta)_0\), we obtain the expression

\[
\hbar \omega_n |\psi_n^{\alpha \beta}\rangle + \hbar \omega_L |\psi_n^{\alpha 0\beta}\rangle - \mathcal{H}_0 |\psi_n^{\alpha \beta}\rangle = -\frac{\mu_\beta}{2} |\psi_n\rangle
\]

\[(6.1.51)\]

Multiplying through by \(|\psi_j\rangle\),

\[
\hbar \left[ \omega_{jn} + \omega_L \right] |\psi_j\rangle |\psi_n^{\alpha \beta}\rangle = -\frac{1}{2} |\psi_j\rangle |\mu_\beta |\psi_n\rangle
\]

\[(6.1.52)\]

where \(\omega_{jn} = \omega_j - \omega_n\). From (6.1.46), we see that \(|\psi_j\rangle |\psi_n^{\alpha \beta}\rangle = a_{jn}^{\beta\gamma}\), so that

\[
a_{jn}^{\beta\gamma} = \frac{\langle \psi_j | \mu_\beta | \psi_n \rangle}{2\hbar (\omega_{jn} - \omega_L)}
\]

\[(6.1.53)\]

which allows us to write

\[
|\psi_n^{\alpha \beta}\rangle = \sum_{j \neq n} \frac{\langle \psi_j | \mu_\beta | \psi_n \rangle}{2\hbar (\omega_{jn} - \omega_L)} |\psi_j\rangle.
\]

\[(6.1.54)\]
In an entirely equivalent way, we are able to group the other terms involving individual field components, i.e. \((\bar{E}_\beta)_0, (B_\beta)_0, \nabla_\beta(E_\gamma)_0\) and \(\nabla_\beta(\bar{E}_\gamma)_0\), to yield the other expansion coefficients

\[
|\psi^{b\beta}_n\rangle = \sum_{j \neq n} \frac{\langle \psi_j | \mu_{\beta} | \psi_n \rangle}{2\hbar(\omega_j - \omega_L)} |\psi_j\rangle,
\]

\[
|\psi^{c\beta}_n\rangle = \sum_{j \neq n} \frac{\langle \psi_j | m_{\beta} | \psi_n \rangle}{2\hbar(\omega_j + \omega_L)} |\psi_j\rangle,
\]

\[
|\psi^{d\beta}_n\rangle = \sum_{j \neq n} \frac{\langle \psi_j | m_{\beta} | \psi_n \rangle}{2\hbar(\omega_j - \omega_L)} |\psi_j\rangle,
\]

\[
|\psi^{e\gamma\beta}_n\rangle = \sum_{j \neq n} \frac{\langle \psi_j | \Theta_{\beta\gamma} | \psi_n \rangle}{6\hbar(\omega_j - \omega_L)} |\psi_j\rangle,
\]

\[
|\psi^{f\gamma\beta}_n\rangle = \sum_{j \neq n} \frac{\langle \psi_j | \Theta_{\beta\gamma} | \psi_n \rangle}{6\hbar(\omega_j + \omega_L)} |\psi_j\rangle.
\]

### 6.1.5 The Raman Optical Activity Tensors

Given the form of our perturbed wavefunction, we are now in a position to evaluate expectation values of the multipole moment operators. For example, the dipole moment arising from the perturbation associated with the electromagnetic field in the \(n^{th}\) eigenstate can be expressed as

\[
\langle \psi'_n | \hat{\mu}_\alpha | \psi'_n \rangle - \langle \psi_n | \hat{\mu}_\alpha | \psi_n \rangle = \left[ \langle \psi'_n | \hat{\mu}_\alpha | \psi'^{a\beta}_n \rangle + \langle \psi'^{a\beta}_n | \hat{\mu}_\alpha | \psi'_n \rangle \right] (E_\beta)_0
\]

\[
+ \left[ \langle \psi'_n | \hat{\mu}_\alpha | \psi^{b\beta}_n \rangle + \langle \psi^{b\beta}_n | \hat{\mu}_\alpha | \psi'_n \rangle \right] (\bar{E}_\beta)_0 + ...
\]

(6.1.60)
6.1. GENERAL THEORY

Dealing first with the terms involving kets of perturbed eigenstates, we can make use of (6.1.54) and (6.1.55), in addition to the identities (6.1.43) and (6.1.44)

\[
\langle \psi_n | \hat{\mu}_\alpha | \psi_n^{a\beta} \rangle (E_\beta) + \langle \psi_n | \hat{\mu}_\alpha | \psi_n^{b\beta} \rangle (\bar{E}_\beta) =
\]

\[
= \langle \psi_n | \hat{\mu}_\alpha \sum_{j \neq n} \frac{\langle \psi_j | \hat{\mu}_\beta | \psi_n \rangle}{2\hbar(\omega_{jn} - \omega_L)} | \psi_j \rangle (E_\beta) + \langle \psi_n | \hat{\mu}_\alpha \sum_{j \neq n} \frac{\langle \psi_j | \hat{\mu}_\beta | \psi_n \rangle}{2\hbar(\omega_{jn} + \omega_L)} | \psi_j \rangle (\bar{E}_\beta)
\]

\[
= \sum_{j \neq n} \frac{\langle \psi_n | \hat{\mu}_\alpha | \psi_j \rangle \langle \psi_j | \hat{\mu}_\alpha | \psi_n \rangle}{2\hbar} \left[ \frac{(E_\beta) + (\bar{E}_\beta)}{(\omega_{jn} - \omega_L)} + \frac{(\bar{E}_\beta)}{(\omega_{jn} + \omega_L)} \right]
\]

\[
= \sum_{j \neq n} \frac{\langle \psi_n | \hat{\mu}_\alpha | \psi_j \rangle \langle \psi_j | \hat{\mu}_\alpha | \psi_n \rangle}{\hbar} \left[ \frac{\omega_{jn}((E_\beta) + (\bar{E}_\beta)) + \omega_L((E_\beta) - (\bar{E}_\beta))}{(\omega_{jn}^2 - \omega_L^2)} \right]
\]

(6.1.61)

Similarly, for the eigenbras

\[
\langle \psi_n^{a\beta} | \hat{\mu}_\alpha | \psi_n \rangle (E_\beta) + \langle \psi_n^{b\beta} | \hat{\mu}_\alpha | \psi_n \rangle (\bar{E}_\beta) =
\]

\[
= \sum_{j \neq n} \frac{\langle \psi_n | \hat{\mu}_\alpha | \psi_n \rangle \langle \psi_j | \hat{\mu}_\alpha | \psi_n \rangle}{\hbar} \left[ \frac{\omega_{jn}((E_\beta) + i(\bar{E}_\beta))}{(\omega_{jn}^2 - \omega_L^2)} \right]
\]

(6.1.62)

so we can sum the two and obtain a fully analytical expression for the perturbed dipole moment (momentarily including only the contributions from the real and complex components of the electric field)

\[
\left[ \sum_{j \neq n} \frac{\omega_{jn} \langle \psi_n | \hat{\mu}_\alpha | \psi_j \rangle \langle \psi_j | \hat{\mu}_\beta | \psi_n \rangle}{\hbar(\omega_{jn}^2 - \omega_L^2)} + \sum_{j \neq n} \frac{\omega_{jn} \langle \psi_n | \hat{\mu}_\beta | \psi_j \rangle \langle \psi_j | \hat{\mu}_\alpha | \psi_n \rangle}{\hbar(\omega_{jn}^2 - \omega_L^2)} \right] (E_\beta) +
\]

\[
- \left[ \sum_{j \neq n} \frac{i \langle \psi_n | \hat{\mu}_\alpha | \psi_j \rangle \langle \psi_j | \hat{\mu}_\beta | \psi_n \rangle}{\hbar(\omega_{jn}^2 - \omega_L^2)} + \sum_{j \neq n} \frac{i \langle \psi_n | \hat{\mu}_\beta | \psi_j \rangle \langle \psi_j | \hat{\mu}_\alpha | \psi_n \rangle}{\hbar(\omega_{jn}^2 - \omega_L^2)} \right] (\bar{E}_\beta) + ...
\]

(6.1.63)

Invoking the Hermiticity of the multipole moment operators, i.e.

\[
\langle \psi_n | \hat{\mu}_\beta | \psi_j \rangle \langle \psi_j | \hat{\mu}_\alpha | \psi_n \rangle = \langle \psi_n | \hat{\mu}_\alpha | \psi_n \rangle \cdot \langle \psi_j | \hat{\mu}_\beta | \psi_j \rangle
\]
we can rewrite (6.1.63) as
\[
\left[ \sum_{j \neq n} \frac{\omega_{jn}}{\hbar(\omega_{jn}^2 - \omega_L^2)} \langle \psi_j \mid \hat{\mu}_\alpha \rangle \langle \psi_j \mid \hat{\mu}_\beta \rangle \langle \psi_n \rangle \right] (E_\beta)_0 \\
- \left[ \sum_{j \neq n} \frac{i \langle \psi_n \mid \hat{\mu}_\alpha \rangle \langle \psi_j \mid \hat{\mu}_\beta \rangle \langle \psi_n \rangle}{\hbar(\omega_{jn}^2 - \omega_L^2)} \right] (\dot{E}_\beta)_0 + \ldots
\] (6.1.64)

and invoking the definition of the real and complex parts of a complex number,
\[
z = x + iy
\]
\[
x = \text{Re} \frac{z + \bar{z}}{2} \quad y = \text{Im} \frac{z - \bar{z}}{2i}
\]
we are left with the expression
\[
\langle \psi'_n \mid \hat{\mu}_\alpha \mid \psi'_n \rangle = (\alpha_{\alpha\beta})_{nn} (E_\beta)_0 + \frac{1}{\omega_L} (\alpha'_{\alpha\beta}) (\dot{E}_\beta)_0
\] (6.1.65)

where the factor \(\omega_L^{-1}\) has been introduced to enforce dimensional consistency, and the electric dipole polarisability tensors have been introduced
\[
(\alpha_{\alpha\beta})_{nn} = \frac{2}{\hbar} \sum_{j \neq n} \frac{\omega_{jn}}{\omega_{jn}^2 - \omega_L^2} \text{Re} \langle \psi_n \mid \hat{\mu}_\alpha \rangle \langle \psi_j \mid \hat{\mu}_\beta \rangle \langle \psi_n \rangle
\] (6.1.66)
\[
(\alpha'_{\alpha\beta})_{nn} = -\frac{2}{\hbar} \sum_{j \neq n} \frac{\omega_L}{\omega_{jn}^2 - \omega_L^2} \text{Im} \langle \psi_n \mid \hat{\mu}_\alpha \rangle \langle \psi_j \mid \hat{\mu}_\beta \rangle \langle \psi_n \rangle
\] (6.1.67)

By use of the terms involving the electric field gradient and magnetic field, it can be shown in a similar way that (6.1.65) becomes
\[
\langle \psi'_n \mid \hat{\mu}_\alpha \mid \psi'_n \rangle = (\alpha_{\alpha\beta})_{nn} (E_\beta)_0 + \frac{1}{\omega_L} (\alpha'_{\alpha\beta}) (\dot{E}_\beta)_0 + \frac{1}{3} (A_{\alpha\beta\gamma})_{nn} \nabla_\beta (E_\gamma)_0 \\
+ \frac{1}{3\omega_L} (A'_{\alpha\beta\gamma})_{nn} \nabla_\beta (\dot{E}_\gamma)_0 + (G_{\alpha\beta})_{nn} (B_\beta)_0 + \frac{1}{\omega_L} (G'_{\alpha\beta})_{nn} (\dot{B}_\beta)_0 + \ldots
\] (6.1.68)
6.1. GENERAL THEORY

where we have introduced a number of higher order optical activity tensors

\[
(A_{\alpha,\beta\gamma})_{nn} = \frac{2}{\hbar} \sum_{j \neq n} \frac{\omega_j n}{\omega_j^2 - \omega_L^2} \text{Re} \langle \psi_n | \hat{\mu}_\alpha | \psi_j \rangle \langle \psi_j | \hat{\Theta}_{\beta\gamma} | \psi_n \rangle \tag{6.1.69}
\]

\[
(A'_{\alpha,\beta\gamma})_{nn} = -\frac{2}{\hbar} \sum_{j \neq n} \frac{\omega_L}{\omega_j^2 - \omega_L^2} \text{Im} \langle \psi_n | \hat{\mu}_\alpha | \psi_j \rangle \langle \psi_j | \hat{\Theta}_{\beta\gamma} | \psi_n \rangle \tag{6.1.70}
\]

\[
(G_{\alpha\beta})_{nn} = \frac{2}{\hbar} \sum_{j \neq n} \frac{\omega_j n}{\omega_j^2 - \omega_L^2} \text{Re} \langle \psi_n | \hat{m}_\alpha | \psi_j \rangle \langle \psi_j | \hat{m}_\beta | \psi_n \rangle \tag{6.1.71}
\]

\[
(G'_{\alpha\beta})_{nn} = -\frac{2}{\hbar} \sum_{j \neq n} \frac{\omega_L}{\omega_j^2 - \omega_L^2} \text{Im} \langle \psi_n | \hat{m}_\alpha | \psi_j \rangle \langle \psi_j | \hat{m}_\beta | \psi_n \rangle \tag{6.1.72}
\]

The induced electric quadrupole and magnetic dipole moments can be found in a similar fashion, yielding expressions that resemble (6.1.68)

\[
\langle \psi_n' | \hat{\Theta}_{\alpha\beta} | \psi_n' \rangle = (A_{\gamma,\alpha\beta})_{nn}(E_\gamma)_0 - \frac{1}{\omega_L}(A'_{\gamma,\alpha\beta})_{nn}(\dot{E}_\gamma)_0 + \ldots \tag{6.1.73}
\]

\[
\langle \psi_n' | \hat{m}_\alpha | \psi_n' \rangle = (G_{\beta\alpha})_{nn}(E_\beta)_0 - \frac{1}{\omega_L}(G'_{\beta\alpha})_{nn}(\dot{E}_\beta)_0 + \ldots \tag{6.1.74}
\]

where we have truncated the expressions so as to not introduce any more optical activity tensors that appear in higher order terms.

6.1.6 ROA Intensities

By certain combinations of the optical activity tensors, we are able to construct optical activity tensors that are invariant with respect to the choice of frame of reference. These quantities are denoted the isotropic invariants of the optical activity tensors, and can be combined to yield the ROA intensities[355]. Following the nomenclature used by Polavarapu[356], the numerator and denominator of the
CID expression in (6.0.1) are given by

\[ I^R - I^L = \frac{48\omega_c}{c} \left[ \frac{1}{\omega_L} \gamma_j^2 + \frac{1}{3\omega_L} \delta_j^2 \right] \]  \hspace{1cm} (6.1.75)

\[ I^R + I^L = 2 \left[ 45\alpha_j^2 + 7\beta_j^2 \right], \]  \hspace{1cm} (6.1.76)

where \( c \) is the speed of light. The terms \( \alpha_j^2, \beta_j^2, \gamma_j^2, \delta_j^2 \) are the isotropic invariants of the optical activity tensors, and are defined as various combinations of partial derivatives of the optical activity tensors,

\[ \alpha_j^2 = \frac{1}{9} \left( \frac{\partial \alpha_{xx}}{\partial Q_j} + \frac{\partial \alpha_{yy}}{\partial Q_j} + \frac{\partial \alpha_{zz}}{\partial Q_j} \right), \]  \hspace{1cm} (6.1.77)

\[ \beta_j^2 = \frac{1}{2} \left\{ \left( \frac{\partial \alpha_{xx}}{\partial Q_j} - \frac{\partial \alpha_{yy}}{\partial Q_j} \right)^2 + \left( \frac{\partial \alpha_{xx}}{\partial Q_j} - \frac{\partial \alpha_{zz}}{\partial Q_j} \right)^2 + \left( \frac{\partial \alpha_{yy}}{\partial Q_j} - \frac{\partial \alpha_{zz}}{\partial Q_j} \right)^2 \right\} \]  \hspace{1cm} (6.1.78)

\[ \gamma_j^2 = \frac{1}{2} \left\{ \left( \frac{\partial \alpha_{xx}}{\partial Q_j} - \frac{\partial \alpha_{yy}}{\partial Q_j} \right) \left( \frac{\partial G'_{xx}}{\partial Q_j} - \frac{\partial G'_{yy}}{\partial Q_j} \right) + \left( \frac{\partial \alpha_{xx}}{\partial Q_j} - \frac{\partial \alpha_{zz}}{\partial Q_j} \right) \left( \frac{\partial G'_{xx}}{\partial Q_j} - \frac{\partial G'_{zz}}{\partial Q_j} \right) + \left( \frac{\partial \alpha_{yy}}{\partial Q_j} - \frac{\partial \alpha_{zz}}{\partial Q_j} \right) \left( \frac{\partial G'_{yy}}{\partial Q_j} - \frac{\partial G'_{zz}}{\partial Q_j} \right) \right\} \]  \hspace{1cm} (6.1.79)
\[ \delta_j^\beta = \frac{\omega_L}{2} \left[ \left( \frac{\partial \alpha_{yy}}{\partial Q_j} - \frac{\partial \alpha_{xx}}{\partial Q_j} \right) \frac{\partial A_{xyy}}{\partial Q_j} + \left( \frac{\partial \alpha_{xx}}{\partial Q_j} - \frac{\partial \alpha_{zz}}{\partial Q_j} \right) \frac{\partial A_{yyz}}{\partial Q_j} \right. \\
+ \left( \frac{\partial \alpha_{zz}}{\partial Q_j} - \frac{\partial \alpha_{yy}}{\partial Q_j} \right) \frac{\partial A_{zzx}}{\partial Q_j} + \frac{\partial \alpha_{xy}}{\partial Q_j} \left( \frac{\partial A_{xyy}}{\partial Q_j} - \frac{\partial A_{yyz}}{\partial Q_j} + \frac{\partial A_{yyx}}{\partial Q_j} - \frac{\partial A_{yxz}}{\partial Q_j} \right) \left( \frac{\partial A_{yyz}}{\partial Q_j} - \frac{\partial A_{zzy}}{\partial Q_j} + \frac{\partial A_{xyz}}{\partial Q_j} - \frac{\partial A_{yxz}}{\partial Q_j} \right) \right], \]

where derivatives are performed with respect to the normal coordinates of the system. We qualitatively discuss the computation of these derivatives in Section 6.2.1.

6.2 Previous Computational Work

6.2.1 Algorithmic Developments

The computation of \textit{ab initio} Raman spectra has been possible for a number of years. The electric dipole-electric dipole polarisability tensor, \( \alpha_{\alpha\beta} \), for a molecular system can be expressed quantum mechanically by truncating the infinite sum-over-states of (6.1.66) to finite order[354]. Rewriting (6.1.66) for convenience,

\[ (\alpha_{\alpha\beta})_{nn} = \frac{2}{\hbar} \sum_{j \neq n} \frac{\omega_{jn}}{\omega_{jn}^2 - \omega_L^2} \text{Re} \langle \psi_n | \hat{\mu}_\alpha | \psi_j \rangle \langle \psi_j | \hat{\mu}_\beta | \psi_n \rangle \]

We assume that the incident excitation frequency does not induce resonance within the system, which can be enforced by taking \( \omega_L \ll \omega_{jn} \). Taking this inequality as opposed to \( \omega_L \gg \omega_{jn} \) ensures that we do not promote the system to a higher excitation level. This nonresonant approximation allows us to characterise \((\alpha_{\alpha\beta})_{nn}\)
as independent of the incident excitation, i.e. the “static limit”, where \( \omega_{jn}^2 - \omega_L^2 \approx \omega_{jn}^2 \), and obtain

\[
(\alpha_{\alpha\beta})_{nn} = \sum_{j \neq n} \frac{2}{E_j - E_n} \text{Re} \langle \psi_n | \hat{\mu}_\alpha | \psi_j \rangle \langle \psi_j | \hat{\mu}_\beta | \psi_n \rangle
\]  

(6.2.1)

where \( E_j \) is the energy of eigenstate \( | \psi_j \rangle \). Since the vast majority of biochemical systems (excluding transition metal complexes, which we do not discuss here) do not exhibit strong correlation, the gap between the ground and higher energy states is typically very large[357], which then allows us to truncate the expression (6.2.1) to low order, since excited states do not contribute significantly to the molecular polarisability[358].

Derivatives of \( (\alpha_{\alpha\beta})_{nn} \) with respect to the internal degrees of freedom of the molecular system have historically been evaluated numerically by computing \( (\alpha_{\alpha\beta})_{nn} \) at a number of geometrically displaced molecular geometries[359]. Taking the geometric displacement to be small, one can invoke a finite difference method to obtain numerical derivatives.

As we have seen in Section 3, the normal coordinates form a convenient internal basis with which to define a molecular system. Working with the harmonic approximation for the molecular potential energy surface, the normal coordinates are eigenfunctions of the Hessian. It is possible to work with higher accuracy methods that use anharmonic corrections to the molecular energy. However, anharmonic corrections are computationally cumbersome, and have only been utilised in small molecular systems, such as methylloxi­rane[360]. Indeed, calculation of the Hessian alone is a computationally intensive procedure. A number of techniques are available for avoiding explicit computation of the full Hessian, where only those modes with the highest intensity are evaluated by the “intensity-tracking” technique of Reiher and coworkers[361, 362], but are not widely implemented.
More recent developments have allowed for the incorporation of electric field perturbations into the molecular Hamiltonian\[363\]. Since \((\alpha_{\alpha\beta})_{nn}\) results from the second derivative of the molecular energy with respect to an applied electric field, third order derivatives of the Hamiltonian with respect to the electric field yield the spatial derivatives of \((\alpha_{\alpha\beta})_{nn}\). Analytical expressions for this Hamiltonian are rarely available, and so derivatives with respect to internal degrees of freedom can be found as above, by displacing the molecular geometry and evaluating \((\alpha_{\alpha\beta})_{nn}\) at these displaced geometries, followed by use of a finite differences method. Modern techniques, however, rely on analytical expressions for the spatial derivatives of \((\alpha_{\alpha\beta})_{nn}\) with respect to the internal molecular degrees of freedom\[364\].

It is only more recent developments that have allowed for the calculation of \textit{ab initio} ROA spectra. From (6.1.79) and (6.1.80), we see that in addition to the derivatives of \((\alpha_{\alpha\beta})_{nn}\), we also require derivatives of the electric dipole-magnetic dipole and electric dipole-electric quadrupole polarisability tensors, \(G'_{\alpha\beta}\) and \(A_{\alpha\beta\gamma}\), respectively. Whilst derivatives of the latter involve relatively straightforward extensions to the derivatives of \((\alpha_{\alpha\beta})_{nn}\), it is derivatives of \(G'_{\alpha\beta}\) that have historically required numerical treatment.

The seminal contribution to this field is that of Polavarapu\[356\]. By using an expression for \(G'_{\alpha\beta}\) in the static limit, i.e. frequency-independent form of the tensor\[365\], \(G'_{\alpha\beta}\) can be evaluated at a number of displaced geometries, and numerical derivatives taken at the Hartree-Fock level of theory. Later work extended the approach of Polavarapu to the DFT level of theory\[366\] and to frequency-dependent polarisabilities \[367\]. It is only the relatively recent developments of Liégois et al.\[368\] that have permitted a fully analytical approach to the computation of ROA spectra, where an analytical derivative procedure was introduced at the Hartree-Fock level of theory. DFT methods are also supported in modern
quantum chemical software packages[369].

6.2.2 Solvation Effects

The fact that biochemical systems are stabilised by solvent interactions makes their modelling problematic. Since the solvent plays an important role in dictating the energetics of a molecular system, it becomes imperative that spectroscopic calculations include some level of solvent modelling. It has been shown that simply omitting the solvent from spectroscopic calculations yields poor quality results that severely digress from experiment[370]. Indeed, when dealing with charged species, such as zwitterionic amino acids, the molecules are simply not energetically stable if the solvent is not included[371], and simply optimise to the neutral amine-carboxylic acid species in the absence of constraints.

The two major models for solvation effects in \textit{ab initio} calculations are implicit and explicit solvation. With the former[372], the solvent environment is modelled as a homogeneous dielectric medium, thus increasing the magnitude of electrostatic interactions relative to the system \textit{in vacuo}. The Polarisable Continuum Model[33] (PCM) and Conductor-like Screening Model[34] (COSMO) have been effective in predicting molecular energies[373], but lack any atomic-level properties of the solvent, such as the presence of solvent hydrogen bond donors/acceptors. Implicit solvation methods have been found to offer minimal benefits in computing vibrational frequencies, and therefore vibrational spectra, relative to gas phase calculations[374].

A number of explicit solvation schemes have been investigated. At the lowest end, one can include a small number of solvent molecules and treat the entire solute-solvent system quantum mechanically, perhaps embedded in an implicit
6.2. PREVIOUS COMPUTATIONAL WORK

solvation field. The number of explicit solvent molecules treated in this way is somewhat arbitrary, and their positioning around the solute can be problematic. Solvent molecules can be added in an *ad hoc* fashion around polar (or non-polar if the solvent is non-polar) groups, leading to significant improvement over gas phase spectra[375]. However, a number of problems can also be attributed to this methodology, such as poor geometrical optimisation of the system as a result of its distance from a minimum on the potential energy surface[376].

Another more successful method is the use of QM/MM, where the solvent is treated classically and the solute quantum mechanically. This has proven to be a very adept methodology for calculating spectra. For example, Cheeseman et al.[299] have predicted the methyl−β − d−glucose ROA spectrum to an impressive level of accuracy, reproducing the vast majority of spectral features. More recently, this approach has been further validated on a number of other systems, such as glucuronic acid[377] and β − d−xylose[378]. However, one is still constrained in selecting a conservative number of solvent molecules to include in the calculation for it to remain tractable; simply including hundreds of solvent molecules in the QM/MM calculation is not routinely feasible.

A complexity in the QM/MM method of modelling solvent arises in the geometrical optimisation of the system. For the normal modes of motion of the system to be correctly determined, the system is required to relax to an energetic minimum on the potential energy surface. Geometry optimisation is notoriously difficult for a system including even a conservative number of explicit solvent molecules, as modelled in [377, 378]; the presence of water molecules in these studies flattens the potential energy surface, rendering the location of well-defined minima difficult. Two methods were proposed; OptAll and OptSolute. The OptAll scheme optimises the entire QM/MM system, while the OptSolute method
freezes all solvent molecules and optimises only the solute in the field of the static solvent. The latter method is naturally quicker than the former, but was found to yield poorer quality spectra. Some trade-off between accuracy and computational tractability is to be expected, but without there being an accepted methodology for qualitatively comparing predicted spectra to experimental spectra, it is difficult to ascertain whether the improvement in spectral quality warrants the additional computational efforts.

6.2.3 Quantitative Measures of Similarity

A note is in place on the absence of a quantitative measure of spectral quality in the literature. Typically, the comparison of experimental and computed spectra is reduced to a “dark art”, where experts cast their eye over two spectra and ascertain the prediction quality with vague statements, such as “good agreement”. Spectra are composed of highly complex patterns, and so validation in this crude manner is particularly unsatisfactory. Ideally, one would like to reduce the determination of the quality of prediction to some unambiguous numerical procedure.

The difference in absolute amplitudes between computed and experimental spectra is irrelevant since the amplitude of an experimental spectrum is a function of collection time. As such, one requires that the absolute amplitudes of the experimental and computed spectra be normalised for a direct comparison to be made. Relative amplitudes are, however, informative.

An intuitive approach to quantitatively comparing experimental and computed spectra is to repurpose the Carbó index[379, 380]. The Carbó index was initially proposed in the context of ascertaining the similarity in electron density between two molecules, and subsequently correlating this similarity by a structure-activity
relationship. Whilst heavily cited, the notion of the Carbó index is perhaps unsatisfying as it requires the interpretation of tiny differences ($O(10^{-3})$) in molecular similarity to be related to gross differences in molecular activity. The Carbó index, $S_{fg}$, has functional form

$$\begin{aligned}
S_{fg} &= \frac{\int f(x)g(x)dx}{\sqrt{\int f(x)^2dx \int g(x)^2dx}} \approx \frac{\sum_{i=1}^{N} f(\nu_i)g(\nu_i)}{\sqrt{\sum_{i=1}^{N} f(\nu_i)^2 \sum_{j=1}^{N} g(\nu_j)^2}},
\end{aligned}
$$

where $f(x)$ and $g(x)$ are arbitrary functions. One can immediately interpret $S_{fg}$ based on the numerator, which is the inner product (or “overlap” in quantum mechanical parlance) of two functions, appropriately normalised by the denominator so that $0 \leq S_{fg} \leq 1$, $S_{fg} = 0$ denoting no similarity and $S_{fg} = 1$ denoting equivalence between the functions. Neither experimental or computed spectra have analytical form, so the integrals in the above equation require discretisation. The approximation in parentheses in the above equation denotes the discrete form of the Carbó index, which has been repurposed to denote the similarity between two spectra, $f(\nu)$ (experimental) and $g(\nu)$ (computed), the independent variable being the wavenumber[381, 382].

However, the Carbó index is not immediately applicable in this form. For example, if experimental and computed spectra had the exact same form, but one was shifted along the abscissa relative to the other by some small amount, the Carbó index could classify the spectra as unsimilar. In reality, the two spectra are exactly equivalent but for some shift that could have been resultant from a particular experimental setup.

To remedy this failing, one can introduce a number of free parameters into the evaluation of $f(\nu)$ and $g(\nu)$ at the specified points. Maximising the Carbó index with respect to these free parameters then corresponds to the similarity between two spectra. One free parameter that immediately springs to mind is that intro-
duced by Radom[383], who noticed that computed spectra are overestimated in the abscissa inter alia the harmonic approximation, and so the computed spectrum should be scaled. i.e. $g(\alpha \nu)$, with $0.95 \leq \alpha \leq 1.0$.

Another free parameter could correspond to a shifting, so that the computed spectrum is $g(\alpha \nu + \beta)$, with $\beta$ some global spectrum shift. One could select $\alpha, \beta$ to be anisotropic, so that they scale and shift different regions of the computed spectrum by different amounts. The possibilities are virtually endless, and at some point the optimisation of free parameters simply leads to the computed spectrum becoming infinitely plastic. In this case, the similarity between computed and experimental spectra becomes meaningless.

In our work, we have regrettably avoided use of a similarity measure between computed and experimental spectra since it is not immediately clear how to introduce free parameters without rendering the similarity metric questionable. However, this is certainly an avenue that can be explored in the future, and would be of undoubted use to the community.

### 6.2.4 Large Biomolecular Systems

In the previous section, we have restricted our discussion to small molecular systems. However, biomolecular systems are frequently composed of large polymers, such as polypeptides, whose length frequently spans over 100 amino acid residues in length. It is hugely challenging to model the ROA spectrum of such a large system given the computational complexities associated with calculating ROA spectra. However, a number of groups have approached this problem with a great degree of success.
6.2. PREVIOUS COMPUTATIONAL WORK

Early ROA studies of polymeric systems have included helical decaalanine\[289\] (10 amino acids), valinomycin\[384\] (12 amino acids) and the $\beta$-domain of metallothionein [385] (31 amino acids), where the latter was modelled \textit{in vacuo}. The results obtained were impressive, given that these studies were all performed before analytical derivatives of $G'_{\alpha\beta}$ were available.

A particularly impressive tool is the Cartesian Coordinate Tensor Transfer (CCT) methodology\[386\], which resembles the notion of the Kabsch RMSD that we employ in Section 6.3.3. The basic idea is the fragmentation of a large system into a number of small subsystems. The unitary transformation from the subsystem to the large system is found by minimising the associated RMSD between the two systems. The unitary transformation matrix can then be used to transform quantities computed in the subsystem, such as the Hessian and optical activity tensors, into the large system, exploiting the apparent locality of these quantities. CCT has been used in the calculation of vibrational circular dichroism spectra of oligonucleotides\[387\] and IR spectra of polypeptides\[388\], with results that are almost indistinguishable from the fully \textit{ab initio} spectrum. Similarly impressive results have been obtained for insulin[389] (51 amino acids).

In spite of the excellent results obtained through CCT, there have been a number of studies which refute the validity of CCT for ROA spectra\[390, 391\]. In short, the authors claim that the ROA $G'_{\alpha\beta}$ and $A_{\alpha,\beta\gamma}$ tensors are not properly transformed by CCT. The criticisms have been rebutted in a later work\[392\], where it has been claimed that the fragmentation scheme needs to be carefully selected when dealing with systems involving intramolecular hydrogen bonds, and that CCT yields very good results given an appropriate fragmentation. Indeed, a recent study using CCT has provided excellent results for the ROA spectra of a number of large globular proteins\[393\], such as hen egg-white lysozyme, comprising well over 100
amino acids.

6.3 Technical Details

6.3.1 Zwitterionic Histidine

Histidine is somewhat unique in its ability to uniquely undertake a number of biochemical roles. Traditionally characterised as a polar amino acid, histidine does not substitute particularly well with other amino acids[394]. The nitrogen atoms of the imidazole group have long been known to act as proton shuttles, and grant functionality for a number of enzymatic mechanisms. However, histidine has also been shown to participate in interactions which are typically perceived to involve large aromatic amino acids such as tryptophan and phenylalanine. Noncovalent interactions involving histidine have been shown to exist in the form of $\pi - \pi$ stacking between its imidazole ring and a number of DNA nucleobases[395].

Histidine features in the so-called “catalytic triad” of a number of biochemically important enzymes, such as serine and cysteine proteases, where it acts as a proton shuttle during the process of protein catabolism[396]. The Manzetti mechanism[397] suggests that matrix metalloproteinases, a set well-characterised cancer therapeutic targets, attain catalysis through a pair of histidine residues which coordinate a Zn$^{2+}$ ion. The penta-coordinate Zn$^{2+}$ ion is induced to act as a reversible electron donor, and can hydrolyse the scissile peptide bond. Indeed, research within the field of biocatalysis has found evidence for histidine residues within tripeptide motifs coordinating even more exotic metal ions, such as Pt$^{2+}$ and Au$^{3+}$, and exhibiting chemotherapeutic properties[398].
Five protonation states of histidine exist: \( \text{His}^{2+}, \text{His}^{1+}, \text{His}^0, \text{His}^{1-} \) and \( \text{His}^{2-} \), where the superscript denotes the formal charge state of the species. In addition to these protonation states, the imidazole ring in \( \text{His}^0 \) and \( \text{His}^{1-} \) can exist in one of two tautomeric forms. This originates from the lone pairs of the nitrogen atoms contributing to \( \pi \)-delocalisation around the five-membered imidazole ring. The imidazole nitrogen atoms are labelled \( N_\pi \) or \( N_\tau \), denoting the nitrogen one and two bonds away, respectively, from the alkyl sidechain, as shown in Figure 6.1.

For convenience, we also depict the major torsional degrees of freedom that are typically varied for conformational studies of histidine; \( \chi_1 \) and \( \chi_2 \) are the side chain dihedrals.

Tautomeric forms are denoted by specifying the protonated nitrogen, i.e. \( N_\pi \text{H} \).
or N$_r$H. Using this nomenclature, we can uniquely label the seven protonation/tautomeric states of histidine: His$^{2+}$, His$^{1+}$, His$^0$[N$_r$H], His$^0$[N$_r$H], His$^{1-}$[N$_r$H], His$^{1-}$[N$_r$H] and His$^{2-}$. These states are represented pictorially in Figure 6.2.

Figure 6.2: Zwitterionic histidine protonation states and tautomers.

In the following, we compute the Raman and ROA spectra for the His$^0$ and His$^{1+}$ charge states, as these are by far the most biologically prevalent forms. In forming the His$^0$ spectra, we include contributions from both tautomers of histidine. A further point worth consideration is that charge states transition in response to a pH. However, pH is a macroscopic quantity, and so cannot be properly accounted for in single molecule studies. We introduce the fractional distribution of charge states in a solution at a given pH, i.e. the proportion of the species in a given charge state at a given pH. We present the fractional distributions of the histidine charge states as a function of pH in Figure 6.3.

At physiological pH, we see that roughly 94% of histidine is the His$^0$ form, with small amounts of His$^{1-}$ and His$^{1+}$. If we are to compare our single charge state spectra to experimental spectra, we must ensure that the experimental conditions coincide with the maxima of the relative fractional distribution curves. Alternatively, if we were attempting to compute the spectra from previously acquired
Figure 6.3: Fractional contribution of the various histidine charge states as a function of pH. Blue, green, red and yellow curves correspond to the fractional contributions of the His$^{2+}$, His$^{1+}$, His$^{0}$ and His$^{1-}$ states, respectively. We have limited the pH range between 0 and 12, and so omit the fractional contribution of the His$^{2-}$ state.

Experimental results, we would have to consider the effects of additional charged species, and include their contributions into the final spectrum.

We have prepared samples of histidine at a pH of 7.8 and 4.2, both at a concentration of 0.26M. Spectra were collected over a period of 7 hours, using an incident wavelength of 532nm, a laser power of 700mW and a 1.47 second exposition time. All spectra have been normalised so that they can be directly compared with the computed spectra. The absolute amplitudes are therefore irrelevant in our analysis.
6.3.2 Computing Spectra from a Number of Conformers

An optical spectrum of a molecular system is a function of the internal degrees of freedom of a system. Two vastly differing conformers of a system possess distinct spectra; while the major features are conserved between the two, e.g. peptide bond signal, a number of the finer details are subject to change. Therefore, when we compute an optical spectrum, we require some means for including the contribution to the spectrum of all conformers that make up the conformational space of the molecular system.

Since we are limited by computational resources, it is simply not feasible to exhaustively compute spectra for every possible conformer of a system and combine them to yield the total spectrum. Instead, we are reduced to computing the spectra of a small number of conformers, and using these to form the total spectrum. Each conformer spectrum has an associated weight, \( w_k = \exp[-\beta \Delta G(x_k)] / \sum_{k=1}^{N} w_k \), where \( \beta \) is the thermodynamic beta (= \( 1/k_B T \), where \( k_B \) is the Boltzmann constant and \( T \) is the temperature), \( \Delta G(x_k) \) is the free energy of the \( k^{th} \) conformer, and the weights are normalised over all conformer weights. Then, given the spectrum associated with the \( k^{th} \) conformer, \( f_k(\nu) \), the total spectrum is given by

\[
f(\nu) = \sum_{k=1}^{N} w_k f_k(\nu) .
\]

Introducing some nomenclature, each conformer spectrum is Boltzmann weighted. The Boltzmann weight of a spectrum is linked to the proportion of time spent in the given conformer. By analysis, we see that the lower the free energy of a conformer, the higher its Boltzmann weight, implying that the amount of time the system spends in a given conformational state is a function of its energy.

From the definition of the total spectrum, it is apparent that if we are to select
only a small number of conformers with which to compute a spectrum, we would like these conformers to be as low in energy as possible to maximise their Boltzmann weights. Therefore, optimisation of the conformers is required. However, we also require as diverse a conformer set as possible. The collection time for an experimental spectrum is of the order of hours, which means the system is able to sample all regions of conformational space, i.e. barriers between energetic minima do not constrain the system to a single conformational state.

The issue of selecting as diverse a conformational sample as possible is tackled in Section 6.3.3, while the issue of optimising the system is dealt with in Section 6.5.

### 6.3.3 Filtering of Similar Geometries

For the following discussion, the $3 \times N$ matrix $\mathbf{X} = \begin{bmatrix} x_1 & \ldots & x_N \end{bmatrix}$, $N$ being the number of atoms in a molecule, is used to denote a conformation, where $x_i$ is the Cartesian position vector for the $i^{th}$ atom.

Perhaps the most ubiquitous form of geometrical comparison between two structures, $\mathbf{X}$ and $\mathbf{Y}$, is the root mean square deviation (RMSD), which we denote $R(\mathbf{X}, \mathbf{Y})$,

$$R(\mathbf{X}, \mathbf{Y}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |x_i - y_i|^2}. \quad (6.3.1)$$

However, in the form of (6.3.1), a number of pitfalls can be encountered when dealing with molecular conformations. The most significant issue is the fact that $R(\mathbf{X}, \mathbf{Y})$ is not invariant with respect to rigid transformations. For example, if $\mathbf{X}$ and $\mathbf{Y}$ possess the same internal geometry, but are rigidly translated relative to one another, then (6.3.1) represents the two geometries as being dissimilar. The same result holds for the case of rigid rotations relative to one another. Molecules
(in homogeneous fields) are entirely equivalent after having undergone rigid translations, and so the RMSD in its current incarnation is of no value for our purposes.

We can render (6.3.1) in a useful form by finding $\mathcal{R}(X, Y)$ such that it is minimised with respect to all rigid rotations and translations, i.e. by minimising the function

$$E(X, Y) = \frac{1}{N} \sum_{i=1}^{N} |Ux_i - y_i|^2,$$

where $U$ is an orthonormal matrix, i.e. $U^T U = I$, the identity matrix, and can therefore be interpreted as a matrix defining a rigid transformation [399]. Taking the square root of $E(X, Y)$ therefore corresponds to the minimised RMSD (mRMSD) between the two molecular conformations. A further step is required to ensure our definition of the mRMSD is robust; if the orthonormal transformation matrix $U$, does not constitute a right-handed system ($\det(U) < 0$), then it corresponds to an improper rotation, i.e. a transformation that can be represented by some combination of a rotation about an axis and subsequent reflection in a plane. Thus, we require computation of the determinant of $U$ to ensure we deal only with proper rotations.

The method for minimising (6.3.2) with respect to $U$ can be undertaken in a number of ways. Historically, the orthonormality conditions of $U$ have been added to (6.3.2) as Lagrange multipliers. Formal solutions were found by Kabsch[400, 401], after whom the algorithmic formalism for solution of (6.3.2) is named. We proceed in deriving the form of $U$ in a more physically intuitive way, elaborated upon in a number of more recent sources [402, 403].

Conveniently, (6.3.2) can be made invariant with respect to global translations before seeking the form of $U$. Defining the centroids of $X$ and $Y$ as

$$x_C = \frac{1}{N} \sum_{i=1}^{N} x_i \quad y_C = \frac{1}{N} \sum_{i=1}^{N} y_i,$$

(6.3.3)
we can shift each molecular conformation to its centroid-centred frame of reference by \( \mathbf{x}_i = \mathbf{x}_i - \mathbf{x}_C \) and \( \mathbf{y}_i = \mathbf{y}_i - \mathbf{y}_C \), a convention which we adopt herein. Allowing the centroids \( \mathbf{x}_C \) and \( \mathbf{y}_C \), to coincide can be shown to be equivalent to optimally aligning the two molecular conformations\[404]. Thus, we are free to continue with the rotational alignment of the two conformations.

Rewriting (6.3.2),

\[
N \mathcal{E}(X, Y) = \sum_{i=1}^{N} |x'_i - y|^2 ,
\]

where we have used the shorthand \( x'_i = U \mathbf{x}_i \). Alternatively, we can use a matrix representation

\[
N \mathcal{E}(X, Y) = \text{Tr} \left[ (X' - Y)^\top (X' - Y) \right] ,
\]

where \( X' = U \mathbf{X} \). Expansion of the right-hand side of (6.3.5) can be performed by invoking the identity

\[
\text{Tr}(X' - Y)^\top (X' - Y) = \text{Tr}(X'^\top X') + \text{Tr}(Y^\top Y) - 2 \text{Tr}(Y^\top X') .
\]

Recognising\(^6\) that \( \text{Tr}(X'^\top X') = \sum_{i=1}^{n} x_i^2 \) and \( \text{Tr}(Y^\top Y) = \sum_{i=1}^{n} y_i^2 \),

\[
N \mathcal{E}(X, Y) = \sum_{i=1}^{N} \left[ x_i^2 + y_i^2 \right] - 2 \text{Tr}(Y^\top X') .
\]

Since the summation is not dependent upon \( U \), the minimisation of \( \mathcal{E}(X, Y) \) is equivalent to maximising \( \text{Tr}(Y^\top X') = \text{Tr}(Y^\top U \mathbf{X}) \). Since the trace of a product of matrices is commutative, the quantity we wish to maximise can be rewritten as \( \text{Tr}(\mathbf{X} \mathbf{Y}^\top U) \).

We can write \( \mathbf{X} \mathbf{Y}^\top \) as a singular value decomposition (SVD), i.e. \( \mathbf{X} \mathbf{Y}^\top = \mathbf{V} \Sigma \mathbf{W}^\top \), where \( \mathbf{V}, \mathbf{W} \) are orthonormal matrices of left and right eigenvectors, respectively.

\(^6\)These components do not need to be primed since the rotation matrix \( U \) does not modify the length of the \( \mathbf{x}_i \).
and $\Sigma$ is a diagonal matrix of ordered eigenvalues, i.e. $\sigma_{11} > \sigma_{22} > ...$. Once again invoking the commutativity of the trace of a product of matrices

$$\text{Tr}(T^T X') = \text{Tr}(V \Sigma W^T U) = \text{Tr}(\Sigma W^T U V) = \sum_{i=1}^{3} \sigma_{ii} w_i^T U v_i, \quad (6.3.7)$$

where $w_i$ and $v_i$ are the $i^{th}$ columns of $W$ and $V$, respectively. Introducing $T = W^T U V$,

$$\text{Tr}(Y^T X') = \sum_{i=1}^{3} \sigma_{ii} t_{ii} \leq \sum_{i=1}^{3} \sigma_{ii}, \quad (6.3.8)$$

where the inequality results from the fact that $T$ is orthonormal since it is formed from a product of orthonormal matrices. The orthonormality of $T$ also guarantees $\det(T) = \pm 1$. By analysis, we see that this quantity is maximised when $T$ is equal to the identity matrix. Following from this, $T = W^T U V = I$, and since $W, V$ are orthogonal, $WIV^T = WW^T U V V^T$, implying

$$WV^T = U, \quad (6.3.9)$$

where $W^T$ and $V$ are the right- and left-eigenvectors of the covariance matrix $XY^T$.

Our final consideration centres on whether $U$ defines a proper or improper rotation, as we have previously alluded to. We can account for this by specifying that if $\det(XY^T) > 0$, then $d = 1.0$, otherwise $d = -1.0$, allowing us to give the final expression for $U$,

$$U = W \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{bmatrix} V^T. \quad (6.3.10)$$

We are now in a position to give the algorithmic workflow for ascertaining the Kabsch RMSD between two molecular geometries in Algorithm 7.
6.4. COMPUTATIONAL DETAILS

1. Build the geometry matrices $X, Y$
2. Transform $X$ and $Y$ to their respective centroid-centred frames
3. Compute the covariance matrix, $XY^T$
4. Form the SVD of the covariance matrix, $XY^T = V\Sigma W^T$
5. Calculate $d = \det[XY^T]$
6. Compute the Kabsch rotation matrix, $U$, from (6.3.10)

Algorithm 7: Kabsch RMSD Algorithm

We propose a workflow for the implementation of the Kabsch RMSD for computing a set of maximally dissimilar molecular conformations, the “ensemble”, from a larger set. To this end, we propose the use of a greedy algorithm similar to that introduced in Section 5.2.2, the MaxMin algorithm. Given some initially large set of molecular conformations, we implement Algorithm 2, until an ensemble of the required size has been constructed. This ensemble then corresponds to the set of structures that are maximally dissimilar from the provided larger set.

6.4 Computational Details

6.4.1 Molecular Dynamics

The arrangement of solvent molecules is typically accomplished by use of molecular dynamics. The system is solvated and the resultant trajectory is parsed into a number of snapshot configurations. These configurations are subsequently used as the conformational ensemble for spectral calculations. Recent work has selected snapshots entirely at random, with very rough (qualitative) estimates used to maximise the conformational diversity of the ensemble. The snapshots are then energetically minimised, and those snapshots with a high Boltzmann weight form an ensemble which are considered representative of the dominant system conforma-
tions. Since these conformations are dominant, it is presumed that they contribute most to the spectrum of the system.

We have conducted 100ns molecular dynamics trajectories for His$^0[\text{N},\text{H}]$, His$^0[\text{N},\text{H}]$ and His$^{1+}$, solvated in boxes comprising 469 water molecules, with a chloride ion added in the latter case to ensure the system remained electrically neutral. Throughout, we will use the GROMACS molecular dynamics package[405] in conjunction with the OPLS-aa force field[406].

Initial energy minimisation of these systems was undertaken with the steepest descent method, until convergence of the total system energy was attained. Following energy minimisation, we have conducted two separate equilibration phases. The first was a 1 ns NVT equilibration, using the Berendsen thermostat to maintain a temperature of 300 K, whilst the second was a 1ns NPT equilibration using the Parrinello-Rahman barostat to maintain a temperature and pressure of 300 K and 0.1 MPa, respectively. The particle mesh Ewald methodology was used to treat long-range electrostatics. A cutoff distance of 10 Å was chosen for Coulombic and van der Waals interactions. Both stages of equilibration were verified as being properly equilibrated by checking the convergence of the temperature for the NVT equilibration and the density of the system for the NPT equilibration. Both showed convergence after roughly 100 ps, and so our equilibration was assumed sufficient.

Our final step involved a 100 ns production molecular dynamics run in the NPT ensemble with temperature and pressure of 300 K and 0.1 MPa (standard conditions), respectively. To this end, we have used the Berendsen thermostat coupled with the isotropic Parrinello-Rahman barostat. The dynamical timestep used for our simulations was 0.5 fs, with a snapshot geometry output every 5ps, resulting in 20,000 snapshot geometries. Coulombic and van der Waals interactions were
treated as we have described in the preceding paragraph.

### 6.4.2 Geometric Filtering

Past work\cite{407} on computing the Raman and ROA spectra of zwitterionic histidine has implemented a simplistic conformational selection tool. Six major *in vacuo* conformational preferences for histidine were proposed to arise from the $\chi_1, \chi_2$ torsional degrees of freedom: trans-plus ($\chi_1 = 90^\circ, \chi_2 = 180^\circ$); trans-minus ($\chi_1 = -90^\circ, \chi_2 = 180^\circ$); gauche-plus-plus ($\chi_1 = 90^\circ, \chi_2 = 60^\circ$); gauche-plus-minus ($\chi_1 = -90^\circ, \chi_2 = 60^\circ$); gauche-minus-plus ($\chi_1 = 90^\circ, \chi_2 = -60^\circ$); and gauche-minus-minus ($\chi_1 = -90^\circ, \chi_2 = -60^\circ$). These conformers were subsequently solvated in an *ad hoc* manner and energetically minimised. However, the solvation of *in vacuo* energetic minima does not yield a physically realistic conformational ensemble. Intramolecular hydrogen bonds stabilise *in vacuo* structures, which are not as preferable when the solute is solvated. Under explicit solvation, interactions with solvent are more favourable than the intramolecular hydrogen bonds\cite{408}. Not only this, but zwitterionic amino acids are not stable *in vacuo*. This instability necessitates the introduction of non-physical constraints on energetic optimisation to maintain the zwitterionic form.

For each system, we have utilised the Kabsch selection methodology outlined in Section 6.3.3. For each trajectory comprising 20,000 snapshots, we have parsed the 20 snapshots whose solute geometries are the most mutually diverse (as outlined in Section 6.3.3). Solvent geometries were not included in the Kabsch filtering. We present a conformer analysis of the solute geometries obtained by the Kabsch filtering in Figure 6.4.

We have highlighted in grey those regions of the $(\chi_1, \chi_2)$ plot in Figure 6.4 that
Figure 6.4: Sidechain dihedral angles ($\chi_1$, $\chi_2$) taken by conformers within the three histidine ensembles obtained by a Kabsch filtering of the MD trajectories; His$^{1+}$ (blue circles), His$^0[N_\pi H]$ (green diamonds) and His$^0[N_\tau H]$ (red triangles). Grey hatched regions correspond to those conformers used in the work of Deplazes et al.
6.4. COMPUTATIONAL DETAILS

correspond to the \textit{in vacuo} minima used in \cite{407}, $\pm 30^\circ$. Concerning the His$^{1+}$ conformers, the ($\chi_1 = 60^\circ, \chi_2 = -60^\circ$) region is not sampled at all. Analysis shows this to result from an unfavourable ammonium-imidazole $N\pi H$ contact. Concerning the His$^0$ conformers, when $\chi_2 = 90^\circ$, a favourable ammonium-imidazole $N\pi$ interaction stabilises the His$^0[N\tau H]$ tautomer.

The poor sampling of the $\chi_1 = -60^\circ$ region for the His$^0[N\tau H]$ tautomer arises from an inability to form a stabilising carboxylate-imidazole $N\pi H$ interaction. In contrast, the protonated $N\pi$ in the His$^0[N\pi H]$ and His$^{1+}$ forms permits this stabilising interaction.

Importantly, we see the significant levels of sampling in the $\chi_1 = \pm 180^\circ$ regions for all forms of histidine. These conformations correspond to an extended imidazole group, where no intramolecular interactions involving the zwitterionic groups are formed. In these conformations, solvent molecules are able to favourably interact with both the peptide groups and the imidazole, in keeping with the findings of \cite{408}. We have therefore demonstrated the inadequacy of using \textit{in vacuo} minima for the conformational sampling of solvated systems.

For our microsolvation studies, we have decided to use water clusters comprising 5, 10, 15 and 20 explicit water molecules. To this end, we selected the closest \{5, 10, 15, 20\} water molecules to the centre of mass of the zwitterionic histidine from each conformer. The Raman and ROA spectra for each of these was computed and compared to experiment, in the expectation that the increasing number of water molecules would yield spectra more similar to experiment.
6.4.3 Calculation of Spectra

All \textit{ab initio} calculations are undertaken using the Gaussian09 software package, with a two-layer ONIOM\cite{409} method. Histidine is modelled in the high layer, and solvent is modelled in the low layer. The high layer is treated with the B3LYP/6-31G(d) level of theory, and the low layer with the AMBER99SB force field including TIP3P parameters for the water molecules. Electronic embedding is used, so that the low layer electrostatics are incorporated into the quantum mechanical Hamiltonian. When appropriate, \textit{ab initio} optimisations are performed with the Berny algorithm in the Cartesian basis, and we specify that no microiterations are used for electronic embedding.

For calculation of the molecular optical activity tensors, we invoke the analytical twostep formalism, or $n+1$ algorithm\cite{410}, in which the harmonic frequency and ROA tensor calculations can be separated into differing levels of theory. For calculation of the normal coordinates and harmonic frequencies, we have used the B3LYP/6-31G(d) level of theory, while for the frequency-dependent ROA tensors\cite{411}, we have used the B3LYP/rDPS level of theory\cite{412}. rDPS is a rarified basis set that is based on the 3-21++G basis set with semi-diffuse $p$ functions on all hydrogen atoms\cite{413}. rDPS has been shown to provide a similar level of accuracy in the resulting spectra to much larger Dunning basis sets, such as aug-cc-pVDZ.

An excitation wavelength of 532nm was used for calculation of the ROA tensors, in keeping with the conventional experimental setup. We have obtained scattered circular polarisation backscattered (SCP-180) ROA intensities. Individual conformer spectra are Boltzmann weighted, and those with a Boltzmann weight exceeding 0.5% are used to form the spectrum (using a Lorentzian bandwidth of $10\text{cm}^{-1}$),
6.5. **CONFORMER OPTIMISATION**

for comparison with experimental spectra.

When referring to peptide Raman/ROA spectra, there are a number of spectral windows that are prominent enough to merit individual names. The Amide I band is situated at roughly 1650 cm$^{-1}$, and primarily comprises carbonyl stretching modes. The Amide II band is situated at roughly 1550 cm$^{-1}$ and comprises equal proportions of C-N-H bending and C-N stretching modes. Finally, the Amide III band is situated near 1300 cm$^{-1}$, and is dominated by the same modes as for the Amide II band. We will refer to these regions by their respective names herein.

### 6.5 Conformer Optimisation

The optimisation of solvated systems is notoriously difficult owing to the absence of well-defined minima on the PES. If the PES is significantly undulant, conventional minimisation algorithms struggle, as low order spatial derivatives do not suffice in describing the local topology of such functions[291], and cannot direct the system to appropriate regions on the PES. One is frequently reduced to sequentially optimising the system at a number of increasingly complex levels of theory to obtain convergence of the maximum atomic forces and displacements. Whilst tedious, it is also no guarantee of convergence, as we now outline.

To circumvent having to compute (time-consuming) spatial derivatives of the energy at every step of the geometry optimisation, updating methods are typically invoked[414, 415]. Whilst one can in principle explicitly compute the spatial derivatives of the energy at each optimisation step, for large systems this is not feasible. As such, the Hessian is explicitly computed at the start of the geometry optimisation, and subsequently updated as a function of the displacement of the system.
from this initial geometry by use of low order derivative information. By approxi-
mating the Hessian at each step of the geometry optimisation, one can save a
great deal of time. Indeed, such Hessian updating schemes have proven to be very
accurate for a number of systems[416, 417].

However, consider now the case where the system begins in a conformation which
is far from the energetic minimum. More steps are required for the geometry
optimisation to reach an energetic minimum. Given that the potential energy
surface is highly undulant, the Hessian updating scheme leads to increasingly poor
approximations to the Hessian as the optimisation proceeds further from the initial
geometry. As such, the geometry optimisation fails to converge. One is then
required to adopt some intermediate approach, where the Hessian is explicitly
calculated, updated for a number of timesteps, followed by explicit recomputation.
The frequency with which one performs the explicit recomputation of the Hessian
is then a matter for striking a balance between efficiency and accuracy.

An alternative approach to optimising the entire system has been alluded to in
Section 6.2.2. In the work of Zielinski[378] and Mutter[377], two levels of optimi-
misation were considered; OptAll and OptSolute. The former optimises the
entire system, and the latter only the solute in the field of the unoptimised solvent.
Noting that, in these studies, the optimisations were performed on snapshots from
MD trajectories, the system is initially far from the minimum energy conforma-
tion, rendering the optimisation all the more difficult. Hence, if one is able to “get
away with” a simpler optimisation, the savings will be substantial.

The reasonable agreement of OptSolute with OptAll in the referenced works
indicates that one does not necessarily require the entire system to be optimised
to obtain informative spectra. Considering the solute-solvent system as a whole,
the OptSolute methodology calculates the spectra of unoptimised systems; it is
6.5. CONFORMER OPTIMISATION

only a subsystem (the solute) that is optimised. This then begs the question of the value of energetic minima for these calculations, and whether strict geometry optimisation is an unnecessary burden.

The OptSolute model is appealing owing to the reduced computational complexity of the resultant optimisation. However, its use raises three pertinent questions; (1) How should the individual conformer spectra be Boltzmann weighted; (2) If the OptSolute scheme is valid, can we simplify the calculations further by adopting an even less rigorous optimisation criterion, and; (3) Are alternative approximate optimisation schemes viable? We deal, and elaborate upon, each of these questions in turn in the following sections. For our test case, we take microsolvated zwitterionic histidine with five explicit water molecules. The conformers have been taken from the snapshots obtained with the methodology outlined in Section 6.4.

6.5.1 Boltzmann Weighting

If the solute is optimised in the field of the unoptimised solvent, we query whether it is proper to use the energy of the entire system for Boltzmann weighting the conformation. We make the (reasonable) assumption that the dominant spectral features arise from the solute, particularly in the higher wavenumber regions, where librational modes do not feature. Consider the effects of optimising the solute while freezing the solvent: the solute will adopt an energetically preferable conformation, while the solvent remains unoptimised. The Boltzmann weight of the system could be relatively low owing to a particularly unfavourable solvent conformation. However, if the solute is in a comparatively low energy conformation, the low Boltzmann weight is not a proper reflection of the contribution of the conformer’s spectrum to the overall spectrum.
In light of this argument, we propose two means for Boltzmann weighting a the spectrum of a given conformation. The first takes the energy of the entire solute-solvent system (“system-weighting”), while the second takes the energy of just the optimised solute, without the solvent (“solute-weighting”). We evaluate the effects of the two weighting schemes on the conformers of microsolvated histidine. The Boltzmann weights of the conformers from the two schemes are compared with one another in Table 6.1.

<table>
<thead>
<tr>
<th>Conformer Number</th>
<th>System Weight (%)</th>
<th>Solute Weight (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31.6</td>
<td>14.7</td>
</tr>
<tr>
<td>2</td>
<td>22.3</td>
<td>21.6</td>
</tr>
<tr>
<td>11</td>
<td>6.7</td>
<td>16.3</td>
</tr>
<tr>
<td>25</td>
<td>20.1</td>
<td>20.6</td>
</tr>
<tr>
<td>27</td>
<td>18.0</td>
<td>23.6</td>
</tr>
<tr>
<td>35</td>
<td>1.3</td>
<td>3.2</td>
</tr>
</tbody>
</table>

Table 6.1: Conformer Boltzmann weights exceeding 1% for the 35 microsolvated histidine systems using the system-weighting and solute-weighting schemes.

From Table 6.1, we see that for both Boltzmann weighting schemes, the same six conformers dominate the resultant spectrum, albeit in different proportions. The first conformer is dominant with system-weighting, while the Boltzmann weight of the first conformer is roughly three times lower with solute-weighting. We can conclude that, in this case (and we have no reason to suspect that the result would not be general), both weighting schemes yield the same dominating conformers, but their contributions to the resulting spectrum differ rather significantly. To evaluate whether the resultant spectra are affected by the differing Boltzmann weights, we present the Raman and ROA spectra generated from the two schemes, in addition to the experimental spectrum, in Figures 6.5 and 6.6, respectively.
Figure 6.5: Raman spectra of the different Boltzmann weighting schemes. The top pane shows the Raman spectrum resulting from the system-weighting and the middle pane from the solute-weighting. The bottom pane is the experimental Raman spectrum of zwitterionic histidine.
Figure 6.6: ROA spectra of the different Boltzmann weighting schemes. The top pane shows the ROA spectrum resulting from system-weighting and the middle pane from the solute-weighting. The bottom pane is the experimental ROA spectrum of zwitterionic histidine.
From the Raman spectra in 6.5, there appear to be no discernible difference between the two weighting schemes. A number of subtle exceptions exist, but certainly none which render one spectrum superior to the other when compared with the experimental spectrum. From the ROA spectra in 6.6, we see that the two Boltzmann weighting schemes yield spectra that are again quite similar to the experimental spectrum. This similarity suggests that the Boltzmann weighting scheme does not massively impact on the spectra. However, there are a number of slight differences between the two. The 1350-1450 cm\(^{-1}\) region is the dominant region of the experimental spectrum, with a well-defined +ve/-ve/+ve profile. We see that the first positive region is recovered by both of our calculated spectra, but is arguably better modelled with the solute-weighting.

Indeed, it is interesting to note that this first peak is slightly blue-shifted relative to the experimental spectrum for the system-weighting. With the solute-weighting, this peak coincides quite well with the experimental spectrum. The subsequent -ve/+ve profile is present and well-recovered by both, but again the final positive peak is better recovered with the solute-weighting. The triplet of peaks in the 900-1100 cm\(^{-1}\) region of the experimental spectrum is also a well-defined feature, reproduced by both of our computed spectra. However, it is difficult to characterise which of the computed spectra models this region better.

From the results we have given, we have concluded that the difference between the two Boltzmann weighting schemes is small. In the absence of a definitive result, we have chosen to use the solute-weighting scheme owing to the small improvements we have alluded to in the 1350-1450 cm\(^{-1}\) region. We also believe that considering only solute energies makes intuitive sense in the context of the optimisation schemes introduced in Section 6.5.3.
CHAPTER 6. RAMAN OPTICAL ACTIVITY

6.5.2 Effects of Level of Optimisation

We now assess whether we can circumvent the need to perform stringent geometry optimisations on our system. We have used three levels of optimisation for the comparison; no optimisation (conformers taken straight from the MD trajectory), the “Loose” (Maximum force ≤ 2.5 × 10^{-3} au ; Maximum displacement ≤ 1.0 × 10^{-2} Å), and “Regular” (Maximum force ≤ 4.5 × 10^{-4} au ; Maximum displacement ≤ 1.8 × 10^{-3} Å) optimisation schemes available in the gaussian package. The resultant spectra are given in Figures 6.7 and 6.8. A fourth level of optimisation, “Tight” (Maximum force ≤ 1.5 × 10^{-5} au ; Maximum displacement ≤ 6.0 × 10^{-5} Å), was also attempted, but the system proved difficult to converge, and so we have not pursued this case.

The spectra obtained from the completely unoptimised snapshots differ significantly from the other spectra formed from optimised geometries. This poor modelling is indicative of the fact that the MD snapshots are very far from energetic minima, and so are not representative of the conformational ensemble of the system. The ultimate arbiter is the comparison of the calculated spectrum with the experimental spectrum, and we see that agreement is poor between the unoptimised and experimental spectra. There appear to be no features of the experimental spectrum that are reproduced by the unoptimised spectrum.

Turning our attentions to the optimised spectra, we see that the Raman spectra are very similar to one another, and reproduce a number of the features present in the experimental spectrum. The series of strong peaks in the 1200-1600 cm\(^{-1}\) region are largely accounted for in our computed spectra. However, we are unable to characterise one level of optimisation as superior to the other from the Raman spectra.
Figure 6.7: Raman spectra of zwitterionic histidine computed from a number of levels of geometrical optimisation. The top pane contains the Raman spectrum computed directly from unoptimised MD snapshots, while the middle two contain Raman spectra computed from conformers which have been optimised with the “Loose” and “Regular” convergence criteria. The bottom panel contains the experimental Raman spectrum of histidine.
Figure 6.8: ROA spectra of zwitterionic histidine computed from a number of levels of geometrical optimisation. The top pane contains the ROA spectrum computed directly from unoptimised MD snapshots, while the middle two contain Raman spectra computed from conformers which have been optimised with the “Loose” and “Regular” convergence criteria. The bottom panel contains the experimental ROA spectrum of histidine.
The ROA spectra do, however, differ to some degree. The spectral features in the 800-1000 cm$^{-1}$ window appear to be poorly characterised by the regular optimisation spectrum, but are relatively strong in the loose optimisation spectrum. It could be that the amplitude of the peak at roughly 1100 cm$^{-1}$ dwarfs these features in the regular optimisation, since it appears to be stronger than in the loose optimisation. We note that this peak is not particularly strong in the experimental spectrum. Assessing the 1350-1450 cm$^{-1}$ region, both the loose and regular optimisation schemes reproduce the +ve/-ve/+ve signal, with there being no notable difference in their qualities relative to the experimental spectrum.

Based on these findings, we have deemed the loose optimisation to be sufficient in reproducing the spectral features of the experimental spectrum. The regular optimisation offers little improvement, if at all, in the calculated spectra relative to the loose optimisation. Considering the convergence level is roughly an order of magnitude more stringent for the regular optimisation, it seems to be an unnecessary additional effort to optimise the systems so thoroughly. Indeed, the loose optimisation requires roughly half the amount of time to compute relative to the regular optimisation for this system. Ensuring the system is at least somewhat close to an energetic minimum appears to be sufficient for computing spectra. However, we reiterate that if the system is far from an energetic minimum, as we understand the unoptimised MD snapshots to be, one obtains extremely poor computed spectra.

### 6.5.3 Alternative Optimisation Schemes

Given that the optimisation of a subsystem leads to good quality spectra, the final point we wish to address is whether alternative subsystem optimisation schemes
yield equally informative spectra. The OptAll scheme has been found to produce the highest quality spectra, and so this is the case we use as a benchmark. However, we can also formulate several alternative subsystem optimisation schemes;

(a) OptSolute

The solute is optimised in the field of the unoptimised solvent. We have already alluded to this approach and its adoption in previous work, yielding spectra that are comparable to OptAll.

(b) OptSolvent

The converse approach to OptSolute, where the solvent is optimised in the field of the unoptimised solute. This approach is potentially all the more appealing, since the optimisation takes place in the low layer of the QM/MM, and is therefore faster.

(c) OptSolvent→OptSolute

The OptSolvent methodology is invoked, followed by OptSolute. By optimising the layers separately, we hope that we can recreate the optimisation resulting from OptAll.

(d) OptSolute→OptSolvent

The OptSolute methodology is invoked, followed by OptSolvent. We do not imagine this to differ from the previous approach, but is included for the sake of completeness.

The first two methods, (a) and (b), are concerned with obtaining crude approximations for a large decrease in computational cost. We expect these methods to yield spectra of far poorer quality than OptAll, but to be far quicker to compute. Contrastingly, the last two methods, (c) and (d), look to replicate the quality of
6.5. **CONFORMER OPTIMISATION**

spectrum obtained by the OptAll scheme, while saving some level of computational time in the process. In the following, we will refer to each optimisation scheme by the letter it has been listed with in the above enumeration. When referring to OptAll, we will use (e).

From the Raman spectra of Figure 6.9, we immediately see that (b) performs particularly poorly. All three Amide I-III regions are poorly represented, and the intensity of the low wavenumber regions is not in agreement with the experimental spectrum. However, the other four optimisation schemes yield spectra that are in good agreement with the experimental spectrum. The Amide I-III regions are well-recovered, while the low wavenumber regions are modelled equally well. If we are to be fastidious, we may question the Amide I regions of (a) and (d), where the signal does not possess the same intensity as that in the experimental spectrum. The best agreement with the experimental spectrum is arguably (c), which appears to recover the vast majority of spectral features featured in (e).

Turning our attention to the ROA spectra of Figure 6.10, we notice that the poor performance of (b) is continued, and virtually no experimental features are recovered. We also draw attention to the poor performance of (d) in the low wavenumber regions. In turn, this poor modelling of the low wavenumber regions obscures the finer details of the high wavenumber regions, and the spectrum as a whole deteriorates. Similarly, (a) appears to be a poor approximation to the experimental spectrum; the coarse details of the spectrum appear to be present, but the high wavenumber regions are poorly modelled. In contrast, (c) is in excellent agreement with both the spectra obtained through experiment and (e). With the exception of a few features in the low wavenumber regions, (c) and (e) are almost identical.

Offering some explanation for the results we have obtained, the optimisation of the solute appears to be the key factor in guaranteeing the quality of computed spectra.
Figure 6.9: Raman spectra resulting from the optimisation methodologies we have outlined. Enumeration corresponds to the labels we have used in the main text.
Figure 6.10: ROA spectra resulting from the optimisation methodologies we have outlined. Enumeration corresponds to the labels we have used in the main text.
CHAPTER 6. RAMAN OPTICAL ACTIVITY

The poor results from (b) and (d) presumably derive from the optimisation of the solvent being the final optimisation prior to spectrum calculation, and so the solute is in an unoptimised state. To recreate the finer details of the spectrum, it also appears as if the solvent requires some level of optimisation, particularly with the ROA measurements. Hence why (c) significantly outperforms (a). The Raman spectra seem to be robust with regards to the level of solvent optimisation.

In conclusion, we feel that (c) and (e) offer the best recovery of experimental spectral features. However, we have found that the computational cost associated with (c) more than fourfold less on average relative to (e). Relative to the computational cost associated with (a), (c) is roughly 50% more expensive computationally. Therefore, it seems as if the OptSolvent → OptSolute methodology strikes an ideal balance between accuracy and compute time. We proceed with this methodology through the remainder of our work.

6.6 Microsolvation

We define microsolvation to be the explicit solvation of a system using a small number of water molecules, typically not extending beyond the second solvation shell of the solute. The major question we wish to answer is whether computed spectra from microsolvated systems are able to model experimental spectra, without having to resort to performing calculations on large explicitly solvated systems. If we can answer this query positively, then it facilitates the computation of solvated Raman and ROA spectra\(^7\).

To summarise those methodological choices we have made from the results in

\(^7\)In fact, it is reasonable to generalise these results to the majority of vibrational spectroscopies.
Section 6.5, we have used the “Loose” optimisation setting to optimise all geometries, subject to the OPTSOLVENT $\rightarrow$ OPTSOLUTE optimisation methodology. All spectra are reconstituted based on the Boltzmann weights of the solute molecule, independent of the energetic contribution of the solvent.

### 6.6.1 Neutral Histidine

For the Raman spectra in Figure 6.11, we find that the Amide I band at around 1600 cm$^{-1}$ is well recovered by each of the microsolvated systems. Since the Amide I band is largely indicative of peptide C=O stretching modes, it is initially surprising that its modelling appears to be independent of the level of microsolvation, where solvent interactions presumably damp the stretching mode. However, upon closer inspection of the microsolvated conformers used for the spectra, we find that the peptide C=O is solvated in each conformer, rendering its modelling independent of the degree of microsolvation. Similarly, we recover the Amide II triplet at $\sim$1450-1550 cm$^{-1}$, albeit blue-shifted relative to the experimental spectrum. The triplet signature appears to be most profound with a solvation shell comprising 15 water molecules, where the central peak is dominant. This result is to be expected since a solvation shell of 15 water molecules appears to coincide with the number of water molecules comprising the first solvation shell for zwitterionic histidine. Since the Amide II region is representative of peptide N-H bending and C-H stretching modes, we expect this region to be highly dependent on solvent interactions. It is, nevertheless, surprising that the quality of our computed spectra deteriorates with 20 water molecules.

The Amide III region is slightly less well-modelled by our microsolvated systems. In the experimental spectrum, we observe a well-defined quadruplet of peaks, span-
Figure 6.11: Raman spectra of microsolvated histidines. From the top pane to the bottom pane, we show histidine microsolvated by 5, 10, 15 and 20 explicit water molecules. The final pane contains the experimental histidine Raman spectrum.
6.6. MICROSOLVATION

ning 1200-1350 cm⁻¹. This region is dominated by C-N stretching and N-H bending modes, and we would expect this region to be as solvation-dependent as the other two regions we have discussed. We note that we do not manage to definitively recover the clear quadruplet seen in the experimental spectrum, although the profile does seem to become more enhanced as we increase the degree of microsolvation. Indeed, when 20 water molecules are included, we are able to discern four clear peaks in the Amide III region.

Regarding the impact of the tautomeric forms of histidine, Ashikawa and Itoh[418] have found that the breathing motions for His⁰[NπH] and His⁰[NτH] can be related to Raman peaks at 1304/1260 cm⁻¹, respectively. In the same work, two additional marker regions characterising imidazole stretching modes between the tautomeric forms of histidine were proposed; 1568/1585 cm⁻¹ and 1090/1105 cm⁻¹. Later work by Toyama et al.[419] has suggested an additional marker region at 1320/1354 cm⁻¹.

Unfortunately, the majority of these marker regions inhabit the strong amide regions of the Raman spectrum, making them difficult to discern. However, the 1090/1105 cm⁻¹ region occupies a low intensity region of the Raman spectrum, and appears to be characterised in both our experimental and computed spectra, at each microsolvation level excluding the spectrum with 5 waters.
Figure 6.12: ROA spectra of microsolvated histidines. From the top pane to the bottom pane, we show histidine microsolvated by 5, 10, 15 and 20 explicit water molecules. The final pane contains the experimental histidine ROA spectrum.
The experimental ROA spectrum is not quite as well-recovered as the microsolvated spectra, as depicted in Figure 6.12. This is not to say that the ROA spectra are of poor quality, but simply suffer by comparison with the high quality of the Raman spectra. The dominant feature across both experimental and calculated spectra is the $+ve/-ve/+ve$ signal at 1300-1500 cm$^{-1}$, and is well-recovered at all levels of microsolvation. However, it is worth pointing out that the calculated spectra appear to be blue-shifted by $\sim 50$ cm$^{-1}$ relative to the experimental spectrum. This blue-shifting is somewhat surprising. Note that we have not undertaken abscissa-scaling by 0.96, as prescribed by Radom[383]. Doing so would blue-shift the spectrum further, and so would result in the deterioration of our calculated spectra.

We note that as the level of microsolvation increases, the peak at 1050 cm$^{-1}$ is amplified. A number of peaks exist in this region of the experimental spectrum, but none are as prominent as that in the microsolvated spectrum. Assessing the normal modes of motion around this region, we find that the majority are dominated by sidechain dihedral torsional degrees of freedom. This observation then suggests that the sidechain torsional degrees of freedom are more flexible in the microsolvated systems than in the experiment. Indeed, it is predominantly the zwitterionic groups and imidazole ring that form interactions with the solvent molecules in the microsolvated systems, and so it is to be expected that the alkyl sidechain be poorly solvated, and so not well recovered by the microsolvated spectra.

### 6.6.2 Protonated Histidine

The literature regarding the calculation of Raman and ROA spectra of formally charged species is sparse. Indeed, the ab initio modelling of formally charged
species is in itself difficult, since the basis sets require both diffuse functions and some description of polarisation, the former being of particular importance for anions. To assess the quality of Raman and ROA spectra of a formally charged species, we have selected cationic histidine. The fact that no tautomers exist for this species makes its modelling simpler.

A formally charged species is presumed to undergo stronger interactions with solvent than the neutral species, and so we hypothesise that the degree of microsolvation will significantly influence the quality of the computed spectrum. Since the formal charge results from the protonation of the imidazole ring, we anticipate that the higher wavenumber regions will be poorly reproduced at the low levels of microsolvation since the explicit water molecules tend to primarily interact with the zwitterionic groups.

The computed Raman spectra of Figure 6.13 are in relatively good agreement with the experimental Raman spectrum. The strong peaks at \( \sim 1200 \text{ cm}^{-1} \) and \( \sim 1300 \text{ cm}^{-1} \) become increasingly prominent as the degree of microsolvation increases. The peak at \( \sim 1500 \text{ cm}^{-1} \) is well-recovered by each of the microsolvated spectra. The resolution of the doublet of peaks at roughly 800 cm\(^{-1}\) is similarly improved as the degree of microsolvation increases. Interestingly, the general agreement with the experimental spectrum deteriorates when 10 explicit water molecules are used for microsolvation. However, those spectra including 5, 15 and 20 explicit water molecules are very similar.

We offer a potential reason for the deterioration in spectrum quality when 10 explicit water molecules are included. When 5 explicit water molecules are included, the conformers are largely solvated around the backbone amide and carboxyl groups of the zwitterionic histidine. When 15 and 20 explicit water molecules are included, the conformers are completely solvated, i.e. both the backbone and
Figure 6.13: Raman spectra of microsolvated cationic histidine. From the top pane to the bottom pane, we show cationic histidine microsolvated by 5, 10, 15 and 20 explicit water molecules. The final pane contains the experimental histidine Raman spectrum.
Figure 6.14: ROA spectra of microsolvated cationic histidine. From the top pane to the bottom pane, we show cationic histidine microsolvated by 5, 10, 15 and 20 explicit water molecules. The final pane contains the experimental histidine ROA spectrum.
sidechain groups of the zwitterionic histidine. However, when 10 explicit water molecules are included, only a couple of water molecules solvate the imidazole ring of the zwitterionic histidine. As such, we suggest that the partial solvation of the imidazole results in a biasing of certain vibrational frequencies originating from the imidazole-solvent interactions. The spectrum as a whole deteriorates since these imidazole-solvent frequencies then dominate the resultant spectrum. In summary, it is tempting to think of partial solvation as detrimental to the correct prediction of the vibrational spectra.

The computed ROA spectra in Figure 6.14 are, however, not quite as successful in reproducing the major features of the experimental spectrum. The strong negative band at $\sim 1400 \text{ cm}^{-1}$ is recovered in each of the computed spectra, and the positive doublet in the 1300-1350 cm$^{-1}$ region appears to become clearer as the level of microsolvation increases.

From previous, unpublished work on anionic adenosine triphosphate, we have found that the calculation of ROA spectra for formally charged species is rather difficult. One speculative reason for the poorer quality of the ROA spectra relative to the Raman spectra is the way one computes the intensity of spectral bands for the two spectroscopies. For Raman spectra, the intensity is given by the sum of the intensities of the right- and left-circularly polarised scattered light. For ROA spectra, the intensity is given by the difference in the intensities of the right- and left-circularly polarised scattered light. Therefore, ROA band intensities are more prone to error than the Raman band intensities. In other words, the ROA band intensities are smaller in magnitude than the Raman band intensities, and so the errors are amplified. It is this sensitivity to error that makes ROA a “gold-standard” spectroscopic technique, whereas Raman is somewhat more forgiving.
6.7 Conclusion

We have investigated and presented the results for a number of novel methodologies that can be used in the calculation of \textit{ab initio} Raman and ROA spectra of solvated systems. We are suggestively vague in our use of the term “solvated systems”, since we see no reason why our methodology cannot be applied to systems outside of the domain of peptides, such as DNA bases\cite{420, 421}. A novel conformational sampling methodology allows for the unambiguous extraction of a set of mutually diverse conformers from an MD trajectory of arbitrary length. We have shown that conformers do not require strict optimisation to obtain high quality spectra, and that a two-step subsystem optimisation (i.e. optimising the solvent while keeping the solute frozen, and subsequently optimising the solute while keeping the solvent frozen) yields spectra of the same quality as entire system optimisations. The microsolvation of both formally charged and neutral zwitterionic histidine species can yield calculated spectra that converge towards the experimental spectra.
Chapter 7

Conclusion and Future Work

7.1 General Conclusions

We have presented a number of developments aimed at improving the construction and implementation of the QCTFF. The work in this thesis has focused primarily on conformational sampling, and how it can be used to construct optimal training sets. We have also demonstrated the applicability of the QCTFF to carbohydrates for the first time. Finally, we have undertaken a rigorous study on computing Raman and ROA spectra for charged, microsolvated species, as well as introduced various novel techniques for improving both the computational speed and accuracy of the ROA calculations.

In Chapter 3, we presented a conformational sampling methodology that is both computationally tractable, and in principle, more representative of the sampling one may obtain from the \textit{ab initio} PES than one constructed by a classical force field. The TycHE methodology is ideal for the construction of conformers with
which a machine learning model can be constructed. It naturally implements a 
form of importance sampling in the form of transitioning between seeding geomet-
tries based on Boltzmann weights. It also provides both large-scale conformational 
sampling, in addition to finer sampling of the higher frequency degrees of freedom. 
These properties make it an ideal conformational sampling tool for the construc-
tion of machine learning models.

It may perhaps be feasible to approximately update the Hessian of the system with 
respect to a distortion. Hessian updating methods are frequently invoked during 
*ab initio* geometry optimisations[291], and for small displacements are relatively 
accurate. By use of these Hessian-updating methodologies, one could re-evaluate 
the normal modes of the system once it has been displaced from the seeding geo-
metry, and subsequently vibrate along these new normal modes. However, it is 
anticipated that such a methodology would significantly increase the computa-
tional power required for sampling.

In Chapter 4, we demonstrated that the atomic multipole moments of a set of 
carbohydrates are amenable to the machine learning technique kriging. Whilst 
this has been done in the past for a variety of chemical species including naturally 
occurring amino acids, this is the first foray into the field of glycobiology.

Kriging is able to capture the conformational dependence of the multipole moments 
and make predictions, such that the error in the electrostatic energy relative to that 
derived from *ab initio* data is encouraging, given the popular aim is to obtain errors 
below 4 kJ mol\(^{-1}\). Indeed, the presented methodology is immediately extensible to 
any term arising in an energetic decomposition of a system. If some quantity is 
conformationally dependent, then the dependence can be modelled by kriging.

As such, an entire force field can be parameterized by the current methodology,
reproducing \textit{ab initio} quantities for use in classical MD. This route is preferable to the computationally intensive approach of \textit{ab initio} MD.

Chapter 5 presented a number of methodological developments for the selection of an optimal training set for kriging. The greedy subset selection algorithms have been reformulated to use a novel selection function. This novel selection function results in improved accuracy of the kriging model, relative to one constructed by a random subset selection. However, work is still required to optimise this methodology and yield “ideal” training sets.

A novel subset selection algorithm, the Iterative Voronoi algorithm, has been presented. However, this methodology is only viable for low-dimensional cases, and so cannot be implemented for our purposes. In spite of this, if it were possible to reduce the dimensionality of the systems in some way, it may be feasible to implement the Iterative Voronoi algorithm in the future and establish whether it is a successful subset selection methodology for larger systems.

We have presented a novel means for defining the Domain of Applicability of a machine learning model. By approximating the “restricted affine hull” from a set of points by use of a greedy algorithm to construct the largest $d$-polytope on the points, it has been shown that one can approximately determine whether a point lies within the training set of a machine learning model in low-dimensional spaces. Further work is, however, required, to make the methodology more successful in higher dimensional spaces. The approximation of the DoA of a kriging model is of utility in our work since it allows one to ascertain whether a kriging model is predictive in those relevant regions of conformational space.

Finally in Chapter 6, we investigated and presented the results for a number of novel methodologies that can be used in the calculation of \textit{ab initio} Raman and
ROA spectra of solvated systems. We are suggestively vague in our use of the term “solvated systems”, since we see no reason why our methodology cannot be applied to systems outside of the domain of peptides, such as DNA bases[420, 421]. A novel conformational sampling methodology allows for the unambiguous extraction of a set of mutually diverse conformers from an MD trajectory of arbitrary length. We have shown that conformers do not require strict optimisation to obtain high quality spectra, and that a two-step subsystem optimisation (i.e. optimising the solvent while keeping the solute frozen, and subsequently optimising the solute while keeping the solvent frozen) yields spectra of the same quality as entire system optimisations. The microsolvation of both formally charged and neutral zwitterionic histidine species can yield calculated spectra that converge towards the experimental spectra.

7.2 Future Work

It is anticipated that the QCTFF will be ready for commercial release in the near future. Indeed, soon to be published data shows that geometry optimisation of a number of non-trivial molecules with QCTFF results in optimisation to energetic minima that are within 1 kJ mol\(^{-1}\) of the \textit{ab initio} minima. A local version of \textsc{dl\_poly} 4.0 has been modified to read and use kriging models over the course of a MD simulation, so there is no obstacle to prediction of dynamical quantities and other quantities associated with dynamical trajectories.

In spite of the success of geometry optimisation, a great deal of work will need to be conducted in using the kriging models as efficiently as possible. We have spoken in some detail about the memory requirements associated with a force field comprising kriging models in Appendix B. However, actually \textit{invoking} the kriging
models over the course of a MD simulation is computationally-intensive relative to a conventional classical force field, where one can access and manipulate a database of force field parameters in high-speed memory. Therefore, some strategy must be devised to handle both the transferral of kriging models to high-speed memory, and subsequently make predictions from the kriging models. Considering the variety of high-performance computing (HPC) architectures, a number of solutions are available to efficiently use the kriging models. For instance, one could take advantage of the massively parallel architectures of HPC machines[422]. Future work could also look towards employing more advanced hardware, such as graphical processing units (GPUs)[423], field-programmable gate arrays (FPGAs)[424] or application-specific integrated circuits (ASICs)[425], which can vastly outperform standard HPC systems by use of, for instance, enhanced pipelining capabilities.

It is worth noting that a number of varieties of kriging exist. The version that we have embellished upon in Section 2.4.2 is referred to as “Ordinary Kriging”[426], and is the simplest implementation available. However, a number of alternatives do exist. For instance, “Noisy Kriging”[427] is used when there is a known level of uncertainty in the data points used to construct the training set. The kriging model subsequently accounts for the uncertainty in the data points by not restricting the kriging model to pass directly through training points. A number of alternatives do exist, such as “Nugget Kriging”[428] and “Universal Kriging”[429].

A number of conceptual problems remain and require resolution. The first of these has been alluded to throughout this work; we require some means of defining an atom type so that the kriging models can be used for more than a single atom in a system. Work has previously been undertaken within our group to find some numerical criterion for assessing the transferability of a kriging model[430]. It was thought that if the kriging hyperparameters were similar between two kriging
models, the models would be transferable. However, this line of analysis has proven complex, and is the subject of continual research.

A particularly desirable feature of *ab initio* MD is that the associated dynamics are reactive. Since the electrons are explicitly modelled, chemical bonds can be broken and formed, and the molecular graph can be altered over the course of the simulation. Whilst there exists a small literature on reactive classical force fields[431], if we are to make the QCTFF a valid competitor to the *ab initio* MD schemes, some level of reactivity must be permitted. One could perhaps do so by constructing kriging models for each of the distinct molecular graphs of a system along the reaction coordinate. However, when transition states are unknown, constructing kriging models for these species is not feasible. As an alternative, it may be possible to construct kriging models for only the reactants and products of a particular reaction scheme, and smoothly interpolating “between” the kriging models over the course of the reaction to obtain approximations to atomic quantities during the intermediary stages of the reaction. We have not investigated such a methodology, but some work is certainly warranted.

In keeping with this, we anticipate that the modelling of open shell and excited state systems could perhaps be a stumbling block to the QCTFF. Popular IQA software suites do not offer compatibility with post-Hartree Fock or multireference wavefunctions. Obtaining valid energetic quantities for atoms in these states could be problematic, but this is speculation on the author’s part. The IQA methodology is being extended so that a number of other wavefunction-based methods can be handled[432], so perhaps no problems will be encountered when generalising further. It is also worth noting that IQA decompositions can be undertaken for the likes of transition metals and heavy elements[433, 434], but the effects of including pseudopotentials in wavefunction construction has unclear effects on the resultant
IQA calculations, and is not particularly well-documented.

There is nothing that restricts the training of the QCTFF to biomolecular systems. So long as the software exists for treating the molecular system with QCT and IQA, then kriging models can be constructed and subsequently used in MD. Therefore, there are technically no barriers to the construction of a generally applicable force field. It is hoped that the force field will prove to be of a great deal of use to the computational chemistry community in the near future.
Bibliography


[66] Celia Fonseca Guerra, Jan-Willem Handgraaf, Evert Jan Baerends, and F Matthias Bickelhaupt. Voronoi deformation density (vdd) charges: As-


[82] Richard FW Bader and PM Beddall. Virial field relationship for molecular


[212] Timothy L Fletcher, Shaun M Kandathil, and Paul LA Popelier. The prediction of atomic kinetic energies from coordinates of surrounding atoms


[290] Frank M DiCapua, S Swaminathan, and David L Beveridge. Theoretical evidence for water insertion in helix bending: molecular dynamics of


[299] James R Cheeseman, Majeed S Shaik, Paul LA Popelier, and Ewan W Blanch. Calculation of raman optical activity spectra of methyl-β-d-glucose


[346] LD Barron, L Hecht, EW Blanch, and AF Bell. Solution structure and


[371] Emadeddin Tajkhorshid, KJ Jalkanen, and Sandor Suhai. Structure and vibrational spectra of the zwitterion l-alanine in the presence of explicit


[384] Shigeki Yamamoto, Hitoshi Watarai, and Petr Bouř. Monitoring


[441] Yan Zhao and Donald G Truhlar. The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals. Theoretical Chemistry Accounts, 120(1-3):215–241, 2008.
Appendix A

Density Functional Theory

A.1 Thomas-Fermi Model

The vast majority of (time-independent) \textit{ab initio} methods use the molecular wavefunction of an $N$ electron system, $\Psi(r_1, ..., r_N)$ as a central quality. When one has an appropriate wavefunction, expectation values of operator quantities associated with physical observables can be evaluated. As opposed to working with the wavefunction of a system, the Thomas-Fermi model\cite{435} uses the electron density, $\rho(r)$, as a central quantity. The Thomas-Fermi energy of a system, $E_{TF}[\rho(r)]$, is a functional of the electron density, and can be written

$$E_{TF}[\rho(r)] = C_F \int dr \rho(r)^{5/3} + \int dr \rho(r) v(r) + \frac{1}{2} \int dr \int dr' \frac{\rho(r)\rho(r')}{||r - r'||}.$$  \hspace{1cm} (A.1.1)

The first term of this quantity corresponds to the kinetic energy of a homogeneous electron gas, the second term is the Coulombic attraction between electrons and nuclei and the third term is the Coulombic electron-electron repulsion. $v(r)$, the
“external potential”, has analytical form

\[ v(r) = -\sum_{I=1}^{N_N} \frac{Z_I}{|r - \mathbf{R}_I|} , \]  

where \( \mathbf{R}_I \) is the position vector of the \( I^{th} \) nucleus and \( Z_I \) its atomic number. Note that we have treated the system classically, and so have no way to account for the quantum mechanical nature of the electrons, e.g. fermion exchange.

We seek the electron density which minimises the Thomas-Fermi energy given in (A.1.1), subject to the constraint that the number of electrons remains constant, i.e. \( \int d\mathbf{r} \rho(\mathbf{r}) = N_e \). The minimisation of a quantity subject to a constraint is readily achieved by use of Lagrange multipliers, \( \lambda \),

\[ \frac{\delta E_{TF}[\rho(\mathbf{r})]}{\delta \rho(\mathbf{r})} - \lambda \left[ \frac{\delta}{\delta \rho(\mathbf{r})} \int d\mathbf{r} \rho(\mathbf{r}) - N_e \right] = 0 . \]  

Taking the functional derivative of each term in the Thomas-Fermi energy with respect to the electron density is straightforward, and combining with (A.1.3) yields

\[ \frac{5}{3} C_F \rho(r)^{2/3} + v(r) + \int d\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} - \lambda = 0 , \]  

and is readily solved to find the electron density of the minimum energy state of the system. Unfortunately, the Thomas-Fermi formalism is a poor representation of a quantum mechanical system, owing to the classical description we have invoked. In addition, the approximation of the kinetic energy as in (A.1.1) is particularly crude. As a result, the Thomas-Fermi model fails to predict even qualitatively correct results\cite{436}. For instance, the Thomas-Fermi model is unable to predict atomic bonding, and subsequently the majority of chemistry. Xα-theory\cite{437} was a later correction to the Thomas-Fermi model, but was also largely of little use.
A.2. HOHENBERG-KOHN THEOREMS

A.2 Hohenberg-Kohn Theorems

A.2.1 First Theorem

Given a normalised wavefunction associated with a given external potential, $\psi_v$, which satisfies the Schrödinger equation

$$H_v \psi_v = E_v \psi_v,$$

(A.2.1)

where the Hamiltonian operator can be further decomposed into a sum of kinetic and potential energy operators, as well as the external potential, $H = T + U + v(r)$. We assume there are two potentials, $v_a$ and $v_b$, with an identical wavefunction (to within some phase factor, $e^{i\phi}$), such that

$$H_a \psi_a = E_a \psi_a$$
$$H_b \psi_b = E_b \psi_b$$

(A.2.2)

Multiplying through the second equation by $e^{i\phi}$, and solving the equations simultaneously,

$$\left( H_a - H_b \right) \psi_a = \left( v_a(r) - v_b(r) \right) \psi_a = \left( E_a - E_b \right) \psi_a,$$

(A.2.3)

leaving us with the relation

$$v_a(r) - v_b(r) = E_a - E_b = c,$$

(A.2.4)

where $c$ is a constant. The above implies that wavefunctions differing by a phase factor have potentials that differ by a constant term,

$$\psi_a = e^{i\phi} \psi_b \iff v_a(r) = v_b(r) + c,$$

(A.2.5)

and the reverse holds (potentials differing by a constant term have wavefunctions differing only by a phase factor), hence the relationship is bijective.
Now, consider the two wavefunctions are associated with different potentials, $\psi_a$ and $\psi_b$. By use of the variational principle,

$$E_a < \langle \psi_b | H_a | \psi_b \rangle = \langle \psi_b | T + U + v_a(r) | \psi_b \rangle = E + \int dr \rho_b(r)v_a(r)$$

$$< \langle \psi_b | T + U | \psi_b \rangle + \int dr \rho_b(r)v_b(r) - \int dr \rho_b(r)v_b(r) + \int dr \rho_b(r)v_a(r)$$

$$< E_b + \int dr \left[ v_a(r) - v_b(r) \right] \rho_b(r)$$

$$E_b < E_a + \int dr \left[ v_b(r) - v_a(r) \right] \rho_a(r).$$

Adding the two inequalities, we arrive at

$$E_a + E_b < E_a + E_b + \int dr \left[ v_a(r) - v_b(r) \right] \rho_b(r) + \int dr \left[ v_b(r) - v_a(r) \right] \rho_a(r),$$

which, if $\rho_a(r) = \rho_b(r)$, leads to

$$E_a + E_b < E_a + E_b,$$

which is obviously absurd. As such, to satisfy the above, we require that two distinct potentials, up to an additive constant, lead to two distinct densities, i.e.

$$\rho_a(r) \neq \rho_b(r) \iff v_a(r) \neq v_b(r) + c,$$

which is the Hohenberg-Kohn first theorem. As an aside, Runge and Gross[438] provided the rather extraordinary result that for the time-dependent case, the above only requires a slight modification to

$$\rho_a(r,t) \neq \rho_b(r,t) \iff v_a(r,t) \neq v_b(r,t) + c(t).$$

### A.2.2 Second Theorem

The second Hohenberg-Kohn theorem specifies that the ground state energy of a system can be obtained through the variation of a trial density[439]. This varia-
A.2. HOHENBERG-KOHN THEOREMS

Analogous to the conventional Rayleigh-Ritz variational principle, where the ground state energy of a system is obtained through variation of a trial wavefunction,

$$E_0 = \min_{\Psi} \langle \Psi | H | \Psi \rangle \quad \text{(A.2.12)}$$

The minimisation with respect to the trial density is accomplished by the constrained two-step search methodology of Levy and Lieb[440], where the electron density, $\rho(r)$ is kept constant in the first step. We denote a trial wavefunction, $\Psi^{(\alpha)}$, where $\{\alpha\}$ denote the set of variational parameters to optimise. We begin writing the minimisation as,

$$E_0[\rho(r)] = \min_{\{\alpha\}} \langle \Psi^{(\alpha)} | H | \Psi^{(\alpha)} \rangle$$

$$= \min_{\{\alpha\}} \langle \Psi^{(\alpha)} | T + V + v(r) | \Psi^{(\alpha)} \rangle \quad \text{(A.2.13)}$$

$$= F[\rho(r)] + \int dr v(r) \rho(r) ,$$

where the universal function, $F[\rho(r)]$ is defined as

$$F[\rho(r)] = \min_{\{\alpha\}} \langle \Psi^{(\alpha)} | T + U | \Psi^{(\alpha)} \rangle , \quad \text{(A.2.14)}$$

and optimising the universal functional with respect to $\{\alpha\}$ is the first step of the constrained optimisation. The second step of the optimisation involves the minimisation with respect to the density, such that

$$E_0[\rho(r)] = \min_{\rho(r)} \left[ F[\rho(r)] + \int dr v(r) \rho(r) \right] , \quad \text{(A.2.15)}$$

and we arrive at the Hohenberg-Kohn second theorem, which proposes the existence of a variational principle with respect to the density.
A.3 Kohn-Sham Equations

We consider the universal functional of (A.2.14), $F[\rho(r)]$, where the potential energy operator has been split into two separate components

$$F[\rho(r)] = \mathcal{T}[\rho(r)] + \mathcal{U}[\rho(r)]$$

$$= \mathcal{T}[\rho(r)] + \frac{1}{2} \int dr' \int dr \frac{\rho(r) \rho(r')}{||r - r'||} + E_{xc}[\rho(r)] \quad \text{(A.3.1)}$$

and we hope to capture the exchange-correlation energetics that are omitted in mean-field approaches (such as Hartree-Fock) in the exchange-correlation energy, $E_{xc}[\rho(r)]$. The total energy of the system is written as

$$E[\rho(r)] = \int dv(r) \rho(r) + \mathcal{T}[\rho(r)] + \frac{1}{2} \int dr' \int dr \frac{\rho(r) \rho(r')}{||r - r'||} + E_{xc}[\rho(r)]. \quad \text{(A.3.2)}$$

The forms of $\mathcal{T}[\rho(r)]$ and $E_{xc}[\rho(r)]$, the kinetic and exchange-correlation energies respectively, are in general unknown. $\mathcal{V}_{ee}[\rho(r), \rho(r')]$ is the familiar electrostatic interaction between two charge densities. Now, by use of the constraint that the number of electrons, $\int dr \rho(r) = N$ is a conserved quantity, we are able to form a Lagrangian of the total system energy as given in (A.2.15),

$$\mathcal{L} = F[\rho(r)] + \int dv(r) \rho(r) - \lambda \left( \int dr \rho(r) - N \right), \quad \text{(A.3.3)}$$

where $\lambda$ is a Lagrange multiplier. By the second Hohenberg-Kohn theorem, we are able to take the functional derivative of this quantity to optimise the electron density,

$$\delta \mathcal{L} = \delta \left[ \int dv(r) \rho(r) + F[\rho(r)] - \lambda \left( \int dr \rho(r) - N \right) \right] = 0$$

$$= v(r) + \frac{\delta \mathcal{T}[\rho(r)]}{\delta \rho(r)} + \frac{1}{2} \int dr' \frac{\rho(r) \rho(r')}{||r - r'||} + E_{xc}[\rho(r)] - \lambda = 0 \quad \text{(A.3.4)}$$

$$= \frac{\delta \mathcal{T}[\rho(r)]}{\delta \rho(r)} + V_{KS}[\rho(r)] - \lambda = 0.$$
In the above, we have introduced a number of shorthand notations. The exchange-correlation potential, $V_{xc}[\rho(\mathbf{r})]$, is defined as

$$V_{xc}[\rho(\mathbf{r})] = \frac{\delta E_{xc}[\rho(\mathbf{r})]}{\delta \rho(\mathbf{r})}, \quad (A.3.5)$$

and the Kohn-Sham potential, $V_{KS}[\rho(\mathbf{r})]$ is given by

$$V_{KS}[\rho(\mathbf{r})] = v(\mathbf{r}) + \frac{1}{2} \int d\mathbf{r}' \frac{\rho(\mathbf{r}')}{||\mathbf{r} - \mathbf{r}'||} + V_{xc}[\rho(\mathbf{r})]. \quad (A.3.6)$$

From (A.3.4), we see that the final equation resembles a system of non-interacting particles in an external potential, $V_{KS}[\rho(\mathbf{r})]$, and so the ground state electron density is given by solving a series of one-electron Schrödinger equations. Recall that the electron density is given by

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} f_{i} \psi_{i}^*(\mathbf{r}) \psi_{i}(\mathbf{r}), \quad (A.3.7)$$

where $f_{i}$ is the filling of the $i^{th}$ orbital, $\psi_{i}(\mathbf{r})$. The kinetic energy functional is then typically written as

$$\mathcal{T}[\rho(\mathbf{r})] = -\frac{1}{2} \sum_{i=1}^{N} f_{i} \int d\mathbf{r} \psi_{i}^*(\mathbf{r}) \nabla^2 \psi_{i}(\mathbf{r}). \quad (A.3.8)$$

As such, we obtain the Kohn-Sham equations

$$\left[ -\frac{1}{2} \nabla^2 + V_{KS}(\mathbf{r}) \right] \psi_{i}(\mathbf{r}) = \epsilon_{i} \psi_{i}(\mathbf{r}), \quad (A.3.9)$$

forming an eigensystem, and can be solved by a myriad of techniques.

It has been proved that the exchange-correlation energy, $E_{xc}$, has an exact analytical form. However, the exact form is unknown, and so while DFT is a formally exact theory, the necessity of approximating $E_{xc}$ renders it inexact. A number of strategies exist for approximating $E_{xc}$, which we briefly outline. By invoking
the Local Density Approximation (LDA), we approximate $E_{xc}$ as the integral of a functional of the electron density over all space,

$$E_{xc}^{LDA} = \int d\mathbf{r} f_{xc}[\rho(\mathbf{r})],$$  \hspace{1cm} (A.3.10)

where $f_{xc}$ denotes a general functional of the electron density. One can invoke more accurate, and consequently more time-consuming, functionals by the inclusion of higher order derivatives of the electron density. By invoking the so-called “Generalised Gradient Approximation” (GGA), $E_{xc}$ is written as a functional of both the electron density and its spatial derivatives

$$E_{xc}^{GGA} = E_{xc}^{GGA} [\rho(\mathbf{r}), \nabla \rho(\mathbf{r})],$$  \hspace{1cm} (A.3.11)

and can vastly improve upon the accuracy offered by the LDA. As a natural extension to the GGA functionals, the meta-GGA functionals (m-GGA) (such as the popular Minnesota functionals[441]) include second order spatial derivatives of the electron density,

$$E_{xc}^{m-GGA} = E_{xc}^{m-GGA} [\rho(\mathbf{r}), \nabla \rho(\mathbf{r}), \nabla^2 \rho(\mathbf{r})].$$  \hspace{1cm} (A.3.12)

One can also generalise each of the functionals to account for electron spins. Various combinations of the above functional types also exist, and are referred to as hybrid functionals, such as the common B3LYP functional.
Appendix B

Storage of Kriging Models

In Section 2.4.2, we saw that the prediction of the response at a point, \( y(\mathbf{x}^*) \), is given by (2.4.11). Thus, the prediction of a response requires \textit{a priori} knowledge of, \( \mathbf{R}^{-1}, \hat{\mathbf{\theta}}, \hat{\mathbf{p}}, \hat{\mu}, \hat{\sigma}^2, \mathbf{y} \) and \( \mathbf{r} \). Assuming that all of this data is stored in double precision ASCII format, for a training set size of 1000 points, where the inputs are 60–dimensional (e.g. a single amino acid), we would require

<table>
<thead>
<tr>
<th>Term</th>
<th>Number of Elements</th>
<th>Memory Required (bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbf{R}^{-1} )</td>
<td>( n \times n )</td>
<td>( 8n^2 )</td>
</tr>
<tr>
<td>( \hat{\mathbf{\theta}} )</td>
<td>( d )</td>
<td>( 8d )</td>
</tr>
<tr>
<td>( \hat{\mathbf{p}} )</td>
<td>( d )</td>
<td>( 8d )</td>
</tr>
<tr>
<td>( \hat{\mu} ) and ( \hat{\sigma}^2 )</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>( \mathbf{y} )</td>
<td>( n )</td>
<td>( 8n )</td>
</tr>
<tr>
<td>( \mathbf{r} )</td>
<td>( n )</td>
<td>( 8n )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( 8n(n + 2) + 16(d + 1) )</td>
</tr>
</tbody>
</table>
Since \( n \gg d \) (typically), we find that our storage requirements scales as roughly \( O(n^2) \), so that for a system with \( n = 1000 \), we require approximately 8 MB of storage. As discussed in Section 2.5.2, we require 28 kriging models for each atom type. As such, each atom type will require approximately \( 28 \times 8 = 224 \) MB of memory. As an (admittedly crude) estimation for the number of atom types required, the first parameterisation of the AMBER force field possessed 56 distinct atom types. Accounting for each of these atom types would require \( 224 \times 56 \approx 12.5 \) GB of memory.

Whilst certainly within the realms of what is feasible given modern hardware, having to store 12.5 GB of data in RAM is not an ideal situation. One could sequentially load the kriging models into RAM over the course of an MD simulation when certain atom types are encountered during the energy evaluation stage (which is amenable to massively parallel calculations), one would subsequently be crippled by I/O, and so we would ideally like to compress the data in the files as much as possible.

We primarily notice that since \( \mathbf{R} \) is symmetric, so is its inverse. As such, we need only store the upper triangular form of \( \mathbf{R}^{-1} \), which immediately reduces our storage costs from \( 8n^2 \) to \( 4n(n+1) \). Taking this into account, we reduce our estimated memory requirements from 12.5 GB to roughly 6.3 GB. Whilst this necessitates a computational overhead of symmetrising \( \mathbf{R}^{-1} \) when it is required, this can be minimised with matrix multiplication algorithms for symmetric matrices. A further immediate saving can be made based on kriging being a stochastic process. As such, one can argue that double precision accuracy is not necessary. Use of single precision floats reduces our memory costs by a further factor of two, leading to a requirement of roughly 3.2 GB of storage space.

Our next memory saving comes from a prudent selection of training points for
each kriging model. In doing so, the number of training points required can be
decreased significantly from our earlier estimate of 1000 training points. Speaking
somewhat optimistically, we can approximate the number of training examples
required to roughly 600. This reduces our memory cost to roughly 2.0 GB.

So far, we have assumed that all of our data is stored in ASCII format. In RAM,
data is stored in binary form, which constitutes a threefold reduction in memory
requirements. We can pass on this threefold reduction to our hard disk storage re-
quirements as well by storing all of our data in binary format. In FORTRAN90/95,
the default write to file is record-based, i.e. for each string of double precision
numbers stored, 4 bytes are placed on either side of the string, containing the size
of the string.

Stream-based writing is a FORTRAN2003 functionality, and removes the buffers of
record-based writing. So, for a string of three double precision values, 24 bytes +
8 bytes = 32 bytes of memory are required with record-based writing, in contrast
to 24 bytes required for stream-based I/O. By use of stream-based I/O, we are
able to reduce the hard disk cost from 2.0 GB to just over 700 MB. This quantity
is certainly feasible given modern computational resources.
Addendum
Multipolar electrostatics

Salvatore Cardamone,ab Timothy J. Hughesab and Paul L. A. Popelier*ab

Atomistic simulation of chemical systems is currently limited by the elementary description of electrostatics that atomic point-charges offer. Unfortunately, a model of one point-charge for each atom fails to capture the anisotropic nature of electronic features such as lone pairs or $\pi$-systems. Higher order electrostatic terms, such as those offered by a multipole moment expansion, naturally recover these important electronic features. The question remains as to why such a description has not yet been widely adopted by popular molecular mechanics force fields. There are two widely-held misconceptions about the more rigorous formalism of multipolar electrostatics: (1) Accuracy: the implementation of multipole moments, compared to point-charges, offers little to no advantage in terms of an accurate representation of a system’s energetics, structure and dynamics. (2) Efficiency: atomistic simulation using multipole moments is computationally prohibitive compared to simulation using point-charges. Whilst the second of these may have found some basis when computational power was a limiting factor, the first has no theoretical grounding. In the current work, we disprove the two statements above and systematically demonstrate that multipole moments are not discredited by either. We hope that this perspective will help in catalysing the transition to more realistic electrostatic modelling, to be adopted by popular molecular simulation software.

1. Introduction

Atomistic simulations of large systems over long time scales can only be achieved by using energy potentials, rather than by solving the Schrödinger equation on-the-fly. The question is then how to best represent an atom such that it interacts with other atoms in a realistic manner. A convenient and trustworthy way to answer this question is to start from the electron density, because from the first Hohenberg–Kohn theorem we know that a system’s total energy can be obtained just from its electron density. The original question can then be rephrased as to how one should represent the electron density of a given atom while it is part of a system. Surprisingly, the current and predominant view is to think of an atom in a system as being spherical. This picture corresponds to representing the atomic

---

**Salvatore Cardamone** obtained a 1st class BSc in biochemistry from the University of Sheffield. He has recently moved to the University of Manchester to complete a PhD in theoretical chemistry under the supervision of Prof. Popelier. His research focuses on structural sampling of molecular species and the parameterisation of a novel force field for use with carbohydrates. Other research interests include mathematical formalism of physical systems. Outside of academia, Salvatore is a competitive ballroom and latin dancer.

**Tim Hughes** graduated with a 1st class BSc (chemistry) from The University of Manchester in 2012. He is currently working towards his PhD in computational chemistry under Prof. Popelier developing the novel “Quantum Chemical Topological Force Field” (QCTFF), with particular interest in the non-covalent interactions between molecules. In his spare time he enjoys playing the drums and engaging in team sports such as football.
electrostatic potential as being generated by an atomic point-charge. This means that a single number (the point-charge) is associated with the atom’s nucleus while assuming that this number summarises the complexity of the atomic electron density sufficiently well in order to predict its electrostatic interaction behaviour.

A simple example shows that this view cannot be right. In Fig. 1 we consider the global energy minimum of the water dimer. This case serves to illustrate an essential argument that also applies to hydrogen bonding in general, π–π stacking and halogen bonding, which will be discussed in detail much later.

A typical ab initio calculation on the water dimer will produce a “flap angle” $\alpha$ of about 45°, while a point-charge model (i.e. single charge for each atom) will generate an $\alpha$ angle about 25°. Disregarding the irrelevant details of the level of theory used or the exact nature of this point-charge-model, it is clear that the latter cannot predict the required tilt in the water molecule at the right hand side. Only when extra off-nuclear point-charges are added to oxygen does the geometry prediction improve. Equally, if point multipole moments are added to the oxygen then the prediction improves substantially. This example shows that the currently ubiquitous treatment of electrostatic interaction cannot be correct. This simply case study is relevant because it is generally acknowledged that medium-strength hydrogen bonds can be properly described by the electrostatic interaction.

In summary, a point-charge is spherically symmetric (or isotropic) and therefore it has no directional preference while interacting with another point-charge. However, a point multipole moment on a nuclear site prefers another point multipole to be oriented in a certain way, in order to lower the multipole–multipole interaction energy (for examples see Fig. 3.2, 3.3 and 3.4 in ref. 3). This anisotropy makes a multipole moment directional.

This perspective brings together contributions that highlight the shortcomings of point-charges. It will argue, based on clear and consistent evidence, that the point-charge model is inherently limited in terms of accuracy, provided one introduces only one charge per atom. If more off-nuclear charges are introduced then the accuracy improves but this perspective focuses on a mathematically more elegant solution, which is that of multipole moments. As this perspective delivers the evidence for the superiority of multipole moments over point-charges it aspires that the status quo of the use of point-charges will change.

We can ask, however, if the inadequacy of current point-charge force fields actually matters over very long time scales, when energy errors can perhaps cancel each other. Might these errors become irrelevant fluctuations drowning in the large scale (space and time) phenomena the molecular simulator is interested in? Apparently not, if one reads a very recent statement published in the Conclusions of 100 μs molecular dynamics simulations on 24 proteins. For most of the 24 proteins studied, the simulations drifted away from their native structure (initiated from homology models). The authors stated that “In our view, it is probably more beneficial in the long run to focus on the development of better force fields than on the development of sophisticated methodologies for scoring structures realized in simulation”.

Multipole moments have been introduced decades ago in the field of atomistic energy potentials but they are still not part of the mainstream theoretical treatment of electrostatic interaction, its applications nor concomitant software. Yet, multipole moments arise naturally and rigorously in the treatment of interactions governed by an inverse distance (or 1/r) dependence. In the following we focus on the essence of what a multipole expansion achieves while omitting mathematical details that can be found elsewhere. Fig. 2 schematically shows two interacting charge distributions (left and right). To simplify matters we put the origin inside the left charge distribution and set the origin of the right distribution at $\mathbf{R}$. The position vector $\mathbf{r}$ describes the left distribution by sweeping its volume, while the vector $\mathbf{r} + \mathbf{R}$ does the same for the right distribution while being based at $\mathbf{R}$.

One can think of an infinitesimal bit of electronic charge density, located at $\mathbf{r}$, interacting with an infinitesimal bit of charge density located at $\mathbf{r}' + \mathbf{R}$. These two interacting charge bits are separated by a distance $|\mathbf{r} - (\mathbf{r}' + \mathbf{R})|$. If one wants to know the total interaction between the two charge densities then one needs to sum over all the possible pairs of interacting
interacting multipole moments on site A rank L the multipolar expansion) can be bundled by the interaction. The various terms of the electrostatic energy (appearing in add more detail in their description of the electron distribu-

moment or \( l \) acting electron distributions. Higher rank moments (up to \( l = 1 \), quadrupole moment or \( l = 2 \), etc.) successively add more detail in their description of the electron distribution. The various terms of the electrostatic energy (appearing in the multipolar expansion) can be bundled by the interaction rank \( L \). This rank is defined as the sum of the ranks of the interacting multipole moments on site A and B incremented by one, or \( L = l_A + l_B + 1 \). This interaction rank is the inverse power appearing in the expression \( R^{-L} \), in which \( R \) is the distance between the respective sites at which the interacting multipole moments are centred. The lowest possible rank of \( L = 1 \), which corresponds to the interaction between point-charges, which is the longest possible range of electrostatic interaction. It should not come as a surprise that truncating the multipolar series already at the very first term (\( L = 1 \)) harms the proper descrip-

tion of the intricacy of a given electron distribution.

In summary, many mathematical details and technical issues have been omitted in order to focus on the main points. Interested readers can find them in a recent review. However, this discussion has set the scene and simply introduced a few important concepts that will recur. We should point out that older literature has been omitted in order to make the perspective more timely or because a more recent case study makes the same point as an older one. For example, a paper by Ritchie and Copenhaver published in 1995 compared the electrostatic potential generated by an atom-centered multipole expansion (up to \( l = 3 \)) with that generated by potential-derived charges surrounding some natural and synthetic nucleic acid bases. The multipolar electrostatics always improved the \( \text{rms} \) error, by at least 10% to 30%, resulting in differences as large as 15 kJ mol\(^{-1}\). Such conclusions are reminiscent of later work by Slipchenko, Krylov, Gordon and co-workers, as discussed in Section 2.1.4.

This perspective is organised as follows. The discussion starts, in Section 2, by addressing the increased accuracy of atomic multipole moments over point-charges when modelling the electrostatic interactions between molecules. The application of multipolar electrostatics to both polar systems (water, hydrogen bonding, halogen bonding, biomolecules and solvation) and non-polar systems is addressed. Section 2 also focuses on crystal structure prediction of organic molecules in a separate subsection. Subsequently, Section 3 addresses the efficiency of the implement-
ation of atomic multipole moments in the context of both molecular simulation and in the transferability of multipole moments. Finally, Section 4 focuses on currently used multi-
polar methodologies are briefly discussed, in particular AMOEBA, SIBFA and NEMO. A rather poignant conclusion briefly summarises the current state-of-affairs.

2. Accuracy

2.1 Polar systems and intermolecular interactions

2.1.1 Water. Early electrostatic potentials for water consisted of atomic point-charges fitted to reproduce the bulk properties of liquid water. Examples include the simple point-charge (SPC) model and the TIP3P potential. These potentials are still used today in spite of both suffering from the same known pitfalls, such as accurately reproducing the experimentally observed radial distribution function (RDF) for O \( \ldots \) O (i.e. \( g_{oo}(r) \)), or a reliable dependence of liquid density on temperature. Attempts at improving the description of water involve additional charge sites, intended to represent the oxygen lone pairs. The TIP4P and TIP5P potentials are of this type. Despite an improved representation of the dielectric constant of bulk water and \( g_{oo}(r) \) over TIP4P and TIP3P, TIP5P still
poorly reproduces properties such as the heat capacity and the density versus temperature profile.

The anisotropic site potential for water (ASP-W)$^{14}$ uses an atom-centred Distributed Multipole Analysis (DMA)$^{15}$ expansion, with multipole moments up to quadrupole on oxygen and dipole on the hydrogens, computed at MP2 level. When ASP-W was compared with the point-charge potentials CKL$^{16}$ and NCC$^{17}$ and the multipolar potential PE$^{18}$, only PE provided comparably accurate minimum energy geometry for the water dimer. The ASP-W potential has been further improved to ASP-W2 and ASP-W4$^{19,20}$. The atomic multipolar expansions for ASP-W4 is now truncated at the hexadecapole level, and all interaction terms included up to rank $L = 5$. The ASP-W2/4 potentials give a more detailed description of the potential energy surface (PES) of the water dimer than that of many other water potentials. This potential was also used by Saykally and co-workers$^{21}$ in the interpretation of their terahertz laser vibration-rotation-tunneling spectra and mid-IR laser spectra$^{22}$ of water clusters from the dimer to the hexamer. Over a temperature range of 373–973 K, ASP-W2/4 gave values for the second virial coefficient, $b(T)$, close to the experimental values, which is an improvement over TIPnP ($n = 3, 4$ or 5) point-charge models.

More recently, a novel non-polarisable, multipolar water potential was published$^{23}$, with atomic multipole moments up to hexadecapole moment on all atoms (here called “QCTwater”). Multipole moments of so-called topological atoms were introduced, defined by Quantum Chemical Topology (QCT)$^{24-26}$. QCT is a generalisation of the Quantum Theory of Atoms in Molecules$^{27}$, which defines atoms as natural subspaces in the electron density using the minimal concept of the gradient path$^{28}$. Molecular dynamics simulations were run on 216 water molecules under periodic boundary conditions using QCTwater in order to test the reproduction of bulk thermodynamic and structural properties. QCTwater predicted the maximum density to be at 6°C, in good agreement with the experimental value of 4°C. Monte Carlo simulations using TIP3P and SPC did not reproduce a maximum density at all (within $[-50 \, ^\circ C, 100 \, ^\circ C]$). TIP4P and SCF/E predicted maximum densities at $-15 \, ^\circ C$ and $-38 \, ^\circ C$, respectively. At a temperature of 300 K and pressure of 1 atm, QCTwater recorded a density of 996 kg m$^{-3}$, only 0.5 kg m$^{-3}$ below the experimental value. Upon increasing the pressure, the experimentally observed increase in oxygen coordination number from 5 to $\sim 7.5$ was also reproduced by QCTwater.

QCTwater also outperformed TIP5P when predicting bulk thermodynamic properties such as the diffusion coefficient, thermal expansion coefficient and the isobaric heat capacity of liquid water. QCTwater was also able to reproduce both the experimental O - O RDF and the plot of the experimental diffusion coefficient versus temperature to high accuracy. Due to the inclusion of atomic multipole moments, QCTwater produced a more organised, directional hydrogen-bonded network in the first and second hydration shell compared to TIP4P and SPC.

Because they have parameterised for the reproduction of the bulk properties of liquid water, most point-charge potentials poorly describe ice surfaces and small clusters. The ‘induction model’ for water$^{30}$ models each water molecule by a centre-of-mass multipolar expansion. A comparison to ab initio calculations of the electric field inside a vacancy in ice showed that 70% of the electric field is dipolar and that a hexadecapole was needed.

TIP4P and ASP-W4 were also used to model the behaviour of water adsorbed onto a NaCl surface.$^{21}$ The experimental adsorption isotherm for water on NaCl showed four distinct regions: a low coverage region, a transition region, a high coverage region and a presolution region.$^{12}$ Monte Carlo simulations of the low coverage and high coverage regions were performed using both water potentials. At high coverage, only ASP-W4 predicted a more ordered structure, with three distinct layers of water due to interactions between water molecules with the Na$^+$ and Cl$^-$ ions, while TIP4P did not reproduce this layering.

2.1.2 Hydrogen bonding. Hydrogen bond interactions are not only strong, but are also observed to be highly directional. In many cases this directionality is due to anisotropic features in the electron density, most often as the lone pairs of the acceptor atom.$^{33-35}$ Isotropic atomic point-charges are unable to accurately reproduce experimental bonding geometries for a range of molecules.$^{16-42}$ The Buckingham–Fowler model,$^{43}$ which combines DMA’s multipolar electrostatics with a simple hard-sphere repulsive potential, provides several examples$^{44}$ where point-charges fail, either by leading to a spurious energy minimum, or giving quite misleading electrostatic energies. The multipolar electrostatics of this model also successfully predicted the geometries of a great variety of van der Waals complexes.$^{45}$

Efforts to model the directionality of hydrogen bonding within a point-charge framework either: (i) apply additional functions only to hydrogen bonding atoms, or (ii) add partial charges, typically at the positions of lone pairs. Allinger and Lii$^{46,47}$ implemented a directionality term into the hydrogen bonding potential of the MM3 force field, improving agreement with the ab initio MP2/6-31G** values. Kollman et al. developed$^{48}$ a methodology for deriving additional lone pair point-charges for use within a revised version of the AMBER force field. The new potentials showed that the additional sites reproduced much of the directionality observed in MP2 calculations. The additional point-charges also led to improved molecular dipole moments, in turn leading to more accurate thermodynamic properties upon molecular simulation.

Kong and Yan$^{49}$ showed that multipole moments correctly describe both the directionality and strength of hydrogen bonding for many systems. A minimum interaction rank of $L = 3$ was required to reproduce the bent structures of the dimers of the hydrates of N, O, F, S and Cl. ‘Bending’ forces arising from dipolar and quadrupolar interactions played a key role in determining intermolecular bond angles. Similar results were found by Shaik et al.$^{50}$ where at least $L = 5$ was needed to reproduce the optimised ab initio structures for water clusters and the hydrated amino acids serine and tyrosine. Again, in models where only point-charges were included (i.e. $L = 1$), pseudo-planar ring geometries were predicted that ended up too ‘flat’, i.e. the hydrogen atoms did not enough stick out of the (approximate) plane formed by the oxygen nuclei. However, as the number of water molecules in the cluster increased,
models including only lower order moments made better predictions than for smaller clusters. This is due to two effects: (i) for larger clusters there is an increase in the number of long-range interactions, which are well described by low rank terms, and (ii) water molecules in larger clusters are locked into more rigid hydrogen-bonded networks. This conclusion agrees with the observed success of many point-charge potentials capable of describing ‘bulk’ properties, despite their inability to provide reliable results when implicit water molecules are present.

Ponder et al.51 also came across the superiority of multipolar electrostatics in their work on hydrogen bonding. They calculated the hydrogen bond association energy of the formaldehyde-water complex as a function of the O-H···O=C angle, using both their own multipolar force field AMOEBA,52 the point-charge force field OPLS-AA, and MP2/aug-cc-pVTZ. OPLS-AA is incapable of reproducing the energy minima at ~100° and ~260°, while AMOEBA showed a similar shape to the MP2 curve.

In comparisons such as the one above, one should keep in mind the concept penetration of energy. At short range, even when the multipole expansion still converges and were taken to infinite order, the multipolar energy is in error by an amount called the penetration energy.3,53 For a typical hydrogen bond of 20 kJ mol⁻¹, the penetration energy is about 8 kJ mol⁻¹, which amounts to about 40% of the bond energy. As a result, improvements in the multipole expansion are of limited value without simultaneous improvements in the penetration energy. A simple analytic calculation of the electrostatic interaction between a proton and a hydrogen-like atom of nuclear charge Z shows that the electrostatic potential V(r) in a point at a distance r from the origin is not −1/r. Instead one finds that V(r) = −1/r + exp(−2Zr)(Z + 1)/r. After trivial rearrangement one can write V(r) = −1/r[1 − exp(−2Zr)(rZ + 1)], where the latter correction factor is called a damping function. This function becomes unity at long range and tends to zero at short range. Damping functions54 specifically for the electrostatic interaction appeared as late as 2000. The origin of the penetration energy is the fact that the proton probe at whose position the electrostatic potential is evaluated, lies within the electronic charge cloud that generates the potential.55 It should be emphasised that topological atoms (see QCT) do not need a correction for penetration energy because their finite volume makes it possible for a given point to lie completely outside the (topological) atom (that generates the electrostatic potential).

Secondly, the ultimate reliability of a force field is only as high as its overall balance of energy contributions. In other words, the quality of the multipolar electrostatics needs to be matched by a high-quality representation of the non-electrostatic terms, as well as the treatment of polarisation. The latter receives much attention in this article but this should not give the false impression that the other terms are not important. This high exposure to polarisation is because the main topic of this article is the electrostatic treatment in force fields and polarisation is tightly intertwined with it. In summary, one should recognise that a force field using multipole moments may be successful more because of better parameterisation of the exchange-repulsion, for example, than because of the multipole moments themselves.

Indeed, van der Waals and exchange-repulsion energies can introduce errors of similar sizes as the penetration energy.

Inspired by earlier multipolar simulations56 on liquid HF, Shaik et al.57 ran simulations on liquid imidazole (a heterocyclic aromatic ring) where the electrostatics are described by atomic multipole moments up to hexadecupole. Compared to both OPLS-AA and AMBER simulations, QCT predicted a greater quantity of hydrogen-bonded imidazoles and a lower quantity of stacked imidazoles. This is a consequence of higher order multipolar electrostatics reproducing the directionality of the hydrogen bond, organising the molecules to form a more hydrogen-bonded network. QCT showed strong agreement with the experimental densities, whereas AMBER predicted densities consistently much lower than experiment.

The same authors also performed simulations at room temperature and pressure for aqueous imidazole solutions at different concentrations from 0.5 M to 8.2 M. The density of the solutions in QCT simulations depended on concentration, in very good agreement with experiment up to 5 M, after which QCT started underestimating experiment. The AMBER potential consistently underestimated the solution's density for all concentrations by almost 0.02 g cm⁻³. The QCT system recovered the diffusion coefficient for pure water. In contrast, AMBER predicted a significantly overestimated diffusion coefficient for pure water. The two potentials generated notably different local environments, as seen by RDFs and spatial distribution functions (SDFs).

In 2014, the same group published a dual study on the hydration of serine: (i) static level, i.e., by geometry optimisation via energy minimisation of a microhydrated cluster of serine and (ii) dynamic level, i.e., or by the molecular dynamics simulation and RDF/SDF. At static level, multipolar electrostatics best reproduces the ab initio reference geometry. At dynamic level, multipolar electrostatics produces more structure than point charge electrostatics does, over the whole range. The SDF shows that only multipolar electrostatics shows pronounced structure at long range. Even at short range there are many regions where waters appear in the system governed by multipolar electrostatics but not in that governed by point charges.

Fig. 3 shows the distribution of water molecules in an aqueous imidazole solution from the point of view of the nitrogen atom in imidazole to which a hydrogen is bonded. This atom is referred to as N_H and each coloured dot (red or green) represents the position of a water’s oxygen atom. This N_H···O···N_OH

![Fig. 3 Comparison of QCT (green) and AMBER (red) in terms of Spatial Distribution Functions (SDFs) of N_H···O (isovalue = 2.0 (left) and 3.0 (right)). The carbon atoms are shaded in light blue. [Source: J. Phys. Chem. B, 2011, 115, 11389.]](image-url)
SDF shows that the distribution of oxygen atoms adjacent to N_H (AMBER, red) is asymmetrical at lower isovalues, such as 2.0 (Fig. 3, left panel). At the higher isovalue of 3.0 (right panel), the distribution becomes more circular and its centre coincides with the N-H bond axis. In contrast, the distribution of neighbouring oxygen atoms in the QCT simulations (green) is always symmetrical and centred on the N-H bond axis. This case study is a clear example of the qualitative difference in predictions made on solute–solvent structure by point charges versus multipole moments. Based on a dual RDF and SDF analysis (beyond Fig. 3) this work also revealed pronounced differences in the imidazole dimer in water.

A “weak hydrogen bond”60 is one where the donor atom is not a strongly electronegative atom. Typical examples include C-H - N/O61 or C-H - C-N62,63 These interactions can be of significance for the chiral recognition of a substrate by proteins and also for stabilising the conformations adopted by important biomolecules.64 Simulations utilising classical point-charge force fields do pick up on such interactions to some extent. However, the work of Westhof et al. showed that the cutoff distance for electrostatic interactions must be large in order for weak hydrogen bonds to be observed.65 DMA quadrupole and octopole moments are necessary to find the full range of observed structures of aromatic heterocycles interacting with water compared to when only monopole and dipole moments were used.66 Obviously, the widely used point-charge models such as AMBER, CHARMM and OPLS are currently unable to account for such interactions.

2.1.3 Halogen bonding. There is a growing literature describing what has been termed the ‘halogen bond’, where the halogen atom acts as an electrophile and interacts with a nucleophilic partner in a linear fashion. These linear halogen bonds can both as strong as hydrogen bonding, ranging from ~4 to 160 kJ mol⁻¹, and as directional. Because of this directionality halogen bonds can also influence the structure of a nucleophilic partner in a linear fashion. These linear halogen bondings can be both as strong as hydrogen bonding, ranging from ~4 to 160 kJ mol⁻¹, and as directional. Because of this directionality halogen bonds can also influence the structure of a system in a similar fashion to hydrogen bonds. It may therefore be assumed (correctly) that atomic point charges will be insufficient to reproduce halogen bonding. The linear pattern of bonding was first reported by Ramasubbu et al.67 in 1986, who inspected the adopted crystal structures of halogen atoms within the Cambridge Crystallographic Database. Since its discovery, the halogen bond has been the subject of many studies on its origin and nature.58–71

Tortii and Yoshida showed that the quadrupole moment θ_q of halogen atoms, where the z-axis is defined as the direction of the C-X bond, describes a positive region on the surface of the halogen atom “on the opposite side” (or at 180° degrees on the z-axis where 0° is on the C atom). This region is commonly referred to as the σ-hole, and its position accounts for the observed linear bonding to nucleophiles.59 Halogen bonding was proven to be dictated primarily by electrostatic effects through the work of Tsuzuki et al.68 who studied C₆F₆X and C₆H₄X each interacting with pyridine.

Due to the observed anisotropy in the electronic distribution, point-charges fail to correctly model halogen bonding. In an attempt to introduce halogen bonding into the molecular mechanics (MM) force field AMBER, an extra-point (EP) of positive charge was added to the halogen atoms of 27 halogen containing molecules,72 to mimic the position of the σ-hole. The MM interaction energies of complexes of halogens with Lewis bases had a rms error of only 1.3 kcal mol⁻¹ relative to the MP2 energies. The inclusion of the EP charge sites also improved the molecular dipole moment for a range of halogenated molecules. In a medicinal chemistry application of the EP model, a simulation was carried out on 4,5,6,7-tetrachloro-, bromo-, and iodobenzotriazoles in the active site of the enzyme phospho-CDK2/cyclin. The distributions of the halogen bond angles were in good agreement with the known order of strengths of the different halogens in their bonding. When the standard AMBER potentials were used without the EP charge sites, no halogen bonding was observed, with the X-O distances much larger than in the X-ray structures. Compared to EP, a multipolar force field avoids such ad hoc extensions altogether. Until such force fields were readily available, QM/MM calculations were suggested as an alternative to force fields.73 According to very recent work an approach to describe the geometries by electrostatics alone, without allowing for the anisotropy of the exchange repulsion, is likely to be unsuccessful.

2.1.4 Solvation. AMOEBA has been designed to overcome the incapability of AMBER (e.g. ref. 75) and CHARMM of dealing with polarisation, especially that of solvated ions, which create large local electric fields. Each atom in AMOEBA is represented by a permanent partial charge, dipole moment and quadrupole moment, and many-body terms such as polarisation are handled explicitly through a self-consistent dipole polarisation procedure. The AMOEBA force field has been applied to investigate the solvation of many ions in water,76–78 including Cl⁻, Na⁺, K⁺, Mg²⁺ and Ca²⁺. Grossfield et al.77 showed that despite the AMOEBA parameters being derived from calculations of gas-phase molecules, inclusion of polarisation terms allows both accurate and transferable single-ion solvation free energies and also solvation free energies of whole salts in both water and in formamide. The whole-salt free energies of solvation varied from experimental results by only 0.6 kcal mol⁻¹ on average, whereas the OPLS-AA and CHARMM27 force fields deviated from experiment by 9.8 and 6.6 kcal mol⁻¹, respectively. In the RDF of solvent molecules around the K⁺ and Cl⁻ the non-polarisable force fields show overstructuring, a consequence of fixed point-charges, favouring only a limited range of geometries.

2.2 Non-polar systems

The ability to replicate π-interactions rests on toroidal electronic features above and below the electron-poor plane in ring systems. Spherical electrostatic potentials emanating from atomic centres do not account for this type of system. The electrostatic properties of saturated hydrocarbons were modelled by a point-charge, +p, placed on hydrogen sites, and an opposing charge of −2p centred on the carbon. Whilst this assignment allowed for the reasonably accurate prediction of hydrocarbon crystal structures, the model failed for aromatic systems, even qualitatively. Price demonstrated that the use of DMA convincingly exposed deficiencies in this “separated point-charge” model.

A study on aromatic stacking proposed that π-stacking arises from an interaction between the electron-rich toroids out
of the aromatic plane with the electron-poor σ-backbone of another aromatic species. As such, a point-charge of \(-p\) above and below the aromatic plane, in addition to a compensatory \(+2p\) point-charge on each carbon atom in the plane, accounts for these electronic features. This model may recover the preferred parallel-displaced conformation of two aromatic molecules. Given a system of two aromatic complexes, for example \(\text{C}_6\text{H}_5\text{X} \cdots \text{C}_6\text{H}_5\text{Y}\), one may postulate relative interaction energies based upon the identity of \(\text{X}\). An electronegative group seize electronic population from the \(\pi\)-system in \(\text{C}_6\text{H}_5\text{X}\). This effect results in a decreased electrostatic repulsion between the two interacting \(\pi\)-systems, and enhances the net electrostatic interaction. An electropositive group, on the other hand, would contribute towards the interaction. An electropositive group, on the other hand, would contribute towards the interaction. An electropositive group, on the other hand, would contribute towards the interaction.

The subtle role of electrostatics in such small non-polar systems implies that their modelling requires an equally subtle description of underlying electronic properties, where dispersion is also shown to be a key factor.\(^8\)\(^4\)\(^5\) Such distinct electrostatic features may be captured by the implementation of multipole moments, shown in work\(^8\)\(^6\)\(^7\) where three benzenoids with large negative, neutral and large positive quadrupole moments were complexed with a small molecule (HF, \(\text{H}_2\text{O}\), \(\text{NH}_3\) and \(\text{CH}_4\)) geometry optimised. The multipole moments clearly governed the energetically favoured geometries of the various complexes.

Past work\(^8\)\(^6\) demonstrated that a single central multipole moment expansion diverges with increasing expansion rank. However, a distribution of the multipole moments over the atoms, such as in DMA, overcomes this problem. Such a method recovers correct orientations and electrostatic interaction energies. In a different study, electrostatic minima for several van der Waals complexes were located\(^8\)\(^7\) by a point-charge model and a full DMA up to hexadecamole moment. A notable example in this work utilises the Buckingham–Fowler model to predict five minimum energy conformations of the benzene dimer. The point-charge model predicts the global minimum to be the parallel dimer. In contrast, the DMA model yields a conformation that complies with the \(\text{ab initio}\) level calculation, whereby the most favourable conformation is that of the parallel-displaced dimer. In addition to this, relative energies between the five minima were found to be poorly represented by the point-charge model compared to full DMA.

This inability of point-charge electrostatics to reproduce \(\text{ab initio}\) derived conformations of benzene dimers has been reiterated in the work of Koch and Egert.\(^8\)\(^8\) To demonstrate that the inclusion of anisotropic electrostatic features is imperative not only in small, isolated systems, they considered an additional, supramolecular system of a benzene molecule situated within the cavity of a hexa-oxacyclophane host. Here, a T-shaped complex is formed between the benzene and hydroquinone fragments of the cyclophane. A point-charge energy minimisation resulted in a structure where hydroquinone fragments formed parallel-displaced configurations with the benzene molecule. In contrast, the usage of multipole moments recovered a T-shaped conformation.

Such aromatic complexes are dominant in biological systems, forming essential stabilising elements in nucleic acids,\(^8\)\(^9\) proteins,\(^9\)\(^0\) or carbohydrate–protein interactions in immune complexes,\(^9\)\(^1\) to name a few. Biological processes such as molecular recognition and catalysis are frequently stabilised by interactions between the \(\pi\)-density of aromatic systems. Point-charges provide a poor description of the electronic distribution of aromatic systems, and the XED force field aimed\(^7\)\(^9\) at capturing the anisotropy by the addition of extra point-charge sites. It was able to correctly predict the edge-to-face stacking for a range of substituted polyphenyl species, whereas AMBER, OPLS, MM2 and MM3 were not.

The work of Hill et al.\(^9\)\(^2\) also demonstrates the important role of electrostatics in stabilising aromatic stacking interactions due to a degree of cancelling of the attractive correlation dispersion term by exchange repulsion and delocalisation effects. Gordon and co-workers\(^1\)\(^0\) used their own 'Effective Fragment Potential' (EFP) method to investigate the interactions between nucleic acid bases. The EFP method is described as a low cost alternative to \(\text{ab initio}\) calculations, and can be considered as a polarisable multipolar force field without empirically fitted parameters. A DMA was performed on atomic centres and bond midpoints up to octopole moment. The EFP method accurately reproduced the interactions energies between stacked dimers AA and TT, with deviations from MP2 energies within 1.5 and 3.5 kcal mol\(^{-1}\), respectively.

Tafipolsky and Engels implemented an extension to AMOEBA showing a much improved description of stacked aromatic systems,\(^9\)\(^3\) including atomic multipole moments up to hexadecapole, with dipolar reparameterised polarisabilities, and a specific short-range charge penetration term. When compared against AMOEBA, MM3 and OPLS-AA, the new model showed values for the energies of both the stacked and T-shape dimers of benzene closer to accurate symmetry adapted perturbation theory (SAPT)\(^9\)\(^4\) values.

Marshall et al.\(^9\)\(^5\) ran simulations on \(\beta\)-hairpin structures of model polypeptides involving cation–\(\pi\) interactions between cationic (Me)_n\(^+\)\(\text{-Lys}^{+}\) residues and two aromatic tryptophan side chains \(n = 0, 1, 2, 3\). Simulations were run using the multipolar polarisable force field AMOEBA, and OPLS-AA, CHARMM and AMBER. Only AMOEBA reproduced the experimental NOEs distances between the lysine and tryptophan with any consistency, accurately predicting over 80% of the observed NOEs across the four systems (Fig. 4). The point-charge force fields only predicted 40–50% of the observed NOEs in two simulations, and performed worse still for the remaining ten simulations with a prediction success rate of \(~10\%)\).

2.3 Crystal structure prediction
To accurately predict the structure into which a molecule will crystallise, a computational model must provide a rigorous description of both bonded and non-bonded terms, as well as sampling the entirety of the PES. We restrict the discussion to...
how a multipolar description of the non-bonded electrostatic term can improve prediction accuracy relative to point-charges. Factors effecting other contributions are discussed in detail elsewhere.$^9$6

Typical work assumes that a given molecule will adopt a crystal structure with the lowest possible lattice energy. The corresponding ranking criterion was used by the Cambridge Crystallographic Data Centre (CCDC), who encouraged groups to participate in a series of five blind tests.$^{97-100}$ These tests were organised as competitions in which participants were invited to predict a range of unknown crystal structures as seen in Fig. 5 and 6. In each competition, the participating groups used a range of computational methods by including point-charge, multipolar and statistical approaches. A summary of the results of the five blind tests can be seen in Table 1. At first glance, the results of the early tests called CSP1999, CSP2001 and CSP2004 suggested that methods with a multipolar description of the electrostatics provided no greater reliability for predicting the correct crystal structure relative to point-charge models. For example, the point-charge electrostatics of Verwer and Leusen’s MSI-PP101,102 method outperformed the multipolar computer program DMAREL$^{103}$ method of Price et al. in the CSP1999 test. Post-competition analysis revealed that the searching algorithm was to blame rather than the multipolar force field. This conclusion turned out to be the recurring message across all three early tests. The test set of small, rigid molecules containing only C, H, N and O were generally predicted correctly (with multipolar electrostatics providing a slight advantage over point-charges). However, molecules with a high degree of conformational flexibility were not being sampled thoroughly and as a result, the experimental

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I Rigid</td>
<td>IV Rigid</td>
<td>VIII Rigid</td>
<td>XII Rigid</td>
</tr>
<tr>
<td>Polymorphic P2₁/c Stable 0/11</td>
<td>P2₁/c 2/15</td>
<td>Unfortunately not blind C2/c 4/14</td>
<td>Pbc a 4/14</td>
</tr>
<tr>
<td>Polymorphs found later</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II Rigid</td>
<td>V Rigid</td>
<td>IX Rigid</td>
<td>XIII Rigid</td>
</tr>
<tr>
<td>P2₁/h 1/8</td>
<td>P2₁/h 2-4/15</td>
<td>P2₁/c 1/10</td>
<td>P2₁/c 4/14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>III Flexible</td>
<td>VI Flexible</td>
<td>X Flexible</td>
<td>XIV Flexible</td>
</tr>
<tr>
<td>P2₁/c 1/11</td>
<td>P2₁/c 9/11</td>
<td>P2₁/c 0/12</td>
<td>P2₁/c 3/12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H₂C = CH₃</td>
<td>H C H</td>
<td>XI Replacement Rigid</td>
<td>XV Co-crystal</td>
</tr>
<tr>
<td>VII Flexible</td>
<td>P2₁/c 1/4</td>
<td>P2₁/c 0/13</td>
<td>P2₁/h 2/12</td>
</tr>
</tbody>
</table>

Fig. 5 The structures used in the first four blind tests on crystal structure prediction (CSP). Source: Int. Rev. Phys. Chem., 2008, 27, 541.
structures were not identified. The results of the fourth blind test, \(^{100}\) CSP2007, showed that with the implementation of improved searching algorithms, the multipolar electrostatic method of Price et al. \(^{100}\) consistently outperformed methods with point-charge electrostatics.

Following the success of the CSP2007, a fifth blind test \(^{104}\) was organised named CSP2010. The test molecules used in this study can be seen in Fig. 6. The improved results of CSP2007 led to the introduction of two new categories of molecule: a larger, highly flexible molecule and a hydrate with multiple polymorphs (four polymorphs tested for prediction), leading to a total of nine crystal structures to be tested. Three participating groups (Day \(^{105-107}\) van Eijck, \(^{108}\) and Price et al. \(^{105,109}\)) used atomic multipole moments to describe the electrostatic contribution to the crystal lattice energy. Multipole moment methods clearly outperformed the point charge methods (see Table 1), with multipolar methods correctly predicting four of the nine structures, compared to only one correct prediction by the point charge methods. It is interesting that for molecule XIX van Eijck switched to point charges rather than multipole moments and was able to predict the correct structure. This highlights the importance of factors other than the electrostatic description when predicting crystal structures. The most consistent method was GRACE of Neumann et al. \(^{110,111}\) which used plane-wave DFT to calculate the electrostatic energy, which one would expect to outperform even multipolar methods.

Day et al. \(^{112}\) compared two electrostatic schemes, one an atomic point-charge scheme and the other including multipole moments, for their ability to predict the 64 experimentally observed crystal structures of 50 organic molecules. The multipolar scheme reproduced 44 of the experimental structures to be within the top five most stable crystal structures for each given molecule, whereas the point-charge scheme was able to find only 36 structures. Multipolar electrostatics also correctly predicted 32 of the compounds to have structures within 0.5 kJ mol\(^{-1}\) compared to only 23 predicted by point-charges. In a response to the poor results of the CSP1999 blind test, Mooij and Leusen combined multipole moments with the Dreiding force field, and compared the predictive capabilities of the new model to point-charges. \(^{113}\) Multipole moments were able to correctly predict three out of the five experimental crystal structures as the most stable crystal polymorph, compared to only one by point-charges.

Day et al. \(^{114}\) observed that for 50 organic molecules with many polymorph crystal structures, lattice energy minimisation using atomic point-charges were considerably less accurate for molecules with hydrogen bond donor–acceptor groups than for those without. The point-charge descriptions within the FIT, \(^{114,115}\) W99, \(^{116-118}\) DREIDING, \(^{119}\) CVFF95 \(^{120-122}\) and COMPASS \(^{123}\) force fields used were described as being too simplistic to describe strong, highly directional bonds that guide crystal formation. The presence of strong hydrogen bonding leads to higher energy barriers between different minima on the PES, and acts to trap crystals in the local “metastable” states. An atomic point-charge description flattens these barriers resulting in structures moving to lower energy minima during relaxation stages in the lattice energy calculation. For example, point-charges were unable to predict the experimental “stepped sheet” structure of 2-aminoo-3-nitropyrimidine due to the crystal relaxing into the energy well of another polymorph.

Sometimes multipole moments do not appear to offer any clear advantage over point-charges although, generally, it is found that factors other than the electrostatic potential are responsible for the observed inferiority of multipole moments. A novel electrostatic potential built for the MM3 force field was tested on the crystal structures of oligothiophenes \(^{124}\) and atomic point-charges outperformed multipole moments for all but one case, namely that of \(z\)-perfluorosexithiophene (PFT4). The crystal structure for PFT4 was the structure most influenced by electrostatic interactions, an instance where one should not be surprised that multipolar electrostatics were superior. Brodersen et al. \(^{125}\) compared five electrostatic models including ESP derived point-charges and tested multipole moments in the prediction of 48 crystal structures, again using the DREIDING force field. Due to
strong dependence on intramolecular terms in the force field, such as angle bends, bond stretches and torsion angles, the use of multipole moments did not improve the accuracy of the predicted crystal structures for flexible molecules. They did, however, greatly improve the prediction of rigid molecule crystal structure, where the bonded terms are of less importance.

3. Computational efficiency of multipolar electrostatics

3.1 Transferability

The idea of an atom type is inextricably linked with that of transferability. Whilst complex definitions of an atom type have been proposed, this area remains a source of debate and competing methods. It is, however, widely regarded as a necessary measure to define electrostatic properties as pre-defined parameters for large scale molecular simulation to be truly viable.

The generation of a transferable set of multipole moments is a far more delicate operation than trying to find a corresponding set of partial charges. Whilst a monopole moment is relatively transferable, higher order multipole moments are less so due to their increasing directional dependencies. The latter make it more difficult to obtain a generic set of higher order multipole moments for a given atom type.

Many molecular and group properties are tractable when attempting to demonstrate transferability. One finds that experimental heats of formation, for example, may be reproduced for a generic hydrocarbon CH₃(CH₂)₃CH₃, by fitting to a linear relationship ΔHᵢ = 2A + xB. Here, A and B represent the respective energies of methyl and methylene groups. Indeed, this function is equally applicable to SCF single-point energies for equivalent systems, such that $E = 2E(CH₃) + xE(CH₂)$ is accurate to approximately 0.06 kcal mol⁻¹. Based on this additivity of single point energies, the concept is easily extended to imply the additivity of electron correlation energies. Because the correlation energy is a functional of a group’s electron density, it implies that electronic properties must additionally follow this transferability scheme.

Armed with this, the demonstration that multipole moments possess some amenability to atom typing should follow. In one case study, a set of small molecules composed of the functional groups present in proteins underwent DMA at HF/3-21G level, and the multipole moments of each atom were assessed. Atom typing by atomic number or hybridisation state was seen to be ineffective, but atom typing by bonding to specific functional groups proved to be more successful. Two transferable schemes were developed: ATOM and PEPTIDE. The former utilised the average multipole moments for specific atom types generated from the data set mentioned previously. The PEPTIDE model features a single multipole moment expansion centre for each distinct amino acid. As such, the local environment for each of these expansions centres is conserved for a given amino acid. The usage of ATOM resulted in substantial deviations from the ab initio electrostatic potential while the PEPTIDE model gave far more satisfactory results. Extending from this, a grossly distorted cyclic undecapeptide (a derivative of the immunosuppressive cyclosporine) was analysed by the above two models. The authors compared the electrostatic potentials generated by these models with one generated from DMA. Again, the PEPTIDE model exhibited lower average errors than ATOM.

Many years later, Mooij et al. focused on the generation of an intermolecular potential function implemented in dimers and trimers of methanol. Using a fitted electrostatic term in the intermolecular potential resulted in relatively favourable results: 0.2 kcal mol⁻¹ and 1.6 kcal mol⁻¹ deviations in the dimer and trimer energies, respectively, from counterpoise-corrected MP2/6-311+G(2d,2p) calculations. This is still more impressive than similar studies on other less complex systems that have attempted to parameterise point-charge electrostatics.
Mooij et al.\textsuperscript{131} also worked on a methanol-water and a methane-dimethylether complex. Each of these molecules was assigned a set of atom-centred multipole moments. Equally impressive results were obtained, with all interaction energies replicated to within \(\sim 0.2\) kcal mol\(^{-1}\) of the corresponding \textit{ab initio} calculations. As such, it was concluded that atom-centred multipole moment expansions are indeed transferable between the same molecules in differing environments.

There are several ways of allocating molecular electronic charge to atoms (\textit{e.g.} DMA,\textsuperscript{133} CAMM\textsuperscript{134} or QCT partitioning\textsuperscript{135}). Considering our group’s research interests, we focus here on QCT-based techniques. Focusing on energy, Bader and Beddall\textsuperscript{136} demonstrated that:

1. The total energy of a molecule is given by a sum over the constituent atomic energies.
2. If the distribution of charge for an atom is identical in two different systems, then the atom will contribute identical amounts to the total energy in both systems.

Although these conclusions are given in terms of energy, they hold for any property density of an electronic distribution over an atomic basin. In light of this fact, it was shown by Laidig\textsuperscript{137} that multipole moments, under certain constraints, adhere to the above conclusions, and so exhibit transferability.

The property of transferable multipole moments was successfully adopted by Breneman and co-workers,\textsuperscript{138} in a method denoted Transferable Atom Equivalents (TAEs).\textsuperscript{139} Primarily, a library was generated consisting of atom-based electron density fragments generated from a QCT decomposition of a set of molecules. DMA was subsequently performed on each of these fragments. These TAEs may then be geometrically transformed into a novel system for which the electrostatic potential is required. The fact that atomic property densities are additive in QCT implies that this recombination of TAEs is sufficient to reproduce an electrostatic potential of the system to a quasi-\textit{ab initio} level of theory. It should, however, be noted that the transferability of atomic basins is approximate, and so this method will necessarily carry a small error.

The efficacy of this methodology was subsequently demonstrated through three “peptide-capped” molecules: alanine, diglycine and triglycine. The analytical electrostatic potentials were computed on 0.002 au isodensity surfaces. Equivalent electrostatic potentials were also generated from TAE-reconstructed systems and Gasteiger point-charges for the extended (open) and \(a\)-conformations. The TAE multipole analysis (TAE-MA) reproduced the features of the electrostatic potential generated at \textit{ab initio} level much better than the Gasteiger point-charges did. A point-charge electrostatic model is unable to accurately predict extremes in electronic features.

In later work carried out in this group, all 20 naturally occurring amino acids and their constituent molecular fragments were rigorously assessed\textsuperscript{140} using QCT. A set of 760 distinct topological atoms were generated and cluster analysis identified a set of 42 atom types in total (21 for C, 7 for H, 6 for O, 2 for N and 6 for S). The trivially assigned atom types implemented in AMBER were either too fine-grained (\textit{e.g.} too many for atom types for N) or too coarse-grained (\textit{e.g.} C atom types not diverse enough). Later, an extensive investigation\textsuperscript{141} was carried out for atom typing by atomic electrostatic potential rather than atomic multipole moments as in the previous study. A retinal-lysine system was considered, a prominent feature in the mechanism of bacteriorhodopsin. This study focused on the aldehyde and terminal amino groups of retinal and lysine, respectively. The electrostatic potentials generated by these groups occurring in the full system were compared with those of smaller derivatives of the system. The electrostatic potential of lysine surrounding the terminal amino group was relatively conserved for all derivatives in which two (methylenic) carbon atoms were maintained along the amino acid sidechain. However, the aldehyde group of retinal was more responsive to more distant environmental effects.

This work has recently been further developed,\textsuperscript{142} where the concept of a “horizon sphere” is proposed. This sphere contains all the atoms that a given atom, at the sphere’s centre, “sees” in terms of their polarisation of the electron density on the central atom. An \(x\)-helical segment of the protein crambin was studied. The electrostatic energy was probed by considering the multipole moment expansion (up to rank \(l = 4\)) centred at a \(C\alpha\). A new set of multipole moments for \(C\alpha\) was calculated for each structure dictated by the growing horizon sphere. The interaction energy between \(C\alpha\) and a set of probe atoms was evaluated, leading to the conclusion that formal convergence of this interaction energy is attained at a horizon sphere radius of \(\sim 12\) Å. More work is underway to scrutinise the validity and generality of this conclusion.

Crystallographers who strive for the generation of transferable atomic electron densities,\textsuperscript{143} find qualms with the reconstruction of molecular electron densities from these QCT-derived atomic densities. This is due to the mismatch in interatomic surface topologies between transferred atoms. As such, they believe that it becomes very difficult to generate a continuous electron density from these atomic fragments. Work has therefore been directed towards the generation of pseudoatom databanks that may be utilised to reconstruct experimental electron densities from previously elucidated structures. From this approach follows a natural output in the form of atomic multipole moments. It is important to point out that the aforementioned mismatch can be countered by accepting that the interaction energy between atoms is what ultimately matters, not the perfect construction of gapless sequences of topological atoms. With this premise in mind we have shown that the machine learning method kriging captures,\textsuperscript{144–146} within reasonable energy error bars, the way a QCT atom changes its shape in response to a change in the positions of the surrounding atoms. Jelsch et al.\textsuperscript{147} for example, demonstrated the capacity of transferring experimental density parameters for small peptides, based upon the Hansen–Coppens formalism,\textsuperscript{148} and subsequently built a databank of pseudoatoms. The refinement of high resolution X-ray crystallographic data by referral to this databank has been demonstrated.\textsuperscript{149} A more computationally-orientated route has been developed in parallel to the one above,\textsuperscript{143} whereby the experimental density parameters for a set of pseudoatoms were derived from \textit{ab initio} electron densities of tripeptides. This method showed an enhanced amenability to transferability compared to its experimentally-derived counterpart.
More recently this pseudoatom database has been built upon. Atom types were defined by grouping atoms with the same connectivity and bonding partners, while the atom type properties were defined by averaging over all constituent “training set” pseudoatoms. Single point calculations were initially carried out on a test set of amino acid derivatives at B3LYP/6-31G** level. The geometry of each species was taken directly from the Cambridge Structural Database (CSD). Multipole moment expansions for non-hydrogen atoms were truncated at ranks $l = 4$ and $l = 2$ for the hydrogens. These multipole moments were subsequently averaged and standard deviations defined for the dataset. In terms of performance, the database model appears to give a slightly more pronounced electrostatic potential surrounding oxygen atoms in carboxylate and hydroxyl groups of Ser, Leu and Gln, compared to the more extensive ab initio calculations. Further work showed that the database does relatively well in the prediction of most atomic multipole moments. Exceptions take the form of higher order multipole moments, most particularly for oxygens and nitrogens. Finally, we mention that somewhat poorer results are obtained when considering total intermolecular electrostatic energies in dimers. The errors are of the same magnitude as those obtained from AMBER99, CHARMM27 and MM3. The authors ascribe these results to the implementation of a Buckingham-type approximation, whereby non-overlapping electron densities are assumed. This results in the underestimation of short-range interactions, which is in keeping with the sign of $AE$ in the above calculations. The authors report much-reduced discrepancies in these energies by use of their own refined method, which accounts for this discrepancy.

However, in spite of the issues of the reconstruction of crystal structures by use of QCT, the technique remains amenable to the elucidation of electrostatic properties. Woiniska and Dominikia have given a thorough elaboration on the transferability of atomic multipole moments based on various density partitions, most notably directly comparing the Hansen–Coppens formalism to both QCT and Hirshfeld partitioning. In their study, multipole moments (up to $l = 4$) were assigned to each atom in a set of biomolecular constructs, ranging from single amino acids to tripeptides. Atom types were subsequently defined from this molecule set based on criteria resembling those used by a similar study. By averaging the multipole moments for given atom types in differing chemical environments, a standard deviation from this average value was obtained. A lower standard deviation is indicative of a high degree of transferability, and vice versa. A QCT analysis of ab initio wavefunctions results in highly non-transferable lower-order multipole moments ($l = 0, 1, 2$). Secondly, for the higher-order Hansen–Coppens multipole moments ($l = 3, 4$) are particularly unstable. The atom types found to be non-transferable from the QCT analysis are generally carbons connected to two electron-negative atoms (oxygen or nitrogen), or members of aromatic systems. The decline in the level of transferability for higher-order multipole moments for the Hansen–Coppens method is relatively widespread throughout atom types, most prominently carbon and nitrogen. It is, however, strange to note that for both of these points, the poor level of transferability for QCT and Hansen–Coppens pseudoatoms is constrained to specific atom types; for the rest, these techniques generally give rise to the most transferable multipole moments.

Whilst the lower-order QCT multipole moments are largely non-transferable, they tend to be far more stable when derived from crystallographic data. In spite of this, they are still the least transferable multipole moments in the set, with both Hirshfeld partitioning and the Hansen–Coppens pseudoatom formalism yielding somewhat more stable multipole moments. The authors conclude that the most transferable multipole moments result from Hirshfeld partitioning. QCT discretely partitions electron density into distinct basins and so is vulnerable to numerical issues when undertaking mathematical operations such as integration over the basin. The Hansen–Coppens formalism, on the other hand, suffers from problems with localisation: distant electron density may be incorrectly assigned to a given nucleus. However, an exhaustive study of standard deviations from average multipole moments for given partitioning methods does little to confirm the dominance of one scheme over another. Transferability matters little if the multipole moments in question are incorrectly defined; their subsequent variances over a dataset are inconsequential. In fact, the difficulty in assigning transferable multipole moments to given atom types may equally be indicative of poor atom type definition, or the sheer inability to define a transferable atom type in terms of multipole moments with any great stability. We make a final note in that the atom types defined in this work have been tailored for pseudoatom usage, and so may not be useable with a discrete partitioning scheme (QCT), relative to the so-called ‘fuzzy’ decompositions (Hirshfeld and Hansen–Coppens). In fact, this is concomitant with an analysis of dimer energies obtained from these three techniques. When one uses multipole moments obtained by a QCT decomposition as opposed to a Hirshfeld partitioning, the electrostatic energy obtained more closely resembles that obtained from a Morokuma–Ziegler energy decomposition scheme, by as little as 10%.

### 3.2 Simulation

A recent tour de force regarding the feasibility of biomolecular simulation have seen the computation of time trajectories of systems such as the ribosome. However, this study implemented techniques not optimised for the output of particularly accurate results, using a highly parallelisable CHARMM++ interface in conjunction with the AMBER force field and the NAMD molecular dynamics package, i.e. a partial charge approximation.

Biomolecular simulation requires the implementation of periodic boundary conditions to accurately model the environment in which a system resides. Moreover, the electrostatic energy of a system is slowly convergent. Many solutions to this problem have been proposed over the years. It should be noted that the interaction involving ‘higher order’ multipole moments ($l \geq 1$) is more short-range than that between monopole moments. As such, the problem of slowly convergent long-range interactions is shared by both isotropic point-charges and multipolar electrostatics because the latter encompass point-charges.
We wish to raise the issue of the conformational dependence of electronic properties. In reality, this problem is not unique to higher order multipole moments. Conventional force fields, which employ partial charges, choose to reside in a pseudo-reality of an invariable electrostatic representation, and so rarely encounter conformational dependencies. It is rather more difficult to simply ignore the obvious reality of molecules as flexible entities when using multipole moments, and has been emphasised in analyses using both DMA and CAMM algorithms.\textsuperscript{161,162} Use of the electrostatic properties for one conformation correctly reproduces its corresponding electrostatic potential. However, use of this parameterisation in an alternative conformation results in highly unfavourable energies. It should be noted that this insufficiency is equally prominent when using a conserved set of partial charges between conformers. As such, since the issue of flexible molecules is a computational complexity that pervades all electrostatic approximations, it would be unfair to regard this as an additional burden specifically for multipole moments. Instead, it is a hurdle that both techniques are required to overcome in enhancing the accuracy of simulation.

Evaluating multipole moments as a function of a conformational parameter is appealing. For example, the multipole moments for both atoms in CO may be described analytically as a function of the interatomic distance in the molecule.\textsuperscript{163} However, scaling this idea up to systems with far more conformational degrees of freedom, such as an amino acid, is an appreciably more difficult task. An “analytical compromise” has been proposed in the past,\textsuperscript{164} whereby the multipole moments of an atom in ethanol, glycine and acrolein are represented by a Fourier series truncated at third order, whose free variables correspond to the dihedral angles of the molecule.

Instead of analytical methods, machine learning methods can be used to interpolate between a set of multipole moments defined for different molecular conformations.\textsuperscript{146} As such, one may then predict the multipole moments of an arbitrary conformation that is not present within the initial training set, which corresponds to a true external validation. Fig. 7 demonstrates the errors for (double peptide-capped) histidine obtained in following such a scheme.

Alternative methods have been developed but the literature on these techniques appears to be relatively sparse. For example, it has been proposed\textsuperscript{165} that one may average the atomic multipole moments over all conformers that are sampled during a simulation. This has been done for alanine and glycine by shifting the higher order ($l = 1$) atomic multipole moment expansions to a smaller number of expansion sites distributed throughout the molecule. An additional method, previously tested for energy minimisations of crystal structures, revolves around periodically recalculating the atomic multipole moments for the molecule.\textsuperscript{166} This proved to give substantially better results than the implementation of then-current methodologies, particularly for systems whose structures are dictated by strong hydrogen bonding.

Forces (and torques) must be calculated for the molecular translational and rotational degrees of freedom to be sampled during the course of a simulation. These may be formulated directly by first and second derivatives of interactions energies, by use of translational and rotational differential operators. If one considers a molecule as a rigid body, the individual atomic multipole moments of the molecule are invariant relative to their stationary local axis systems. As such, the derivative of the interaction energy between two molecular species is satisfied by the derivative of the interaction tensor only. This has been demonstrated in the spherical tensor formalisms\textsuperscript{167,168} and its application (e.g. ref. 169). A simulation package that allows for this rigid-body approximation in conjunction with multipolar electrostatics exists\textsuperscript{170} and is called DL\_MULTI.

The invariance of atomic multipole moments with respect to a local axis system no longer holds when abandoning the rigid-body approximation in favour of a realistic flexible-body protocol. As such, differentiation of the interaction energy function requires the derivatives of multipole moments in addition to the interaction tensor. Whilst this requires a more involved series of calculations, it is still an attainable requirement.\textsuperscript{171} Somewhat more problematic is the fact that the local axis systems in which the atomic multipole moments are referenced evolve over the course of a simulation. Due the flexible nature of the molecule, neighbouring atomic positions that make up a local axis system change with respect to time. This results in the subsequent net rotation of the atomic local frames. In the context of QCT and the machine learning method kriging, analytical forces can be calculated for “flexible, multipolar atoms”, although this is not trivial and will be published in the near future.\textsuperscript{172}

Literature quotations of CPU time differences between a molecular dynamics simulation using point-charges versus multipole moments (for a given number of nanoseconds, of a given biomolecule with a given number of water molecules surrounding it), are virtually non-existent. However, Sagui \textit{et al.}\textsuperscript{173}
reported a representative ratio of 8.5 in favour of point-charge electrostatics (as implemented in AMBER 7) when most of the calculation is moved to the reciprocal space (via the PME method) with multipolar interactions up to hexadecapole–hexadecapole being included. The only way a point-charge model can ever match the accuracy of a nucleus-centred multipolar model is via the introduction of extra off-nuclear point charges. What is rarely stated is that these additional charges create an enormous computational overhead in a typical system of tens of thousands of atoms because charge–charge interactions are longer range (1/r-dependence) than any interaction between multipole moments.

4. Implementation of multipolar electrostatics

4.1 AMOEBA

Arguably one of the most successful next-generation force fields is AMOEBA (Atomic Multiple Optimised Energies for Bio-molecular Applications).\(^5^2\) AMOEBA has been proven effective in a variety of biomolecular simulations, ranging from solvated ion systems\(^5^3\) to organic molecules\(^5^6\) and peptides.\(^5^7\)–\(^5^9\) The electrostatic energy component of the force field is broken down into two terms. The first term is concerned with permanent atomic multipole moments (expansions truncated at \(l = 2\)), whilst the second term corresponds to induced multipole moments as a result of polarisation effects. The permanent atomic multipole moments are generated by DMA of a set of small molecules such that atom types may be defined. When implemented during a simulation, these atomic multipole moments may be rotated into various fixed local axis systems within the molecule. AMOEBA proves to be competitive, even with \textit{ab initio} level calculations. All levels of theory tested perform in a uniform manner: the average \(\Delta E\) values across all conformations at the MP2/TQ, \(\text{B97/LP}\), \(\text{B3LYP/Q}\) and AMOEBA levels of theory are 3.73, 3.15, 3.64 and 3.30 kcal mol\(^{-1}\), respectively. These are impressive values, but one must remain aware that they correspond to total energies as opposed to those arising specifically from the electrostatic component of the force field.

A true demonstration of the benefits corresponding to multipole moments arises from a direct comparison between AMOEBA and the various force fields that employ point-charges. Kaminsky and Jensen,\(^1^8^0\) for example, sampled the number of energetic minima of glycine, alanine, serine and cysteine one recovers at MP2 level. The number of minima and their relative energies were subsequently compared to those recovered by use of AMOEBA and seven other point-charge force fields. The results for serine and cysteine are outlined in Table 2, where the \textit{ab initio} data suggests 39 and 47 minima, respectively. We see that AMOEBA consistently outperforms the large majority of point-charge force fields in terms of the mean absolute deviation (MAD) of energies relative to the MP2 results. AMOEBA additionally outperforms all other force fields in terms of the number of the minima it predicts for each amino acid. Note that the latter result gives rise to an artificially large MAD value relative to the other force fields. Considering the aforementioned more favourable MAD corresponding to AMOEBA, this only emphasises the predictive capacity of AMOEBA.

Another study that has directly compared AMOEBA to a variety of other conventional force fields (AMBER, MM2, MM3, MMFF and OPLS) is that of Rasmussen et al.,\(^1^8^1\) where relative conformational energies were approximated. A set of minima were generated for several molecules, each with intermediary electrostatic properties ranging from entirely non-polar to zwitterionic. The ability of the various force fields to predict relative energies of the minima was probed, in addition to three separate AMOEBA parameterisation schemes, differing in atoms typing or level of theory. All force fields performed extremely well for the non-polar molecules, largely due to the minor electrostatic contribution to the conformational energy of non-polar molecules. As such, the level at which electrostatics were calculated is essentially irrelevant. However, as the molecular species become more polar in nature, the point-charge force fields begin to display their erroneous nature relative to the AMOEBA parameterisations, which demonstrate a more uniform predictive capacity. The zwitterionic species were modelled well by several of the point-charge force fields. This can be explained by the fact that full charges are properly represented by a spherical electrostatic potential as the charge is highly localised. Thus, point-charge implementations of electrostatics can model such a case with relative ease.

<table>
<thead>
<tr>
<th>Serine</th>
<th>AMBER94</th>
<th>MM2</th>
<th>MM3</th>
<th>MMFFs</th>
<th>OPLS_2005</th>
<th>AMBER99</th>
<th>CHARMM27</th>
<th>AMOEBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of correct minima</td>
<td>21</td>
<td>19</td>
<td>15</td>
<td>27</td>
<td>23</td>
<td>21</td>
<td>19</td>
<td>34</td>
</tr>
<tr>
<td>Number of erroneous minima</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Percentage erroneous</td>
<td>4.8</td>
<td>10.5</td>
<td>20.0</td>
<td>22.2</td>
<td>13.0</td>
<td>33.3</td>
<td>10.5</td>
<td>20.6</td>
</tr>
<tr>
<td>MAD [kJ mol(^{-1})]</td>
<td>8.9</td>
<td>10.7</td>
<td>14.0</td>
<td>7.4</td>
<td>4.1</td>
<td>10.9</td>
<td>11.1</td>
<td>4.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cysteine</th>
<th>AMBER94</th>
<th>MM2</th>
<th>MM3</th>
<th>MMFFs</th>
<th>OPLS_2005</th>
<th>AMBER99</th>
<th>CHARMM27</th>
<th>AMOEBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of correct minima</td>
<td>23</td>
<td>23</td>
<td>21</td>
<td>28</td>
<td>25</td>
<td>29</td>
<td>21</td>
<td>44</td>
</tr>
<tr>
<td>Number of erroneous minima</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Percentage erroneous</td>
<td>4.3</td>
<td>4.3</td>
<td>14.3</td>
<td>25.0</td>
<td>12</td>
<td>17.2</td>
<td>23.8</td>
<td>2.3</td>
</tr>
<tr>
<td>MAD [kJ mol(^{-1})]</td>
<td>6.3</td>
<td>10.0</td>
<td>13.9</td>
<td>5.4</td>
<td>4.6</td>
<td>5.4</td>
<td>6.6</td>
<td>3.1</td>
</tr>
</tbody>
</table>
The accurate reproduction of the properties of water has long plagued simulation. Being able to account for explicit binding of water molecules, in addition to accurate modelling of bulk properties such as the dielectric constant are necessary features if one wishes to accurately simulate solved systems. AMOEBA attempts to account for the lack of a universally acceptable water model by specifically parameterising the water molecule.\(^{182,183}\) Much like the generic AMOEBA force field, atomic multipole moment expansions up to \(l = 2\) are generated using DMA. Polarisation is accounted for via induced atomic dipoles, and van der Waals interactions are modelled by a buffered 14–7 LJ potential. To the credit of the developers, this model is continually improved upon and reparameterised. Most recently,\(^{184}\) atomic multipole moments were generated for a water model at MP2 level with various basis sets in order to probe the reproduction of hydration free energies for a set of small molecules. Whilst the aug-cc-pVTZ basis set was found to give the best results, 6-311++G(2d,2p) gave a comparable accuracy at a much lower computational cost, and so is recommended for larger simulations.

A direct comparison between an AMOEBA water model parameterised at MP2/6-311++G(2d,2p) level and a widely used point-charge water model reveals the benefits of atomic multipole moment-parameterised electrostatics. A popular choice for explicit solvation is the TIP3P model, which assigns a single point-charge to each atomic centre and implements a 12–6 LJ function. Since the LJ functions differ between the two models, a “TIP3P-like” model was generated, which used the AMOEBA water model, but removed all multipole moments (static and induced), replacing them with point-charges. TIP3P and TIP3P-like models were shown to be equivalent by comparison of RDFs and bulk simulation properties. Deviation of the computed hydration free energies from experimental benchmarks for a set of small molecules are given in Table 3 for both AMOEBA and TIP3P-like models. AMOEBA outperforms the TIP3P-like model quite spectacularly, with an RMSD (AMOEBA) ~3 times smaller than RMSD (TIP3P-like).

### 4.2 SIBFA

The *Sum of Interactions Between Fragments Ab initio*, or SIBFA force field\(^{185}\) is another force field that has gained esteem within the scientific community. In a similar vein to AMOEBA, the electrostatic term of the force field is split into two distinct components: permanent and inducible. However, subtle differences arise in the computation of permanent multipole moments, whereby a DMA protocol is called upon to generate multipole moment expansions (truncated at \(l = 2\)) localised to atomic centres and bond barycentres in the procedure developed by Vigné-Maeder and Claverie.\(^{186}\) While AMOEBA localises inducible dipole moments at atomic centres to account for polarisation effects, SIBFA implements the procedure of Garmer and Stevens\(^{187}\) to account for polarisabilities at hetero-atom lone pairs and bond barycentres.

Several impressive demonstrations of the performance of SIBFA relative to *ab initio* calculations are present in the literature. For example,\(^{188}\) dimerisation energies of formamide and glycol-dipeptide have been established at the SCF/MP2 level of theory, and decomposed into constituent energetic components by use of the Kitaura–Morokuma (KM) procedure. We specifically present energies corresponding to the electrostatic interaction between monomers in Table 4. Equivalent calculations with SIBFA were carried out, with MMs parameterised at the SCF/MP2 level of theory using the Gaussian-type basis set derived by Stevens *et al.*,\(^{189}\) in excellent agreement with the KM results. Analysis of the additional components of the total intermolecular interaction demonstrated the Coulombic portion to be dominant in defining the preference of formamide dimerisation relative to that of glycol-dipeptide. Decomposition of the Coulombic interaction predicted by SIBFA additionally shows that the purely monopole–monopole interaction predicts the opposite preference. In fact, it is the monopole–dipole and monopole–quadrupole portions that recover the correct electrostatic interaction energy. A similar set of experiments was also performed between *cis*-NMA and alanyl-dipeptide, resulting in equivalent conclusions (not shown).

Many of the studies for which SIBFA is utilised appear to focus mainly upon the solvation of metal ions\(^{190,191}\) or the interaction energies of metal ions with biomolecular ligands.\(^{192,193}\) It should be noted that this approach is highly extensible to more complex biomolecular systems, such as metalloenzymes. This is demonstrated in a study that characterised the reasoning behind differential binding energies of a variety of ligands to thermolysin.\(^{194}\)

### Table 3

<table>
<thead>
<tr>
<th>Ethylbenzene</th>
<th>p-Cresol</th>
<th>Isopropanol</th>
<th>Imidazole</th>
<th>Methylthethyl sulfide</th>
<th>Acetic acid</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMOEBA</td>
<td>−0.73</td>
<td>−7.26</td>
<td>−5.58</td>
<td>−10.11</td>
<td>−1.78</td>
<td>−5.69</td>
</tr>
<tr>
<td>TIP3P-like</td>
<td>−0.89</td>
<td>−10.72</td>
<td>−5.29</td>
<td>−10.87</td>
<td>−2.94</td>
<td>−7.46</td>
</tr>
<tr>
<td>Experiment</td>
<td>−0.70</td>
<td>−6.10</td>
<td>−4.70</td>
<td>−9.6</td>
<td>−1.50</td>
<td>−6.70</td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th></th>
<th>Formamide</th>
<th>Glycol dipeptide</th>
<th>Difference, (\delta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(E_{\text{Coulomb}}(\text{SCF}))</td>
<td>−22.1</td>
<td>−14.3</td>
<td>7.8</td>
</tr>
<tr>
<td>(E_{\text{Coulomb}}(\text{SIBFA}))</td>
<td>−20.9</td>
<td>−13.2</td>
<td>7.7</td>
</tr>
<tr>
<td>(E_{\text{d}}(\text{SIBFA}))</td>
<td>−16.2</td>
<td>−20.0</td>
<td>−3.8</td>
</tr>
<tr>
<td>(E_{\text{d}}(\text{SIBFA}))</td>
<td>−1.6</td>
<td>5.8</td>
<td>7.4</td>
</tr>
<tr>
<td>(E_{\text{d}}(\text{SIBFA}))</td>
<td>0.5</td>
<td>−0.6</td>
<td>−1.1</td>
</tr>
<tr>
<td>(E_{\text{d}}(\text{SIBFA}))</td>
<td>−3.1</td>
<td>1.8</td>
<td>4.9</td>
</tr>
<tr>
<td>(E_{\text{d}}(\text{SIBFA}))</td>
<td>0.1</td>
<td>−1.0</td>
<td>−1.1</td>
</tr>
<tr>
<td>(E_{\text{d}}(\text{SIBFA}))</td>
<td>−0.5</td>
<td>0.8</td>
<td>1.3</td>
</tr>
</tbody>
</table>
However, this study primarily focused on the effects of polarisation and charge transfer. As such, despite the impressive results, we merely point it out as a demonstration of the power of SIBFA. Instead, we focus upon the lower scale simulations of these systems in which the role of electrostatics is explicitly defined.

The cation $\text{Zn}^{2+}$ is the second most common transition state metal utilised in biocatalysis, and so the characteristics of its interaction with protein subunits are obviously of importance. Focusing upon the interaction between $\text{Zn}^{2+}$ and the most basic amino acid, glycine, Rogalewicz et al.\textsuperscript{195} characterised two low-energy isomers of the system at MP2/6-31G* level for all non-zinc atoms. The lowest energy isomers correspond to the metal ion interacting with the carboxylate portion of the zwitterionic glycine, whilst those of higher energy are characterised by the metal ion’s chelation of the amino nitrogen and carbonyl oxygen of neutral glycine. The ability of SIBFA to reproduce the relative energies of the seven isomers formed was probed by parameterisation of the multipole moments and polarisabilities by two differing approaches. The first of these, SIBFA-1, corresponds to extracting the multipole moments directly from Hartee–Fock wavefunctions of glycine or its corresponding zwitterion in their entirieties. The second (SIBFA-2) decomposed glycine into two fragments (methylamine and formic acid for the neutral form, protonated methylamine and formate for the zwitterionic), followed by the generation of multipole moments from HF wavefunctions and subsequent matrix rotation into equilibrium geometries. This latter approach is obviously used to probe the transferability of SIBFA multipole moments. The result that appears to be most prominent is the particularly poor performance of the SIBFA-2 scheme at predicting Coulombic energies. However, this is by virtue of the fact that these energies have been derived from fully charged species. As such, the Coulombic attraction between these fragments is not realistic. Upon integration into a larger system, the intensity of the various multipole moments will decrease significantly due to interaction between the formate moiety and the methylammonium species. Analysis of $E_{\text{tot}}$ for this scheme demonstrates the recovery from these overly emphasised multipole moments by compensating through a smaller polarisation energy for these fully charged species. However, it should be pointed out that all isomer relative energies are correctly recovered by both schemes, with the exception of the Coulombic energy of one isomer as predicted by SIBFA-1. Nevertheless, analysis of the total energies reveals that SIBFA performs far better than conventional non-polarisable force fields.

Classical force fields have attempted to cling onto life by giving the illusion of the accounting for anisotropic electronic features by the addition of off-centre partial charges. The authors of SIBFA appear to have hit the final nail in the coffin of these classical force fields. These off-centre partial charges are empirically localised, i.e. placement on expected lone pair sites. However, these sites appear directly as a consequence of the implementation of multipolar electrostatics.\textsuperscript{196} This is most evident when considering halogen bonding. Energy Decomposition Analysis (EDA) was performed on a system of halobenzenes interacting with two possible probes (the divalent cation $\text{Mg}^{2+}$ and water) with the Reduced Variational Space Analysis (RVS) and the aug-cc-pVTZ(-f) basis set. Fig. 8 demonstrates the angular preference for the $\text{C} - \text{X} \cdot \cdot \cdot \text{P}$ interactions for the various halobenzenes (where $\text{X} = \text{F, Cl, Br or I}$ and $\text{P} = \text{Mg}^{2+}$, $\text{H}$ or $\text{O}$). Analysis of the Coulombic portion of the EDA shows that the cation preferentially interacts with ‘out of bond’-axis electronic features in both the chloro- and bromobenzene species as a result of the $\sigma$-hole. It is also immediately evident that the multipolar electrostatics implemented in SIBFA directly superimposes on these curves. As such, the effect of

---

Fig. 8  Energetic profiles as a function of the $\text{C} - \text{X} \cdot \cdot \cdot \text{P}$ angle for (from top left, clockwise) fluorobenzene-, bromobenzene-, and chlorobenzene-$\text{Mg}^{2+}$ systems. Energy from the SIBFA force field is marked in red while Coulombic energy from the Reduced Variational Space (RVS) scheme is in blue. Energetic minima that are not situated at 180° are hallmarks of halogen bonding. [Source: J. Compt. Chem., 2013, 34, 1125.]
the σ-hole can be accounted for by multipole moments, without the need for fictitious empirical partial charges. Further decomposition of the SIBFA electrostatic energy was conducted, whereby it was found that whilst the monopole–monopole interaction favours a simple $\theta = 180^\circ$ angular conformation, it is largely due to the monopole–quadrupole that this conformation is favoured. Preference for the quasi-perpendicular conformations is dictated by the monopole–dipole interaction. The authors suggest that for a perfect superposition of the two curves, higher-order multipole moments may be required.

4.3 NEMO

The final model we consider is that of the NonEmpirical MOdel, NEMO,$^{197}$ used specifically to approximate intermolecular potentials. We only discuss the electrostatic portion of this potential, which is calculated by expanding Hartree–Fock SCF molecular wavefunctions as a sum of atomic natural orbitals. A monopole moment between two atoms is defined by calculation of an overlap integral between the basis functions assigned to the atoms. Doing this for each pairwise interaction of atoms in the system, one generates a “monopole moment matrix”, the diagonal elements of which correspond to local atomic monopole moments. Higher-order multipole moments may also be assigned by replacing the overlap integral aforementioned with a corresponding multipole moment integral. A more complete description of this technique is given elsewhere.$^{198}$ However, for flexible molecules, molecular charge distributions evolve as a function of internal coordinates. The authors overcome this by defining this charge distribution as a function of the native molecular charge distribution and corresponding atomic polarisabilities. Using this scheme, dimer energies and geometries for the four dominant minima on a dimethoxymethane (DME)-water complex were calculated and compared to SCF calculations. The nomenclature for the DME conformations correspond to the orientation of the three dihedral bonds, with $a$, $g$ and $g'$ representing antiperiplanar, gauche and gauche' geometries, respectively. Raman spectroscopy of DME in water predicts an $aga > agg' > aag > aaa$ order of stability. This diverges from that predicted by NEMO, but the authors point out the non-equivalence of solvated DME and a DME-water complex. This is a valid point since NEMO energies generally agree well with the SCF energies.

More recently, higher level calibration of the NEMO potential was carried out based on CCSD(T) benzene dimer energies.$^{199}$ The authors found an excellent agreement with experimental benzene trimer geometries. However, they neglected to compare quantitative data with other computational work, citing their calculations to have been carried out at too high a level of theory for direct comparison with other lower levels of theory. Nevertheless, they reported a general agreement with previous theoretical calculations, without mentioning specific details. More recent work$^{200}$ has further developed upon the NEMO formalism, and improved the level of theory for parameterisation in addition to reporting improved performance of the model.

5. Conclusions

In general, the electrostatic interaction between atoms cannot be described accurately when using only one partial charge per atom. Nevertheless, the “point-charge paradigm” continues to dominate contemporary molecular simulation, with the vast majority of practitioners either ignoring or being unaware of the scientific repercussions of this paradigm. It is important that the computational community has the will to progress beyond this paradigm, especially because a solution is available: multipolar electrostatics. In fact, this solution has been available for a long time but an irreversible embrace of it remains absent, sadly. Journal editors should also help overcoming this unacceptable status-quo.

Currently availability computing power offers the opportunity to finally make a step change in the modelling of electrostatics at a time when, certainly in the area of biomolecular simulation, the trust of experimentalists towards computational predictions needs to be gained. Errors of a few kilojoules per mole can already be enough to draw the wrong qualitative conclusion from a calculation. How the use of point-charges can then be perpetuated at short and medium range is baffling. We hope that this perspective has made a convincing case by collecting and reporting the evidence against point-charges. The message should be clear but it now remains to be seen if a combination of powerful computers and scientific goodwill will finally realise a long overdue transition to multipolar electrostatics.

Acknowledgements

We thank Dr MS Shaik for his help in the preparation of a very early draft of this manuscript. Gratitude is due to BBSRC for providing a PhD studentship for SC, and we also thank EPSRC for funding a PhD studentship for TJH with sponsorship from AstraZeneca Ltd.

References


The conformational flexibility of carbohydrates is challenging within the field of computational chemistry. This flexibility causes the electron density to change, which leads to fluctuating atomic multipole moments. Quantum Chemical Topology (QCT) allows for the partitioning of an “atom in a molecule,” thus localizing electron density to finite atomic domains, which permits the unambiguous evaluation of atomic multipole moments. By selecting an ensemble of physically realistic conformers of a chemical system, one evaluates the various multipole moments at defined points in configuration space. The subsequent implementation of the machine learning method kriging delivers the evaluation of an analytical function, which smoothly interpolates between these points. This allows for the prediction of atomic multipole moments at new points in conformational space, not trained for but within prediction range.

In this work, we demonstrate that the carbohydrates erythrose and threose are amenable to the above methodology. We investigate how kriging models respond when the training ensemble incorporating multiple energy minima and their environment in conformational space. Additionally, we evaluate the gains in predictive capacity of our models as the size of the training ensemble increases. We believe this approach to be entirely novel within the field of carbohydrates. For a modest training set size of 600, more than 90% of the external test configurations have an error in the total (predicted) electrostatic energy (relative to \( ab \) \textit{initio} of maximum 1 kJ mol\(^{-1}\)) for open chains and just over 90% an error of maximum 4 kJ mol\(^{-1}\) for rings. © 2015 The Authors. Journal of Computational Chemistry Published by Wiley Periodicals, Inc.

DOI: 10.1002/jcc.24215
anomic effect is not entirely clear,[3] but the energetics that arise from it must be captured by a force field which deals with carbohydrates. Similarly, the exo-anomeric effect, which deals with substituents linked to an anomic oxygen, forces a separate conformational preference. It is not within the scope of this work to deal with the anomic and exo-anomeric effects in any great detail, and so we refer the interested reader to a more rigorous overview.[4]

4. Similar to the anomic effect, the gauche effect can arise within a number of carbohydrate species and bias rotamer preferences.[3] This effect has an ambiguous origin. Research has suggested it to arise from hyperconjugation or solvent effects. In short, it is the preference of a gauche rotamer over an antirotamer, where the latter would be stereoelectronically preferable. Work by Kirschner and Woods[6] proposed that the gauche effect results from solvent effects, which are not important in our work since we are explicitly dealing with gas phase molecules. However, for a carbohydrate force field to be of use, this effect must be accounted for.

The conformational freedom of carbohydrates renders them somewhat troublesome for experimentalists, as they prove to be highly difficult to characterize by conventional high-resolution structural determination techniques,[7] particularly X-ray crystallography. To be precise, carbohydrates tend to be difficult to crystallize, which is problematic because X-ray diffraction techniques are a valuable source of structural information. As such, the structural characterization of carbohydrates rests with a few experimental techniques, and subsequent validation by computational means. This required harmony between experiment and computation is vastly important, and has been recently explored,[8,9] and so proves to be a fruitful avenue for development.

Classical force fields such as OPLS-AA, CHARMM, GROMOS, and AMBER appear to have characterized carbohydrates as “secondary molecular species” relative to their peptide counterparts. As such, these parameterizations resemble “bolt-on” components. However, force fields that are specifically tailored for carbohydrates do exist and have proven successful. GLYCAM[10] is perhaps the most prominent of these force fields, and has been ported to AMBER. More recently, the advent of DL_FIELD has facilitated the use of GLYCAM parameters within DL_POLY 4.0.[11] GLYCAM has undergone extensive validation in an attempt to demonstrate its efficacy. Several studies have focused upon its ability to reproduce conformer populations in explicitly solvated molecular dynamics (MD) simulations,[6,12,13] which is obviously important owing to the massive conformational freedom of carbohydrates. The applicability of GLYCAM to larger, more biologically relevant structures, such as the binding of endotoxin to recognition proteins[14] or the dynamics of lipid bilayers,[15] has also been demonstrated.

GLYCAM has attempted to break the paradigm of deriving partial charges based on a single molecular configuration. Instead, it has been developed such that the partial charges are averaged over the course of a MD simulation, thus (albeit simplistically) accounting for the dynamic nature of electronic properties.[16] However, it must be emphasised that GLYCAM resides within the partial charge approximation to electrostatics, thus severely limiting its predictive capacity. Sugars are particularly amenable to hydration, yet a partial charge approximation to electrostatics cannot recover the directional preferences of hydrogen bond formation without the addition of extra point charges at non-nuclear positions. The isotropic nature of partial charge electrostatics is readily overcome by use of a multipole moment description of electrostatics, which naturally describes anisotropic electronic features such as lone pairs. The benefits of such a multipole moment description over their partial charge equivalents has been systematically demonstrated over the past 20 years in many dozens of papers, recently reviewed.[17] These benefits are not necessarily outweighed by the common misconception that multipole moment implementations are computationally expensive relative to their point charge counterparts. The long-range nature of point charge electrostatics, $\epsilon(r^{-1})$, relative to higher order multipole moments [dipole-dipole interactions, for example, die off as $\epsilon(r^{-3})$,] means this is not strictly true. Point charges require a larger interaction cutoff radius relative to higher order multipole moments, and, therefore, form the bottleneck in electrostatic energy evaluation. Given proper handling (e.g. parallel implementation), the computational overheads associated with multipole moment electrostatics can be managed.

In the remainder of this article, we shall demonstrate a novel means for modelling electrostatics by use of a multipole moment expansion centred upon each atomic nucleus. The techniques we present will inherently capture the conformational dependence of these multipole moments.

Methodology

Atomic partitioning

The development of molecular orbital theory largely caused a decline in chemical understanding of the theoretical description of molecular systems. The fact that each electron occupies a molecular orbital dispersed across the spatial extent of the molecule gave rise to a valid query: why do functional groups impart some property, such as reactivity, to the molecule, when the distribution of electrons throughout the system is essentially no different to the inert species? Surely, there must be some localization of electrons to the functional group, which permits subsequent functionality. If this is not the case, then even the most fundamental chemical concepts, such as those of nucleophiles and electrophiles, have no theoretical grounding. These terms are used to denote a property of an atom in a molecule, which is not recovered by molecular orbital theory. For example, if one assesses butanol by means of molecular orbital theory, the electrons “belonging” to the hydroxyl group are dispersed throughout the entirety of the molecule. If this is truly the case, then it becomes particularly problematic when one attempts to explain why the presence of the functional group imparts reactivity (an electronic phenomenon), when the electrons are not localized.

Bader and co-workers[18] went some way to address this problem by performing an appealing partitioning of three-dimensional space into atomic basins, which pictorially defines an “atom in a molecule,” an approach called the Quantum
Theory of Atoms in Molecules (QTAIM). The latter is the first segment of a broader approach called Quantum Chemical Topology (QCT),\textsuperscript{[19–22]} which analyzes quantum mechanical functions other than the electron density and its Laplacian. The QTAIM partitioning has been demonstrated to have several advantages compared to other partitioning schemes,\textsuperscript{[23–25]} and enjoy excellent transferability compared to other schemes.\textsuperscript{[26]} By use of Born’s interpretation of quantum mechanics, one generates a physical electron density, \( \rho(r) \), from an \textit{ab initio} wavefunction, \( \psi(r) \), obtained entirely from a first principle calculation. This electron density is then be completely partitioned by use of the gradient operator

\[
V = \frac{\partial^2 \rho}{\partial x^2} + \frac{\partial^2 \rho}{\partial y^2} + \frac{\partial^2 \rho}{\partial z^2}
\]

where \( i, j, k \) are unit vectors along the \( x, y, \) and \( z \) axes, respectively, to generate the vector \( \nabla \rho(r) \). The evaluation of the gradient of a scalar field results in a vector field, the vectors of which are directed along the path of greatest increase in a function. As such, the vectors that define the field \( \nabla \rho(r) \) point toward the greatest increase in the scalar field \( \rho(r) \). If one were to map \( \rho(r) \) by use of a contour plot, such that each contour represented the encasing of a surface with constant electron density, termed an isosurface, then each vector in \( \nabla \rho(r) \) would intersect each isosurface orthogonally.\textsuperscript{[27]}

Points in the vector field defined such that \( \nabla \rho(r) = 0 \) are termed critical points. Within the scalar field \( \rho(r) \) they represent a maximum, minimum or saddle point (mixture of minimum or maximum depending on direction). The identity of each critical point is revealed by assessing the curvature of \( \rho(r) \) at each point, achieved by evaluation of the Hessian of \( \rho(r) \)

\[
H(\rho) = \left[ \begin{array}{ccc}
\frac{\partial^2 \rho}{\partial x^2} & \frac{\partial^2 \rho}{\partial x \partial y} & \frac{\partial^2 \rho}{\partial x \partial z} \\
\frac{\partial^2 \rho}{\partial y \partial x} & \frac{\partial^2 \rho}{\partial y^2} & \frac{\partial^2 \rho}{\partial y \partial z} \\
\frac{\partial^2 \rho}{\partial z \partial x} & \frac{\partial^2 \rho}{\partial z \partial y} & \frac{\partial^2 \rho}{\partial z^2}
\end{array} \right]
\]

This Hessian matrix is a real symmetric matrix and hence Hermitian. Therefore, its eigenvalues are real and they express the magnitude of the curvature along each of the principal axes, which are marked by the direction of the corresponding eigenvectors. The nature of the critical point in question is then given by two easily evaluated parameters: the rank \( (\omega) \) and signature \( (n) \) of the critical point, where the former is defined as the number of nonzero eigenvalues of \( \rho(r) \), and the latter as the sum of the signs of the eigenvalues.

A fundamental result in the topology of \( \nabla \rho(r) \), of great importance in the following, is the partitioning of a molecular system into topological atoms. A key feature necessary to achieve this result is the gradient path. An easy way to grasp what this is to think of a succession of very short gradient vectors, one after the other and constantly changing direction. In the limit of infinitesimally short gradient vectors, one obtains a smooth and (in general) curved path, which is the gradient path. A gradient path always originates at a critical point and terminates at another critical point. Bundles of gradient paths form a topological object depending on the signature of the critical points that the object connects. All possibilities have been exhaustively discussed before\textsuperscript{[28]} but three ubiquitous possibilities are specified as follows: (i) the topological atom is a bundle of gradient paths originating at infinity and terminating at the nucleus, (ii) the bond path (or more generally atomic interaction line) is the set of two gradient paths, each originating at a bond critical point and terminating at a different nucleus, and (iii) the interatomic surface (IAS), which is a bundle of gradient paths originating at infinity and terminating at a bond critical point.

An interatomic surface obeys the following condition

\[
\nabla \rho(r) \cdot n(r) = 0 \quad \forall r \in IAS
\]

where \( n(r) \) is defined as the vector normal to the IAS. By finding all surfaces that obey this condition, the molecule is completely partitioned into topological atoms \( \Omega_i \), where the subscript denotes the atomic basin associated with the \( i^{th} \) atom in a molecule. All key topological features of \( \nabla \rho(r) \) are summarized in Figure 1.

Integration over these atomic basins allows atomic properties \( P_i(\Omega) \) to be defined and calculated. The universal formula from which all atomic properties can be calculated is

\[
P_i(\Omega) = \int_{\Omega} d\mathbf{r} f(\mathbf{r})
\]

where integration with respect to \( d\mathbf{r} \) denotes a triple integration over all three Cartesian coordinates, confined to the atomic volume \( \Omega_i \), and \( f(\mathbf{r}) \) denotes a property density. For example, if \( f(\mathbf{r}) \) equals the electron density \( \rho(\mathbf{r}) \) then the corresponding atomic property is the electronic population of the topological atom. If \( f(\mathbf{r}) = 1 \), then we obtain the atomic volume and when \( f(\mathbf{r}) = \rho(\mathbf{r})R_{\text{m}}(\mathbf{r}) \) the topological atom’s multipole moments,\textsuperscript{[29]} where \( R_{\text{m}}(\mathbf{r}) \) is a spherical tensor\textsuperscript{[30]} of rank \( \ell \) and \( m \). Others have shown\textsuperscript{[31]} the better agreement with reference electrostatic potentials of topological multipole moments compared to CHELPG charges. A further advantage of QCT is that the finite size and nonoverlapping nature of the topological atoms avoids the penetration effect, which may otherwise appear in the calculation of intermolecular interaction energies.

Kriging

We can only outline kriging here, for more technical details the reader is referred to our work on histidine.\textsuperscript{[32]} In general, a machine learning method is trained to find a mapping between an input and an output. The machine learning method kriging\textsuperscript{[33–35]} can also be seen as an interpolative technique able to predict the value of a function at an arbitrary \( d \)-dimensional point, \( \mathbf{x}^* \), given the value of the function at \( n \) different points, \( \{x_1, x_2, \ldots, x_n\} \), in this \( d \)-dimensional space.
Kriging is apt at modeling high-dimensional function spaces, and so is particularly stable when considering the conformational space of large molecules. A cornerstone of kriging is its heart. This function enables kriging to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space. Instead, only the inner products between the images of all data pairs in feature space need to be computed. The kernel that we evaluate is a kernel method in view of the kernel function at its core.

The function values \( \{f(x_1), f(x_2), \ldots, f(x_n)\} \) from which \( f(x') \) is composed form a basis set that is necessarily only complete if each point within \( d \)-dimensional space has been sampled, that is, as \( n \to \infty \). Hence, kriging only gives an approximate value for any predicted point \( f(x') \), and the accuracy of its prediction increases as \( n \to \infty \). An adequate kriging model typically requires at least \( 10 \times d \) data points, which is well within reach of modern day computational power for high-dimensional functions, and so kriging is a feasible machine learning method for our purposes.

There are several intricacies involved with the evaluation of a kriging model, in particular the manner by which multidimensional functions are dealt with. For a \( d \)-dimensional vector in function space, \( x \), each dimension, or feature, is assigned a parameter \( \theta_h \), which maps the variance of \( f(x) \) with respect to a change in the \( h \)th feature. Consider, for example, the function mapped in Figure 2. In relative terms, whilst \( f(x, y) \) changes significantly in response to a change in \( x \), it is essentially invariant with respect to a change in \( y \). As such, \( y \) is relatively insignificant when assessing the correlation between two points in function space, i.e. \( f(x, y) \) is much more dependent on \( x \) than on \( y \). As a result, \( x \) is assigned a higher “importance” than \( y \), which is reflected in the value of \( \theta_x \) where \( \theta_x > \theta_y \).

Kriging is a kernel method in view of the kernel function at its heart. This function enables kriging to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space. Instead, only the inner products between the images of all data pairs in feature space need to be computed. The kernel that we evaluate when obtaining a kriging model is a function of the correlation between two points in feature space, \( x' \) and \( x'' \), such that the correlation between the points, \( R(x', x'') \), is given by

\[
R(x', x'') = \exp \left[ -\sum_{h=1}^{d} \theta_h |x'_h - x''_h|^{p_h} \right]
\] (5)

Brief analysis of this function shows that, if \( x'_h \) and \( x''_h \) are situated closely together for many features \( h \), then the argument of the exponential tends toward zero, leading the correlation between the two points to tend toward one. Note that if the \( h \)th feature is relatively unimportant, it will be assigned a low \( \theta_h \) value. As a result, the \( h \)th term in the sum becomes smaller if \( x'_h \) and \( x''_h \) are relatively far apart, leading to an increased correlation between the two points, which demonstrates that \( f(x'_h) \) and \( f(x''_h) \) are similar.
Given a set of \( n \) points in feature space, an \( n \times n \) correlation matrix, \( R \), is defined, whose elements are defined as the correlation between the \( i^{th} \) and \( j^{th} \) points (and as such is symmetric). The task at hand is then to minimize the mean squared error of prediction of the kriging estimator. It can be shown that this is equivalent to maximizing the likelihood function \( L \), which is given by

\[
L = \frac{1}{(2\pi)^{\frac{n}{2}} |R|^{\frac{1}{2}}} \exp \left[ -\frac{(y-\mu)^T R^{-1} (y-\mu)}{2\sigma^2} \right],
\]

where \( \mu \) is a column vector of ones, \( t \) denotes the transpose, \( \sigma^2 \) is the process variance, and \( \mu \) is a constant term that models the global trend (i.e. “background”) of the column vector \( y \) of observations. This formula arises from the definition of a Gaussian process and is not discussed here. For our purposes, it is more convenient to maximize the natural logarithm of this function, which is done analytically (see Supporting Information of Ref. 36) by differentiation with respect to \( \sigma^2 \) and \( \mu \) and setting the respective derivatives to zero. When these optimal values for \( \sigma^2 \) and \( \mu \) are substituted back into eq. (6), one obtains the “concentrated” log-likelihood function\(^{[32]} \), or

\[
\log L = -\frac{n}{2} \log (\hat{\sigma}^2) - \frac{1}{2} \log(|R|)\]

where \( L \) is the likelihood and \( \hat{\sigma} \) is the process variance (a constant). The parameters \( \theta \) and \( p \), which are \( d \)-dimensional vectors containing the individual feature parameters mentioned previously, must be optimized, which is equivalent to maximizing the “concentrated” likelihood function. From eqs. (5) and (7) it is clear that \( \log L \) is a function of \( \theta \) and \( p \). The function \( \log L \) is the quantity that needs to be maximized, which is done by another machine learning method called particle swarm optimization\(^{[37]} \) We can then make a prediction of the output at a new point \( x' \) with the optimized kriging parameters \( \theta \) and \( p \), using the formula

\[
\hat{y}(x') = \hat{\mu} + \sum_{i=1}^{n} a_i \cdot \phi(x' - x_i).
\]

where \( a_i \) is the \( i^{th} \) element of the vector \( a = R^{-1}(y-\mu) \) where \( \mu \) is the (known) maximized mean, \( \phi(\cdot) \) is calculated\(^{[32]} \) via eq. (5).

In our implementation of machine learning, each atom within a molecule is termed a “kriging center”, with a respective multipole expansion (up to the hexadecapole moment) centered on the nucleus. Higher rank multipole moments are highly sensitive to the change in conformation of the molecule due to a fluctuating electric field. As such, we define each multipole moment as a function of the 3N-6 degrees of freedom of the molecular system, pertaining to each kriging center.

The molecule is distorted by means of energy input into each of its normal modes (discussed in section 2.3), and the multipole moments of each kriging center elucidated for a given training set size. These data are used to construct separate kriging models for each kriging center. From this, the kriging model is then able to, given an arbitrary point in conformational space, predict the associated multipole moments which accompany such a position. This method has been developed and tested substantially within our group, and gives very agreeable results for a number of distinct chemical species\(^{[32,38]} \) but is nevertheless necessarily a subject of intense ongoing refinement.

**Conformational sampling**

Here, we present a conformational sampling methodology that utilizes the normal modes of a molecular system as a means for dynamically evolving the system. This methodology has been used before in our lab for amino acids\(^{[32,39,40]} \) and small molecules\(^{[41]} \) but this is the first time we report it in great detail. Each normal mode has a corresponding frequency that is calculated by diagonalization of the mass-weighted Hessian, \( \mathbf{H} \), the details of which are incorporated in the Supporting Information. With these frequencies, a system of equations of motion is obtained, which permits for the conformational evolution of the molecular system in time. These equations take a harmonic form, and are elaborated upon in the Supporting Information.

Expressing the system in a basis of internal coordinates results in six of the 3N Cartesian degrees of freedom possessing a frequency of zero (these correspond to the global translational and rotational degrees of freedom), and so we need only evaluate the 3N-6 “vibrational” equations of motion\(^{[42]} \). We refer to these \( N_{vib} = 3N-6 \) degrees of freedom in the internal coordinate basis as “modes” of motion. The transformation from a mass-weighted cartesian coordinates, \( \mathbf{q} \), to the set of internal coordinates, \( \mathbf{s} \), is attained by evaluating the 3N x 3N transformation matrix, \( \mathbf{D} \), which satisfies

\[
\mathbf{s} = \mathbf{D} \mathbf{q}
\]

Note that this transformation retains the mass-weighting of \( \mathbf{q} \). Construction of \( \mathbf{D} \) is undertaken by defining six orthogonal vectors corresponding to the global translational and rotational degrees of freedom of the system, as defined by the Sayvetz conditions. To implement these, the system must be specified in a global reference frame (sometimes termed the Eckart frame), the origin and axes of which coincide with the centre of mass and principal axes of inertia, respectively. Sufficient to say that these conditions dictate the system possesses no net angular momentum relative to the Eckart frame, which rotates with the system.

The above leads to the generation a set of six orthogonal vectors, which are invariant under global translational and rotational motion. These vectors correspond to the first six columns of \( \mathbf{D} \). Since the internal coordinates form a mutually orthogonal basis, the resultant \( N_{vib} = 6 \) columns are generated by means of a Gram-Schmidt orthonormalization procedure, whereby the projection of \( \mathbf{D}_i \) on \( \mathbf{D}_j \), \( P_{ij} \), is given by
\[ P_i = \frac{D_j \cdot D_i}{D_j \cdot D_j} \] (10)

Note that the columns of \( D \) are also normalized by this process. If \( D_j \) and \( D_i \) are orthogonal, then \( P_i = 0 \). If \( P_i \neq 0 \), then \( P_i \) is subtracted from \( P_j \) and the process iterated until \( P_i = 0 \).

Generalizing to account for our \( N_{\text{vib}} \) columns,

\[
\begin{align*}
D_1 &= D_1 - \sum_{j=1}^{N_{\text{vib}}} P_{1j} \\
D_2 &= D_2 - \sum_{j=1}^{N_{\text{vib}}} P_{2j} \\
&\vdots \\
D_n &= D_n - \sum_{j=1}^{N_{\text{vib}}} P_{nj}
\end{align*}
\] (11)

where \( 1 \) is column vector of ones. As before, this procedure is iterated until \( P_i = 0 \) for all \( i \) and \( j \), which results in the \( \{D_j\} \) forming a mutually orthonormal set. In computational terms, the threshold value of the \( P_{ij} \), which we require before considering the \( \{D_j\} \) to be mutually orthonormal, is \( O(10^{-8}) \).

The mass-weighted Hessian \( H \), outlined in Supporting Information, is transformed into the internal coordinate basis, by use of \( D \)

\[ H_i = D^\dagger H_0 D \] (12)

where the subscripts denote the basis in which these quantities are expressed, and \( D^\dagger \) denotes the transpose. To evaluate the frequencies of the various modes of motion, we require diagonalization of \( H_0 \)

\[ \mathbf{E}^{-1} H_0 \mathbf{E} = \mathbf{I} \lambda \] (13)

where \( \mathbf{E} \) denote the eigenvectors of \( H_0 \) and \( \mathbf{I} \) is the identity matrix. In our protocol, this is achieved by tridiagonalizing the Hessian by the Householder algorithm, followed by a QR decomposition of the tridiagonal Hessian,\(^{[43]}\) yielding a diagonal Hessian, as required. The resultant eigenvalues, \( (I\lambda)_i = \lambda_i \), are related to the mode frequencies, \( \nu_i \), by

\[ \nu_i = \sqrt{\frac{\lambda_i}{4\pi^2c^2}} \quad \forall i=1, \ldots, 3N \] (14)

where \( c \) is a factor which incorporates the speed of light, \( c \), and the conversion from atomic units to reciprocal centimeters.

Of course, six of these frequencies correspond to the global translational and rotational degrees of freedom of the system and are zero, thus yielding \( N_{\text{vib}} \) non-zero frequencies. The reduced masses and force constants corresponding to the modes with \( \nu_i \neq 0 \) are given by similar manipulations of these quantities. The reader is again directed to Ref. \([42]\) for a discussion of their calculation. The amplitude of the \( j^{th} \) mode, \( A_j \), is given by rearrangement of the familiar expression for the energy of a simple harmonic oscillator, \( E = k_i A_i^2 / 2 \)

\[ A_i = \sqrt{\frac{2E}{k_i}} \] (15)

where \( k_i \) is the force constant of the mode of motion, and \( E \) is the energy available to it. We now have all quantities required to evolve the modes of motion and replicate the vibrational dynamics of the system. The total energy available to the system is given by the expression for thermal energy, \( E = N_{\text{vib}} k_B T / 2 \), and is stochastically distributed throughout the modes. The phase factors of the modes, \( \phi_i \), are also randomly assigned: if \( \phi_i = 0 \) for all modes, then they oscillate in unison, which corresponds to a photonic single frequency excitation. Instead, we assume the modes to resonate out of phase with one another, as energy transfer to each mode from an external heat bath will be predominantly decoherent.

The sole remaining issue is the choice of a dynamical time-step with which to evolve the various modes of motion. Our choice is based on the desire to ensure a single oscillation of a mode is sampled uniformly, i.e. we do not want to bias our sampling toward specific regions of the period function that describes the evolution of the mode. We obtain the time period of the mode as \( T_j = 1 / \nu_j \) and subsequently ensure that the sampling methodology permits \( n_{\text{cycle}} \) points to be evaluated along a single oscillation. In this case, the dynamical time-step for the \( j^{th} \) mode, \( \Delta t_j = 1 / n_{\text{cycle}} \cdot \nu_j \) is left as a user-defined input, and is set to \( n_{\text{cycle}} = 10 \) in the following work. Additionally, the distribution of the total energy throughout the modes is considered a dynamic quantity, and so for every \( n_{\text{reset}} \) samples that are output, the energy is randomly redistributed throughout the system. The phase factors are also redefined at the same frequency. Again, \( n_{\text{reset}} \) is left as a user-defined parameter, and is set as \( n_{\text{reset}} = 2 \) in the following.

We wish to clarify an issue in order to avoid misinterpretation. The above methodology is not meant as an exact technique for the exploration of the molecular potential energy surface (PES). By truncating the Taylor series of the potential energy at second order, we essentially model the local PES as a harmonic well, which is obviously a simplification. However, we believe the above process to be a computationally efficient means for generating molecular conformers. Moreover, the important alternative method of MD to generate conformers is not necessarily more realistic. whilst the success of MD is not in question, the validity of the force fields that are currently implemented is not guaranteed.

**Computational Details**

The workflow proposed below essentially takes an ensemble of configurations as input, and outputs kriging models for the variation in the atomic multipole moments as a function of the configuration of the system:

1. The test system is sampled in accordance with the methods outlined in section 2.3. The general idea is to sample as much of configuration space as would form an ensemble for the true physical system along to the course of a dynamical trajectory. This subsequently allows for the formation of a kriging model that will be used in a purely interpolative context.
2. Single-point calculations are performed on each sample and the resultant ab initio electron density is partitioned by QCT software. The multipole moments of the topological atoms
are subsequently obtained. This allows a kriging model to evaluate a functional form corresponding to the evolution of the various multipole moments as a function of conformation.

3. The sample set is split into a nonoverlapping training set and test set. The training set is utilised for training of our kriging models, i.e. these are the points that the kriging function must pass through. The test set is not trained for, but is used after the construction of the kriging models to evaluate the errors associated with their predictions.

4. A kriging model is built for each multipole moment of each atom, which allows for the generation of a smooth interpolative function, mapping the evolution of the multipole moment against the conformational parameters of choice. By use of particle swarm optimization, we optimize our kriging parameters, \( \{ \theta, p \} \) to obtain an optimal kriging model.

5. The kriging models are assessed by making them predict the multipole moments for each atom in a system whose configuration has not been used for training of the kriging model. However, we do possess the ab initio multipole moments for this configuration. As such, we evaluate the energy associated with all 1–5 (i.e. two nuclei separated by four bonds) and higher \((1–n, n > 5)\) order interatomic electrostatic interactions as given by the predicted multipole moments from the kriging models, and the equivalent energy as given by the (exact or original) ab initio atomic multipole moments. We subsequently assess the deviation of the kriging predictions from the ab initio electrostatic energies.

The choice of 1–n \((n \geq 5)\) interactions over the conventional 1–4 serves a twofold purpose: (i) avoiding any potential divergence in the electrostatic energy between two atom-centered multipole moment expansions, (ii) avoiding any issues from a coupling of torsional and electrostatic energetics. In other work from this lab, to be published soon, we show that short-range electrostatic energy 1–n \((n < 5)\) can be satisfactorily kriged. Nonelectrostatic energy contributions can be calculated within the QCT context and again adequately kriged, a result that will be published elsewhere.

Note that the electrostatic energy \(^{38}\) is the final arbiter in the validation of the kriging models, rather than the atomic multipole moments themselves, which are the kriging observations. The molecular electrostatic energy is calculated by a well-known multipolar expansion \(^{39}\) involving a multitude of high-rank atomic multipole moments \(^{41}\). This expansion is truncated to quadrupole-quadrupole \((L = 5)\) and rank-equivalent combinations (dipole-octopole and monopole-hexadecapole). Second, the interatomic contributions to the total molecular electrostatic energy are limited to 1–5 and higher.

Whilst somewhat indirect, the validation through energy rather than multipole moment has a twofold purpose. Primarily, the energy is the quantity that will be used for dynamical simulations, and so is the ultimate descriptor that we wish to evaluate correctly. Second, the alternative would be to assess the predictive capacity of each individual kriging model. For a system with any sizeable number of atoms, where each atom has 25 individual multipole moment kriging models, the data analysis obviously becomes overwhelming. However, this analysis is unnecessary owing to the uniqueness of the Taylor expansion from which the multipole moments arise. Since the electrostatic energy is computed from two such unique series, then if the electrostatic energy is correctly predicted, the multipole moments must also be correct by deduction. Note that this consideration is valid for a single atom-atom interaction.

In order to gauge the models’ validity, we plot a graph colloquially termed an “S-curve” owing to its typical sigmoidal shape but of course it is really a cumulative distribution function. The S-curve plots the absolute deviation of the predicted energy from the ab initio energy, predicted from the ab initio multipole moments, after having evaluated the multipole moment interactions. Put more precisely, the predicted multipole moments form an energy that is subtracted from an energy obtained from the ab initio moments. Then the absolute value of this difference is taken. Hence compensation of errors is not allowed because first the difference is taken and then the absolute value. These energetic deviations are plotted against the percentile of test configurations that fall on or below the given energetic deviation.

Our aim is then twofold: the first is to reduce the upper tail of the sigmoid such that the 100th percentile error is convergent at as low an error as possible. This corresponds to the predictions being uniformly good across the test set with no spurious predicted interactions. Our second aim is to shift the S-curve as far down the abscissa (i.e. to the left) as possible, which ensures the average error associated with our predictions is as low as possible. The first goal is achieved by certifying that the training points used for the construction of the kriging models form the boundaries of configurational space with respect to our sample set. This boundary checking guarantees that the kriging model is being asked to interpolate from training data. Boundary checking has not yet been implemented, but we propose a simple means by which this could be accomplished. The initial geometry from which we start sampling may be approximated as occupying the center of the sampling domain. The Euclidean metric in the (3N-6)-dimensional conformational space decides which sample points form the boundaries of the sampling domain by their distance to the initial geometry in configurational space. The second goal is attained by consistent improvement of the kriging engine, and making sure that the test points are uniformly close to the training data, allowing for efficient interpolation. Note that we do not necessarily choose test points that are close to the training points. In other words, the training and test sets are constructed independently. By ensuring that the training data is uniformly distributed throughout the sampling domain, the average “distance” between an arbitrary test point and a training point will equal that between some other arbitrary test point and another training point. This guarantees no spurious predictions in under-trained regions of configurational space. Of course, it is prudent to invoke some form of importance sampling, which yields a greater sampling density in more “important” regions of configurational space, but this issue has not been explored.
We work on the tetrose diastereomers erythrose and threose, the smallest carbohydrates that adopt open chain and furanose forms. The particular conformations studied are given in Figure 3. Energetic minima were provided by Prof Alkorta, who had previously conducted a PES scan of these species, and are reported more thoroughly elsewhere. Briefly, the origin of the ALF is the nuclear frame. This procedure makes sure that the kriging focuses on the variation of atomic multipole moments within the molecule. Otherwise, when referring to the global frame (rather than the ALF), the three components of an atomic dipole moment, for example, vary upon rigid rotation of the whole molecule. Training for such a variation is useless. The same principle applies to atomic multipole moments of rank $\ell \geq 2$. The details of the atomic local frame chosen for our work are outlined elsewhere.

The issue of coordinate frames needs further clarification because the Cartesian MD frame coordinate system is not the same as the local coordinate frame within which we have evaluated the atomic multipole moments (ALF). Prior to invoking a kriging model for the evaluation of multipole moments, the

![Figure 3. Comparison of erythrose and threose in the open chain (topological atoms and molecular graph) and in the ring configurations of $\alpha$ and $\beta$ furanose (in traditional ball-and-stick representation). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.](Image 443x759 to 556x785)
conformational state of the system must be converted from a Cartesian frame of reference to an ALF. From here, atomic multipole moments can be evaluated corresponding to the current state of the system. Previous work has derived the forces arising from the interactions of atomic multipole moments within the Cartesian frame of reference,[54] which requires the partial derivatives of the multipole moments with respect to the ALF degrees of freedom. These terms have an analytical functional form, and so computation of the forces in the Cartesian frame of reference can be performed explicitly. These forces can subsequently be utilized by the standard MD procedure.

An embryonic workflow has been integrated within the MD package DL_POLY 4.0. Currently, an atomic kriging model is loaded into memory and the relevant data stored, followed by removal of the kriging model from memory. In this way, the dynamic memory requirements have not yet exceeded roughly 200 Mb, and are thus well within the capabilities of modern computational resources. Of course, memory management is crucial to the speed of the proposed methodology, and so will require a great deal of fine-tuning. However, much speedup can undoubtedly be accomplished by a number of techniques, e.g. caching of regularly used quantities and parallel implementation.

Finally, it is too early to extensively comment on the computational cost of the current approach. It would be naïve to directly compare the flop count of the current force field with a traditional one without appreciating that (i) extra non-nuclear point charges are needed to match the accuracy of multipole moments and the former propagate over long range, (ii) multipolar interactions drop off much faster than 1/r, depending on the rank of the interacting multipole moments, which depends on the interacting elements themselves (see extensive testing in the protein crambin[55]), (iii) the efficiency of the multipolar Ewald summation[56] that is being implemented in DL_POLY 4.0, (iv) the dominance of monopolar interactions at long range (vast majority of interactions) and the (v) outstanding fine-tuning of the kriging models at production mode. The current force field may well be an order of magnitude slower than a traditional force field. This estimate and the fact that the current force field contains electronic information invite one to compare its performance with on-the-fly ab initio calculations instead.

Results and Discussion

Single minimum

The lowest energy conformer for each system was chosen as an input structure for sampling. S-Curves were subsequently generated for each of these training sets by the methodology outlined in the previous section. The S-curves are given in Figure 4, and the accompanying mean errors in Table 1.

The first point to notice is that the open chains for both erythrose and threose are modeled by kriging to a significantly better standard than the furanose forms. We may, however, immediately attribute this to the number of 1-5 and higher interactions occurring in these systems. For both open chains, 25 interactions are required to be evaluated for comparison to the energies produced from the ab initio multipole moments. The numbers of interactions requiring evaluation for the furanose forms comes to 39, which is virtually twice the amount evaluated in the open chain forms. We would subsequently expect a proportional relationship between the number of interactions required for evaluation and the mean error attributed to the kriging model. Whilst we see this to be roughly true when comparing the errors on the threose open and ω-furanose forms, the errors appear disproportionately higher for the other systems.

Kriging is an interpolative technique, and so is not suited for extrapolation. However, we point out that the kriging engine is still predictive for extrapolation- in this case, the prediction falls to the mean value of the function. Obviously this is not ideal for highly undulatory functions. However, considering how the atomic multipole moments do not fluctuate over vast ranges, the mean will often represent a respectable prediction to the function value. The kriging model can be refined in an iterative fashion, whereby extrapolation points are added to

| Table 1. Mean errors associated with the S-curves given in Figure 4. |
|-----------------|-----------------|-----------------|
| Mean error (kJmol⁻¹) | Open chain      | α-Furanose      | β-Furanose      |
| Erythrose        | 0.27            | 1.32            | 1.30            |
| Threose          | 0.34            | 0.83            | 1.47            |

Figure 4. S-curves corresponding to all erythrose (left) and threose (right) systems studied. The open chain forms are systematically better predicted than the corresponding furanose forms. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
studies which attempt to implement neural networks to pre-
machine learning techniques, and have been encountered in
mational space. In fact, these problems are ubiquitous to
which is currently being explored.

Within our methodology, the iterative protocol is a technique
field of machine learning. Whilst not currently implemented
the training set. This is a commonly used technique in the
field of machine learning. Whilst not currently implemented
within our methodology, the iterative protocol is a technique
which is currently being explored.

Regardless of the problem encountered in the above, we
see that it is easily remedied by strategic sampling of confor-
mational space. In fact, these problems are ubiquitous to
machine learning techniques, and have been encountered in
studies which attempt to implement neural networks to pre-
dict a PES. In this field, the problems have been solved to
some extent by making the prediction engine issue a warning
to the user that a point is being predicted which lies outside
of the training set. This proves to be advantageous as the user
may then recognize that the point should be included within
the training set as it obviously lies within an accessible portion
of conformational space. The point can then be included
within the training set, and one can generate refined models
by undergoing this process iteratively.

Training set dependency

We start by discussing the effects of increasing the training set
size on the prediction error for a kriging model. For this pur-
pose, we use the erythrose open chain system owing to its
higher conformational flexibility, which we assume amplifies
the effects of training set size. Kriging models were generated
for this system with training sets ranging from 700 to 1500
sampling points, in increments of 100. The same test set (of
200 points) was reserved for prediction by all models. The
S-curves for this are given in Figure 5.

As expected, the prediction errors of the S-curves in Figure
5 systematically decrease as the training set size is increased.
In other words, the S-curves move to the left with increasing
training set size, although this is not true for all parts of the
S-curves because they clearly intersect in many places. Overall
the uniform increments of 100 in training set size are not
matched by equal uniform strides of improvement in S-curve
shape and position. An alternative way to gauge the improve-
ment in prediction with increasing training set size is monitor-
ing the average prediction error for each S-curve. This value
cannot be read off for an S-curve in Figure 5 but can be easily
calculated.

Figure 6 plots the average prediction error for each S-curve
against increasing training set size: red for “Old FEREBUS” and
blue for “New FEREBUS,” a development version of our kriging
engine, which differs in a number of ways to the “Old
FEREBUS.” We include both “Old FEREBUS” and “New FEREBUS”
data to establish whether any functional forms of average pre-
diction error against increasing training set size are conserved
with respect to improvements in the engine. The “Old
FEREBUS” data show a plateau in the average error (left pane)
at a training set size of about 1200, after an initial decrease in
this error. This plateau would be rather problematic, as it
implies some maximum efficiency of the kriging engine,
beyond which there is no reward for an extension of the train-
ing set. However, this is not the case for the New FEREBUS
data (right pane).

Learning theory states that for a machine learning method
of this type (kriging), the mean prediction error should
decrease asymptotically toward zero, with functional form
\[ A + B/n \] or \[ C + D/\sqrt{n} \], where \( n \) is the training set size, and \( A, B, \]
\( C, \) and \( D \) are fitted constants. Figure 6 plots these asymptotes,
where \( (A_{\text{old}} = 0.105; B_{\text{old}} = 348.11) \) and \( (C_{\text{old}} = -0.293; \)
\( D_{\text{old}} = 22.06) \), as determined by regression analysis against the
Old FEREBUS data, each with \( R^2 \) coefficients of 0.93. Similarly,
for the New FEREBUS data, constants of \( (A_{\text{new}} = 0.097; \)
\( B_{\text{new}} = 293.38) \) and \( (C_{\text{new}} = -0.193; \) \( D_{\text{new}} = 18.60) \) were
obtained. These fitted asymptotes both possess \( R^2 \) values of
0.98. As such, we conclude that the decay of the mean predic-
tion error of our machine learning method possesses, as yet,
inconclusive functional form.

The results in Figure 6 are consistent with the behavior seen
in similar interpolation methods: for an infinite training set
size, the mean prediction error will asymptote to zero. How-
ever, for the methodology to remain computationally feasible,
some finite training set size will of course be required.

So, for example, we find that for a mean prediction error of
0.3 \( \text{kJ mol}^{-1} \), the training set would require about 1450 sam-
ple points for either functional form taken as the decay of the
prediction error, i.e. \[ A + B/n \] or \[ C + D/\sqrt{n} \].

A comment on the nature of the average error is in place
here. In principle, the prediction error consists of the sum
of the estimation error and the approximation error. From
learning theory, one expects the estimation error only to go to
zero. The bias–variance decomposition of a learning algo-
ri thm’s error also contains a quantity called the irreducible
error, resulting from noise in the problem itself. This error has
been investigated some time ago in the context of tests on
kriging of ethanol multipole moments and is caused by the
small noise generated by the integration quadrature of the
atomic multipole moments. Second, any bias caused by an
inherent error in the \( ab \) \initio method used, compared to the

![Figure 5. S-curves for erythrose open chain at various training set sizes. Note the progression of the S-curves towards the lower prediction errors as the training set size increases. However, owing to the logarithmic abscissa, this does not correspond to a uniform enhancement of a kriging model given a consistently larger training set size. (Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.)](Image 5)
best method available (e.g. CCSD(T) with a complete basis set), is not relevant in our error considerations. The reason is that we always assess the performance of kriging training against the (inevitably approximate) ab initio at hand, which we refer to the source of “the” ab initio data.

Multiple minima

The amount of conformational space available to molecular systems reaches levels which are entirely unfeasible for systematic sampling as the number of atoms increases. As such, it becomes all the more prudent to obtain an efficient sampling scheme for our purposes. As we have mentioned, our sampling methodology is limited to local conformational exploration about some given input geometry, since the PES about that point is approximated as a harmonic well. As such, to thoroughly explore conformational space, our methodology requires the usage of a number of such starting geometries.

Then, the molecular PES is approximated by a number of harmonic wells. If the input geometries are sufficiently close to one another, the wells will overlap, and the PES may be explored seamlessly. Non-equilibrium normal mode conformational sampling has also been demonstrated in a recent publication.[59] This advance will facilitate a more thorough sampling of these higher energy parts potential energy surfaces. The validation of this methodology is presented in Part B of Supporting Information.

For the open chain form of erythrose, 174 energetic minima were found by an exhaustive search of conformational space. Figure 7 plots the S-curves obtained for samples which have been generated from different numbers of up to 99 minima. The S-curves display increasingly poor prediction results as the number of starting minima increases. The actual mean errors for these S-curves are summarized in Table 2. This trend has a logical interpretation. As the number of seeding structures increases, the sampled conformational space grows in size. Given a fixed kriging model size, the sampling density therefore decreases. The kriging model then deviates from the true analytical function, and the results from predictions deteriorate.

Of course, thorough sampling of conformational space is an issue for parameterizing any force field, and by no means one that is resultant from our methodology. We may overcome this issue in two ways. The first is the ongoing improvement of our kriging engine to deal with larger training sets comprising more molecular configurations. The second is by undertaking sampling with only a subset of the energetic minima that are available. This is all the more valid an approach if most of the minima are very high in energy relative to the lowest-lying minima. These

<table>
<thead>
<tr>
<th>Number of minima</th>
<th>1</th>
<th>20</th>
<th>40</th>
<th>60</th>
<th>80</th>
<th>99</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean prediction error (kJ mol⁻¹)</td>
<td>0.27</td>
<td>0.85</td>
<td>0.9</td>
<td>1.23</td>
<td>1.30</td>
<td>1.37</td>
</tr>
</tbody>
</table>

Note the “bunching” of prediction error when 60 energetic minima or higher are used as seeds for the conformational sampling.
regions of conformational space will be accessed very infrequently during the course of a MD simulation, and so may be sampled much more coarsely. This selective sampling is quite readily employed, and has been discussed at length in the literature. For example, Brooks and Karplus[60] found that a comprehensive sampling of conformational space for bovine pancreatic trypsin inhibitor could be achieved by evolving only the lowest frequency normal modes of motion. Needless to say, this is readily accomplished by our sampling methodology.

Conclusion

We have demonstrated that the atomic multipole moments of a set of carbohydrates are amenable to the machine learning technique kriging. Whilst this has been done in the past for a variety of chemical species including naturally occurring amino acids, this is the first foray into the field of glycobiology. Kriging is able to capture the conformational dependence of the multipole moments and make predictions, such that the error in the electrostatic energy relative to that derived from \textit{ab initio} data is encouraging, given the popular aim is to obtain errors below 4 \textit{kJ mol}^{-1}. Indeed, the presented methodology is immediately extensible to any term arising in an energetic decomposition of a system. If some quantity is conformationally dependent, then the dependence can be modelled by kriging. As such, an entire force field can be parameterized by the current methodology, reproducing \textit{ab initio} quantities for use in classical MD. This route is preferable to the computationally intensive approach of \textit{ab initio} MD.

Acknowledgment

The authors thank the BBSRC for funding a DTP PhD studentship.

Keywords: quantum theory of atoms in molecules  carbohydrate quantum chemical topology conformational sampling kriging electrostatics multipole moments

How to cite this article: S. Cardamone, P. L. A. Popelier. J. Comput. Chem. 2015, 36, 2361–2373. DOI: 10.1002/jcc.24215

Additional Supporting Information may be found in the online version of this article.


Realistic Sampling of Amino Acid Geometries for a Multipolar Polarizable Force Field

Timothy J. Hughes,[a,b] Salvatore Cardamone,[a,b] and Paul L. A. Popelier*[a,b]

The Quantum Chemical Topological Force Field (QCTFF) uses the machine learning method kriging to map atomic multipole moments to the coordinates of all atoms in the molecular system. It is important that kriging operates on relevant and realistic training sets of molecular geometries. Therefore, we sampled single amino acid geometries directly from protein crystal structures stored in the Protein Databank (PDB). This sampling enhances the conformational realism (in terms of dihedral angles) of the training geometries. However, these geometries can be fraught with inaccurate bond lengths and valence angles due to artefacts of the refinement process of the X-ray diffraction patterns, combined with experimentally invisible hydrogen atoms. This is why we developed a hybrid PDB/nonstationary normal modes (NM) sampling approach called PDB/NM. This method is superior over standard NM sampling, which captures only geometries optimized from the stationery points of single amino acids in the gas phase. Indeed, PDB/NM combines the sampling of relevant dihedral angles with chemically correct local geometries. Geometries sampled using PDB/NM were used to build kriging models for alanine and lysine, and their prediction accuracy was compared to models built from geometries sampled from three other sampling approaches. Bond length variation, as opposed to variation in dihedral angles, puts pressure on prediction accuracy, potentially lowering it. Hence, the larger coverage of dihedral angles of the PDB/NM method does not deteriorate the predictive accuracy of kriging models, compared to the NM sampling around local energetic minima used so far in the development of QCTFF.

Introduction

The rapid but accurate evaluation of potential energy for biomolecular simulation continues to be a challenge. Next generation force fields, which could eventually replace the traditional force fields, continue to be developed. Among the former are AMOEBA,[1] XED,[2] SIBFA,[3] and ACKS2,[4] which all advocate multipolar electrostatics,[5,6] absent in classical architectures.[7,8] The Quantum Chemical Topological Force Field (QCTFF)[9,10] shares this approach to improved electrostatic energy prediction but, on top of this, introduces machine learning to handle electron density fluctuations in response to changes in nuclear configuration. QCTFF aims at capturing the end result of this polarization process rather than the process itself. The machine learning models that QCTFF depends on need to be properly trained with a sufficient number of configurations, but perhaps more importantly, with relevant configurations. The work presented here deals with this problem, and does so in the context of real protein structures.

Machine learning focuses on algorithms that can learn from data, in this case properties (multipole moments and energies) of topological atoms. Machine learning proposes computational methods that generate predictive models that map an output variable to a set of input variables. Models are then built through a training procedure using a set of input values with known output. QCTFF, which continues to be developed in our lab, is an innovative approach to predicting the energy of a molecular system much faster than first principle calculations can. For that purpose, QCTFF captures atomically partitioned first principle information of the system trained for. QCTFF achieves this by relying on a machine learning method called kriging,[11–13] which is increasingly being used[14–19] in the community of force field and potential design.

Traditional force fields approximate energy through bonded and nonbonded contributions that incorporate often loosely defined atom types with their own set of experimentally or computationally obtained parameters. QCTFF operates outside this traditional framework: its architecture does not distinguish between bonded and nonbonded interactions, and atom types do not need to be defined. Instead, QCTFF focuses directly on how atoms interact, allowing for a spectrum of covalency rather than a bonded/nonbonded dichotomy. QCTFF maps atomic properties (the output variables) to molecular coordinates (a set of input variables) using kriging. Therefore, a QCTFF atom will be endowed with a number of kriging
models, each describing how an atomic property changes as a function of the coordinates of the molecular system.

To build a QCTFF kriging model, example molecular geometries must be obtained to train the model. QCTFF development targets the simulation of biomolecules, in particular proteins, hence amino acids are molecules of key interest. When sampling amino acid geometries as input for kriging models, the sampled geometries must include all the conformations that one may reasonably expect to occur during the simulation of a protein. Our current paradigm for the sampling of molecular geometries is to use a NM sampling approach. To do this, a small number of stationary points on the potential energy surface of a given molecule of interest are located, and the NM at each stationary point (or local energy minimum) are calculated. Energy is then put randomly into the NM to distort the molecule, and “snapshots” are taken to obtain distorted geometries. The minimum energy conformations of all 20 naturally occurring amino acids have been reported in a comprehensive study, all obtained at the same level of theory. Kriging models built from NM sampled geometries have been used to predict successfully the atomic multipole moments of a range of molecules. These include small organics, amino acids, and hydrogen bonded dimers. Recently, the electronic kinetic energy of QCT atoms (see Quantum Chemical Topology section) has been successfully incorporated into kriging models for methanol, NMA, glycine and triglycine. Intra-atomic terms such as the (electronic) kinetic energy are not explicitly incorporated in classical force fields but to gain an appreciation of chemical phenomena, such as steric hindrance, intra-atomic terms have been proven important and therefore should be included in QCTFF. Some interesting work quantifies the steric effect, still within QCT, but in the context of experimental electron densities, conceptual DFT, and energy decomposition analysis. The only other alternative sampling approach investigated draws snapshots from a molecular dynamics simulation, which has been done for liquid water. In the current work, a third sampling method is investigated, one that is pivotal for a realistic sampling of amino acid conformations and one that incorporates experimental information (X-ray structures).

Amino acids are typically described as consisting of two units: a back bone and a side chain. The conformational preference of the backbone unit is dictated by the secondary structure of the proteins and is well understood. The dihedral angles denoted $\Phi$ and $\Psi$ describe the backbone using Ramachandran plots. These plots relate the values of $\Phi$ and $\Psi$ to a particular secondary structure. Different amino acids display preferences for different regions of the Ramachandran plot, and a thorough investigation of the preferences for all 20 naturally occurring amino acids has been performed. The side chain of an amino acid may exist as a number of different rotamers depending on the side chain dihedrals. Extensive work has been undertaken by other groups to understand the relative populations of the different rotamers occupied by each amino acid, and this has led to a number of rotamer libraries being constructed. A rotamer library is a comprehensive guide, drawn from molecular dynamics simulation or protein crystallography, detailing the statistical populations and frequencies of the dihedral angles adopted by amino acid side chains. These libraries may then be used to predict, build, design and solve new protein structures. Torsional energy terms are so important that they receive special attention in force field design, see Ref. [42] for a recent example.

Normal modes sampling has proved successful at sampling conformational space around an input energetic minimum or stationary point. However, one must consider whether the gas phase minimum energy geometries of an amino acid accurately mimic the amino acid structures found in proteins. We note that, in more general terms, the biases induced by datasets that are restricted to stationary or only little deformed structures were also discussed within the context of DFT. It is accepted that amino acids and polypeptides have an intrinsic propensity for specific molecular configurations, and that this preference can differ depending on whether the amino acid exists in a folded protein tertiary structure or a disordered, solvated state. Ramos and coworkers performed ab initio calculations on all 20 natural amino acids using both gas phase and PCM solvation. Of the 323 chemical bonds and 469 angles present, they found mean unsigned errors of less than 0.02 Å and 3° between the PCM and gas phase bonds and angles, respectively. However, the environment of a globular protein is different to that of a hydrated polypeptide due to a number of factors such as intrareside hydrogen bonding and steric considerations that have an effect on the amino acid conformation.

The work of Jha et al. clearly shows the effect of the environment on the backbone angles $\Phi$ and $\Psi$. They compared the geometric preferences of all 20 amino acids using data from two protein coil libraries: one including residues in structural motifs, and the other only those residues in disordered sections of the proteins. The ratios of structures found in the $\beta$-sheet, PPII, and $\alpha$-helical regions were clearly different between the two libraries. To further demonstrate the effect of environment on the structural preferences of amino acids, the distribution of structures obtained from both coil libraries also differed significantly from those obtained experimentally for the central residue of Gly-X-Gly tripeptides (where X is a naturally occurring amino acid). It has been shown, both experimentally (using NMR J couplings) and computationally, that disordered amino acid residues favor specific regions of the Ramachandran plot (typically $\beta$-sheet and PPII regions) in contrast to the conformational populations found in ordered protein secondary structures. It has also been shown that the side chain rotamer preference of an amino acid is related to the secondary structure of the polypeptide in which it resides, and this relationship between environment and structure has been used successfully in rotamer libraries to predict side chain conformations. In the long term, these results imply that gas phase energy minima of single amino acids used to sample geometries from, are insufficient to sample all important chemically relevant structures.

The efficient sampling of molecular geometries is a challenging problem due to the rapid increase in the available conformational space as molecules grow in size. A systematic search
of conformational space to find low energy structures is impractical and inefficient. A number of efficient approaches have been presented in the literature including the use of molecular dynamics, Monte Carlo, transition path sampling, and metadynamics. Additionally, fragment based approaches may be used to improve a systematic approach by reducing the number of conformations searched through elimination processes. An example of such an approach is that of Luo and coworkers, where, by fragmenting the Gly-Tyr-Gly-Arg tetrapeptide, they reduced 19.6 billion possible candidates for the global minimum conformation down to only 5760.

An alternative to computational sampling approaches for finding important amino acid geometries is to source them from protein crystal structures. Unfortunately, crystal structures cannot be used directly as input into kriging models for several reasons. First, only heavy atoms are detectable by X-ray crystallography and so the hydrogen atom coordinates are dependent upon the refinement process used. Second, removing an amino acid from a crystal structure breaks the peptide bonds at either end of the backbone, which drastically changes the chemical environment and results in incomplete valence of the terminal atoms. Therefore, some post-Protein Databank (PDB)-extraction modifications to the sampled amino acids are required before input to QCTFF. Thirdly and finally, the resolution of the atomic coordinates varies from one crystal structure to another, and sometimes unrealistic bond lengths and angles may be present within a crystal structure. To address the above concerns, a novel sampling approach is presented here. This approach samples amino acids from the PDB, relaxes bond lengths, and valence angles by an NM method while preserving the dihedral angles, and then performs nonstationary NM sampling around each sampled amino acid. This approach is termed PDB/NM and the details of both sampling approaches are explained in the following sections.

Background and Methods

Because many of the technical points concerning QCTFF have been described in detail in previous work of our lab, we only give a brief overview of the key concepts here. A comprehensive introduction to kriging and how it features in QCTFF is given in Ref. [19] while Refs. [24,25] provides the most up-to-date detail on the overall training procedure of QCTFF, now called GAIA. Additional descriptions of the machine learning method are also provided in Refs. [17,26].

Quantum chemical topology

Underpinning the development of QCTFF is Quantum Chemical Topology (QCT), which embraces all work in quantum chemistry that uses the topological language of dynamical systems (e.g., attractor, basin, homeomorphism, gradient path, separatix, critical points). QCT contains the “quantum theory of atoms in molecules” as a special case where this topological language is applied to the electron density $\rho$ and its Laplacian. A topological atom $\Omega_A$ is a bundle of gradient paths (i.e., trajectories of steepest ascent through $\rho$), terminating at a maximum critical point, which typically coincides with the nucleus A. Topological atoms are defined in a parameter-free manner, and they are nonoverlapping and sharply bounded (at the inside of the molecule) by so-called interatomic surfaces.

It is a good idea to expand the $1/r_2$ expression occurring in the equation for the Coulomb energy between two electron densities. A popular and compact expansion introduces spherical harmonics, which in turn lead to atomic multipole moments. Multipole moments are able to describe the anisotropy of the electron density, in contrast to (isotropic) point charges used by popular force fields such as AMBER and CHARMM. The charge of an atom is the zero-order term of the multipolar expansion, and it is only by including higher-order terms that the anisotropy of the electron density is described. There is considerable evidence, as collected in a recent review, of the advantages of multipolar electrostatics over point charges. QCTFF incorporates multipolar electrostatics, and in the current work it is the atomic multipole moments that are the topological property of interest, that is, they are the output that kriging is tasked to predict.

The Coulomb interaction between two topological atoms $\Omega_A$ and $\Omega_B$ is given by

$$E_{AB}^{\text{Coul}} = \sum_{l,m} Q_{lm}^{A} T_{lm}^{AB} Q_{lm}^{B}$$

where $Q_{lm}^{A}$ is a multipole moment and $T_{lm}^{AB}$ is the interaction tensor between two multipole moments. A convenient concept when dealing with the electrostatic interaction between two multipole orders of $l_A$ and $l_B$ is the interaction rank, $L$, given by

$$L = l_A + l_B + 1$$

It has been shown that interaction rank $L=5$ provides a satisfactory description of the electrostatics acting in system. Note that $L=5$ requires all atomic multipole moments up to and including hexadecupole (fourth order multipole moments, $\ell=4$) to be calculated, resulting in 25 multipole moments for each atom.

Atomic properties other than multipole moments may be obtained from QCT. The interacting quantum atoms (IQA) method is a well-developed topological energy decomposition scheme based on the calculation of the exact nonexpanded topological Coulomb energy. IQA decomposes a molecular system in a combination of both intra-atomic (“self”) and interatomic energy terms. Details of the decomposition scheme are beyond the scope of this article but QCTFF is currently incorporating the non-Coulomb terms by the same kriging treatment as the atomic multipole moments in the current work.

The atomic local frame and kriging

QCTFF uses kriging, also known as Gaussian process regression, which is a method of capturing the changes in atomic multipole moments as a function of molecular geometry. A detailed description is provided in earlier work so only a brief description is provided here. As the coordinates of
an atomic system evolve, for example when bonds stretch and angles bend, the topological properties of the atoms involved will change, e.g. example their atomic charges (or monopole moments). Using kriging, it is possible to build models capable of predicting changes in an atomic property by evaluating the molecular coordinates. In the present work, kriging models are built for the first 25 atomic multipole moments (up to, and including, hexadecapole moment) of each atom in the amino acids alanine (Ala) and lysine (Lys). By treating the atomic multipole moments in this way, both polarization and charge transfer effects are captured.

A chemical system may be defined by a minimum of \(3N-6\) internal coordinates. In the language of machine learning, the \(3N-6\) coordinates around an atom are referred to as features, and it is these features that a multipole moment is mapped to. In QCTFF an atomic local frame (ALF) is defined to describe the \(3N-6\) coordinates around a central atom. Consider a central atom, denoted A. First, the Cahn–Ingold–Prelog rules are used to determine the two atoms of highest priority bonded to A, and these atoms are termed X and Y in order of priority. The distances \(R_{AX}\) and \(R_{AY}\), and the angle \(\theta_{AXY}\) define the three ALF coordinates. Subsequently a right-handed coordinate system is stabilized using the XAY plane. All other atoms in the system can then be described by three polar coordinates, \(R_{AX}, \phi_{AX},\) and \(\theta_{AX}\). One therefore obtains \(N-3\) sets of three spherical polar coordinates each, which combined with the aforementioned ALF coordinates make up the \(3N-6\) coordinates required, that is, \(3(N-3)+3 = 3N-6\).

Returning to kriging, the change in a given multipole moment is smooth with respect to a change in the ALF coordinates. Therefore it is safe to interpolate the atomic multipole moments of an unknown molecular geometry existing inside a set of known geometries. Kriging is used to build models capable of accurate interpolation of the atomic multipole moments by mapping an input (nuclear coordinates) to an output (a multipole moment). To achieve this, a training set of molecular geometries, which is strictly not part of the training set. For each test molecule, we predict all the multipole moments of all the atoms in the system, and then calculate all electrostatic interactions between atoms separated by a minimum of three covalent bonds (i.e., 1, 2 and \(n > 3\) interactions). Each predicted interaction energy (between two atoms A and B) is then compared to the original (i.e., not kriged) interaction energy obtained from the original (i.e., not kriged) atomic multipole moments. Then the errors of all the aforementioned interactions within one molecular geometry are summed. The absolute value of this summed error (for each test geometry) will be plotted against percentile (i.e., % of test geometries) to obtain a called S-curve. Each point on such a curve corresponds to this final absolute error (i.e., \(|\Delta E_{\text{system}}|\) in eq. (9)). The S-curve will be described later when one is obtained. The complete description of errors just mentioned is expressed is eq. (9),

\[
|\Delta E_{\text{system}}| = |E_{\text{system}}^{\text{original}} - E_{\text{system}}^{\text{predicted}}| = \sum_{AB} E_{AB}^{\text{original}} - \sum_{AB} E_{AB}^{\text{predicted}}
\]

where \(x'\) and \(x'\) are training points composed of \(d\) features. The parameters \(\theta_h\) \((0 \leq \theta_h \leq 2)\) describe the importance of each feature \(h\) and may be written as the \(d\)-dimensional vectors \(\theta\) and \(p\). A large value of \(\theta_h\) corresponds to a feature being highly correlated to the output multipole moment. The parameter \(p_h\) describes the smoothness of the function, and is often close to 2.

A second crucial concept underpinning kriging is the so-called concentrated (or reduced) log-likelihood function \(\hat{L}\), defined as

\[
\hat{L}(\theta, p) = -\frac{\bar{n}}{2} \log(\hat{\sigma}^2) - \frac{1}{2} \log(|R|)
\]

where

\[
\hat{\sigma}^2 = \frac{(y - \hat{\mu})^T \hat{R}^{-1} (y - \hat{\mu})}{\bar{n}}
\]

and

\[
\hat{\mu} = \frac{1^T \hat{R}^{-1} y}{1^T \hat{R}^{-1} 1}
\]
PDB sampling method

PDB sampling is performed by the in-house (scripting) code MOROS and is used to extract all seed geometries of a particular amino acid from a set of crystal structures. A list of the 260 PDB crystal structure codes sampled from is provided in Part A of the Supporting Information. Hydrogen atoms were added to all protein crystal structures using the HAAD code of Li et al.\(^7^7\) The HAAD algorithm was developed to add accurately hydrogen atoms by analyzing the positions of nearby heavy atoms, following the basic rules of orbital hybridization and through optimization of steric and electrostatic parameters. HAAD was found to outperform the popular software CHARMM and REDUCE\(^7^8\) with the RMSD of predicted hydrogen atom positions decreased by 26% and 11%, respectively, when compared to high resolution X-ray and neutron diffraction structures. MOROS returns as output “capped” amino acids meaning that \(\text{H}_3\text{CC}(=\text{O})\) and \(-\text{N}(\text{H})\text{CH}_3\) are appended at the N and C termini of the sampled amino acid, respectively. These atoms are included so that the peptide bonds remain intact, and thereby yield a more realistic representation of an amino acid while present in a protein. The capping groups are built by extracting the atomic coordinates from the residues preceding and following the residue of interest. Figure 1 shows the atoms extracted by MOROS including the amino acid of interest (blue box), and also atoms that make up the caps (red box).

In preparation for nonstationary NM treatment, the sampled amino acid geometries are then allowed to partially geometry-relax, that is under the restriction of fixed dihedral angles. This stage is important as it removes some of the outlying bond lengths originally present due to the poor quality crystal structure resolution.

The next step in PDB sampling is to perform a frequency calculation on each amino acid geometry, by first obtaining the Hessian of the potential energy on that point of the surface, for input for the non-stationary NM sampling of the geometry. A choice must be made regarding the number of PDB-sampled amino acid geometries to use as input for non-stationary NM, as this choice influences the number of geometries sampled using NM. This choice is investigated in Results and Discussion section, and unless otherwise stated, 300 random PDB-sampled amino acid geometries are input to the nonstationary NM. The combined PDB and nonstationary NM sampling method will henceforth be referred to as PDB/NM.

“Normal Modes” sampling

Typical normal mode analysis is conducted at an energetic minimum (or stationary point) on the molecular potential energy surface. However, the mathematics leading to NM does not restrict their use only at stationary points. A simple generalization of the derivation of the molecular NM enables their evaluation at nonstationary points on the potential energy surface. This derivation is provided in Part B of the Supporting Information. In the following, we present a conformational sampling methodology, which uses these “non-stationary point normal modes” as a means for distorting a molecule, that is, sample its configurations. By diagonalization of the mass-weighted Hessian, \(\mathbf{H}\), the frequency of each of the \(N_{\text{vib}} = 3N - 6\) NM is evaluated. These \(N_{\text{vib}}\) NM are orthogonal and form a complete basis within which internal molecular motions can be described. With the mass-weighted force vector, \(\mathbf{F}\), a set of \(N_{\text{vib}}\) harmonic equations of motion is obtained. These equations of motion allow us to distort the molecular geometries, and perform a sampling of conformational space.

We now discuss the computational means utilized to obtain the various parameters required to evolve the NM. This subsequently permits us to obtain a set of geometries we consider representative of realistic vibrational states of a molecular system. What follows is a brief paraphrase of the excellent explanation given by Ochterski\(^7^9\) Beginning with the transformation from the mass-weighted Cartesian coordinates, \(\mathbf{q}\), to the set of \(N_{\text{vib}}\) internal coordinates, \(s\), we construct the \(3N \times 3N\) transformation matrix, \(\mathbf{D}\), satisfying

\[
\mathbf{s} = \mathbf{Dq}
\]

Outlining the construction of \(\mathbf{D}\) is beyond the scope of this article. Suffice to say that six orthonormal vectors occupy the first six columns of \(\mathbf{D}\), and correspond to the global translational and rotational motions of the system (as given by the Sayvetz conditions). The remaining \(N_{\text{vib}}\) harmonic vectors are generated by means of a Gram-Schmidt orthonormalization procedure.

The mass-weighted force \(\mathbf{F}\) and the mass-weighted Hessian \(\mathbf{H}\), both outlined in Part B of the Supporting Information, are transformed into the internal coordinate basis, by use of \(\mathbf{D}\)

\[
\mathbf{F}_i = \mathbf{D} \mathbf{F}_q \quad \mathbf{H}_i = \mathbf{D}^\top \mathbf{H}_q \mathbf{D}
\]

where the subscripts denote the basis in which these quantities are expressed and \(^\top\) denotes the transpose. To evaluate the frequencies of the various modes of motion, we diagonalize \(\mathbf{H}_i\)

\[
\mathbf{E}^{-1} \mathbf{H}_i \mathbf{E} = \mathbf{I}
\]

where \(\mathbf{E}\) denote the eigenvectors of \(\mathbf{H}_i\), and \(\mathbf{I}\) is the identity matrix. The resultant eigenvalues, \(\langle \Omega_i \rangle \equiv \lambda_i\), are related to the mode frequencies, \(v_i\), by

\[
v_i = \sqrt{\frac{\lambda_i}{4\pi^2 c^2}} \quad \forall i = 1, \ldots, 3N
\]

where \(c\) is a factor comprising the speed of light and the conversion between atomic units and \(\text{cm}^{-1}\). Of course, six of these frequencies correspond to the global translational and rotational degrees of freedom of the system, thus yielding \(N_{\text{vib}}\)
nonzero frequencies. The reduced masses and force constants, corresponding to the modes with nonvanishing frequency, are given by similar manipulations of these quantities. The reader is again directed to Ochterski\textsuperscript{[29]} for a discussion of their calculation.

The amplitude of the $i$th mode, $A_i$, is given by rearrangement of the familiar expression for the energy of a simple harmonic oscillator

$$A_i = \sqrt{\frac{2E}{k_i}}$$

where $k_i$ is the force constant of the mode of motion, and $E$ is the energy available to it. We now have all quantities required to evolve the modes of motion and replicate the vibrational dynamics of the system. The total energy available to the system is given by the expression for thermal energy, $E = N k_B k T / 2$, and is stochastically distributed throughout the modes. A temperature of 298 K was used throughout this work. The phase factors of the modes, $\phi$, are also randomly assigned: if $\phi = 0$ for all modes, then they oscillate in unison, which is physically unrealistic. Instead, we assume the modes to resonate out of phase with one another, as energy transfer to each mode from an external heat bath will be strongly decoherent.

Let us note that the average thermal energy available to each mode will comply with a standard equipartition of energy for a physically realistic sampling methodology. The energy available to each mode is then subjected to small stochastic fluctuations. However, one deduces from the above description of our own methodology that we did not follow the route of equipartition. The driving force for this decision was to increase the domain of conformational space, which is then accessible to our sampling methodology. As explained above, we have chosen to distribute the total thermal energy stochastically through all modes. Given a standard equipartition of thermal energy, the $i$th mode, $q_i$, is limited to the domain $q_i^0 - A_i / 2 \leq q_i \leq q_i^0 + A_i / 2$, where $q_i^0$ is the reference state of the mode and $A_i$ is given in eq. (14). However, by stochastically distributing the thermal energy through the modes, the energy available to the $i$th mode, $E_i$, can then take any value in the range $0 \leq E_i \leq n k_B T / 2$, as long as the sum of the $E_i$ is $n k_B T / 2$. In this sense the currently applied methodology is more general than that of the equipartition. If $E_i$ takes the value of $n k_B T / 2$ for all modes, then the sampling domain coincides with the sampling domain of a standard equipartition of energy. However, all other combinations of the $E_i$ have different sampling domains. The sampling domain that is accessible to our stochastic distribution of thermal energy through the modes is then the union of all sampling domains that arise from all possible combinations of the $E_i$. We therefore obtain the largest sampling domain possible for our methodology, which is necessary for the construction of a widely applicable kriging model.

Two issues arise with stochastically distributing the thermal energy through the modes, one methodological and one conceptual. The methodological concern is that there is a non-negligible probability for a significant proportion of the available thermal energy being placed into one mode. If this mode is strongly linked to the motion of a bond length or valence angle, then there is the potential for sampling nonphysical geometries. We have implemented a filtering procedure that prevents the output of such nonphysical geometries. Consider a bond between atoms A and B, of length $r_{AB}$ within a seed geometry. If $r_{AB}$ exceeds a value of $k_{BOND}$ multiplied by the sum of the atomic covalent radii, ($r_A + r_B$), then the geometry is considered nonphysical and rejected. Similarly, if $r_{AB}$ is lower than the inverse of $k_{BOND}$ multiplied by the sum of the atomic covalent radii, the bond is considered too short and rejected. In other words, every bond length must obey the inequality $(1/k_{BOND})(r_A + r_B) \leq r_{AB} \leq k_{BOND}(r_A + r_B)$. Valence angles undergo a similar treatment, so that given any valence angle of the seed geometry, $\alpha_0$, the corresponding valence angle of the sampled geometry, $\alpha$, must obey the inequality $\alpha_0/k_{ANGLE} \leq \alpha \leq k_{ANGLE}\alpha_0$. In the following work, the “stretching” parameters, $k_{BOND}$ and $k_{ANGLE}$, were both set to 1.20. The conceptual concern that we mentioned is that distributing the thermal energy stochastically throughout the modes is nonphysical in terms of equilibrium thermodynamics. For our purposes we are more interested in sufficiently large sampling domain.

The sole remaining issue is the choice of a dynamical time step with which to evolve the various modes of motion. We ensure that a single oscillation of a mode is sampled uniformly. In other words, for a complete cycle of the $i$th harmonic equation of motion, the time period of the mode is $T_i = 1 / \omega_i$. A parameter, $n_{cycle}$, defines the number points to be evaluated along a single cycle of the harmonic equation of motion. From this, we define the quantity $\Delta t = T_i / n_{cycle}$, which is the dynamical timestep for the equation of motion. The quantity $n_{cycle}$ is left as a user-defined input, and is set to $n_{cycle} = 10$ from now on. Additionally, the distribution of the total energy throughout the modes is considered a dynamic quantity, and so for every $n_{reset}$ samples that are output, the energy is randomly redistributed throughout the system. The phase factors are also redefined at the same frequency. Again, $n_{reset}$ is left as a user-defined parameter, and is set as $n_{reset} = 2$ in the following. A further justification for the way we sample is given in Part C of the Supporting Information.

**Computational details**

Sampling of amino acids from the crystal structures was performed by the in-house code MOROS while the in-house FORTRAN code TYCHE distorted the geometries according to NM. The fully automated GAIA code (formerly named AUTOLINE in previous work) was used to build the training and test sets of molecular geometries (Fig. 2). An expanded flow chart of the GAIA procedure is given by Fletcher et al.\textsuperscript{[24]} Once the sampled amino acid geometries were obtained from either PDB/NM or NM, the molecular wave function for each geometry was obtained at the B3LYP/aug-cc-pVDZ level using GAUSSIAN09.\textsuperscript{[28]} The FORTRAN program AIMAll\textsuperscript{[81]} obtained the atomic multipole moments. The parameters $\alpha\beta = \text{auto}$ and $\omega\alpha = \text{high}$, which are standard in GAIA, because $\omega\alpha = \text{high}$ has been seen in the past as a good compromise between accuracy and speed.
Kriging models were built and then tested using the in-house codes FEREBUS and NYX, respectively. All kriging models were built using $N_{\text{train}} = 1000$ training geometries and were tested on 400 randomly selected geometries from the remaining 1000. Experience has shown that kriging models deteriorate in prediction quality as the standard integration error (i.e., the familiar Lagrangian $L$ of atom $\Omega$ or $L(\Omega)$) increases. Hence it is best to set $L(\Omega)$ as low as possible but this norm causes an increasing number of integrations to have to be discarded. A good compromise is allowing a maximum integration error of $L(\Omega)=0.001$ a.u. This value was enforced throughout this work, which keeps the number of discarded atoms reasonable but not nil, explaining the surplus of sampled geometries at the outset.

Results and Discussion

Kriging models were built for the two amino acids alanine (Ala) and lysine (Lys) using geometries sampled from four different sampling approaches: PDB_NO_OPT, PDB_OPT, NM and PDB/NM. These four methods are described in Table 1.

Table 1. An overview of the four sampling approaches.

<table>
<thead>
<tr>
<th>Sampling Approach</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB_OPT</td>
<td>Molecular geometries sampled directly from crystal structure coordinates and H atoms added by the HAAD program. GAUSSIAN fully optimizes bond lengths and valence angles but all dihedral angles remain fixed.</td>
</tr>
<tr>
<td>PDB_NO_OPT</td>
<td>Molecular geometries taken directly from PDB coordinates and H atoms added by HAAD. Single-point GAUSSIAN calculations without any geometry relaxation.</td>
</tr>
<tr>
<td>NM</td>
<td>Standard NM sampling procedure using TYCHE to sample molecular geometries from a number of local energy minima in the gas phase. The local energy minima themselves are not included in either training or test sets.</td>
</tr>
<tr>
<td>PDB/NM</td>
<td>300 randomly selected PDB “seed geometries” sampled with PDB_OPT, each acquiring 7 geometries generated from the nonstationary NM. The “seed geometries” themselves are not included in either training or test sets.</td>
</tr>
</tbody>
</table>

Alanine was chosen because it is the smallest amino acid with a (nontrivial) side chain. Because there is only one side chain dihedral angle ($\chi_1$), as opposed to the four dihedral angles ($\chi_1$, $\chi_2$, $\chi_3$, $\chi_4$) controlling the side chain of lysine, the $\phi$ and $\psi$ angles dominate the dihedral motion of alanine. Lysine has the most flexible side chain of all 20 naturally occurring amino acids, and therefore has been chosen as a rigorous test of the performance of kriging when dealing with highly flexible molecules. Figure 3 shows the four side chain dihedrals in lysine around C–C bonds or $\chi_1$, $\chi_2$, $\chi_3$, and $\chi_4$.

Testing the PDB/NM sampling approach

Kriging models were built for the amino acids Ala and Lys using the four sampling strategies defined in Table 1. Ramachandran plots for the sampled alanine geometries by each of the sampling methods are shown in Figure 4. The dihedral angles are fixed to the same values in both the PDB_OPT and PDB_NO_OPT approach, which is why Figure 4 assigns the same color (blue) to the distribution of $\psi$ and $\phi$ angles of their geometries. As expected, the PDB-sampled Ramachandran plots for both Ala and Lys display a sampling bias toward the $\alpha$-helix and $\beta$-sheet regions with additional clusters of geometries in the left-handed helix region. The green Ramachandran plots display the sampled geometries obtained by the NM method. A number of islands of geometries around the gas-phase energy minima are observed. Several islands are clearly disconnected but some may overlap, such as the long island in lysine (bottom box) at the bottom right of the whole cluster of islands. Because there are regions of conformational space populated by the PDB sampling approaches but not the NM approach, we conclude here that NM sampling from gas phase energy minima is inadequate for building kriging models to be used in biomolecular simulation. This is most noticeable in the case of Lys, where the NM Ramachandran plot appears sparsely populated compared to both the other sampling methods and the Ala NM Ramachandran plot. This is because the side chain of lysine is very flexible, and for each of the nine actual islands in the Ramachandran plot, there are multiple overlapping energy minima with different side chain conformations. This explains why the 39 input minima only...
appear as nine islands on the Ramachandran. The orange Ramachandran plots, containing the Ala and Lys geometries sampled by the PDB/NM approach, strongly resemble the plots of both PDB_OPT (blue) and PDB_NO_OPT (blue) but with fewer points in regions away from the \( \alpha \)-helix and \( \beta \)-sheet region. This is because the 300 "seed" geometries used as input for the NM sampling were randomly selected from the PDB_OPT sampled geometries and, statistically, they are most likely to be sampled from these well populated \( \alpha \)-helix and \( \beta \)-sheet regions. The benefit of PDB/NM (orange) is that, on top of realistic distributions of dihedral angles, bond lengths and angles are more realistic and they are both varied.

Figure 5 shows so-called spider plots of the side chain dihedral angles sampled by each of the sampling approaches. In a spider plot, each of the four axes (meeting at the origin) corresponds to all values that each of the four side chain dihedrals \( \chi_n \) (where \( n = 1, 2, 3, \) or 4) can adopt, that is, from \(-180^\circ\) to \(180^\circ\). Each sampled geometry then corresponds to a quadruplet of dihedral values \( (\chi_1, \chi_2, \chi_3, \chi_4) \), each marked by a point on each of the four corresponding axes. These four points are then linked by four colored lines, which form a (typically lozenge-like) pattern. From the density of these patterns one obtains an instant glimpse of the conformational diversity (or lack thereof) of the side chain geometries.

Clearly, the NM sampling approach (green) samples a very limited range of side chain geometries and does not return the regions of high sampling frequency obtained by the PDB_OPT and PDB_NO_OPT (blue) approaches. For example, the gauche\(^-\) (\(-60^\circ\)) conformation of \( \chi_1 \) is the most sampled conformation in the protein crystal structures but this conformation is not at all present in NM. The preference of \( \chi_1 \) to be in the gauche\(^-\) conformation in proteins is a well-documented
phenomenon\cite{25} and thus NM sampling’s shortcomings are highlighted. The PDB/NM spider plot (orange) shows a better sampling of side chain dihedral angles than that of NM. However, the former shows a sparser sampling of the less populated combinations of dihedral angles compared to PDB_OPT and PDB_NO_OPT (blue).

Table 2 presents a summary of the relative performance of each sampling approach and the resulting kriging model accuracy for both amino acids. The range in the B3LYP/aug-cc-pVDZ energy of the Ala and Lys geometries sampled by each of the four methods is also included in Table 2. For both amino acids the NM sampled geometries show the smallest range. This is because the NM sampling method uses the lowest energy gas phase conformations as the input minima, and hence all sampled geometries from this method are distortions of these low energy geometries. Therefore, large deviations from the various energy minima cannot occur because the distorted geometries are confined by their respective well. This situation is different to that found in PDB geometries. Here, the lysine geometries sampled by the PDB/NM method have the largest range in ab initio energy, 421 kJ mol$^{-1}$, which is much larger than found in any other sampling approach. This is expected as the PDB/NM geometries undergo substantial dihedral sampling, as well as bond length and angle distortions caused by the nonstationary NM sampling.

Table 2 also lists the average bond length range for all bonded atom pairs in the sampled Ala and Lys geometries, calculated for each sampling method. For both Ala and Lys, PDB_OPT yields the lowest average bond length range, 0.02 Å, due to the relaxation of the bonds to their optimal lengths and nonstationary NM sampling algorithms in TYCHE.

Kriging models were built for both Ala and Lys using 1000 molecular geometries obtained from each of the four sampling approaches and were tested on 400 previously unseen (i.e., external and not trained for) molecular geometries obtained by the corresponding sampling approach. For example, kriging models built using geometries sampled using the PDB_NO_OPT method were tested on PDB_NO_OPT geometries, PDB/NM kriging models were tested on PDB/NM geometries, etc. Figure 6 shows the S-curves for all four sampling methods. As an example of how to read such an S-curve: 88% of geometries in the external test set for alanine’s PDB_NO_OPT kriging models (top, red curve) have an error of maximum 4 kJ mol$^{-1}$, which is much larger than found in any other sampling approach. This is expected as the PDB/NM geometries undergo substantial dihedral sampling, as well as bond length and angle distortions caused by the nonstationary NM sampling.
the more accurate the model that it describes. The error displayed by an S-curve corresponds to that given by eq. (9), that is, \[
\sum_{i<j} E_{AB}^{\text{predicted}} - E_{AB}^{\text{original}}.
\]
As such, each point on an S-curve corresponds to the absolute value of the sum of the errors of all predicted Coulombic interactions between pairs of atoms in one test molecular geometry, relative to the original interaction energies. This value is referred to as both the “total absolute error” and also the “S-curve error.”

In connection with the information shown in Figure 6, note that Table 2 also reports the average absolute total error and S-curve error for all four sampling approaches for Ala. The correlation between bond length and average S-curve error \(\frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} |\Delta E_{\text{system}}|\) is fairly strong, with an \(R^2\) value of 0.90 (see Fig. 7). To illustrate this point further, the difference in average total error (S-curve error or \(|\Delta E_{\text{system}}|\) between PDB/NM and NM is 0.6 kJ mol\(^{-1}\) (see Table 2), although the PDB/NM approach samples a much larger range of dihedral conformational space than NM. In contrast to this, PDB_OPT, which has a much larger range of dihedral conformational space, has an average total error lower than that of NM. This observation is a result of the following effect. Under the assumption of an identical dihedral sampling (as is the case for PDB_NO_OPT and PDB_OPT), increasing the range of bond lengths increases the volume of configurational space that the kriging models have to describe. This increase results in a more difficult kriging problem leading to increased prediction errors. It also is observed that changing a bond length has a dominant effect on the multipole moments of the atoms involved. This is illustrated in Supporting Information Figures S1–S3 where plots of \(C_\alpha\) charge against both N–C\(_\alpha\) bond length and backbone \(\psi\) angle are provided for the Ala geometries sampled by the PDB/NM, PDB_OPT and NM approaches, respectively. In both the PDB/NM and NM sampled plots, the \(C_\alpha\) charge shows correlation with the N–C\(_\alpha\) bond length but not with the \(\psi\) angle. It is only in the plots obtained from the PDB_OPT geometries (where the N–C\(_\alpha\) bond length range is significantly reduced as a result of partial geometry relaxation) that any correlation between \(C_\alpha\) charge and \(\psi\) can be seen. In summary, the correlation

Figure 6. Errors in the predicted total electrostatic interaction energies (1–4 and higher) of alanine (top) and lysine (bottom) for kriging models trained with molecular geometries obtained by: PDB_OPT (blue), PDB_NO_OPT (red), NM (green), and PDB/NM (orange). The dashed purple lines mark the 1 kcal mol\(^{-1}\) threshold. (Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.)
patterns above prove the dominance of bond length variation over dihedral sampling in posing a challenge to kriging.

The same conclusions may be drawn from the Lys S-curves as from the Ala S-curves: average bond length deviation is the most important factor dictating the average S-curve error (Fig. 7), and although larger dihedral sampling increases the average error, it does so to a lesser extent than a large average bond length deviation. PDB_OPT has the lowest average S-curve error (Lys: 1.6 kJ mol$^{-1}$ and Ala: 0.7 kJ mol$^{-1}$) due to the optimized bond lengths having the lowest average deviation (0.02 Å for both ALa and Lys). The PDB/NM S-curve has the highest average error due to having the largest average bond length deviation and also a large dihedral sampling. PDB_NO_OPT has the largest maximum S-curve error but, unlike the high error PDB_NO_OPT point on the Ala S-curve, there is no clear structural reason behind the highest energy geometry. This could indicate that the geometry lies outside of the configurational space of the training set. The overall shape of an S-curve may be related to the quality of the test geometries and the range of conformational space. For example, the NM S-curve (green) is steep with only a small bend at the top. This is a result of the relatively small set of seed geometries causing the sampled geometries to be clustered close together. Therefore all test geometries are close to a training geometry within the kriging model and the errors remain constant throughout. In contrast, the PDB_NO_OPT (red) geometries are not clustered together and therefore the test geometries can be further away from the nearest training set geometry leading to larger errors. This gives rise

Figure 7. Average bond length deviation against average total (S-curve) error for the different sampling approaches of Ala (left) and Lys (right): PDB_OPT (blue), PDB_NO_OPT (red), NM (green), and PDB/NM (orange). All data taken from Table 2. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 8. Individual intramolecular interaction prediction errors in Ala against interaction distance obtained for models built using the four sampling approaches: PDB_OPT (blue), PDB_NO_OPT (red), NM (green), and PDB/NM (orange). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
to the less steep climb of this S-curve and its longer tail toward the 100% ceiling.

Each point on the S-curve is a sum of all 1,4 and higher intramolecular interaction prediction errors within a single test geometry \( (\sum_{AB} E_{AB}^{\text{original}} - E_{AB}^{\text{predicted}}) \) from eq. (9)). Because of the sum, potential cancellation of positive and negative interaction errors is included within the S-curve. To increase the transparency of the results we now focus on the construction of the S-curve. Figure 8 shows all interaction errors for all Ala test geometries plotted against interaction distance for each sampling approach. The maximum absolute interaction error \( (\max |E_{AB}^{\text{original}} - E_{AB}^{\text{predicted}}|) \) and average absolute interaction error \( (\text{average } |E_{AB}^{\text{original}} - E_{AB}^{\text{predicted}}|) \) for each approach is included in Table 2. Supporting Information Figure S4 shows a plot analogous to Figure 8 but for the sampled Lys geometries. The average absolute interaction errors follow the same trend as the total S-curve error (PDB/NM \( \approx (\text{NM} > \text{PDB_NO_OPT} > \text{PDB_OPT}) \)). For all sampling approaches used, the largest average absolute interaction error was only 0.4 kJ mol\(^{-1}\) (NM and PDB/NM sampled geometries). The correlation between average absolute interaction error and total error is very high with an \( R^2 \) of 0.97 for Ala and 0.99 for Lys. The plots of the average interaction prediction error versus the total error can be seen in Supporting Information Figure S5.

The standard deviation of the interaction errors for each method is provided in Table 3 for both Ala and Lys. Both PDB_OPT and PDB_NO_OPT have significantly larger standard deviations for Lys (0.5 kJ mol\(^{-1}\) and 0.8 kJ mol\(^{-1}\), respectively) than for Ala (0.2 kJ mol\(^{-1}\) and 0.4 kJ mol\(^{-1}\), respectively) as is expected by comparison of the blue and green plots in Figures 8 and Supporting Information S4. The PDB/NM interactions in Lys also have a larger standard deviation (0.7 kJ mol\(^{-1}\)) than the PDB_NM interactions in Ala (0.6 kJ mol\(^{-1}\)). Larger standard deviations emerge for Lys because it is a larger, more flexible molecule than Ala and so the kriging problem for PDB sampled geometries is much harder. Thus the kriging model is unable to find as good a solution for Lys than for Ala.

**Optimum ratio of input geometries to sampled geometries for the PDB/NM sampling approach**

The hybrid PDB/NM sampling approach has been presented as a means of sampling chemically relevant amino acid geometries for kriging models, taking advantage of the benefits afforded by both PDB and NM sampling whilst avoiding the problems associated with either method. The ratio (denoted \( 1:n \)) of PDB-seed geometries (set to 1) to nonstationary NM sampled geometries (set to \( n \)) will now be discussed. The maximum dihedral sampling corresponds to a 1:1 ratio of PDB sampled “seed geometries” to NM sampled geometries. However, this ratio is computationally expensive because each PDB-sampled amino acid seed geometry then needs to be partially geometry-relaxed. Conversely, a ratio smaller than 1:1 (i.e., 1:n where \( n > 1 \)) requires fewer geometry optimizations, but decreases the sampling of (dihedral) conformational space. A smaller number of sampled geometries per PDB-seed geometry will also affect the difficulty of the kriging problem as the sampling of conformational space will increase (assuming a constant training set size).

Training sets have been built, using the PDB/NM sampling approach, for ratios of seed geometries to NM-sampled geometries of 1:20, 1:10, 1:4, 1:2, and 1:1, always with a total of 1200 NM-sampled geometries in each case. These geometries were randomly reshuffled and then kriging models were built using 800 training geometries, and were tested on 400 (external) geometries.

---

**Table 3. Standard deviation of interaction prediction errors for both Ala and Lys from kriging models built from geometries sampled from the four sampling approaches (kJ mol\(^{-1}\)).**

<table>
<thead>
<tr>
<th>Sampling</th>
<th>Ala</th>
<th>Lys</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB_OPT</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>PDB_NO_OPT</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>NM</td>
<td>0.7</td>
<td>0.5</td>
</tr>
<tr>
<td>PDB/NM</td>
<td>0.6</td>
<td>0.7</td>
</tr>
</tbody>
</table>

---

![Figure 9](image9.png) **Figure 9.** Errors in the predicted total 1–4 and higher electrostatic interaction energies of lysine by kriging models trained with molecular geometries obtained by the PDB/NM approach with different numbers of PDB-seed geometries (see key on graph, 1200 corresponds to the 1:1 ratio in the main text). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

![Figure 10](image10.png) **Figure 10.** Average total error versus the number of PDB seed geometries for kriging models of lysine obtained from the PDB/NM sampling methodology. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
Figure 9 shows the total energy S-curve obtained for each training set. Increasing the number of PDB-seed geometries does not significantly reduce the quality of the kriging model obtained. The average values of the S-curve energies have been plotted against the number of input minima in Figure 10. There is a trend for a larger number of PDB-seed geometries to have a higher average S-curve error, but not dramatically so. The range of errors is only ~0.6 kJ mol\(^{-1}\), between a 1:20 ratio of PDB-seed geometries to sampled geometries (average error of 3.8 kJ mol\(^{-1}\)) and a 1:1 ratio (average error of 4.4 kJ mol\(^{-1}\)).

Conclusions

The topological force field QCTFF contains a machine learning component that handles polarization and charge transfer (in a unified way). The machine learning method used, called kriging, needs a data set of molecular geometries to train on. Here we focus on obtaining a more realistic and relevant training set for amino acids. Before the current study, we sampled the training set by distorting the local energy minima of (peptide-capped) amino acids (in the gas phase) according to NO at those stationary points. Using the Protein Data Bank (PDB) we show here that these gas phase stationary points miss a number of important amino acid geometries that are present in a folded protein.

We present a new sampling approach that combines sampling of amino acid geometries from the Protein Data Bank (PDB) with nonstationary NM (NM) distortion. To the best of our knowledge the latter technique has not been attempted before. This hybrid approach is called PDB/NM and is tested on alanine and lysine, the most flexible amino acid of all. The use of the PDB greatly expands the sampling in the space of dihedral angles, both in range and density. Does this expansion lead to worse kriging models, given the larger variation and diversity in dihedral angles? The answer is negative because it turns out that the range in bond lengths is actually the prime factor in determining the difficulty and hence the predictive accuracy of the kriging models. As a result, the new PDB/NM sampling method (which is more “informed”) performs as well as the original “gas phase energy minimum” NM sampling. All kriging models lead to very good electrostatic energy prediction errors where more than 60% of external test geometries have a value of less than 4 kJ mol\(^{-1}\). Within the PDB/NM paradigm, the quality of the kriging models is not compromised much even if the training set consists of PDB-sampled geometries only, which corresponds to maximum coverage of conformational space. In summary, the good news is that realistic dihedral angles can safely be combined with realistic bond lengths and angles into a single successful kriging model.

Further work utilizing rotamer libraries to guide the construction of training sets is planned to create training sets that do not depend on the crystal structures sampled from, but still mimic the structures expected in real proteins.

Keywords: quantum theory of atoms in molecules · quantum chemical topology · conformational sampling · kriging · electrostatics · protein data bank

How to cite this article: T. J. Hughes, S. Cardamone, P. L. A. Popelier. J. Comput. Chem. 2015, 36, 1844–1857. DOI: 10.1002/jcc.24006

Additional Supporting Information may be found in the online version of this article.
The prediction of topologically partitioned intra-atomic and inter-atomic energies by the machine learning method kriging

Peter Maxwell¹² · Nicodemo di Pasquale¹² · Salvatore Cardamone¹² · Paul L. A. Popelier¹²

Received: 24 May 2016 / Accepted: 11 July 2016 / Published online: 27 July 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract The construction of a novel protein force field called FFLUX, which uses topological atoms, is founded on high-rank and fully polarizable multipolar electrostatics. The machine learning method kriging successfully predicts multipole moments of a given atom with as only input the nuclear coordinates of the atoms surrounding this given atom. Thus, trained kriging models accurately capture the polarizable multipolar electrostatics of amino acids. Here we show that successful kriging models can also be constructed for non-electrostatic energy contributions. As a result, the full potential energy surface of the (molecular) system trained for can be predicted by the corresponding set of atomic kriging models. In particular, we report on the performance of kriging models for each atom’s (A) (1) total atomic energy ($E_{A_{\text{QCT}}}$), (2) intra-atomic energy ($E_{A_{\text{intra}}}$) (both kinetic and potential energy), (3) exchange energy ($V_{A_{\text{XC}}}$) and (4) electrostatic energy ($V_{A_{\text{el}}}$) of atom A with the rest of the system ($A'$), and (5) interatomic energy ($V_{A_{\text{inter}}}$). The total molecular energy can be reconstructed from the kriging predictions of these atomic energies. For the three case studies investigated (i.e. methanol, N-methylacetamide and peptide-capped glycine), the molecular energies were produced with mean absolute errors under 0.4, 0.8 and 1.1 kJ mol⁻¹, respectively.

Keywords Quantum chemical topology (QCT) · Interacting quantum atoms (IQA) · Quantum theory of atoms in molecules (QTAIM) · Kriging · Machine learning · Amino acids · Force field

1 Introduction

There is a consensus that traditional force fields, which are instrumental in the vast majority of modern molecular simulations, need further improvement. Their limiting accuracy is regularly pointed out in the literature, as a cause for discrepancies between a given force field’s predictions and experimental results. Indeed, if the sampling during the simulation is adequate, then only the potential can be blamed if predictions fail to be trustworthy. In order to improve their energy prediction, popular force fields such as AMBER and CHARMM have been modified on several occasions. Fairly recent modifications were extensively tested [1] in 2011 with the most powerful dedicated molecular simulation hardware in the world. The four force fields tested in this protein folding work were Amber ff03, Amber ff99SB*-ILDN, CHARMM27 and CHAMRM22*. It was found that the folding mechanism and the properties of the unfolded state depended substantially on the choice of force field. Another extensive and more recent study [2] concluded, from a millisecond of simulations on intrinsically disordered proteins, that eight well-known force fields generate unexpectedly huge differences in chain dimension, hydrogen bonding and secondary structure content. In fact, discrepancies are so serious that changing the force field has a stronger effect on secondary structure content than changing the entire peptide sequence. Such comparisons are quite rare but precious because they clearly demonstrate, while eliminating any sampling issues or hardware
limitations, that much more work needs to be done. The question is which type of work.

Over the last decade our strategy has been to re-examine and challenge the core architecture of classical force fields. The type of work that accompanies such a bold strategy can be characterized as arduous, and above all, systematic. At the outset of this long term project, the ubiquitous atomic point charges (one on each nucleus) were replaced by nucleus-centred atomic multipole moments. This step is shared by other next-generation force fields such as AMOEBA [3], XED [4], SIBFA [5] and ACKS2 [6] and is driven by a clearly justified desire towards increasingly accurate electrostatics [7, 8].

The current approach embraces so-called topological atoms as the “entity of information” from which any system (molecular or ionic) is built. Much work has been carried out [9–17] in order to obtain a deep understanding of the convergence behaviour and accuracy of the electrostatic interaction between topological atoms, as well as the electrostatic potential they generate. Topological atoms are defined by the quantum theory of atoms in molecules (QTAIM) [18–21] as finite-volume fragments in real 3D space. As quantum atoms [20, 22], topological atoms are deeply rooted in quantum mechanics [23]. These atoms result from a parameter-free partitioning of the electron density, introducing sharp boundaries whose shape responds to any variation in nuclear geometry. The finite size of topological atoms prevents penetration effects and the associated correction in the form of damping functions. Topological multipolar electrostatics proved to be successful in the description of electrostatic interaction in proteins [24].

The next step in the construction of a topological force field is the inclusion of electrostatic polarization. In principle, the multipole moments of any given atom are influenced by all the atoms surrounding it, but this influence typically decreases the further away the surrounding atoms are. In order to be able to handle the full complexity of this influence, we invoked machine learning early on. Initially we used neural networks [25] and applied it to water clusters [26]. In 2009 it turned out [27] that a completely different machine learning method called kriging performed more accurately than neural networks. Although kriging was computationally more expensive, it coped better with the larger number of molecules surrounding the atom of interest. Kriging [28], which is also known as Gaussian regression analysis [29], originates in geostatistics but has been used in very different application areas, including the prediction [30] of atomic properties when inside molecules.

The essence of our kriging approach is the establishment of a direct mapping between an atomic multipole moment (output) and the nuclear coordinates (input) of the surrounding atoms. This mapping is obtained after training to a training set of (molecular or cluster) geometries, and is able to make a prediction to a previously unseen geometry of the surrounding molecules. For this purpose we take advantage of the renowned interpolative power of kriging. Moreover, when using kriging models in an extrapolative way, the model returns the average value of the atomic property of interest observed over the training set. Finally, it must be pointed out that a kriging model is not returning an atomic polarizability but the atomic multipole moment itself, after the polarization process is complete. This strategy has an important advantage when the kriging models are used during a molecular dynamics simulation: the atomic moments do not need to be computed (typically iteratively) from the polarizabilities. Instead, the multipole moments are predicted “on the fly”, directly from the nuclear coordinates of the surroundings at any given time step.

Most attention has been devoted to modelling the electrostatic interaction at long-range by means of kriged atomic multipole moments. As this procedure is understood and works well [31–37], the next step is how to combine this electrostatic energy with the non-electrostatic energy contributions. Preliminary and unpublished work expressed the latter in the traditional manner, i.e. with Hooke-like potentials reinforced with anharmonic extensions. The parameterization of these potentials, in the presence of kriged electrostatics, turned out to be inadequate. For this reason, a more satisfactory and elegant alternative strategy was carried out, which is to combine kriged electrostatics with kriged non-electrostatics. In this streamlined procedure the machine learning method is trained for energy quantities that are obtained from the same topological energy partitioning [38] that yields the atomic multipole moments. In 2014, the atomic kinetic energy was successfully kriged [39] as the first non-electrostatic energy contribution. That work presented proof-of-concept based on four molecules of increasing complexity (methanol, N-methylacetamide, glycine and triglycerine). For all atoms tested, the mean atomic kinetic energy errors fell below 1.5 kJ mol⁻¹, and far below this value in most cases.

In the current article, we go further and deliver proof-of-concept for the kriging of non-electrostatic atomic energy contributions. For that purpose we have adopted the interacting quantum atoms (IQA) scheme proposed by Blanco et al. [40]. This is a topological energy partitioning scheme, inspired by early work [11] on atom–atom partitioning of intramolecular and intermolecular Coulomb energy. In IQA, the kinetic energy is subsumed in the intra-atomic energy (or sometimes called “self energy”), which also contains the potential energy of the electrons interacting with themselves and with the nucleus, both within a given atom. This intra-atomic energy plays a pivotal role in stereo-electronic effects, including intramolecular Pauli-like repulsion. Furthermore, IQA can calculate the electrostatic
interaction between two atoms that are so close that their multipolar expansion diverges. The IQA method achieves this goal by using a variant of the six-dimensional volume integration (over two atoms) proposed in Ref. [11], which avoids the multipolar expansion altogether. Similarly, IQA does not multipole-expand the (inter-atomic) exchange energy, although this can be done [15]. However, this route is not followed by our topological force field. This energy contribution expresses covalent bonding energy. Within the Hartree–Fock ansatz, these three energy contributions (self, Coulomb and exchange) complete an IQA partitioning. However, post-Hartree–Fock methods introduce a fourth (non-vanishing) contribution that is associated with electron correlation. In the current work, we use a version of IQA [41] that is compatible with DFT with an eye on including electron correlation effects. We invoke the use of DFT level with the largest of systems studied here, capped glycine.

Three molecules were chosen here for this proof-of-concept investigation: methanol, \textit{N}-methylacetamide (NMA) and glycine, which is capped by a peptide bond both at its C and N terminus. These systems were chosen to represent a progressive sequence of molecular complexity, while being relevant to biomolecular modelling: methanol features as the side residue in the amino acid serine, NMA is the smallest system modelling a peptide bond, while capped glycine represents an amino acid in an oligopeptide. This work is the first report of combining machine learning with a full topological energy partitioning.

2 Methodological background

2.1 The interacting quantum atoms (IQA) approach

Figure 1 shows examples of topological atoms appearing in \textit{N}-methylacetamide, which were generated by in-house software [42, 43]. QTAIM essentially defines a topological atom as a three-dimensional subspace determined by the bundle of gradient paths (of a system’s electron density) that are attracted to the atom’s nucleus. This partitioning idea also features in other topological approaches [44], such as that in connection with the electron localization function (ELF). The topological energy partitioning method IQA is a third approach that uses the central idea of the so-called gradient vector field to extract chemical information from a system. Quantum chemical topology (QCT) [45] is a collective name to gather all topological approaches (10 so far, see Box 8.1 in Ref. [20]) that share the abovementioned central idea. The acronym QCT resurfaces in QCTFF, the force field under construction here [22], which uses topological atoms. The new name for QCTFF is FFLUX, for which a very recent and accessible perspective [46] can be consulted. We also note here that it has been shown before [47] that atom types that can be computed using the atomic properties of topological atoms in amino acids.

It is clear from Fig. 1 that QCT partitions a molecule into well-defined non-overlapping atoms [48]. Moreover, these topological atoms do not show any gaps between them; in other words, they partition space exhaustively. It is important to pause and briefly discuss the full consequence of this property. Exhaustive partitioning infers that each point in space belongs to a topological atom: all space is accounted for. In principle, this property must have repercussions [49] for docking studies, as will become clear when QCT starts being used at this larger molecular scale. Classical drug design (e.g. [50]) thinks of both ligand and the protein’s active site as bounded by artificial surfaces (e.g. solvent accessible surface) based on standard van der Waals radii and an image of overlapping hard spheres. This view necessarily introduces "gaps of open space", which belong neither to the ligand nor to the protein. However, quantum mechanically we know that electron density resides in those gaps, no matter how small or thin they are. Electron density generates an electrostatic potential, and hence, also generates electrostatic interaction energy contributions. If a gap is not accounted for, then energy will be missing, which interferes with the energy balance during the docking process. However, if there is no gap, as in QCT, then all energy is properly accounted for.

In brief, IQA quantitatively describes the total energy of an atom, even if the system is not at a stationary point in the potential energy surface. In other words, unlike in orthodox QTAIM, there is no need to invoke the atomic virial theorem.
The intra-atomic energy can be further partitioned, into the atomic energies, one for each atom \( A \), denoted \( E_{\text{IQA}}^A \), followed by its breakdown into intra-atomic (or ’self’) and interatomic interaction energies,

\[
E_{\text{IQA}}^\text{molec} = \sum_A E_{\text{IQA}}^A = \sum_A E_{\text{intra}}^A + \frac{1}{2} \sum_A \sum_{B \neq A} V_{\text{AB}}^\text{inter}
\]

where \( E_{\text{intra}}^A \) and \( V_{\text{AB}}^\text{inter} \) are the intra-atomic (of atom \( A \)) and inter-atomic (between atoms \( A \) and \( B \)) energies, respectively. The intra-atomic energy can be further partitioned,

\[
E_{\text{intra}}^A = T^A + V_{\text{en}}^{AA} + V_{\text{en}}^{A}\n
\]

where \( T^A \) is the kinetic energy of the electrons associated with atom \( A \), \( V_{\text{en}}^{AA} \) is the (repulsive) electron–electron potential energy, and \( V_{\text{en}}^{A}\n \) is the (attractive) electron–nuclear potential energy. Together, these three energies comprise the intra-atomic energy possessed by a single atom.

The interatomic energy attributed to a pair of atoms can also be further partitioned,

\[
V_{\text{inter}}^{AB} = \left(V_{\text{cl}}^{AB} + V_{\text{Coul}}^{\text{en}} + V_{\text{X}}^{\text{cor}}\right) + V_{\text{ee}}^{AB}
\]

where \( V_{\text{cl}}^{AB} \), \( V_{\text{Coul}}^{\text{en}} \) and \( V_{\text{X}}^{\text{cor}}\n \) were described above but now with the ordering of the subscript components being allied to the ordering of the atoms in the superscript. For example, subscript ’\text{en}’ and superscript ’\text{AB}’ refers to the electrons of atom \( A \) and the nucleus of atom \( B \). In addition to the electronic energy components, \( V_{\text{nm}}^{AB} \) is the repulsive nucleus–nucleus potential energy.

The electron–electron energy \( V_{\text{ee}}^{AB} \) can be even further partitioned to give the components in Eq. (4),

\[
V_{\text{ee}}^{AB} = V_{\text{Coul}}^{AB} + V_{\text{X}}^{\text{en}} + V_{\text{X}}^{\text{cor}}
\]

where \( V_{\text{Coul}}^{AB} \) represents the Coulombic interaction between the electrons in atoms \( A \) and \( B \), \( V_{\text{X}}^{\text{en}} \) represents the inter-electron exchange potential energy and \( V_{\text{X}}^{\text{cor}}\n \) the inter-electron correlation potential energy. Combining the bracketed terms in Eq. (3) with the Coulomb energy only, leads to the total electrostatic energy between two atoms, or \( V_{\text{elec}}^{AB}\n \), which is often written as \( V_{\text{cl}}^{AB}\n \) because of the “classical” nature of the electrostatic potential energy (devoid from any purely quantum mechanical exchange energy),

\[
V_{\text{cl}}^{AB} = \left(V_{\text{nm}}^{AB} + V_{\text{en}}^{AB} + V_{\text{ne}}^{AB}\right) + V_{\text{Coul}}^{AB}
\]

We have extensively studied this energy \( V_{\text{cl}}^{AB}\n \) in terms of its multipolar convergence behaviour. The quantity \( V_{\text{cl}}^{AB}\n \) incorporates the widely reported electrostatic multipole moments’ contribution of the long-ranged electrostatic energy, in addition to the short-range electrostatic contribution obtained from IQA. Equation (3) can be rewritten as Eq. (6). This is done by first substituting the bracketed expression by \( V_{\text{cl}}^{AB} - V_{\text{Coul}}^{AB}\n \) as obtained from Eq. (5), and then by substituting \( V_{\text{ee}}^{AB}\n \) using Eq. (4), such that after cancellation of \( V_{\text{Coul}}^{AB}\n \) we obtain,

\[
V_{\text{inter}}^{AB} = V_{\text{cl}}^{AB} + V_{\text{X}}^{\text{en}} + V_{\text{X}}^{\text{cor}} = V_{\text{cl}}^{AB} + V_{\text{X}}^{\text{inter}}
\]

The new expression separates the interatomic energy into the interplay of ionic-like \( V_{\text{cl}}^{AB}\n \), covalent \( V_{\text{X}}^{\text{en}}\n \) and correlation \( V_{\text{X}}^{\text{cor}}\n \) energies. Note that it is often convenient to combine exchange and correlation in one term. These three energies along with the intra-atomic energy compose the four primary energies that FFLUX is built upon.

Until recently [41] the inclusion of any computationally affordable correlation energy has been lacking because IQA is incompatible with both perturbation theory and density functional theory (DFT) methods. Indeed, the methods that are compatible with the original IQA (i.e. full configuration interaction (FCI), complete active space (CAS), configuration interaction with single and double excitations (CISD) or coupled cluster with single and double excitations (CCSD) levels of theory) demand much greater computational expense. Neither perturbation theory nor standard DFT methods provide a well-defined second-order reduced density matrix, and hence IQA cannot be straightforwardly applied to them. Together with Dr TA Keith, the main author of the QCT computer program AIMAll [52], a DFT-based IQA method that incorporates at least some correlation was validated [41] by our group. The solution involved incorporating the explicit B3LYP atomic exchange functional in order to correctly calculate an atom’s total atomic exchange, thereby recovering the ab initio energy of the whole molecule. However, the fact that the functional cannot be used to calculate interatomic exchange (see Ref. [41] for details) led us to calculate the interatomic component using the Hartree–Fock-like expression but then with Kohn–Sham orbitals inserted. The remaining intra-atomic exchange–correlation is then calculated as the difference between the atomic exchange–correlation directly obtained from the B3LYP functional and the Hartree–Fock-like interatomic exchange.

In this investigation, we will again make use AIMALL. This program is able to return a useful quantity, denoted \( V_{\text{inter}}^{AA}\n \), which is defined as follows:

\[
V_{\text{inter}}^{AA} = \sum_{B \neq A} V_{\text{inter}}^{AB}
\]
where $A'$ represents every atom other than atom $A$. Note that Eq. (7) defines the atom-centred interatomic energy contribution $V_{AA'}_{\text{inter}}$. In other words, it summarizes how an atom interacts in full with all other atoms. Note that the quantity $V_{AA'}_{\text{inter}}$ is obtained computationally cheaper than by summing over the individual pair-wise atomic contributions $V_{AB}_{\text{inter}}$. However, the computational advantage of $V_{AA'}_{\text{inter}}$ is offset by a reduction in the chemical insight that we obtain from being able to inspect each atom pair individually. This loss occurs over and above that caused by lumping together the electrostatic, exchange or correlation energy contributions (see Eq. 6). Instead, the interaction energy is defined in terms of a given atom $A$ experiencing the entire surrounding molecular environment. Equally, $V_{AA'}_{\text{inter}}$ can be decomposed into $V_{AA'}_{\text{intra}}$ and $V_{AA'}_{\text{cl}}$ components, much like the pair-wise $AB$ energies. For our current purpose, we are satisfied with this formulation in spite of the reduced chemical insight it gives because our prime motivation is to predict atomic energies, and not to predict chemical insight.

Returning to the IQA formalism, we obtain Eq. (8), which expresses three ways (“approaches”) to break up the molecular energy into atomic contributions. Approach $A$ was already present in Eq. (1) while substituting $V_{AA'}_{\text{inter}}$ of Eq. (7) into Eq. (1) leads to Approach $B$. Finally, Approach $C$ follows from Eq. (6) and now applying the idea behind Eq. (7) to $V_{AB}_{\text{cor}}, V_{AB}_{\text{intra}}$ and $V_{AB}_{\text{cl}}$ yielding

$$E^{\text{molec}}_{\text{IQA}} = \sum_A E^A_{\text{IQA}} = \sum_A E^A_{\text{intra}} + \frac{1}{2} \sum_A V_{AA'}_{\text{inter}}$$

Approach $A$

$$= \sum_A E^A_{\text{intra}} + \frac{1}{2} \sum_A V_{AA'}_{\text{cl}} + \frac{1}{2} \sum_A V_{AA'}_{\text{cl}}$$

Approach $B$

$$= \sum_A E^A_{\text{intra}} + \frac{1}{2} \sum_A V_{AA'}_{\text{cl}} + \frac{1}{2} \sum_A V_{AA'}_{\text{cl}}$$

Approach $C$

which is the key equation for the analysis in this paper. Note that $V_{AA'}_{\text{inter}}$ is always halved when used in Eq. (8), attributing half of the energy to a single atom $A$, in order to prevent double-counting of the interatomic energy in the molecule.

We aim for a greater understanding of both the quantitative nature of these five types of energy ($E^A_{\text{IQA}}, E^A_{\text{intra}}, V_{AA'}_{\text{inter}}, V_{AA'}_{\text{intra}}, V_{AA'}_{\text{cl}}$ and $V_{AA'}_{\text{cl}}$) and their suitability in FFLUX after being kriged. As suggested in Eq. (8), the molecular energy can be recovered through three different approaches, each incorporating the use of different IQA energies. Approach $A$ uses only the total atomic energy of an atom, denoted $E^A_{\text{IQA}}$. The atomic energy $E^A_{\text{IQA}}$ is a sum of the intra- and inter-atomic energy and hence expresses their resulting “trade-off” by the single quantity that it is. This final energy, $E^A_{\text{IQA}}$, suffices by itself for FFLUX being able to predict the structure and dynamics of a system, because the latter only depend on the total atomic energy, not its breakdown. Approach $B$ exposes the separation of the intra-atomic and interatomic energies for insight into how an atom itself experiences the environment it is in. Finally, Approach $C$ takes the separation one step further, using the individual exchange and electrostatic energies in the interatomic description of an atom.

In order to clarify the strategy for the complete treatment of energy contributions in FFLUX a comment about $V_{AB}_{\text{corr}}$ in Eq. (4) is necessary. This energy contribution covers dynamic correlation and hence dispersion. Our preferred route is to treat $V_{AB}_{\text{corr}}$ in exactly the same way as $V_{AA'}_{\text{cl}}$ and $V_{AB}_{\text{cl}}$. This approach, for which proof-of-concept has been reached in our lab, will guarantee a seamless integration of dispersion in the FFLUX ansatz. This strategy will thereby avoid the typical problems (e.g. the need for damping functions) that alternative dispersion methods introduce.

For a more exhaustive description of the IQA partitioning scheme including additional formulae, its capabilities and previous applications, we refer to the original literature [40, 53–58].

2.2 Kriging (Gaussian regression analysis)

As a machine learning method, kriging has its roots in geostatistics where it has been used to predict the location of precious material after being taught these locations [28]. Within FFLUX, kriging is used to map geometrical change within a molecule, obtained from nuclear coordinates, to a corresponding atomic property, which can be an IQA energy or atomic multipole moment. The atomic property is the machine learning output and the coordinates are the input. Although the full details are given elsewhere [33, 34] we explain here how these coordinates are constructed. It is advantageous that the coordinates are internal in nature (so only $3N - 6$ for a nonlinear $N$-atom system). On each nucleus we install a so-called atomic local frame (ALF), which enables the definition of a polar angle and an azimuthal angle to describe the position of each nucleus in the system, except for the three nuclei required in defining the ALF. The distance between the ALF’s origin and the nucleus completes the triplet of (spherical polar) coordinates for a given nucleus in the system (other than that on which the ALF is installed). Machine learning language calls these coordinates features, as they are the input variables to kriging in this case. Finally we note that the first three features of the vector of $3N - 6$ features (necessary to describe unambiguously a molecular geometry) consists of (1) the distance between the origin and the first nucleus, which fixes the ALF’s $x$-axis, (2) the distance between the origin and the second nucleus (fixing the ALF’s $xy$-plane), and (3) the angle suspended by the first nucleus, the origin and the second nucleus.
Before introducing the mathematics involved, it is useful to understand the kriging procedure qualitatively. Before a kriging model can predict a quantity, it must first be trained for using a number of molecular geometries with corresponding atomic properties. These data form the training set while the test set will consist of data that do not belong to the training set. The external character of the test set makes the assessment of the kriging predictions meaningful, because predictions of the training set data would be exact anyway (with the type of kriging used here). The molecular geometries used to build each set are obtained from sampling a molecular energy well that surrounds an atomic property of choice. For a more comprehensive description of the kriging protocol, the reader is invited to refer to our previous publications [33, 35–37, 59, 61].

The next paragraph provides a very brief summary. With regard to using the key kriging formula in Fig. 2 (after training), previously unseen features \( F = \{f_k\} \) are inserted, returning the atomic property \( f(F) \). It is clear that the argument of the exponential is a distance function, which is not necessarily Euclidean (i.e., \( p_k \neq 2 \)).

In terms of the optimization procedure, first the concentrated log-likelihood is calculated analytically. The function is then maximized by a different machine learning method because this cannot be achieved analytically. We have successfully used particle swarm optimization (PSO), the mathematical details of which can be found in Ref. [60]. The optimization of the parameters \( \theta_k \) and \( p_k \) via PSO is the most computationally expensive step in the overall kriging process. However, optimizing these parameters allows the user to obtain the highest possible concentrated log-likelihood function, ensuring that the best possible model is obtained. The time for the PSO optimization is proportional to the number of geometries in the training set and the number of atoms in the molecule. The result is an analytical formula (see Fig. 2) linking the geometrical features of a molecule and the atomic property of choice. For a more comprehensive description of the kriging protocol, the reader is invited to refer to our previous publications [33, 35–37, 59, 61].

### 2.3 Sampling of distorted geometries

The selection of training examples with which to construct a kriging model is of great importance. The geometries of the training set should be representative of the physically realistic regions of conformational space. This representation ensures that predictions corresponding to relevant molecular geometries are always made in areas of conformational space that have been trained for in the kriging model. Conventional methods will use some form of
molecular mechanics, parameterized by a classical force field, to generate the training geometries. However, it has been shown that molecular mechanics does not necessarily sample the relevant areas of conformational space over the course of a typical trajectory [62–64].

Our approach attempts to locally approximate the ab initio molecular potential energy surface about a “seeding” conformation, which in this work is the global energetic minimum of the molecule. Whilst more than one seeding conformation can be used, granting a greater exploration of the potential energy surface, we found that for the molecules considered in this work, the Boltzmann weight of the global minimum exceeded 0.75 in all cases. By evaluating the first- and second-order spatial derivatives of the potential energy (Jacobian and Hessian, respectively) at the seeding geometry, one can construct an approximate local potential energy surface through a Taylor expansion. The dynamics on this local approximation to the potential energy surface are then governed by a set of harmonic equations of motion, referred to as the molecular normal modes [65].

Here we outline the major features of our methodology, whilst a more thorough description of is given elsewhere [66, 67]. For an N-atom molecular system, we can define a 3 N × 3 N transformation matrix, D, that converts a mass-weighted Cartesian state vector, q, to an internal coordinate state vector, s. Given D, we can transform the mass-weighted Cartesian Hessian, Hq, to an internal coordinate basis through

\[
D^T H_q D = H_s
\]

The frequencies of the molecular normal modes are then given by diagonalising Hs

\[
\mathcal{E}^{-1} H_s \mathcal{E} = I \lambda
\]

where \( \mathcal{E} \) correspond to the eigenvectors of \( H_s \), I is the identity matrix, and the eigenvalues of \( H_s \) are given by the 3 N diagonal elements \((I_3 \lambda)_{ii} = \lambda_i \). The \( i \)th normal mode frequency, \( v_i \), is related to the \( i \)th eigenvalue through the expression

\[
v_i = \sqrt{\frac{\lambda_i}{4\pi^2 c^2}}
\]

where \( c \) is a conversion factor incorporating the speed of light and a conversion from atomic units to reciprocal centimetres. Six of these normal mode frequencies are equal to zero in an internal coordinate basis, corresponding to the global translational and global rotational degrees of freedom of the molecular system.

The amplitude of vibration of the \( i \)th normal mode, \( A_i \), is given by the standard expression for a simple harmonic oscillator,

\[
A_i = \sqrt{\frac{2E_i}{k_i}}
\]

where \( k_i \) is the force constant of the \( i \)th normal mode (calculable from \( v_i \)), and \( E_i \) is the energy available to it. Each normal mode is allocated an amount of energy given by a standard equipartition, \( E_i = k_B T/2 \), where \( k_B \) is the Boltzmann constant and \( T \) is the temperature at which the simulation is performed. To allow for a little more flexibility, each of the \( E_i \) is subjected to a stochastic Gaussian fluctuation. Given the amplitude and frequency of the oscillator, each normal mode can evolved in discrete time.

The final matter requiring discussion is how time is discretized for our equations of motion. Given the frequency of the oscillator, we can compute its time period, \( T_i = 1/v_i \). We choose a discrete timestep, \( \Delta t_i \), such that for each time period, \( T_i = \Delta t/\text{cycle} \), where \( \text{cycle} \) is a user-defined parameter. After every \( \text{cycle} \) timesteps, we also perturb the energy available to each normal mode by a new Gaussian-distributed number. To reduce the correlation between samples, a final user-defined parameter, \( n_{\text{out}} \) is also defined. We define \( n_{\text{out}} \) to correspond to the number of discrete timesteps that we allow to pass before outputting a sample to the training set. For the work conducted here, we set \( \text{cycle} = 10 \) and \( n_{\text{out}} = 100 \).

### 3 Computational methods

#### 3.1 The GAIA protocol

Three molecules (methanol, NMA and peptide-capped glycine) have been selected as case examples. Initially, the methanol and NMA molecules were generated in Gaussian-View and optimized to a minimum energy geometry using the Gaussian 09 program, at HF/6-31+G(d,p) level of theory. Single-point energy calculations were performed on the resulting structures, outputting the respective molecular wavefunction and Hessian of the potential energy, calculated at the same level. For the capped glycine molecule, we selected the global minimum conformation from the nine known energetic minima described in a previous publication [68]. For glycine, the calculations were performed at B3LYP/apc-1 [69] level of theory, in-keeping with the level of theory used in previous research [68]. Working at B3LYP level complements a recent publication [41] that validates the extension of the IQA approach to the B3LYP density functional. Prior to such work, the typical IQA partitioning restrictions demanded a well-defined second-order density matrix, thus ruling out correlation-inclusive and approximate Hamiltonian theories, including the density functional theory (DFT) functionals [40, 70–72].

After obtaining the molecular wavefunctions, the process of sampling, performing the energy partitioning and building the kriging models was achieved using the in-house pipeline software, known as the GAIA protocol. The
GAIA protocol is outlined in Fig. 3, which displays the sequence of steps, flowing left to right, starting with development of sample geometries and terminating with analysing the models.

GAIA automates the passing of information between two in-house Fortran 90 programs (TYCHE and FEREBUS—orange boxes in Fig. 3) and two commercially available programs (GAUSSIAN09 [73] and AIMAll [52]—green boxes). The output data from one step subsequently forms the input for the following step, until a seeding geometry (or set of seeding geometries) has been converted into a fully trained kriging model. Each program’s role within GAIA can be summarized in a few lines:

1. TYCHE: distorts an input seed geometry, using the molecular normal modes, to create a broad range of sample geometries that collectively describe a local patch on the molecular potential energy surface (around the seed).

2. Gaussian 09 [73]: performs single-point energy calculations and outputs the wavefunction of each molecule.

3. AIMAll [52] (version 15.09.12): starts from the wavefunction of a molecule to obtain the IQA energy partition values: $E_{\text{IQA}}$, $E_{\text{intra}}$, $V_{\text{AA}}$, $V_{\text{XC}}$ and $V_{\text{cl}}$.

4. FEREBUS: uses a training set of molecular geometries to build kriging models of atomic energy (any of the five types above). FEREBUS then validates each model by predicting a test set and comparing the models’ predicted value to the known true value.

The GAIA protocol outlined here is a slight deviation to that reported [22] before for the parameterization procedure of FFLUX. The deviation is a result of the current exclusion of atomic multipole moments but incorporation of the IQA atomic energy components instead. Thus, Fig. 3 represents the protocol tailored to this investigation only.

A set of 4000 initial samples were generated for each molecule by TYCHE from the distortion of a single energy minimum, at a user-defined temperature of 450 K. After single-point energy calculations and wavefunctions were obtained from Gaussian 09 for every sample, IQA energy contributions were obtained from AIMAll (with default quadrature and integration grid options). We set to the value of 3 the AIMAll parameter ‘-encomp’ referring to the IQA energies to be computed. As soon as one atom attains a Lagrangian integration error, $L(\Omega)$, greater than the user-defined threshold of 0.001 Hartree, then this atom is removed from the training set, as well as all remaining atoms of the molecular geometry in which the offending
atom occurred. This process is known as scrubbing in GAIA. Scrubbing ensures that samples with “noisy” atomic energies (i.e. large ΔΩ value) are excluded from the development of the model. From the samples remaining in the pool after scrubbing, 500 are set aside as the test set and the remaining number of samples (to the nearest hundred) are used for the training set. The resulting training sets were 3400, 3300 and 3000 for methanol, NMA and capped glycine, respectively. These training sets were then employed to generate kriging models for each molecule using the in-house program FEREBUS. The kriging parameter, pₖ, was optimized in the development of all the models. The settings in FEREBUS were as follows: noisy kriging was not requested, tolerance set to 10⁻⁹, convergence to 200 and the swarm-specifier to “dynamic”. Finally, so-called S-curves are produced to illustrate the energies errors on each molecular model. The development and meaning of an S-curve is described in the next section.

### 3.2 Energy error analysis

As announced earlier, each molecule will be modelled using three approaches, resulting in a tiered level of chemical detail (in this investigation), separating the interatomic and ionic-like electrostatic components, and again returns models for each.

- **Approach A**: Modelling the molecule using only the total unpartitioned atomic energy, E_{IAA}.
- **Approach B**: Modelling the molecule using the intra-atomic (E_{intra}⁺) and interatomic (V_{AA}⁻) atomic energies.
- **Approach C**: Modelling the molecule using the intra-atomic energy (E_{intra}⁺) and the two key interatomic energies: exchange–correlation (V_{AA}⁻) and classical electrostatic (V_{cl}⁻).

Approach A provides the fastest (computationally) and simplest model of a molecule, at the atomic level. Approach B offers a chemically intuitive separation of the intra-atomic and overall interatomic energies and provides models for both. Approach C offers the highest level of chemical detail (in this investigation), separating the interatomic interaction energy into the covalent-like exchange and ionic-like electrostatic components, and again returns models for each.

Moving on to the analysis of the models, we should reintroduce how S-curves can be used to fully convey the quality of a kriging model. The S-curve is a cumulative distribution function (up to 100 %) of absolute energy errors for each test point within the test set. An S-curve plots the absolute (energy) error over the whole molecule (x-axis) versus the test set data point (i.e. molecular geometry) (y-axis) as represented as a percentage (100 %/500 data points = 0.2 % per data point). Thus, each test set molecular geometry point corresponds to one point on the S-curve. In order to plot the total molecular energy error (x-axis), it must be calculated the generalized expression appearing in Eq. (13),

$$\Delta E_{\text{Molec}} = \sum_{A} \sum_{Y} |E_{\text{Y},\text{Act}}^A - E_{\text{Y},\text{Pred}}^A|$$

where ‘Y’ is a general notation representing any of five possible IQA energy contributions (E_{IAA, Act}⁺, V_{AA}⁻, V_{cl}⁻ and V_{cl}⁺), and n is the number of atomic energies being used to describe an atom (or the total atomic model). In this work n can be one, two or three only (hence the upper limit n ≤ 3 in Eq. 13). The value of n depends on the modelling approach. In particular, for approach A we have that n = 1 (E_{IAA}), for approach B n = 2 (E_{intra}⁺ and V_{AA}⁻) and approach C leads to n = 3 (E_{intra}⁺, V_{AA}⁻ and V_{cl}⁻). Before summing over the atoms, counted by index A, a sum over n atomic energies must take place in order to obtain the atomic model. When a model is tested, the predictions of the model can be averaged and compared to the average value of the true values. As a result, a model can be said to, on average, slightly over- or under-predict as determined by a positive or negative difference between the averaged predicted and true values. Therefore, summing over these energy models allows for a cancellation of such errors across the models in two possible ways: (1) across the atomic energies that together constitute a single atom, which is atomic cancellation, and (2) across the atomic models that together constitute the molecule, molecular cancellation. The value obtained for ∆E_{\text{Molec}}, as plotted on the S-curve x-axis, represents the final error for the molecular energy.

The mean absolute error (MAE) for the molecular model is calculated according to Eq. (14). The MAE can be used as a simple measure of the model quality and can be calculated for a single energy model or for a collection of models (such as the resulting molecular model, see Eq. 13),

$$\Delta E_{\text{Molec}} = \frac{1}{N_{\text{test}}} \sum_{M=1}^{N_{\text{test}}} \Delta E_{\text{Molec}}$$

where the sum runs over the N_{test} = 500 molecular geometries of the test set (of methanol, NMA or glycine).

A final measure, the MAE percentage error, MAE%, can also be calculated by dividing the MAE by the size of the energy well range of the test set. This error is given in Eq. (15),

$$\text{MAE} = \frac{\Delta E_{\text{Molec}}}{E_{\text{MAX}} - E_{\text{MIN}}}$$

where ‘MAX’ refers to the highest molecular energy in the test set, and ‘MIN’ to the lowest. Note that the Electronic
Supplementary Material reports atomic MAE% values, which are defined in Eq. S1 there, by simply replacing $\Delta E_{\text{MAE}}$ by $\Delta E_{\text{Atom}}$. Converting the error into a percentage allows a transferable measure, independent of the energy range of the sampling well, thereby making the MAEs from different molecules more comparable. The MAEs can also be used to compare the quality of individual atomic energy models (also in Eq. 15), which individually may experience a broad variety of energy ranges.

Finally, this work shows, for the first time, $S$-curves of the complete molecular energy ($\Delta E_{\text{Molec}}$) rather than only the multipolar electrostatic energy or later the kinetic energy. Because multipole moments are not used in this work all atoms can interact with each other electrostatically (without concerns about possible divergence). In other words, the complete electrostatic interaction is subject to kriging here, for the first time, covering all 1,2; 1,3 and 1,4 interactions.

4 Results

4.1 Methanol

The 4000 samples (i.e. molecular geometries) generated by TYCHE (see Fig. 4) were subject to scrubbing in GAIA, followed by 500 samples then set aside as the test set. After rounding down to the nearest hundred, 3400 samples remained and formed the training set. These 3400 training set samples sampled an energy well with an energy range of $\pm 115$ kJ mol$^{-1}$.

Figure 5 plots the $S$-curves for methanol for each of the three modelling approaches ($A = E_{\text{IQA}}^{A}$, $B = E_{\text{intra}}^{A}$ and $V_{\text{inter}}^{AA'}$, $C = E_{\text{intra}}^{A}$, $V_{\text{xc}}^{AA'}$ and $V_{\text{cl}}^{AA'}$). It is clear that over 95% of the test set geometries have $\Delta E_{\text{Molec}}$ energy errors below 1 kJ mol$^{-1}$, across all modelling approaches. Such a low error is a pleasing result and an encouraging start to this analysis. An analysis of each molecular model (i.e. approach) is given in Table 1. Interestingly, the simplest model, approach $A$, performs the best out of the three approaches with a MAE% error of 0.3%. Approaches $B$ and $C$ perform very similarly with an MAE% error of 0.4% each, respectively.

Notably, the maximum absolute error observed for approach $B$ is 2.3 kJ mol$^{-1}$, larger than that for approach $C$, which returns 1.7 kJ mol$^{-1}$. One would expect that the most chemically insightful approach, which is $C$, is the one that accumulates the highest error. This presumption follows from the fact that approach $C$ (which has 18 models, or 3 energies for each of the 6 atoms) has 6 additional models compared to approach $B$ (with 12 models, or 2 energies for each of the 6 atoms), and the extra models accrue additional kriging errors with each model. This matter will be discussed in Sect. 5.2.

At the atomic level, Fig. 6 shows the MAEs for each atomic energy type ($E_{\text{IQA}}^{A}$, $E_{\text{intra}}^{A}$, $V_{\text{inter}}^{AA'}$, $V_{\text{xc}}^{AA'}$ and $V_{\text{cl}}^{AA'}$) for each atom in methanol. At first glance, carbon influences the accuracy of the model. In general, $C1$ has the highest MAEs and is therefore the least accurately modelled atom overall. Following $C1$, $O2$ has the next highest errors, followed by the methyl hydrogens ($H3/4/5$), and finally the most accurately modelled atom, the alcoholic hydrogen $H6$. In assessing how the IQA atomic energy types compare, the story is also clear. Without exception, the following order appears, starting with the least accurate: $E_{\text{intra}}^{A}$, $V_{\text{inter}}^{AA'} < V_{\text{xc}}^{AA'}$ and $V_{\text{cl}}^{AA'}$. The errors for all energies for any atom never exceed 0.8 kJ mol$^{-1}$. The interplay between these energies will be discussed in the Sect. 5.

Only looking at the MAE ignores the range of energy that a particular energy has been subjected to in the sampling stage. The MAE percentage error (MAE%) makes the
MAEs more comparable when assessing the difficulty for the kriging engine. The MAE percentage errors, along with the MAEs and energy ranges, are tabulated in Table S1 in the Electronic Supplementary Material. Approach A corresponds to S1 (a), approach B to S1 (b) and approach C to S1 (c). The analysis for the EA intra model [used in both approaches B and C and given in S1 (b)] is not repeated in S1 (c), as the same model is used for each approach. Figure S1 plots MAE versus atomic energy range in order to observe any correlation between them. Some weak correlation can be seen, but nothing strong enough to validate such a relationship.

**4.2 NMA**

The NMA models were trained using 3300 training set samples (Fig. 7), sampling an energy well with an energy range of ~84 kJ mol\(^{-1}\).

Figure 8 plots the S-curve for each NMA molecular model.

In Fig. 8, almost 95% of the all test set samples have the \(E_{\text{Molec}}^A\) energy correctly predicted within 2.5 kJ mol\(^{-1}\) (across all models). This time there is a clearer separation of the S-curve models, with again approach A (\(E_{\text{IQAn}}^A\)) being the best modelled of the 3 approaches. Interestingly and unexpectedly, approach C (\(E_{\text{intra}}^A, V_{\text{AA'}}^\text{cl}, V_{\text{AA'}}^\text{cl}'\)) performs better than approach B (\(E_{\text{intra}}^A\) and \(V_{\text{AA'}}^\text{inter}\)) for the NMA system. Given this result, we must now consider whether the dual-cancellation (atomic and molecular) allowed in Eq. (13), prevents

---

**Table 1**  Quantitative analysis of the methanol models

<table>
<thead>
<tr>
<th>Measure</th>
<th>(E_{\text{Molec}}^A) (approach A)</th>
<th>(E_{\text{intra}}^A) &amp; (V_{\text{AA'}}^\text{cl}) (approach B)</th>
<th>(E_{\text{intra}}^A) &amp; (V_{\text{AA'}}^\text{cl}) &amp; (V_{\text{AA'}}^\text{cl}') (approach C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum absolute error</td>
<td>1.5</td>
<td>2.3</td>
<td>1.7</td>
</tr>
<tr>
<td>Minimum absolute error</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Absolute error range</td>
<td>1.5</td>
<td>2.3</td>
<td>1.7</td>
</tr>
<tr>
<td>Mean absolute error (MAE)</td>
<td>0.3</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Standard deviation (SD)</td>
<td>0.2</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Test set energy range</td>
<td></td>
<td>86.0</td>
<td></td>
</tr>
<tr>
<td>Training set energy range</td>
<td></td>
<td>114.9</td>
<td></td>
</tr>
<tr>
<td>MAE% error</td>
<td>0.3</td>
<td>0.4</td>
<td>0.4</td>
</tr>
</tbody>
</table>

All energies are given in kJ mol\(^{-1}\). MAE% represents the MAE error with respect to the energy range of the test set.
a straightforward correlation between the number of atomic models composing a molecular model and the quality of the molecular model. Glycine will further aid our understanding of this, and the topic will be discussed further in the Sect. 5.

The MAE percentage errors for approaches A, B and C are 0.6, 1.3 and 0.9 %, respectively (see Table 2). Again, this is a pleasing result considering the molecular complexity has risen from \((3\times6)−6=12\) geometrical features for methanol, to \((3\times12)−6=30\) features for NMA. While the number of geometrical features increased by a factor 2.5, the errors for approaches A and C barely doubled. This favourable behaviour stimulates a further upscaling of features. With the \(S\)-curves being less entangled for NMA, the maximum absolute error falls in line with the shape and position of the respective \(S\)-curve.

Figure 9 is the counterpart of Fig. 6, this time for NMA. Here, we can further confirm that the atomic energy MAEs appear directly related to both the element and energy type being modelled. Initially, the atoms forming NMA can be immediately separated into their elements for the carbon, oxygen and hydrogen atoms, but the nitrogen atoms have very similar errors to the methyl carbons. In NMA, the MAEs also separate the atoms into atom types for carbon (carbonyl carbon and methyl-cap carbons are easily distinguishable). In fact, the oxygen is also modelled so well it is close to being indistinguishable from the hydrogens. Following this, the same trend in energy prediction accuracy is present in NMA as it was in methanol, that is, the sequence \(E^A_{\text{intra}} < V^A_{\text{inter}} < E^A_{\text{IQA}} < V^A_{\text{XC}}\) (most accurate) remains valid.

The MAE percentage errors, MAEs and energy ranges are all reported for each atom in Table S2. Figure S3 once more confirms the weak correlation between MAE and energy range, this time for NMA. Figure S4 is analogous to Fig. 9, but plotting MAE% instead of MAE. In going from MAE to MAE%, the range of the energy is now incorporated. As we can see in Figure S4, the trend previously identified for the MAE \((E^A_{\text{intra}} < V^A_{\text{inter}} < E^A_{\text{IQA}} < V^A_{\text{XC}})\) is no longer true, and no clear trend is seen.

### 4.3 Glycine (Gly)

The capped glycine models were trained using 3000 training set samples (Fig. 10), sampling an energy range of \(-163\) kJ mol\(^{-1}\).

Figure 11 shows the \(S\)-curve for glycine.
The inset glycine conformation in Fig. 11 is the global minimum identified by TYCHE as forming ~75% of the Boltzmann distribution when all 9 energy minima were used as seeds. The predominance of the global minimum determined our choice to sample only around this conformation. Almost 95% of the all test set samples have the $E_{\text{IQA}}$ molecules energy correctly predicted within ~2.9 kJ mol$^{-1}$ (across all models). The $S$-curves show a resemblance to those in Fig. 5 but shifted to a higher prediction error. Here, there is no longer a clear separation of the $S$-curves for modelling approaches B and C. Approach A ($E_{\text{IQA}}^A$) still has the lowest errors of the three approaches. Surprisingly approaches B ($E_{\text{intra}}^A$ and $V_{\text{intra}}^{AA'}$) and C ($E_{\text{intra}}^{AA'}$, $V_{\text{intra}}^{AA'}$, and $V_{\text{cl}}^{AA'}$) is that they again result in the same MAEs of 1.1 kJ mol$^{-1}$. These data are presented in Table 3.

The MAE percentage errors for approaches A, B and C are 0.6, 0.9 and 0.9 %, respectively (see Table 3). Again, this is a very pleasing result considering the molecular complexity has once more risen from $(3 \times 12) - 6 = 30$ geometrical features for NMA, to $(3 \times 19) - 6 = 51$ features for glycine. In spite of a near doubling of the number of geometrical features, there is little change in the MAEs going from NMA to glycine.

Figure 12 reconfirms the trends identified in Fig. 6 and Fig. 9. Indeed, MAEs are related to the element being modelled and IQA energy type, with atom typing appearing even more evident. Within the carbons, three types are present but only two classes are distinguishable: $C_{\text{C}=O} > C_{\text{a}} ≡ C_{\text{methyl}}$. Once more, the nitrogen atoms are fairly indistinguishable from the latter class of carbons. This time, the amino hydrogens (H5 and H11) also appear with slightly higher errors than seen for the aliphatic hydrogens. When looking at the trends in the energies themselves, the trends previously seen for methanol and NMA are once more observed in Gly, where $E_{\text{intra}}^A < V_{\text{intra}}^{AA'} < E_{\text{IQA}}^A < V_{\text{cl}}^{AA'} < V_{\text{XC}}^{AA'}$ (most accurate). A direct comparison of the errors on the atoms present in both NMA and Gly will be given in the Sect. 5.

Table S3 and the corresponding plots of Figures S5 and S6 are analogous to Table S2 and Figures S3 and S4, this time for glycine. The same observation of a weak correlation between energy range and MAE for each IQA energy type is made in Figure S5. Figure S6 displays the MAE% values for capped glycine. It is clear that 13 out 19 atoms show $E_{\text{IQA}}$ standing out as the least accurate energy to model. For previous systems this majority trend was not seen. However, this conclusion can be rationalized by remembering that $E_{\text{IQA}}$ is the total atomic energy, and hence influenced by every type of energy change within the atom. Hence, it would be reasonable for it to be the most sensitive when the energy is considered relative to the energy range.

### 5 Discussion

The discussion is divided into four subsections covering key topics that have either been postulated at the beginning...
of the research or have arisen during the analysis of the results. The first two subsections each correspond to an objective.

5.1 Feasibility of modelling IQA energies

The first objective of this investigation is to assess the suitability and quality of modelling the five IQA energies that can be used in the formulation of FFLUX. In short, we have successfully kriged models and made good molecular energy predictions using three possible combinations of the IQA energies, making them all suitable for use in FFLUX. The resulting molecular models having excellent errors, of less than ±0.4, 1.3 and 0.7 kJ mol\(^{-1}\) for methanol, NMA and Gly, respectively. Qualitatively speaking, Figs. 5, 8 and 11 display behaviour analogous to previous S-curves in previous literature [31, 39], leading to an overall successful prediction of both multipolar electrostatic and non-electrostatic energetics.

In order to quantitatively compare the quality of the results, we need to draw on a previous paper [39] where a component of \(E_{\text{intra}}^A\), namely the kinetic energy \(T^A\), was kriged for every atom in a similar set of systems (methanol, NMA, glycine and triglycine). We decided to compare our results with the kinetic energy results, only for methanol and NMA, because their training set sizes match best. In that work [39], MAEs for the atomic kinetic energy were obtained of 0.8 kJ mol\(^{-1}\) (0.1 %) and 0.7 kJ mol\(^{-1}\) (0.3 %) for a methanol–carbon and the carbonyl–carbon in NMA, respectively. In our work, we have presented differing training set sizes (3400 and 3300 for methanol and NMA, respectively), but it is still useful to compare the results. For \(E_{\text{intra}}^A\), we obtain MAEs of 0.7 kJ mol\(^{-1}\) (0.3 %) and 1.5 kJ mol\(^{-1}\) (0.4 %), respectively, and for \(V_{\text{inter}}^{AA'}\) we have MAEs of 0.5 kJ mol\(^{-1}\) (0.3 %) and 1.0 kJ mol\(^{-1}\) (0.2 %) for the equivalent atoms, respectively. Hence, despite our training sets being larger, the MAE errors remain slightly higher than those observed for the kinetic energy. This confirms an initial suspicion that the summative nature of both \(E_{\text{intra}}^A\) and \(V_{\text{inter}}^{AA'}\) results in a more complicated kriging problem than an example of the subcomponents (\(T^A\)) forming these energies. However, our overall similar performance is still very promising given that we have the complete energy of an atom \(A\) (and thus a molecule when summing over \(A\)) being modelled with comparable errors, albeit using larger training sets. Another advantage of this investigation is the ability to krige only one, two or three energies, yet still capturing the energetic behaviour of the whole molecule. Hence, this design saves substantial computational time by not needing to krige every individual IQA atomic energy (\(T^A, V_{\text{ex}}^{AA'}, V_{\text{ex}}^{AA'}, V_{\text{en}}^{AA'}, V_{\text{en}}^{AA'}, V_{\text{nn}}^{AA'}\) and \(V_{\text{nn}}^{AA'}\)) (should \(V_{\text{ex}}\) still remain unpartitioned). It is also noted that across all atoms in all three systems investigated here, the MAE error never exceeded 1.5 kJ mol\(^{-1}\) (with the majority under 1 kJ mol\(^{-1}\)), or a MAE percentage error over 1.4 %, for any energy.

Another measure of quality to loosely compare our results with are the previously kriged electrostatic multipole moments, which describe the classical electrostatic interaction energy for 1,4 and higher order interactions [33]. Here, the notation ‘1,4’ describes the interaction between atoms separated by 3 covalent bonds. A 1,5 interaction has 4 separating covalent bonds, and so on. For 1,4 and higher interactions (i.e. 1, \(n > 4\)) in a capped histidine system (29 atoms), the MAE for the intramolecular electrostatic energy calculated through kriged multipole moments, was 2.5 kJ mol\(^{-1}\). In our investigation \(V_{\text{cl}}^{AA'}\) and \(V_{\text{inter}}^{AA'}\) never exceed an MAE of 1.4 kJ mol\(^{-1}\). Is this MAE respectable compared to the multipolar electrostatic energy error? Yes, because \(V_{\text{cl}}^{AA'}\) (and \(V_{\text{inter}}^{AA'}\) too) accounts for all electrostatic interactions and the multipolar electrostatic analysis only for 1,4 and higher interactions (due to convergence limitations). Admittedly, a training set of only 600 training set geometries [33] was used for the latter analysis.

Finally, we point out that we are currently investigating a potential reduction in the number of training set samples needed to obtain suitably accurate atomic and molecular models. This research is focussed on the selective building of training sets, and a variety of approaches are currently being investigated to achieve this.

5.2 Cancellation of errors

A second objective of the current investigation is to observe to what extent any cancellation of errors takes place within the summative combination of kriging models described in Eq. 9. As previously described, this approach offers the potential to benefit from the fortuitous cancellation of errors, but also equally the unfortunate accumulation of errors as a result of the machine learning. In particular, an atomic energy component may be, on average, predicted to be more stable than the true energy (i.e. overestimated). Another atomic energy component may be, on average, predicted to be less stable than the true energy (i.e. underestimated). As a result, the summative combination of both an over- and underestimated result allows for some cancellation, resulting in an overall energy prediction being more accurate, when only these two energies are considered. Accumulating these cancellations further across many energy models and across many atoms increases this effect dramatically.

Despite the cancellation appearing to rely on chance, since there is no control for the over- or underestimation of the energy models, the prediction of the kriging models will always be consistent, i.e. the same geometrical features are used to map all atomic energies within a single atom,
and each atom’s geometrical features are related to every other atom in the molecule. Hence, if one energy is overestimated and the other is underestimated with those identical or related features, then this same interplay between the kriging models will always be present. Naturally, the opposite can also occur where, for example, two overestimating models result in a summed higher total error instead of cancelling. However, from the results in Figs. 6, 9 and 12, it is evident that the summative energies ($E_{\text{IQA}}^A$ and $E_{\text{AA}}^A$) generally have lower errors than by simply summing the absolute errors of the components that make up these energies ($E_{\text{intra}}^A + E_{\text{inter}}^{AA}$ and $V_{\text{XC}}^A + V_{\text{cl}}^{AA}$, respectively). Evidence of the molecular cancellation (described in Sect. 3.2) can also be seen, given that atomic MAE often are between 0.5 and 1.5 kJ mol$^{-1}$, but the resulting molecular MAE is always $\leq 1$ kJ mol$^{-1}$ (for NMA).

5.3 $S$-curve analysis

The MAEs (and MAE percentage errors) of the total $E_{\text{IQA}}^\text{Molec}$ molecular energy, as given in Tables 1, 2 and 3 for each respective molecule, are arguably the most important values obtained in this investigation. Therefore, these values are representative of an overall quality check for this investigation. Through averaging the MAE of the three approaches (A, B and C) for each molecule, we obtain a single MAE error for each molecule: 0.3 kJ mol$^{-1}$ (methanol), 0.7 kJ mol$^{-1}$ (NMA) and 0.9 kJ mol$^{-1}$ (Gly). These $<1$ kJ mol$^{-1}$ results become more impressive when the total energy for each system is compared: methanol has a molecular energy of $\sim 302,000$ kJ mol$^{-1}$, NMA $\sim 648,700$ kJ mol$^{-1}$ and Gly $\sim 1,191,500$ kJ mol$^{-1}$.

Of the three approaches, approach A consistently is the most accurate for each molecule. This is a result of the minimal approach incorporating only a single model per atom in the molecule. However, the performance of approaches B and C were less distinguishable or predictable. Either approach was capable of being slightly more accurate than the other. However, the molecular error when calculated using either approach B or C was always within 0.3 kJ mol$^{-1}$ of one another. In conclusion, we consider all three routes as suitable candidates for modelling the molecular energy, each incorporating a different level of chemical insight.

One further point to note in the analysis of our $S$-curves, are the “rogue points” present near the 100 % ceiling of the plots. Few rogue points occur for approach B in methanol, but more noticeably for approaches B and, in particular, approach A for Gly. These points are considered rogue due to the large gap that appears separating these points from the almost continuous S-like shape of the plot. Figure 13 (on capped glycine) sheds lights on how rogue points arise.

In Fig. 13, the glycine geometries that passed GAIA’s scrubbing procedure are plotted according to their molecular energies and separated according to which set they belong to, that is, the training set in blue or the test set in red. The information shown in Fig. 13 is essentially a one-dimensional distribution of molecular energies (y-axis) but spread out in two dimensions by introducing an x-axis that merely counts the 3000 training set samples and the 500 test set samples. The test set remains identical for all three modelling approaches.

A large vertical gap between the blue points in Fig. 13 indicates a lack of training points in that energy region. On the other hand, continuous lines indicate a high density of points, covering well the corresponding energy region. The point in the test set at $-1,198,412.1$ kJ mol$^{-1}$ (encircled green in Fig. 13) is the geometry corresponding to the maximum predicted error ($8.9$ kJ mol$^{-1}$) seen in the $S$-curve of approach A (utmost right point in the red curve in Fig. 11), denoted RTP1 (Rogue Test Point 1). The point in the test set at $-1,198,406.7$ kJ mol$^{-1}$ (encircled orange in Fig. 13) is the geometry corresponding to the maximum error ($7.9$ kJ mol$^{-1}$) seen in the $S$-curve of approach B (utmost right point in blue curve in Fig. 11), denoted RTP2. Both these molecular energies appear to be reasonably well sampled in the training set, with nearby (blue) points of $-1,198,412.5$ and $-1,198,406.9$ kJ mol$^{-1}$. Unusually, there was no problem for approaches B and C in predicting the molecular energy for RTP1, with errors of 3.4 and 1.3 kJ mol$^{-1}$, respectively. Similarly, for RTP2, errors of 1.1 and 3.1 kJ mol$^{-1}$ were obtained for approaches A and C. This suggests that it is generally not a lack of training geometries that are causing the large rogue errors. Instead it could be any one (or combination) of the following three effects: (1) the potential energy surface around these points is undulant (in general for the molecule, or for a particular IQA energy) and the training geometries included are
not enough to fully capture this landscape adequately, (2) the test points are outside of the domain of applicability, defined as the region of conformational space that can be interpolated by the training points of the kriging model. In other words, points lying outside of the domain of applicability correspond to points that lie outside of the training set, and so the kriging model is required to perform an extrapolation to make a prediction [74], (3) when summing across models, the balance of accumulating errors results in reduced cancellation of errors. All these reasons would also be supported by working with the higher-energy geometries where there are fewer samples in the training set, compared to those closer to the energy of the seed minimum.

Another measure that is useful when analysing the cause of a poor prediction within FEREBUS is the mean signed error (MSE) (or mean signed deviation, MSD). In statistics, the MSE is a measure of how close a predicted value matches the true quantity. Having a high MSE for a particular prediction indicates that the model is not well trained for in that region, and is a hallmark of working outside the domain of applicability. Taking glycine’s approach A as an example, the C12 atom stands out as an atom with a particularly poor $E_{\text{IQA}}$ prediction for RTP1 (an error of 5.2 kJ mol$^{-1}$). C12 also has an MSE approximately five times the average across all of the test geometries. Some of the other atoms in glycine also indicate a slightly increased difficulty in predicting $E_{\text{IQA}}$ for this test geometry, but not to the same degree as for C12. Hence, it can be concluded that C12 is the source of the error for RTP1, due to the model operating outside of its domain of applicability. Evidently, as approaches B and C perform well in predicting this molecular energy, this MSE explanation either is irrelevant when using $E_{\text{intra},i}$, $V_{\text{intra}}$, $V_{\text{AA}'}$, and/or $V_{\text{AA}''}$, or the effect is dampened by the cancellation of errors. This type of analysis can be applied to any rogue point on an $S$-curve.

In contrast, the point at $-1.198,380.1$ kJ mol$^{-1}$ (encircled black in Fig. 13) in the test set (TP1) (which is not a rogue point), appears to be the least well sampled in the training set, but the predicted errors for this sample are 0.3 kJ mol$^{-1}$ (approach A), 2.4 kJ mol$^{-1}$ (approach B) and 2.0 kJ mol$^{-1}$ (approach C), thus, not near the maximal points on each of the $S$-curves. The unexpectedly good prediction of TP1 is credited to kriging’s impressive interpolation between two largely spaced training points.

### 5.4 Evidence for atom typing

Figures 6, 9 and 12 illustrated the ‘difficulty’ of modelling each atomic energy, according to the MAE. From this analysis, we learned that atoms belonging to a particular functional group had a MAE that distinguished some from others, independently of the IQA energies being used for this observation. Across our three systems, the carbonyl carbons were the most difficult atoms to model, with a maximum MAE value of $\sim 1.2$–$1.3$ kJ mol$^{-1}$, in both NMA and Gly. The carbonyl carbon was followed by similar maximum MAEs values for the $\alpha$-carbon, the amino nitrogens and methyl carbons of $\sim 0.5$ kJ mol$^{-1}$, in methanol, NMA and Gly. The oxygens were easily distinguishable with a much lower maximum MAE of around $0.1$ kJ mol$^{-1}$, followed by the consistently very accurately modelled hydrogens with maximal MAEs of $<0.1$ kJ mol$^{-1}$. It is interesting to note that the N1 atom in NMA, resulted in similar errors to that of the corresponding atoms N4 and N10 in Gly. The same is true of the NMA atoms carbonylic C2 (with C6 and C8 in Gly) and O3 (with O7 and O9 in Gly), within $\pm 0.2$ kJ mol$^{-1}$. Hence, with atoms having comparable MAEs across multiple molecules, there is some basic evidence of atom typing. However, it is not a rule that can be used in distinguishing all present functional groups, as evidenced by the difficulty in distinguishing $\alpha$-carbon, methyl carbon and the amino nitrogen groups, using only the MAEs. It would be interesting to see which further trends are observed when a broader range of functional groups are studied.

The discussion above is not the first time atom typing has been considered within the study of an energy partitioning. A recent article by Patrikeev et al. [75], investigated the performance of several density functionals in their evaluation of both Kohn–Sham and correlation kinetic energies of topological atoms, and also commented on discriminating atom types through such atomic descriptors. Initial findings for the Kohn–Sham energies indicated a strong link between some of the tested functionals and the atomic number (or element) of an atom. A further finding in the assessment of correlation kinetic energies allowed aromatic and aliphatic hydrogens to be separated. It should also be reiterated that IQA within DFT is a little tricky since the Kohn–Sham approach does not lead to exact correlated reduced density matrices [76].

### 6 A note on dispersion and transferability

The only type of energy contribution that is lacking in the current kriging treatment of all IQA energy contributions is that associated with dispersion. Admittedly, the current treatment includes electron correlation, but because we used B3LYP this electron correlation does not cover dispersion effects. However, soon-to-be-published work of our group successfully kriges the IQA intra-atomic, $E_{\text{intra}}$, and interatomic, $V_{\text{inter}}$, energies at the M06-2X/aug-cc-pVDZ level of theory. This functional describes (or mimics) some mid-range dispersion effects, but the ultimate goal of FFLUX is to invoke a post-Hartree–Fock method (non-DFT) to cover dispersion properly.
A second note concerns the transferability of the obtained model with respect to an exchange-correlation functional other than B3LYP. The only other exchange–correlation functionals implemented in the program AIM-ALL are LSDA and M06-2X. The current investigation has not been mirrored using any other functional from which we could directly compare transferability results. However, there is potential for transferability to be considered in the aforementioned work to be published, where M06-2X was used. Like in the current work, IQA atomic energy predictions made for multiple different systems could be compared. Some understanding of the transferability of another exchange–correlation functional can come from earlier work [77] from our laboratory in which the electrostatic energy, obtained through atomic multipole moments, was kriged at three levels of theory, namely HF, B3LYP and M06-2X. From that work, one would expect that M06-2X will perform similarly to B3LYP.

7 Conclusion

The development of the novel force field FFLUX now moves beyond its machine learning treatment of multipolar electrostatics. We demonstrate that short-range non-multipolar electrostatics can now also be kriged successfully. Moreover, (non-multipolar) exchange energies as well as intra-atomic energies (beyond just the kinetic energy) are now also kriged with promising energy errors. As a result, chemical bonding and stereo-electronic effects are now, by way of principle, incorporated in FFLUX. This achievement is realized within the context of the methodology of interacting quantum atoms (IQA).

Three approaches (A, B, and C), incorporating five IQA atomic energies ($E_{\text{IQA}}^\text{intra}$, $E_{\text{IQA}}^\text{inter}$, $V_{\text{AA}}^\text{intra}$, $V_{\text{AA}}^\text{inter}$, $V_{\text{XC}}$, and $V_{\text{El}}$), were successfully used to develop molecular models offering control in balancing accuracy and chemical insight. The most accurate and least expensive molecular model (approach A) was built using the total atomic energy $E_{\text{IQA}}^\text{A}$, with MAEs of ±0.3, 0.4 and 0.6 kJ mol$^{-1}$ for methanol, NMA and capped glycine, respectively. Interestingly, the more insightful formalisms involving the intra- and interatomic components (approach B), and the interatomic exchange and electrostatic contributions (approach C), resulted in similar MAEs. These errors are on a par with previous literature and are a result of the combination of models benefitting from cancellation of errors. The latter occur both within an atom’s total energy modelling (atomic cancellation), and also when summing across total atomic models (molecular cancellation) to obtain the molecular model.

In summary, the novel strategy and results were a successful proof-of-concept approach, developed to be integrated into FFLUX. Future work will build upon the method presented here employing the models in the application of geometry optimization, initially without, but later with, the incorporation of multipolar electrostatics. Future research is also focussing on creating intelligent training sets, designed to reduce the number of samples used in the building of kriging models.

Acknowledgments We acknowledge the EPSRC for funding through the award of an Established Career Fellowship to P.L.A.P. (Grant EP/K005472), and the BBSRC for studentships to S.C. and P.M.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

The computational prediction of Raman and ROA spectra of charged histidine tautomers in aqueous solution

Salvatore Cardamone, Beth A. Caine, Ewan Blanch, Maria G. Lizio and Paul L. A. Popelier

Histidine is a key component of a number of enzymatic mechanisms, and undertakes a myriad of functionalities in biochemical systems. Its computational modelling can be problematic, as its capacity to take on a number of distinct formal charge states, and tautomers thereof, is difficult to capture by conventional techniques. We demonstrate a means for recovering the experimental Raman optical activity (ROA) spectra of histidine to a high degree of accuracy. The resultant concordance between experiment and theory is of particular importance in characterising physically insightful quantities, such as band assignments. We introduce a novel conformer selection scheme that unambiguously parses snapshots from a molecular dynamics trajectory into a smaller conformational ensemble, suitable for reproducing experimental spectra. We show that the “dissimilarity” of the conformers within the resultant ensemble is maximised and representative of the physically relevant regions of molecular conformational space. In addition, we present a conformer optimisation strategy that significantly reduces the computational costs associated with alternative optimisation strategies. This conformer optimisation strategy yields spectra of equivalent quality to those of the aforementioned alternative optimisation strategies. Finally, we demonstrate that microsolvated models of small molecules yield spectra that are comparable in quality to those obtained from ab initio calculations involving a large number of solvent molecules.

1. Introduction

Raman optical activity (ROA) is a powerful chiroptical technique that can be used to probe a number of molecular properties. In the early days of its inception, ROA was used to deduce stereochemical information of molecular species by comparison with band patterns of related molecules. Unlike a number of other spectroscopies, ROA intensities in the low wavenumber region of the spectrum are rich in information, since these regions correspond to, for example, torsional degrees of freedom and large scale conformational modes of motion. ROA also possesses the capacity to determine the proportions of enantiomers within a sample, which is of great importance in purification processes.

The structural features of biochemical systems make them particularly amenable to ROA spectroscopy. ROA offers a far more robust analysis of biomolecular structure than more common forms of vibrational spectroscopy because of its sensitivity to chirality. The ROA signal of a system is largely dominated by the more rigid and chiral elements of that system, and in the case of proteins and peptides this corresponds to the secondary structure motifs exhibited by the polypeptides. This makes ROA highly useful for structural characterisation of complex biomolecules, in contrast to Raman spectroscopy, where amino acid sidechain signals typically dominate the spectrum. ROA is by no means constrained to peptide systems, and the ROA signals of other biochemical species, such as DNA bases and oligomeric nucleic acids, also prove informative. The utility of ROA in structural analysis has been exemplified over the years by a number of studies on the folding, assembly and dynamical properties of biochemical systems.

Histidine is somewhat unique in its ability to perform a number of biochemical roles. Traditionally characterised as a polar amino acid, histidine does not substitute particularly well with other amino acids. The nitrogen atoms of the imidazole group have long been known to act as proton shuttles, and grant functionality for a number of enzymatic mechanisms. However, histidine has also been shown to participate in...
interactions that are typically perceived to involve large aromatic amino acids such as tryptophan and phenylalanine. Noncovalent interactions involving histidine have been shown to exist in the form of stacking between its imidazole ring and a number of DNA nucleobases.13

Histidine features in the “catalytic triad” of a number of biochemically important enzymes, such as serine and cysteine proteases, where it acts as a proton shuttle during the process of protein catalysis.14 The Manzetti mechanism15 suggests that matrix metalloproteinases, a set of well-characterised cancer therapeutic targets, attain catalysis through a pair of histidine residues that coordinate a Zn$^{2+}$ ion. The penta-coordinate Zn$^{2+}$ ion is induced to act as a reversible electron donor, and can hydrolyse the scissile peptide bond. Indeed, research within the field of biocatalysis has found evidence for histidine residues within tripeptide motifs co-ordinating even more exotic metal ions, such as Pt$^{2+}$ and Au$^{3+}$, and exhibiting chemotherapeutic properties.16

Five protonation states of histidine exist: His$^{2+}$, His$^{1+}$, His$^0$, His$^{-1}$ and His$^{-2}$, where the superscript denotes the formal charge state. In addition to these protonation states, the imidazole ring in His$^0$ and His$^{-1}$ can exist in one of two tautomeric forms. This effect originates from the lone pairs of the nitrogen atoms contributing to π-delocalisation around the five-membered imidazole ring. The imidazole nitrogen atoms are labelled $N_a$ or $N_b$, the former being the nitrogen one bond away from the β carbon in the alkyl sidechain, and the latter being the nitrogen two bonds away from β carbon,17 as shown in Fig. 1. For convenience, we also depict the major torsional degrees of freedom that are typically varied for conformational studies of histidine in Fig. 1; $(\phi,\psi)$ are the conventional Ramachandran dihedrals, while $\chi_1$ and $\chi_2$ are the side chain dihedrals. Tautomeric forms are denoted by specifying the protonated nitrogen, i.e. $N_aH$ or $N_bH$. Using this nomenclature, we can uniquely label the seven microstates of histidine: His$^{2+}$, His$^{1+}$, His$^0[N_aH]$, His$^0[N_bH]$, His$^{-1}[N_aH]$, His$^{-1}[N_bH]$ and His$^{-2}$. We use the nomenclature “microstate” throughout this work in reference to one of these seven forms of histidine.

In this work, we present a number of developments that we have instituted and tested for the accurate computation of both Raman and ROA spectra. We have investigated only the His$^0$ (i.e. both His$^0[N_aH]$and His$^0[N_bH]$) and His$^{1+}$ charge states here, since these are the primary forms that histidine takes in the majority of biochemical systems at physiological pH. The issues of conformer selection, conformer solvation and conformer optimisation are each tackled independently, and their effects are demonstrated on both tautomeric systems and formally charged systems.

2. The problem of conformer generation

2.1 Solvation

The fact that biochemical systems are stabilised by solvent interactions makes their modelling problematic. Since the solvent plays an important role in dictating the energetics of a molecular system, it becomes imperative that spectroscopic predictions include some level of solvent modelling. It has been shown that simply omitting the solvent from spectroscopic calculations yields poor quality results that severely digress from experiment.18 Indeed, when dealing with charged species, such as zwitterionic amino acids, the molecules are energetically only stable if the solvent is included,19 and optimise to the neutral amine-carboxylic acid species in the absence of constraints.

The two major models for solvation effects in ab initio calculations are implicit and explicit solvation. With the former,20 the solvent environment is modelled as a homogeneous dielectric medium, reducing the magnitude of electrostatic interactions relative to the system in vacuo. The Polarisable Continuum Model21 (PCM) and Conductance-like Screening Model22 (COSMO) have been effective in predicting molecular energies,23-25 but lack any atomic-level properties of the solvent, such as the solvent’s ability to form hydrogen bonds with a solute. Implicit solvation methods have been found to offer minimal benefits in computing vibrational frequencies, and therefore vibrational spectra, relative to gas phase calculations.26

A number of explicit solvation schemes have been investigated. At the lowest end, one can include a small number of solvent molecules and treat the entire solute–solvent system quantum mechanically, perhaps embedded in an implicit solvation field. The number of explicit solvent molecules treated in this way is arbitrary, and their positioning around the solute can be problematic. Solvent molecules can be added in an ad hoc fashion around polar (or non-polar if the solvent is non-polar) groups, leading to significant improvement over gas phase spectra.27 However, a number of problems can also be attributed to this methodology, such as poor optimisation of the system geometry owing to its distance from a minimum on the potential energy surface.28

Another more popular method is the use of QM/MM, where the solvent is treated classically and the solute quantum mechanically. This has proven to be a very successful method for calculating spectra. For example, Cheeseman and coworkers29 have predicted the methyl-β-D-glucose ROA spectrum to an impressive level of accuracy, reproducing the vast majority of

Fig. 1 Zwitterionic histidine (His$^0[N_aH]$) where the imidazole nitrogen atoms have been labelled as indicated in the text. The two sidechain dihedral angles, $\chi_1$ and $\chi_2$, are also labelled.
spectral features. More recently, this approach has been further validated on a number of other systems, such as glucuronic acid\(^{30}\) and \(\beta\)-D-xylose.\(^{31}\) However, one is still constrained in selecting a conservative number of solvent molecules to include in the calculation for it to remain tractable; simply including hundreds of solvent molecules in the QM/MM calculation is not routinely feasible.

2.2 Optimisation

Complications arise in the QM/MM method of modelling solvent during the geometrical optimisation of the system. For the normal modes of motion of the system to be correctly determined, the system is required to relax to an energetic minimum on the potential energy surface. Geometry optimisation is notoriously difficult for a system including even a small number of explicit solvent molecules;\(^{30,31}\) the inclusion of water molecules in these studies flattens the potential energy surface, making it difficult to find well-defined minima. Two methods to resolve this problem have therefore been proposed: OptAll and OptSolute. The OptAll scheme optimises the entire QM/MM system (with a less rigorous convergence threshold than is usually implemented), while the OptSolute method freezes all solvent molecules and optimises only the solute in the field of the static solvent. The latter method is naturally much quicker than the former, but was found to yield poorer, although still informative, quality spectra. Some trade-off between accuracy and computational tractability is to be expected, but without there being an accepted methodology for qualitatively comparing predicted spectra to experimental spectra, it is difficult to ascertain whether the improvement in spectral quality warrants the additional computational efforts.

3. Technical details

3.1 Experimental setup

We have prepared samples of histidine at a pH of 7.8 and 4.2, both at a concentration of 0.26 M (equivalent to 40 mg mL\(^{-1}\)). Spectra were collected over a period of 7 hours, using an incident wavelength of 532 nm, a laser power of 700 mW and a 1.47 s exposition time. All spectra have been normalised so that they can be directly compared with the computed spectra. The absolute amplitude is therefore irrelevant in our analysis.

3.2 Molecular dynamics

The arrangement of solvent molecules is typically accomplished by use of molecular dynamics. The system is solvated and the resultant trajectory is parsed into a number of snapshot configurations. These configurations are subsequently used as the conformational ensemble for spectral calculations. Recent work\(^{32}\) has selected the snapshots entirely randomly, with very rough (qualitative) estimates used to maximise the conformational diversity of the ensemble. The snapshots are then energetically minimised, and those snapshots with a high Boltzmann weight form an ensemble that is considered representative of the dominant system conformations.

We have conducted 100 ns molecular dynamics trajectories for His\(^{[\text{N,H},\text{H}]}\), His\(^{[\text{N,H}]}\) and His\(^+\), solvated in boxes comprising 469 water molecules, with a chloride ion added in the latter case to ensure the system remained electrically neutral. Throughout, we have used the GROMACS\(^{33}\) molecular dynamics package in conjunction with the OPLS-AA\(^{34}\) force field.

Energy minimisation of these systems was initially undertaken with the steepest descent method, until convergence of the total system energy was attained. Following energy minimisation, we have conducted two separate equilibration phases. The first was a 1 ns \(NVT\) equilibration, using the Berendsen thermostat to maintain a temperature of 300 K, whilst the second was a 1 ns \(NPT\) equilibration using the Parrinello–Rahman barostat to maintain both a temperature of 300 K and a pressure of 0.1 MPa (1 bar). The particle mesh Ewald methodology was used to treat long-range electrostatics. A cut-off distance of 10 Å was chosen for Coulombic and van der Waals interactions. Both stages of equilibration were verified as being properly equilibrated by checking the convergence of the temperature for the \(NVT\) equilibration and the density of the system for the \(NPT\) equilibration. Both showed convergence after roughly 100 ps, and so our equilibration was assumed to be sufficient.

Our final step involved a 100 ns production molecular dynamics run in the \(NPT\) ensemble at a temperature of 300 K and a pressure of 0.1 MPa. To this end, we have used the Berendsen thermostat coupled with the isotropic Parrinello–Rahman barostat. The dynamical timestep used for our simulations was 0.5 fs, with a snapshot geometry output every 5 ps, resulting in 20,000 snapshot geometries. Coulombic and van der Waals interactions were treated as we have described in the preceding paragraph.

3.3 Filtering of similar geometries

For the following discussion, the \(3 \times N\) matrix \(X = [x_1 \ldots x_N]^T\), \(N\) being the number of atoms in the molecule, is used to denote a molecular conformation, where \(x_i\) is the Cartesian position vector of the \(i\)th atom.

Perhaps the most ubiquitous form of geometrical comparison between two structures, \(X\) and \(Y\), is the root mean square deviation (RMSD), which we denote \(R(X, Y)\),

\[
R(X, Y) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |x_i - y_i|^2}
\]  

(1)

However, in the form of eqn (1), a number of pitfalls can be encountered when dealing with molecular conformations. The most significant issue is the fact that \(R(X, Y)\) is not invariant with respect to rigid transformations. For example, if \(X\) and \(Y\) possess the same internal geometry, but are rigidly translated relative to one another, then eqn (1) represents the two geometries as being dissimilar. The same result holds for the case of rigid rotations relative to one another. Of course, molecules in homogeneous environments are entirely equivalent after having undergone rigid translations, and so the RMSD in its current incarnation is of no value for our purposes.
We can render eqn (1) in a useful form by finding \( R(X, Y) \) such that it is minimised with respect to all rigid rotations and translations, i.e. by minimising the function
\[
E(X, Y) = \frac{1}{N} \sum_{i=1}^{N} \|Ux_i - y_i\|^2
\]
where \( U \) is an orthonormal matrix, i.e. \( U^T U = I \), the identity matrix, and can therefore be interpreted as a matrix defining a rigid transformation.\(^{35}\) Taking the square root of \( E(X, Y) \) therefore corresponds to the minimised RMSD (mRMSD) between the two molecular conformations. A further step is required to ensure our definition of the mRMSD is robust: if the orthonormal transformation matrix \( U \), does not constitute a right-handed system (\( |U| < 0 \)), then it represents an improper rotation, i.e. a transformation that can be represented by some combination of a rotation about an axis and subsequent reflection in a plane. Thus, we require computation of the determinant of \( U \) to ensure we deal only with proper rotations.

The minimisation of eqn (2) with respect to \( U \) can be undertaken in a number of ways. Historically, the orthonormality conditions of \( U \) have been added to eqn (2) as Lagrange multipliers and formal solutions obtained, as given by Kabsch,\(^{36,37}\) after whom the algorithmic formalism for solution of eqn (2) is named. We proceed in deriving the form of \( U \) in a more physically intuitive way, elaborated upon in a number of more recent sources.\(^{38,39}\) For the sake of brevity, we simply present the computable expression for \( U \),
\[
WY^T = U
\]
where \( W \) and \( V \) are the right- and left-eigenvectors of the so-called covariance matrix, \( XY^T \), and the components of \( X \) and \( Y \) have been centred on their respective molecular centroids.

Our final consideration centres on whether \( U \) defines a proper or improper rotation, as we have previously alluded to. We can account for this by specifying that if \( |XY^T| < 0 \), then \( d = 1 \), otherwise \( d = -1 \), allowing us to give the final expression for \( U \),
\[
U = W \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{bmatrix} V^T
\]

We propose a workflow for the implementation of the Kabsch RMSD for computing a set of maximally dissimilar molecular conformations, the “ensemble”, from a larger set. For our work, this larger set corresponds to the MD trajectory, and the ensemble is constructed by parsing snapshots of the MD trajectory into conformers of the ensemble. Throughout, we keep the notion of “trajectory” and “snapshot” distinct from “ensemble” and “conformer”. To this end, we propose the use of a greedy MaxMin heuristic, where points are added to the ensemble by iteratively transferring from the trajectory those snapshots that are most dissimilar to all those currently in the ensemble, one at a time.\(^{40}\) So, given some initial large trajectory of molecular snapshots, we compute the Kabsch RMSD between all snapshots and transfer maximally dissimilar snapshots into the ensemble, until an ensemble of the required size has been constructed. This ensemble then corresponds to the set of conformers that are maximally dissimilar from the trajectory.

### 3.4 Ab initio spectrum calculation

All ab initio calculations were undertaken using the GAUSSIAN09\(^{41}\) software package, with the two-layer ONIOM method.\(^{42}\) Histidine was modelled in the high layer, and solvent was modelled in the low layer. The high layer was treated at the B3LYP/6-31G(d) level of theory,\(^{43}\) and the low layer with the AMBER99SB force field including TIP3P parameters for the water molecules. Electronic embedding was used, so that the low layer electrostatics were incorporated into the quantum mechanical Hamiltonian. The ab initio optimisations were performed with the Berny algorithm in the Cartesian basis, and no micro-iterations were used for electronic embedding.

For calculation of the molecular optical activity tensors, we invoke the analytical two-step formalism, or the \( n + 1 \) algorithm,\(^{44}\) in which the harmonic frequency and ROA tensor calculations are separated into differing levels of theory. For calculation of the normal coordinates and harmonic frequencies, we have used the B3LYP/6-31G(d) level of theory, while for the frequency-dependent ROA tensors, we have used the B3LYP/rDPS level of theory.\(^{45}\) rDPS is a rarefied basis set that is based on the 3-21+G basis set with semi-diffuse p-functions on all hydrogen atoms.\(^{46}\) rDPS has been shown to provide a similar level of accuracy in the resulting spectra to much larger Dunning basis sets,\(^{45}\) such as aug-cc-pVDZ.

An excitation wavelength of 532 nm was used for calculation of the ROA tensors, in keeping with the conventional experimental setup. We then obtained scattered circular polarisation backscattered (SCP-180) ROA intensities. Individual conformer spectra were Boltzmann weighted, and those with a Boltzmann weight exceeding 0.5% were used to form the spectrum (using a Lorentzian bandwidth of 10 cm\(^{-1}\)), for comparison with experimental spectra.

### 4. Conformational ensemble construction

Past work\(^{47}\) on computing the Raman and ROA spectra of zwitterionic histidine has implemented a simplistic conformational selection tool. Six major in vacuo conformational preferences for histidine were proposed to arise from the \( \chi_1/\chi_2 \) torsional degrees of freedom: trans-plus (\( \chi_2 = 180^\circ, \chi_1 = 90^\circ \)); trans-minus (\( \chi_2 = 180^\circ, \chi_1 = -90^\circ \)); gauche-plus–plus (\( \chi_2 = 60^\circ, \chi_1 = 90^\circ \)); gauche-plus–minus (\( \chi_2 = 60^\circ, \chi_1 = -90^\circ \)); gauche-minus–plus (\( \chi_2 = -60^\circ, \chi_1 = 90^\circ \)); and gauche-minus–minus (\( \chi_2 = -60^\circ, \chi_1 = -90^\circ \)). These conformers were solvated in an ad hoc manner (a number of water molecules were randomly added around the histidine in each computational model) and energetically minimised. However, the solvation of in vacuo energetic minima does not yield a physically realistic conformational ensemble. Intramolecular hydrogen bonds stabilise in
structures, which are not as preferable when the solute is solvated. Under explicit solvation, interactions with solvent are more favourable than the intramolecular hydrogen bonds. In addition to this, zwitterionic amino acids are not stable in vacuo. This instability necessitates the introduction of non-physical constraints on energetic optimisation to maintain the zwitterionic form.

For each system, we have utilised the Kabsch selection methodology outlined in Section 3.3. For each MD trajectory comprising 20 000 snapshots, we have parsed the 20 snapshots whose solute geometries are the most mutually diverse by use of a greedy MaxMin heuristic. Solvent geometries were not included in the Kabsch filtering. We present a conformer analysis of the conformational ensemble obtained by this Kabsch “filtering” in Fig. 2.

We have highlighted in grey those regions of the \( \chi_1, \chi_2 \) plot in Fig. 2 that correspond to the in vacuo minima employed by Deplazes and coworkers. Concerning the His\(^{1+} \) conformers, the \( (\chi_2 = -60^\circ, \chi_1 = 60^\circ) \) region is not sampled at all. Analysis shows this fact to result from an unfavourable ammonium–imidazole N\(_p\)H contact. Concerning the His\(^0 \) conformers, when \( \chi_2 = 90^\circ \), a favourable ammonium–imidazole N\(_p\) interaction stabilises the His\(^0[\text{N}_p\text{H}] \) microstate. The poor sampling of the \( \chi_1 = -60^\circ \) region for the His\(^0[\text{N}_p\text{H}] \) microstate arises from an inability to form a stabilising carboxylate–imidazole N\(_p\)H interaction. In contrast, the protonated N\(_p\)H of the His\(^0[\text{N}_p\text{H}] \) and His\(^{1+} \) microstates allow for this stabilising interaction.

Importantly, we see the significant levels of sampling in the \( \chi_1 = \pm 180^\circ \) regions for all ensembles. These conformations correspond to an extended imidazole group, where no intramolecular interactions with the zwitterionic groups are formed. In these regions, solvent molecules are able to favourably interact with both the zwitterionic groups and the imidazole, in keeping with the findings of ref. 48. We have therefore demonstrated the inadequacy of using in vacuo minima for the conformational sampling of solvated systems.

For the His\(^0 \) spectra, we have combined both His\(^0[\text{N}_p\text{H}] \) and His\(^0[\text{N}_t\text{H}] \) microstate ensembles into a single ensemble. For our microsolvation studies, we have decided to use water clusters comprising 5, 10, 15 and 20 explicit water molecules. To this end, we have selected the closest [5,10,15,20] water molecules to the centre of mass of the zwitterionic histidine from each conformer.

5. Weighting and optimisation

The optimisation of solvated systems is notoriously difficult owing to the absence of well-defined minima on the PES. If the PES is significantly undulant, conventional minimisation algorithms struggle, as low order spatial derivatives do not suffice in describing the local topology of the potential energy surface. One is frequently reduced to sequentially optimising the system at a number of increasingly complex levels of theory to obtain convergence of the maximum atomic forces and displacements. Whilst tedious, it is also no guarantee of convergence, as we now outline.

To circumvent having to compute (time-consuming) second order spatial derivatives of the energy at every step of the geometry optimisation, updating methods are typically invoked by commercially available quantum chemistry packages.

---

**Fig. 2** Sidechain dihedral values (\( \chi_1 \) and \( \chi_2 \)) taken by conformers within the three ensembles obtained by a Kabsch filtering of the MD trajectories outlined in Section 3.2. His\(^{1+} \) (blue circles), His\(^0[\text{N}_p\text{H}] \) (green diamonds) and His\(^0[\text{N}_t\text{H}] \) (red triangles). Grey hatched regions correspond to those conformers used in ref. 47.
Whilst one can in principle explicitly compute the second order spatial derivatives of the energy at each optimisation step, for the majority of systems this is not a feasible approach. As such, the Hessian is explicitly computed at the start of the geometry optimisation, and subsequently updated as a function of the displacement of the system from this initial geometry by use of first order derivative information. By approximating the Hessian at each step of the geometry optimisation, one can save a great deal of time. Indeed, such Hessian updating schemes have proven to be very accurate for a number of systems.\textsuperscript{52,53}

However, we consider now the case where a complex system begins in a conformation that is far from the energetic minimum. Given that the potential energy surface is highly undulant, the Hessian updating scheme leads to increasingly poor approximations to the Hessian as the optimisation proceeds further from the initial geometry. As such, the geometry optimisation fails to converge. One is then required to adopt some intermediate approach, where the Hessian is explicitly calculated, updated for a number of time steps, followed by explicit recomputation. The frequency with which one performs the explicit recomputation of the Hessian is then a matter for striking a balance between efficiency and accuracy.

An alternative approach to optimising the entire system has been alluded to in Section 2.2. In the work of Zielinski et al.\textsuperscript{31} and Mutter et al.\textsuperscript{36} two levels of optimisation were considered: OptAll and OptSolute. The former optimises the entire system, and the latter optimises only the solute in the field of the unoptimised solvent. Noting that, in these studies, the optimisations were performed on snapshots from MD trajectories, the system is initially far from the minimum energy conformation, rendering the optimisation all the more difficult. Hence, if one is able to “get away with” a simpler optimisation, the savings in computational time will be substantial.

The reasonable agreement of OptSolute with OptAll in the referenced works indicates that one does not necessarily require the entire system to be optimised to obtain informative spectra. Considering the solute–solvent system as a whole, the OptSolute methodology calculates the spectra of unoptimised systems; technically speaking, it is only a subsystem (in this case the solute) that is optimised. This then begs the question of the value of energetic minima for these calculations, and whether strict geometry optimisation is an unnecessary burden.

The OptSolute model is appealing owing to the reduced computational complexity of the resultant optimisation. However, its use raises three pertinent questions: (1) How should the individual conformer spectra be Boltzmann weighted? (2) If the OptSolute scheme is valid, can we simplify the calculations further by adopting an even less rigorous optimisation criterion? and (3) Are alternative subsystem optimisation schemes viable? We deal with, and elaborate upon, each of these questions in turn in the following sections. As our test case, we take microsolvated zwitterionic histidine with five explicit water molecules. The conformers have been taken from the snapshots obtained with the methodology outlined in Section 4.

### 5.1 Boltzmann weighting

If the solute is optimised in the field of the unoptimised solvent, we query whether it is appropriate to use the energy of the entire system for Boltzmann weighting the conformation? We make the (reasonable) assumption that the dominant spectral features arise from the solute, particularly in the higher wavenumber regions, where librational modes do not feature. Consider the effects of optimising the solute while freezing the solvent: the solute will adopt an energetically preferable conformation, while the solvent remains unoptimised. The Boltzmann weight of the solute–solvent system could be relatively low owing to a particularly unfavourable solvent conformation. However, if the solute is an a comparatively low energy conformation, the low Boltzmann weight is not a proper reflection of the contribution of the conformer’s spectrum to the overall spectrum since the solute dictates the major features of the spectrum.

In light of this argument, we propose two means for Boltzmann weighting the spectrum of a given conformer. The first takes the energy of the entire solute–solvent system (“solute–solvent-weighting”), while the second takes the energy of just the optimised solute, without the solvent (“solute-weighting”). The Boltzmann weights of the conformers from the two schemes are compared with one another in Table 1.

From Table 1, we see that for both Boltzmann weighting schemes, the same six conformers dominate the resultant spectrum, albeit in different proportions. The first conformer is dominant with solute–solvent-weighting, while the Boltzmann weight of the first conformer is roughly two times lower with solute-weighting. We can conclude that, in this case (and we have no reason to suspect that the result would not be general), both weighting schemes yield the same dominating conformers, but their contributions to the resulting spectrum differ rather significantly. To evaluate whether the resultant spectra are affected by the differing Boltzmann weights, we present the Raman and ROA spectra generated from the two schemes, in addition to the experimental spectra, in Fig. 3 and 4, respectively.

From the Raman spectra in Fig. 3, there appear to be no pronounced differences between the two weighting schemes. A number of subtle exceptions exist, but certainly none that render one spectrum superior to the other when compared with the experimental spectrum. From the ROA spectra in Fig. 4, we see that the two Boltzmann weighting schemes yield spectra that are again quite similar to the experimental spectrum. This similarity suggests that the Boltzmann weighting scheme does not massively impact on the spectra. However, there are a

<table>
<thead>
<tr>
<th>Conformer number</th>
<th>Solute–solvent weight (%)</th>
<th>Solute weight (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31.6</td>
<td>14.7</td>
</tr>
<tr>
<td>2</td>
<td>22.3</td>
<td>21.6</td>
</tr>
<tr>
<td>11</td>
<td>6.7</td>
<td>16.3</td>
</tr>
<tr>
<td>25</td>
<td>20.1</td>
<td>20.6</td>
</tr>
<tr>
<td>27</td>
<td>18.0</td>
<td>23.6</td>
</tr>
<tr>
<td>35</td>
<td>1.3</td>
<td>3.2</td>
</tr>
</tbody>
</table>
number of slight differences between the two. The 1350–1450 cm\(^{-1}\) region of the experimental spectrum is dominant, with a well-defined +ve/−ve/+ve profile. We see that the first positive region is recovered by both of our calculated spectra, but is arguably better modelled with the solute-weighting.

Indeed, it is interesting to note that this first peak is slightly blue-shifted relative to the experimental spectrum for the solute–solvent-weighting, whereas with the solute-weighting, this peak coincides quite well with the experimental spectrum. The subsequent −ve/+ve profile is present and well-recovered by both, but again the final positive peak is better recovered with the solute-weighting.

From the results we have given, we have concluded that the difference between the two different Boltzmann weighting schemes is small. In the absence of a definitive result, we have chosen to use the solute-weighting scheme owing to the small improvements we have alluded to in the 1350–1450 cm\(^{-1}\) region. We also believe that considering only solute energies makes intuitive sense in the context of the optimisation schemes, as we have discussed.

### 5.2 Level of optimisation

We now assess whether we can circumvent the need to perform stringent geometry optimisations on our system. We have used three levels of optimisation for the comparison: (i) no optimisation (conformers taken straight from the MD snapshots), (ii) the “Loose” (Maximum force ≤2.5 × 10\(^{-3}\) a.u.; Maximum displacement ≤1.0 × 10\(^{-2}\) Å), and (iii) “Regular” (Maximum force ≤4.5 × 10\(^{-3}\) a.u.; Maximum displacement ≤1.8 × 10\(^{-3}\) Å) optimisation schemes available in the GAUSSIAN09 package. The resultant spectra are presented in Fig. 5 and 6. A fourth level of optimisation, “Tight” (Maximum force ≤1.5 × 10\(^{-5}\) a.u.; Maximum displacement 6.0 × 10\(^{-5}\) Å), was also attempted, but the system proved difficult to converge, and so we have not pursued this level.

The Raman spectra in Fig. 5 obtained from the completely unoptimised conformers differ significantly from the other spectra formed from optimised conformers. This poor modelling is indicative of the fact that the MD snapshots are very far from energetic minima, and so are not representative of the conformational ensemble of the system. The ultimate arbiter is the comparison of the calculated spectrum with the experimental spectrum, and we see that agreement is poor between the unoptimised and experimental spectra. There appear to be no features of the experimental spectrum that are reproduced by the unoptimised spectrum.

Turning our attention to the optimised spectra, we see that the Raman spectra are very similar to one another, and reproduce a number of the features present in the experimental spectrum. The series of strong peaks in the 1200–1600 cm\(^{-1}\) region are largely accounted for in our computed spectra. However, we are unable to characterise one level of optimisation as superior to the other from the Raman spectra.

The ROA spectra given in Fig. 6 do, however, differ to some degree. Surprisingly, the spectral features in the 800–1000 cm\(^{-1}\) window appear to be poorly characterised by the regular
optimisation spectrum, but are relatively strong in the loose optimisation spectrum. It could be that the amplitude of the peak at roughly 1100 cm$^{-1}$ dwarfs these features in the regular optimisation. We note that this peak is not particularly strong in the experimental spectrum. Assessing the 1350–1450 cm$^{-1}$ region, both the loose and regular optimisation schemes reproduce the +ve/−ve/+ve signal, with there being no notable difference between their qualities relative to the experimental spectrum. However, the signal appears to be red-shifted by roughly 50 cm$^{-1}$ with the loose optimisation relative to both the regular optimisation and experimental spectra.

Based on these results, we have deemed the loose optimisation to be sufficient in reproducing the spectral features of the experimental spectrum. The regular optimisation offers little improvement in the calculated spectra relative to the loose optimisation. We do not consider the red-shifting of the loose optimisation relative to the experimental spectrum to be detrimental, particularly since the spectral features are still well-recovered. Considering the convergence level is roughly an order of magnitude more stringent for the regular optimisation, it seems to be an unnecessary additional effort to optimise the systems so thoroughly. Indeed, the loose optimisation requires roughly half the amount of time to compute relative to the regular optimisation for this system. Ensuring the system is at least somewhat close to an energetic minimum appears to be sufficient for computing spectra. We reiterate that if the system is far from an energetic minimum, as we understand the unoptimised MD snapshots to be, one obtains extremely poor computed spectra.

5.3 Alternative optimisation schemes

Given that the optimisation of a subsystem leads to informative spectra, the final point we wish to address is whether alternative subsystem optimisation schemes yield equally informative spectra. The OptAll scheme has been found to produce the highest quality spectra, and so this is the case we use as a benchmark. However, we can also formulate several alternative subsystem optimisation schemes:

(a) OptSolute
The solute is optimised in the field of the unoptimised solvent. We have already alluded to this approach and its adoption in previous work, yielding spectra that are comparable to OptAll.

(b) OptSolvent
The converse approach to OptSolute, where the solvent is optimised in the field of the unoptimised solute. This approach is potentially all the more appealing, since the optimisation takes place in the low layer of the QM/MM, and is therefore faster.

(c) OptSolvent → OptSolute
A “two-step” subsystem optimisation scheme – the OptSolvent methodology is invoked, followed by OptSolute. By optimising the layers separately, we hope that we can recreate the optimisation resulting from OptAll.

(d) OptSolute → OptSolvent
The OptSolute methodology is invoked, followed by OptSolvent. We do not imagine this to differ from the previous approach, but is included for the sake of completeness.
(e) OptAll

The entire system is optimised simultaneously.

The first two methods, (a) and (b), are concerned with forsaking accuracy for a large decrease in computational cost. We expect these methods to yield spectra of poorer quality than OptAll, but to be far quicker to compute. Contrastingly, the last two methods, (c) and (d), look to replicate the quality of spectrum obtained by the OptAll scheme, while saving some level of computational time in the process. In the following, we will refer to each optimisation scheme by the letter it has been listed with in the above enumeration. For example, when referring to OptAll, we will use (e).

From the Raman spectra presented in Fig. 7, we immediately see that (b) performs particularly poorly. All three amide I–III regions (1650 cm\(^{-1}\), 1550 cm\(^{-1}\) and 1300 cm\(^{-1}\), respectively) are poorly represented, and the intensity of the low wavenumber regions is not in agreement with the experimental spectrum. However, the other four optimisation schemes yield spectra that are in good agreement with the experimental spectrum. The amide I–III regions are well-recovered, while the low wavenumber regions are modelled equally well. If we are to be fastidious, we may question the amide I regions of (a) and (d), where the signal does not possess the same relative intensity as that in the experimental spectrum. The best agreement with the experimental spectrum is arguably (c), which appears to recover the vast majority of spectral features featured in (e).

Turning our attention to the ROA spectra of Fig. 8, we notice that the poor performance of (b) is continued, and virtually no experimental features are recovered. We also draw attention to the poor performance of (d) in the low wavenumber regions. In turn, this poor modelling of the low wavenumber regions obscures the finer details of the high wavenumber regions, and the spectrum as a whole deteriorates. Similarly, (a) appears to be a poor approximation to the experimental spectrum; the coarse details of the spectrum appear to be present, but the high wavenumber regions are poorly modelled. In contrast, (c) is in excellent agreement with both the spectra obtained through experiment and (e). With the exception of a few features in the low wavenumber regions, (c) and (e) are almost identical.

Offering some explanation for the results we have obtained, the optimisation of the solute appears to be the key factor in guaranteeing the quality of computed spectra; the poor results from (a) and (c) presumably derive from the unoptimised state of the solute when the spectrum is calculated. To recreate the finer details of the spectrum, it also appears as if the solvent requires some level of optimisation, particularly with the ROA measurements. Hence, this is why (c) significantly outperforms (a). The Raman spectra seem to be robust, or mainly invariant, with regards to the level of solvent optimisation.

In conclusion, we feel that (c) and (e) offer the best recovery of experimental spectral features. However, we have found that the computational cost associated with (c) is reduced by more than fourfold on average relative to (e). Relative to the computational cost associated with (a), (c) is roughly 50% more expensive computationally. Therefore, it seems as if the OptSolvent → OptSolute or (c) methodology strikes an ideal balance between accuracy and computational time. We proceed with this methodology through the remainder of this work.

6. Microsolvation

We define microsolvation to be the explicit solvation of a system using a small number of water molecules, typically not extending beyond the second solvation shell of the solute. The major question we wish to answer is whether computed spectra from microsolvated systems are able to model experimental spectra, without having to resort to performing calculations on large explicitly solvated systems. If we can answer this query positively, then it facilitates the computation of solvated Raman and ROA spectra.

6.1 Neutral histidine

For the Raman spectra in Fig. 9, we find that the amide I band at around 1600 cm\(^{-1}\) is well recovered by each of the microsolvated systems. Since the amide I band is largely indicative of peptide C=O stretching modes, it is initially surprising that its modelling appears to be independent of the level of...
microsolvation, where solvent interactions presumably damp the stretching mode. However, upon closer inspection of the microsolvated conformers used for the spectra, we find that the peptide C=O is solvated in each conformer, rendering its modelling independent of the degree of microsolvation. Similarly, we recover the amide II triplet at \( \sim 1450-1550 \text{ cm}^{-1} \), albeit blue-shifted relative to the experimental spectrum. The triplet signature appears to be most profound with a solvation shell comprising 15 water molecules, where the central peak is dominant. This result is to be expected since a solvation shell of 15 water molecules appears to coincide with the number of water molecules comprising the first solvation shell for zwitter-ionic histidine. Since the amide II region is representative of peptide N–H bending and C–H stretching modes, we expect this region to be highly dependent on solvent interactions. It is, nevertheless, surprising that the quality of our computed spectra deteriorates with 20 water molecules.

The amide III region is slightly less well-modelled by our microsolvated systems. In the experimental spectrum, we observe a well-defined quadruplet of peaks, spanning 1200–1350 cm\(^{-1}\). This region is dominated by C–N stretching and N–H bending modes, and we would expect this region to be as solvation-dependent as the other two regions we have discussed.

We note that we do not manage to definitively recover the clear quadruplet seen in the experimental spectrum, although the profile does seem to become more enhanced as we increase the degree of microsolvation. Indeed, when 20 water molecules are included, we are able to discern four clear peaks in the amide III region.

Regarding the impact of the tautomeric forms of histidine, Ashikawa and Itoh\(^5\) have found that the breathing motions for His\(^0[N\_H] \) and His\(^0[N\_H] \) can be related to Raman peaks at 1304/1260 cm\(^{-1}\), respectively. In the same work, two additional marker regions characterising imidazole stretching modes between the tautomeric forms of histidine were proposed: 1568/1585 cm\(^{-1}\) and 1090/1105 cm\(^{-1}\). Later work by Toyama et al.\(^6\) has suggested an additional marker region at 1320/1354 cm\(^{-1}\). Unfortunately, the majority of these marker regions inhabit the strong amide regions of the Raman spectrum, making them difficult to discern. However, the 1090/1105 cm\(^{-1}\) region occupies a low intensity region of the Raman spectrum, and appears to be characterised in both our experimental and computed spectra, at each microsolvation level excluding the spectrum with 5 waters.

The experimental ROA spectrum is not quite as well-recovered as the microsolvated spectra, as depicted in Fig. 10.
This is not to say that the ROA spectra are of poor quality, but simply suffer by comparison with the high quality of the Raman spectra. The dominant feature across both experimental and calculated spectra is the +ve/-ve/+ve signal at 1300–1500 cm$^{-1}$, and is well-recovered at all levels of microsolvation. However, it is worth pointing out that the calculated spectra appear to be blue-shifted by ~50 cm$^{-1}$ relative to the experimental spectrum. This blue-shifting is somewhat surprising. Note that we have not undertaken abscissa-scaling by 0.96, as prescribed by Radom.$^{46}$ Doing so would blue-shift the spectrum further, and so would result in the deterioration of our calculated spectra.

We note that as the level of microsolvation increases, the peak at 1050 cm$^{-1}$ is amplified. A number of peaks exist in this region of the experimental spectrum, but none are as prominent as that in the microsolvated spectrum. Assessing the normal modes of motion around this region, we find that the majority are dominated by sidechain dihedral torsional degrees of freedom. This observation then suggests that the sidechain torsional degrees of freedom are more flexible in the microsolvated systems than in the experiment. Indeed, it is predominantly the zwitterionic groups and imidazole ring that form interactions with the solvent molecules in the microsolvated systems, and so it is to be expected that the alkyl sidechain be poorly solvated, and so not well recovered by the microsolvated spectra.

### 6.2 Cationic histidine

The literature regarding the calculation of Raman and ROA spectra of formally charged species is sparse. Indeed, the *ab initio* modelling of formally charged species is in itself difficult, since the basis sets require both diffuse functions and some description of polarisation, the former being of particular importance for anions. To assess the quality of Raman and ROA spectra of a formally charged species, we have selected cationic histidine. The fact that no tautomers exist for this species makes its modelling simpler.

A formally charged species is presumed to undergo stronger interactions with solvent than the neutral species, and so we hypothesise that the degree of microsolvation will significantly influence the quality of the computed spectrum. Since the formal charge results from the protonation of the imidazole ring, we anticipate that the higher wavenumber regions will be poorly reproduced at the low levels of microsolvation since the explicit water molecules tend to primarily interact with the zwitterionic groups.

The computed Raman spectra of Fig. 11 are in relatively good agreement with the experimental Raman spectrum. The strong peaks at ~1200 cm$^{-1}$ and ~1300 cm$^{-1}$ become increasingly prominent as the degree of microsolvation increases. The peak at ~1500 cm$^{-1}$ is well-recovered by each of the microsolvated spectra. The resolution of the doublet of peaks at roughly 800 cm$^{-1}$ is similarly improved as the degree of microsolvation increases. Interestingly, the general agreement with the experimental spectrum deteriorates when 10 explicit water molecules are used for microsolvation. However, those spectra including 5, 15 and 20 explicit water molecules are very similar.

We offer a potential reason for the deterioration in spectrum quality when 10 explicit water molecules are included. When 5 explicit water molecules are included, the conformers are largely solvated around the backbone amide and carboxyl groups of the zwitterionic histidine. When 15 and 20 explicit water molecules are included, the conformers are completely solvated, *i.e.* both the backbone and sidechain groups of the zwitterionic histidine. However, when 10 explicit water molecules are included, only a couple of water molecules solvate the imidazole ring of the zwitterionic histidine. As such, we suggest that the partial solvation of the imidazole results in a biasing of certain vibrational frequencies originating from the imidazole–solvent interactions. The spectrum as a whole deteriorates since these imidazole–solvent frequencies then dominate the resultant spectrum. In summary, it is tempting to think of partial solvation as detrimental to the correct prediction of the vibrational spectra.

The computed ROA spectra in Fig. 12 are, however, not quite as successful in reproducing the major features of the experimental spectrum. The strong negative band at ~1400 cm$^{-1}$ is recovered in each of the computed spectra, and the positive doublet in the 1300–1350 cm$^{-1}$ region appears to become clearer as the level of microsolvation increases.
From previous, unpublished work on anionic adenosine triphosphate, we have found that the calculation of ROA spectra for formally charged species is rather difficult. One speculative reason for the poorer quality of the ROA spectra relative to the Raman spectra is the way one computes the intensity of spectral bands for the two spectroscopies. For Raman spectra, the intensity is given by the sum of the intensities of the right- and left-circularly polarised scattered light. For ROA spectra, the intensity is given by the difference in the intensities of the right- and left-circularly polarised scattered light. Therefore, ROA band intensities are more prone to error than the Raman band intensities. In other words, the ROA band intensities are smaller in magnitude than the Raman band intensities, and so the errors are amplified. It is this sensitivity to error that makes ROA a “gold-standard” spectroscopic technique, whereas Raman is somewhat more forgiving.

7. Conclusion

We have investigated and presented the results for a number of novel methodologies that can be used in the calculation of \textit{ab initio} Raman and ROA spectra of solvated systems. We are suggestively vague in our use of the term “solvated systems”, since we see no reason why our methodology cannot be applied to systems outside of the domain of peptides, such as DNA bases.\textsuperscript{57,58} A novel conformational sampling methodology allows for the unambiguous extraction of a set of mutually diverse conformers from an MD trajectory of arbitrary length. We have shown that conformers do not require strict optimisation to obtain high quality spectra, and that a two-step subsystem optimisation (i.e. optimising the solvent while keeping the solute frozen, and subsequently optimising the solute while keeping the solvent frozen) yields spectra of the same quality as entire system optimisations. The microsolvation of both formally charged and neutral zwitterionic histidine species can yield calculated spectra that converge towards the experimental spectra.

Acknowledgements

S. C. and M. G. L. thank the BBSRC for awarding DTP studentships and P. L. A. P. the EPSRC for the award of an Established...
Career Fellowship (EP/K005472). All authors are grateful to Professor P. Bou for access to his ROA spectrometer for performing the experimental measurements.

References

54. I. Ashikawa and K. Itoh, Biopolymers, 1979, 18, 1859.