Validation of the Quantum Chemical Topological Force Field, QCTFF

A thesis submitted to the University of Manchester for the Degree of Doctor of Philosophy in the Faculty of Engineering and Physical Sciences

2015

Timothy James Hughes

School of Chemistry

List of Contents

List of Tables

List of Figures

Abstract

University of Manchester
Timothy James Hughes
Doctor of Philosophy
2015

Validation of the Quantum Chemical Topological Force Field, QCTFF

Until such a time that computers are powerful enough to routinely perform *ab initio* simulation of large biomolecules, there will remain a demand for less expensive computational tools. Classical force field methods are widely used for the simulation of large molecules. However, their low computational cost comes at the price of introducing approximations to the description of the system, for example atomic point charges and Hooke type potentials. The quantum chemical topological force field, QCTFF, removes the classical approximations and uses a machine learning method, kriging, to build models that map *ab initio* atomic properties to changes in the internal coordinates of a chemical system. The atomic properties come from quantum chemical topology, QCT, and include atomic multipole moments and also energy terms from the interacting quantum atoms (IQA) energy decomposition scheme. By using atomic multipole moments, the electrostatic interactions between atoms is described in a more rigorous fashion than most classical force fields, and polarisation is captured through the use of kriging models. In this thesis, the QCTFF approach has been applied to a selection of test cases including small molecular dimers and amino acids. Kriging models are built using a "training set" of molecular geometries, and an investigation of different approaches for sampling amino acids is provided. The concept of the "atomic horizon sphere" is discussed, where the effect on the multipole moments of an atom in an increasingly large environment is investigated. This is an important investigation required to guide the development of future QCTFF training sets. Investigations into the effect of deprotonation of basic and acidic amino acids side chains is provided, as well as a study of the short range repulsion between atoms.

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.

Copyright Statement

Acknowledgements

I would like to thank Prof. Popelier not only for his supervision and guidance throughout my PhD, but also for teaching me the value of a middle name. I would also like to thank AstraZeneca for their funding.

I would like to extend my thanks to my fiancée Kim for her encouragement and (near endless) patience. I would also like to acknowledge the support I have received from my family throughout my time in Manchester: My parents Chris and Ruth, my brother Sam and my sisters, Charlotte and Bernie. I also am grateful to the Intellectual Men for their continuous friendship (*Huzzah!*). Big thanks to the members of the Popelier group past and present who have made my time in the office so enjoyable: Sav, Tim, Pete, Kate, Nicco, James (or is it Luke?), Stuart, Francois, Joe, Mark and Caesar.

Chapter 1

Introduction and Methods

1.1. Background

In silico techniques are widely used throughout chemistry[2]. But why study a chemical system using computation? The computational modelling of experimentally observable chemical systems offers unique insight and analysis that, along with experimental results, provides a more complete understanding of the chemical system than experiment alone is capable of. The better understanding of the system can then act as a guide in the design of future experiments. For example, the development of photovoltaic devices frequently utilises computation to identify the HOMO and LUMO orbitals of the dye molecule and the semiconductor using density functional theory (DFT)[3, 4]. The calculations aid the design of dye molecules that will give efficient photovoltaic cells. Another example of computational methods being used in parallel with experiment is the elucidation of reaction pathways [5-7]. High level *ab initio* data are capable of giving energies not just of reactants and products, but also energies of competing transition states which are unobtainable directly from experiment.

Computational methods are often a faster and cheaper alternative to experiment, and computation is widely used in industry to streamline research and development processes. For example, a pharmaceutical company wanting to design a novel inhibitor for a target protein may save a significant sum of money by screening many possibly drug candidates using computational methods such as docking[8], molecular dynamics (MD) simulation[9], or Monte-Carlo (MC) simulation[10]. These techniques are capable of narrowing down the drug candidates in a fraction of both the time and the cost that would be taken by experimental means. A detailed review of the applications of docking to medicinal chemistry problems is provided in **Appendix A**.

A practical limit on the accuracy of data available to a computational chemist is that on the size of the system being studied. *Ab initio* methods such as DFT, Møller-Plesset 2 (MP2) and coupled cluster techniques (e.g. CCSD(T)) scale rapidly with system size ($O(N^5)$, $O(N^6)$ and $O(N^7)$ for MP2, CCSD and CCSD(T) respectively[11]) but are highly accurate and involve little approximation. Molecular mechanics (MM) force fields offer an alternative when the system of interest is too large for *ab initio* calculation. MM force fields treat chemical systems as a sum of classical potentials and are much faster than *ab initio* methods. The details of MM force fields are described in detail in **Section 1.2**. Until such a time as computers are powerful enough to routinely provide *ab initio* results for large systems such as proteins, the search for the best MM potential is an active area of research.

The aim of this body of work is to present my personal contribution towards the development and improvement of the Quantum Chemical Topological Force Field (QCTFF), a potential developed by the Popelier group for application to biomolecular and medicinal applications. Despite naming it a force field, QCTFF is both fundamentally and theoretically removed from the classical approaches to force field methodology. Instead of using parameterised atom types and classical potentials, QCTFF uses machine learning and *ab initio* quantum chemical topological (QCT) information to build models that provide near *ab initio* accuracy on a MM time scale.

1.2. Force Field Methodology

A force field is a set of equations that may be solved in order to reproduce the total energy of a system. Although difference force fields all differ from one another, they share many features and potential forms. Examples of common force fields are AMBER[12], CHARMM[13, 14], GROMOS[15] and AMOEBA[16]. Because QCTFF takes a radically different approach to force field development, the following description of standard MM potentials is not relatable to QCTFF. However, to appreciate the radical approach that QCTFF takes to biomolecular force field development, it is important to provide the reader with an overview of the standard approach.

An MM force field calculates the total internal energy, U_{tot}, of a system as a sum of bonded and non-bonded terms,

$$U_{tot} = \sum_{atoms} (U_{bonded} + U_{non-bonded}) \tag{1.1}$$

where U_{bonded} and $U_{non-bonded}$ can be further decomposed to give,

$$U_{bonded} = \sum_{bonds(AB)} U_{AB}^{bond\ stretch} + \sum_{angles(ABC)} U_{ABC}^{angle\ bend} + \sum_{coupling(AB,ABC)} U_{AB,ABC}^{cross\ terms} + \sum_{torsions(ABCD)} U_{ABCD}^{torsional} \tag{1.2}$$

$$U_{non-bonded} = \sum_{atom\ pairs(AB)} U_{AB}^{electrostatic} + \sum_{atom\ pairs(AB)} U_{AB}^{Lennard-Jones} \tag{1.3}$$

Bonded terms include bond stretches, angle bends, stretch-bend cross terms, and torsional potentials. The non-bonded terms consists of the electrostatic interaction, the van der Waals dispersion interaction and the Pauli repulsion. The latter two terms are normally described together as a Lennard Jones type potential.

The bond stretch term is obtained by Taylor expansion about the equilibrium separation of two bonded atoms, r_{eq}, between two bonded atoms A and B,

$$U(r)_{AB} = U(r_{eq}) + \frac{dU}{dr}(r_{AB} - r_{eq,AB}) + \frac{1}{2!}\frac{d^2U}{dr^2}(r_{AB} - r_{eq,AB})^2 + \frac{1}{3!}\frac{d^3U}{dr^3}(r_{AB} - r_{eq,AB})^3 + \cdots \tag{1.4}$$

When equation 1.4 is truncated at the first non-zero term, one obtains *Hooke's Law*,

$$U(r_{AB}) = \frac{1}{2}k_{AB}(r_{AB} - r_{eq,AB})^2 \tag{1.5}$$

Equation 1.5 contains two constants, the force constant, k_{AB}, and the equilibrium bond distance, $r_{eq,AB}$. These values are easily obtained by experimental techniques, typically IR spectroscopy.

The angle bend term of a classical MM force field is obtained in an analogous manner to that of the bond stretch term, described by Taylor expansion about the equilibrium bond angle. When truncated after the first non-zero term, one obtains:

$$U(\theta_{ABC}) = \frac{1}{2}k_{ABC}(\theta_{ABC} - \theta_{eq,ABC})^2 \tag{1.6}$$

where the force constant, k_{ABC}, and equilibrium angle, $\theta_{eq,ABC}$, are obtained by experiment.

The above expressions for bond stretches (equation 1.5) and angle bends (equation 1.6) provide good description of stretches and bends so long as the atoms remain near their equilibrium positions. Equation 1.5 fails to model bond stretches when r_{AB} is much larger than $r_{eq,AB}$. In such a situation, the energy of a bond becomes infinitely positive. This is not reasonable, and may be "balanced" by including higher terms in the Taylor expansion to give a potential of the form,

$$U(r_{AB}) = \frac{1}{2}\left[k_{AB} + k_{AB}^{(3)}(r_{AB} - r_{eq,AB}) + k_{AB}^{(4)}(r_{AB} - r_{eq,AB})^2\right](r_{AB} - r_{eq,AB})^2 \tag{1.7}$$

As the third order force constant $k_{AB}^{(3)}$ is negative, its inclusion ensures infinitely positive energies are no longer a concern, but now the energy tends to infinitely negative values. Hence the inclusion of the quartic term in equation 1.7. It is uncommon for higher order terms to be included in bond stretch potentials, but higher order terms are common for angle bend potentials; in fact the MM3 force field includes terms up to and including the 6th order force constant $k_{ABC}^{(6)}$ [17].

The stretching motion of bonds and the bending of angles are not isolated events- such molecular vibrations are coupled with one another. Consider a single molecule of water. As the angle θ_{HOH} falls below the value $\theta_{eq,HOH}$ the increased repulsion between the two δ^+ H atoms will be offset by an increase in the two O-H bond lengths. Such a coupling between two vibrational coordinates can be modelled in a MM force field though means of cross terms. A simple cross term coupling a bond stretch with an angle bend is presented in equation 1.2.8.

$$U(r_{AB}, \theta_{ABC}) = \frac{1}{2} k_{AB,ABC}\left(r_{AB} - r_{eq,AB}\right)\left(\theta_{ABC} - \theta_{eq,ABC}\right) \tag{1.8}$$

In addition to bond stretches, angle bends and their associated cross terms, the bonded potential includes potentials describing rotation about torsion angles, ω_{ABCD}. Torsional terms are periodic in nature and thus the functional form of such terms differ to those describing the bonded terms discussed previously. Rather than Taylor expansion around an equilibrium point, a Fourier series is instead used, of general form:

$$U(\omega_{ABCD}) = \frac{1}{2} \sum_{(j),ABCD} V_{j,ABCD}\left[1 + (-1)^{j+1} \cos\left(j\omega_{ABCD} + \psi_{j,ABCD}\right)\right] \tag{1.9}$$

where (j) is a set of periodicities, $V_{j,ABCD}$ is the amplitude, and $\psi_{j,ABCD}$ is the phase angle.

Cross terms involving torsional potentials also feature in most MM force fields. Similar to the above discussion on stretch-bend coupling, one can imagine that in the eclipsed conformation of ethane the C-H bonds may extend from $r_{eq,CH}$ to lessen steric repulsion between H atoms. Such terms often take a form similar to:

$$U(r_{AB}, \omega_{ABCD}) = \frac{1}{2} k_{AB,ABCD}\left(r_{AB} - r_{eq,AB}\right)[1 + \cos(j\omega + \psi)] \tag{1.10}$$

Moving now to the non-bonded terms within a classical force field, the discussion shall turn to the van der Waals dispersion term. This is an attractive interaction between all atoms. Such an interaction arises due to the correlated movements of electrons giving rise to instantaneous atomic (and molecular) moments that are orientated so as to be attractive. The dominant term of this interaction is that of the induced dipole- induced dipole interaction. This has a $-\frac{1}{r^6}$ dependence so therefore as two atoms approach one another the energy gets increasingly negative, tending to infinitely negative at very small separations.

Clearly this is not what happens in the real world. As electron density of two atoms begins to overlap at short separations, Pauli repulsion between the electrons increases rapidly.

This is typically described by a $\frac{1}{r^{12}}$ dependence. This gives rise to the famous *Lennard-Jones* potential,

$$U(r_{AB}) = 4\varepsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}}\right)^{6}\right] \tag{1.11}$$

where ε_{AB} is the depth of the potential energy well and σ_{AB} is the internuclear separation at which $U(r_{AB})$ becomes positive. While the $-\frac{1}{r^6}$ dependence of the attractive term is theoretically grounded (as stated above), the $\frac{1}{r^{12}}$ dependence of the repulsive term has no such foundation. It is simply chosen because it provides a 'good' description of the energy. Other potential forms are used by different force fields, for example the *AMOEBA* force field of Ponder *et al.*[16] use a 'buffered 14-7 potential'. This takes the form,

$$U(r_{AB}) = \varepsilon_{AB}\left(\frac{1.07}{\rho_{AB} + 0.07}\right)^{7}\left(\frac{1.12}{\rho_{AB}^{7} + 0.12} - 2\right) \tag{1.12}$$

where $\rho_{AB} = r_{AB}/r_{AB}^{0}$ and r_{AB}^{0} is the minimum energy distance between nuclei A and B. Such a potential is chosen to provide a steeper repulsive region at short range, and because it provides a better fit to both *ab initio* gas phase calculations and the properties of liquid noble gasses[18]. Parameters such as the well depth and minimum energy separations vary across force fields, although they are typically obtained by both *ab initio* calculation and experimental results.

The electrostatic component of a classical force field is typically described by Coulomb's law,

$$U(r_{AB}) = \frac{q_A q_B}{\varepsilon r_{AB}} \tag{1.13}$$

where q_A and q_B are the partial charges of atoms A and B respectively. There are a number of ways in which these partial charges are assigned, such as restrained electrostatic potential (RESP), Mulliken population analysis and atoms in molecules (AIM) charges. As previously stated, the use of atomic point charges provides poor description of the electrostatic contribution to the total energy of a system. In particular, point charges fail to correctly model interactions where the electron density is anisotropic, for example lone pairs, delocalised π systems and σ-holes.

One of the core features of QCTFF is its treatment of the electrostatics. Multipole moments up to and including the hexadecapole moments are present on all atoms. How the

multipole moments are obtained is outlined in **Section 1.3**, and a thorough comparison of atomic multipole moments versus atomic point charges is included in **Section 1.6**.

1.3 Quantum Chemical Topology, QCT

Underpinning the development of QCTFF[19] is Quantum Chemical Topology(QCT)[20], which embraces all work[21] in quantum chemistry that uses the topological language of dynamical systems (e.g. attractor, basin, gradient path, critical points). QCT contains the "quantum theory of atoms in molecules" (QTAIM)[22-24] as a special case where this topological language is applied to the electron density ρ and its Laplacian. The theory of interacting quantum atoms, IQA, of Pendas *et al.* [25] is another example of a theory that falls under the umbrella of QCT.

A topological atom Ω_A is a bundle of gradient paths (i.e. trajectories of steepest ascent through ρ), 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 non-overlapping and sharply bounded (at the inside of the molecule) by so-called interatomic surfaces. A gradient path is a trajectory in 3D space, which can be seen as consisting of infinitesimal vectors orthogonal to envelopes of constant electron density ρ.

$$\nabla_{\rho(r)} = \frac{\partial \rho}{\partial x}\boldsymbol{i} + \frac{\partial \rho}{\partial y}\boldsymbol{j} + \frac{\partial \rho}{\partial z}\boldsymbol{k} \tag{1.14}$$

where $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ are unit vectors that maintain the directionality of the x, y, z axis, and $\frac{\partial \rho}{\partial x}$ is the change in electron density with respect to movement along the x-axis. The gradient paths follow the direction of increasing ρ, until terminating at a critical point. The latter is an attractor, which can only be a nucleus (which is mostly the case), a bond critical point, a ring critical point or a cage critical point. **Figure 1.1** shows an example of a QCT partition of the furan molecule. The nature of a critical point is determined by analysis of the eigenvalues of the Hessian at the critical point. The Hessian of the electron density is given by

$$Hessian = \begin{bmatrix} \frac{\partial^2 \rho}{\partial x^2} & \frac{\partial^2 \rho}{\partial x \partial y} & \frac{\partial^2 \rho}{\partial x \partial z} \\ \frac{\partial^2 \rho}{\partial y \partial x} & \frac{\partial^2 \rho}{\partial y^2} & \frac{\partial^2 \rho}{\partial y \partial z} \\ \frac{\partial^2 \rho}{\partial z \partial x} & \frac{\partial^2 \rho}{\partial z \partial y} & \frac{\partial^2 \rho}{\partial z^2} \end{bmatrix} \tag{1.15}$$

of which the diagonal $D_{Hessian}$ is given by

$$D_{Hessian} = \begin{bmatrix} \frac{\partial^2 \rho}{\partial x^2} & 0 & 0 \\ 0 & \frac{\partial^2 \rho}{\partial y^2} & 0 \\ 0 & 0 & \frac{\partial^2 \rho}{\partial z^2} \end{bmatrix} = \begin{bmatrix} \lambda_x & 0 & 0 \\ 0 & \lambda_y & 0 \\ 0 & 0 & \lambda_z \end{bmatrix} = (\lambda_x, \lambda_y, \lambda_z)$$

(1.16)

where the $(\lambda_x, \lambda_y, \lambda_z)$ describes the local curvature of the electron density at any point in space. $(\lambda_x, \lambda_y, \lambda_z)$ are the eigenvalues of $D_{Hessian}$ and can be used to state whether a point in space is a maximum, minimum or saddle point in electron density.

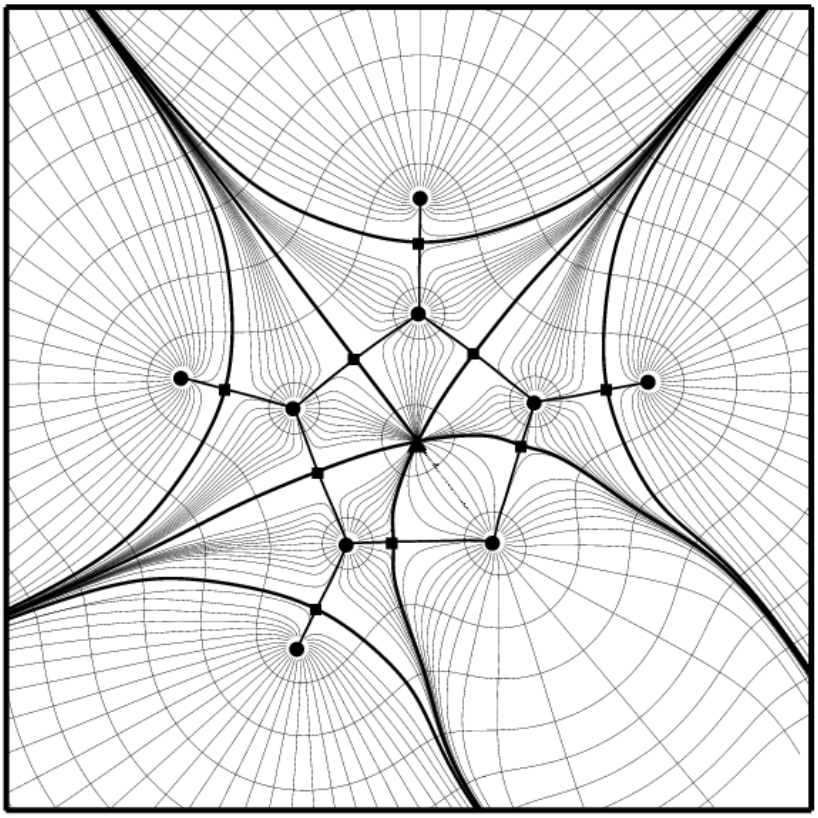


Figure 1.1: The gradient vector field of furan. Topological features such as gradient paths, bond critical points, nuclear attractors and zero-flux surfaces can all be observed.

The four types of critical points are described by two values. The first is the "rank" which is determined by the number of non-zero eigenvalues at the critical point. The second number is the sum of the signs of the eigenvalues for the diagonalised Hessian at the critical point of interest. A nucleus is therefore a (3,-3) critical point as it is a maximum in electron density in the x, y and z directions. A bond critical point is (3,-1) critical point as it is a maximum in two directions and a minimum in one (the axis on which the bond lies). A ring critical point is a (3,1) critical point as it is a maximum in only one direction (orthogonal to the ring) and a minimum in two, and the finally a cage critical point, which is surrounded in three dimensions by atoms, is a (3,3) critical point. Analysis of topological

properties at critical points provides insight into the nature of the chemical systems being studied. Study of QTAIM metrics at critical points such as the electron density, ρ, the Laplacian, $\nabla^2\rho$, and the energy density (the sum of potential and kinetic energy), H, has been successfully used to describe chemical systems such as heavy metal complexes[26, 27], transition metal complexes [28], aromatic molecules [29] and non-covalent interactions[30, 31].

The electrostatic interaction between atoms is the topic of much of the work in this thesis, and QCT allows for a rigorous treatment of this interaction. Partitioning by QCT gives rise to well defined, non-overlapping atoms[23] for which atomic multipole moments may then be obtained. Atomic multipole moments provide an anisotropic description of the electron density around the nucleus of a topological atom, and were one of the key original motivators for the development of QCTFF. The superiority of an anisotropic description of an atomic electron density is discussed in great detail in **section 1.6**. The Coulomb interaction[32] energy between two topological atomic basins Ω_A and Ω_B is given by:

$$E_{AB}^{Coul} = \int_{\Omega_A} d\mathbf{r_1} \int_{\Omega_B} d\mathbf{r_2} \frac{\rho_{tot}(\mathbf{r}_1)\rho_{tot}(\mathbf{r}_2)}{r_{12}}$$

(1.17)

where ρ_{tot} is equal to the sum of the electron density ρ and the nuclear charge density. The expression $1/r_{12}$ in equation 1.17 can be replaced by series expansion involving the spherical harmonics[33, 34] to give:

$$\frac{1}{r_{12}} = \sum_{l_A=0}^{\infty} \sum_{l_b=0}^{\infty} \sum_{m_A=-l_A}^{l_A} \sum_{m_B=-l_B}^{l_B} T_{l_A l_B m_A m_B} R_{l_A m_A}(\mathbf{r_1}) R_{l_B m_B}(\mathbf{r_2})$$

(1.18)

where $R_{lm}(\mathbf{r})$ is a regular spherical harmonic. The interaction tensor T depends upon both the mutual orientation of the two interacting atoms A and B, and their internuclear distance. The simplest interaction term is that two monopole moments (or essentially atomic charges), where T is simply 1/r. Substituting equation1.18 into equation 1.17 gives:

$$E_{AB}^{Coul} = \sum_{l_A l_B m_A m_B} Q_{l_A m_A} T_{l_A l_B m_A m_B} Q_{l_B m_B}$$

(1.19)

where Q_{lm} represents a multipole moment:

$$Q_{lm} = \int_{\Omega} d\mathbf{r}\rho_{tot}(\mathbf{r})R_{lm}(\mathbf{r})$$

(1.20)

that is obtained after a 3D integration over the complicated volume of the topological atom. It is convenient to define an interaction rank L between two multipole moments of order l_A and l_B by:

$$L = l_A + l_B + 1$$

(1.21)

Previous work [35, 36] has shown that an interaction rank of $L = 5$ provides satisfactory description of structural and dynamic characteristics of a system. The value of L is identical to the inverse power in the $1/R^L$ behaviour of an interatomic electrostatic interaction. For example, dipole…dipole interactions behave by the well-known $1/R^3$ law given that $L = 1 + 1 + 1 = 3$. Truncating at $L = 5$ requires monopole, dipole, quadrupole, octopole and hexadecupole moments for each atom. Therefore, in this work all topological atoms are described by $1 + 3 + 5 + 7 + 9 = 25$ multipole moments each.

1.4. The Theory of Interacting Quantum Atoms, IQA

The theory of interacting quantum atoms (IQA) of Pendas et al[25] has recently been incorporated into QCTFF development, as a means of replacing the classical force field terms outlined in **Section 1.2**. IQA is a topological energy decomposition scheme that has successfully been applied to a wide range of chemical systems[37-39]. IQA fits under the umbrella of quantum chemical topology. The total energy of a system, E_{IQA} is given as a sum of atomic self-energies V_{self} and of interaction energies V_{inter},

$$E_{IQA} = \sum_{A}^{n} V_{self}^{A} + \sum_{A}^{n}\sum_{B<A}^{n-1} V_{inter}^{AB}$$

(1.22)

where n is the total number of atomic basins in the system. V_{self} is further decomposed into electron-electron and electron-nucleus interactions, V_{ee}^{AA} and V_{ne}^{AA} respectively, and the electronic kinetic energy of atom A, T^A,

$$V_{self,A} = V_{ne}^{AA} + V_{ee}^{AA} + T^A$$

(1.23)

where V_{ee}^{AA} has a classical, coulombic component V_{cl}^{AA} and an exchange-correlation component V_{xc}^{AA}.

V_{inter}^{AB} is also a sum of contributions,

$$V_{inter}^{AB} = V_{nn}^{AB} + V_{ee}^{AB} + V_{ne}^{AB} + V_{en}^{AB} + V_{xc}^{AB}$$

(1.24)

where V_{nn}^{AB} is the interaction between the two atomic nuclei, V_{ee}^{AB} is the interaction between the electrons of atom A and the electrons of atom B, V_{ne}^{AB} is the interaction between the nucleus of atom A interacting with the electrons of atom B, V_{en}^{AB} is the interaction between the electrons of A with the nucleus of B and V_{xc}^{AB} is the exchange-correlation energy between atoms A and B. The first four components on the right hand side of **Equation 1.24** can be combined and written as the classical interaction energy V_{cl}^{AB} giving

$$E_{IQA} = \sum_{A}^{n} V_{self}^{A} + \sum_{A<B}^{n} (V_{cl}^{AB} + V_{xc}^{AB})$$

(1.25)

The atomic self-energy by itself is a difficult value to interpret. The difference in the atomic self-energy relative to a reference system lends itself to interpretation more easily by observation of changes in the self-energy, ΔV_{self}^{A}. To illustrate this point, an IQA study by Eskandari and Van Alsenoy on the rotational barrier of biphenyl [37] showed that the "steric clash" may be described as a consequence of the ortho-H atoms experiencing an increase in their self-energies when going from a staggered to an eclipsed conformation. This is despite the E_{inter}^{HH} between "clashing" H atoms being most attractive in the eclipsed conformation.

Due to the difficulty in interpreting an atomic self-energy the deformation energy of an atom, V_{def}^{A}, can be calculated. V_{def}^{A} is defined as the difference in the self-energy between the free, unbound atom (or fragment), and the self-energy of the atom in a chemical system,

$$V_{def}^{A} = V_{self,bound}^{A} - V_{self,unbound}^{A}$$

(1.26)

V_{def}^{A} is always positive because the unbound atom is always lower in energy than the bound atom. The deformation energy may be extended to a group of atoms. For example, the deformation energy of a water molecule upon formation of the water dimer would be given by

$$V_{def}^{waterA} = \sum_{A}^{Atoms\ in\ Water\ X} V_{self,in\ dimer}^{A} - V_{self,free\ water}^{A}$$

(1.27)

The short-range repulsion between atoms is the subject of **Chapter 6**, and the IQA interpretation of this interaction is given as the sum of V_{def} for all atoms or fragments and all V_{xc} interactions between the two atoms or fragments y and z, and is named the XRC energy.

$$XRC^{nm} = \sum_{A,B}^{y,z} V_{def}^{A,B} + \sum_{A}^{z}\sum_{B}^{y} V_{xc}^{AB} \tag{1.28}$$

where A is an atom in fragment z and B is an atom in fragment y. Note that at the HF level of theory there is no correlation and the XRC energy is simply the XR energy.

1.5. The Atomic Local Frame and Kriging

QCTFF is designed to be used as a force field for molecular dynamics simulations, and therefore it should be obvious that QCT and IQA properties cannot be calculated at each time step as this would be highly computationally expensive. Classical force fields use parameterised atom types, described by a set of constants that are fed into simple potentials to obtain the energy of the system. QCTFF offers a radically different approach using the machine learning method kriging[40-42], also known as Gaussian process regression[43], as a method of capturing the changes in topological atomic properties as a function of molecular geometry. Therefore, instead of a given atom type being described by a list of parameters, in QCTFF it is described by a collection of kriging models, each describing the changes in a single topological property (such as an atomic multipole moment or the atomic self-energy) with respect the system's coordinates. This approach naturally includes polarisation and charge transfer effects.

In the work presented in this thesis, kriging models have been built for both atomic multipole moments and also the IQA self and interaction energies. 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. 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 **Chapter 2** kriging models are built for the first 25 atomic multipole moments of seven hydrogen bonded dimer complexes and also the IQA atomic self and interaction energies of three weakly bound complexes. In **Chapter 3** 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).

In order to build a kriging model, one must define a coordinate system. 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 in order 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 θ_{XAY} define the three ALF coordinates. Subsequently a right-handed coordinate system is stablished using the *XAY* plane. All other atoms in the system can then be described by three polar coordinates, R_{AK}, ϕ_{AK} and θ_{AK}. One therefore obtains *N*-3 sets of three spherical polar coordinates each, which combined with the aforementioned ALF coordinates make up the 3*N*-6 coordinates required, i.e. 3(*N*-3)+3 = 3*N*-6.

Returning to kriging, the change in a given multipole moment or IQA component is smooth with respect to a change in the ALF coordinates. Therefore it is safe to interpolate the topological properties 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 properties by mapping an input (nuclear coordinates) to an output (a topological property). To achieve this, a training set of molecular geometries with known values of the topological property is required. The sampling of molecular geometries for training kriging models is described in the description of the GAIA protocol later in this chapter. Kriging models calculate topological properties of a new geometry by the following process:

$$\hat{y}(\boldsymbol{x}^*) = \hat{\mu} + \sum_{i=1}^{n} a_i \cdot r_i \tag{1.29}$$

where $\hat{y}(\boldsymbol{x}^*)$ is a topological property at a new set of coordinates $\boldsymbol{x}^*$ and $\hat{\mu}$ is the global (average) value of the property over the whole training set. The factor a_i is the i^{th} element of the vector $\boldsymbol{a} = \boldsymbol{R}^{-1}(\boldsymbol{y} - \mathbf{1}\hat{\mu})$ and r_i is the i^{th} element of $\boldsymbol{r}$, defined by

$$\mathbf{r} = \{cor[\varepsilon(\mathbf{x}^*), \varepsilon(\mathbf{x}^1)], cor[\varepsilon(\mathbf{x}^*), \varepsilon(\mathbf{x}^2)], \ldots, cor[\varepsilon(\mathbf{x}^*), \varepsilon(\mathbf{x}^n)]\}^T \tag{1.30}$$

where *T* marks the transpose.

Kriging treats all topological properties as an error from the global value, and it is the correlation of these errors for a given multipole moment between all n training points that is calculated by kriging. This is achieved by building a $n \times n$ correlation matrix $\boldsymbol{R}$ between all pairs of training points with elements R_{ij}, given by

$$\mathrm{R}_{ij} = cor[\epsilon(\boldsymbol{x}^i), \epsilon(\boldsymbol{x}^j)] = \exp\left[-\sum_{h=1}^{d} \theta_h \left|x_h^i - x_h^j\right|^{p_h}\right] \tag{1.31}$$

where $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$ are training points composed of d features. The parameters θ_h ($\theta_h \geq 0$) and p_h ($1 < p_h \leq 2$) describe the importance of each feature h and may be written as the d-dimensional vectors $\boldsymbol{\theta}$ and $\boldsymbol{p}$. A large value of θ_h corresponds to a feature being highly

correlated to the output topological property. 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

$$\hat{L}(\boldsymbol{\theta}, \boldsymbol{p}) = -\frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2}\log(|\boldsymbol{R}|) \tag{1.32}$$

where

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \mathbf{1}\hat{\mu})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \mathbf{1}\hat{\mu})}{n} \tag{1.33}$$

and

$$\hat{\mu} = \frac{\mathbf{1}'\boldsymbol{R}^{-1}\boldsymbol{y}}{\mathbf{1}'\boldsymbol{R}^{-1}\mathbf{1}} \tag{1.34}$$

where $\boldsymbol{y}$ is a vector of response values for each training point and $\mathbf{1}$ is a vector of 1s. Another (very different) machine learning method called particle swarm optimisation (PSO)[44] then searches for the optimum values of θ and p that maximise the concentrated log-likelihood function.

In **Section 1.5** it was stated that each atom is described by 25 multipole moments, and therefore when building kriging models to describe electrostatics, 25 kriging models, each modelling one multipole moment, are required per atom. When building kriging models to describe the IQA energy of a system, each atom has two kriging models associated with it, one describing V_{self}^{A} and the other describing $V_{inter}^{A,A'}$. The A' in $V_{inter}^{A,A'}$ corresponds to all atoms except A.

The kriging models are tested on an *external test set* of geometries, and this process is described in **Section 1.7** where the GAIA protocol for building and testing QCTFF kriging models is outlined.

1.6. Multipole Moments against Atomic Point Charges

It has been stated that the inclusion of higher order multipole moments should provide an improved description of the electrostatic term in a MM force field. What follows is a review of the literature where multipole moments have been applied. The review first addresses polar systems and the many intermolecular interactions that exist in such systems, and then moves to assess the advantages of multipole moments in the prediction of the crystal structures of organic molecules.

1.6.1. Polar systems and Intermolecular Interactions

Chemistry is overwhelmingly polar. One consequence of this polarity is that chemical systems are dominated by electrostatic interactions between partially charged atoms, both attractive and repulsive. It is beyond the scope of this work to present a comprehensive list

of all the different classes of intermolecular interactions, systematically providing each with a comparison of point charge electrostatics and multipolar methods- not only are there too many classes of interaction to cover them all, but also not all have, as yet, received comprehensive multipolar treatment. What this review does aim to achieve is to present a selection of 'key' interactions and systems, and to provide the reader with enough information to draw their own, informed, conclusions. It is the opinion of the authors that the accurate modelling of polar systems requires multipole moments, preferably with polarization taken into account, and it will be seen that the point charge approximation is unable to outperform the more complete description given by multipole moments. The discussion initially shall address descriptions of water where it will be seen that to accurately model water, a system dominated by hydrogen bonds, one must capture the anisotropy of the electron density through use of multipole moments. The discussion shall then move to show that this is the case for many other types of intermolecular interaction and chemical phenomenon.

1.6.1.1. Water

Water is a highly polar molecule with bulk properties unusual for a molecule of its size; it is a liquid in standard conditions and it is less dense in the solid phase than the liquid. These are known to be a consequence of strong hydrogen bonds between adjacent molecules. As biological processes occur predominantly in aqueous medium, an accurate electrostatic model for water is vital, and this need is represented by the multitude of potentials developed specifically to describe water for use in molecular simulations [45, 46]. Early electrostatic potentials for water consisted of atomic point charges that are fit to reproduce the bulk properties of liquid water. Examples include the simple point charge (SPC) model of Berendesen *et al.* and the TIP3P potential of Jorgensen *et al*. These potentials are still used today, however they both are known to suffer from the same pitfalls- notably they are unable to accurately reproduce the experimentally observed radial distribution function for O…O, $g_{OO}(r)$, and provide poor description of the density vs. temperature profile.

Many attempts at improving the description of water within a point charge approximation involve the use of additional charge sites, away from the nuclear positions. This is to accurately capture the anisotropic electronic density around the molecule, in particular the oxygen lone pairs. This method is analogous to the use of additional charge sites often placed above and below the plane of an aromatic ring to model the delocalised electron clouds. The TIP4P and TIP5P potentials of Jorgensen *et al*. [47, 48] and the ST2 potential of Stillinger and Rahman [49] are of this type, and remain the most widely used in simulations involving water to date [46]. The TIP4P model shifts the negative partial charge of the oxygen towards the centre of mass, whereas the TIP5P model consists of a tetrahedral arrangement of charges with two partial positive charges centred on the hydrogen atoms and two equal and opposite partial charges located at the lone pair sites of

the oxygen atom. 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 vs. temperature profile. The inaccuracies of simple point charge water potentials have led to the development of many new water potentials incorporating multipolar electrostatics [50-52].

The anisotropic site potential for water (ASP-W) of Stone and Millot [53] uses an atom centred distributed multipole analysis (DMA) expansion, with multipole moments up to quadrupole on the oxygen atom and dipole on the hydrogen atoms, computed at the MP2 level. The need for the quadrupole moments on the oxygen atoms in particular are reportedly needed for accurate description of the electronic distribution about a water molecule. When the ASP-W potential was compared with the point charge potentials CKL [54] and NCC [55], and the multipolar potential PE [56], only PE provided comparably accurate minimum energy geometry for the water dimer. The ASP-W potential has been further improved to give the ASP-W2 and ASP-W4 potentials [57, 58]. The atomic multipolar expansions for the ASP-W4 model is now truncated at hexadecupole moments, and all interaction terms up to rank L=5 are included in the model. The ASP-W2/4 potentials were found to give a more detailed description of the potential energy surface of the water dimer than many other water potentials in the literature, although perfect agreement with the high level *ab initio* calculations of Smith *et al* was not achieved [239]. ASP-W2/4 gave values for the second virial coefficient, $B(T)$, close to the experimental values over the temperature range of 373-973K, an improvement over point charge models such as the TIP(X)P potentials as previously discussed.

As a result of being parameterised for the reproduction of the bulk properties of liquid water, most point charge potentials, such as TIPXP, provide a poor description at ice surfaces and for small clusters. Xantheas *et al*. [51] showed the importance of higher order multipoles for the accurate modelling of the electric field of ice surfaces and water clusters. In their work they present the 'induction model', in which each water molecule is modelled by a centre of mass multipolar expansion. The induction model was tested by comparison of the electric field inside a vacancy in ice to first principle calculations (MP2 and the density functional PW91). It was found that 70% of the electric field is dipolar and that a centre-of-mass multipolar expansion up to hexadecupole is needed to obtain good agreement with the *ab initio* calculations. The authors comment that the induction model showed that "*accurate description of the electrostatic interactions of water molecules can be achieved without having to carry out the computationally demanding first principle calculations for large systems*". The TIP4P model predicted the electric field within the studied ice vacancy to be 20% higher than that obtained from first principles, despite the TIP4P model expressing "*acceptable overall properties*".

In another direct comparison of point charge electrostatics against multipolar methods, TIP4P and ASP-W4 were used to model the behaviour of water adsorbed onto a NaCl surface [59]. The experimental adsorption isotherm for water on NaCl shows four distinct regions; a *low coverage* region, a *transition region*, a high coverage region and a *presolution* region [60]. Monte Carlo simulations of the *low coverage* and *high coverage* regions were performed using both water potentials. The results of the *low coverage* simulation for both water potentials showed clustering of water molecules, which agreed with experimental FTIR data. At *high coverage* only the ASP-W4 potential showed a more ordered structure, with three distinct layers of water due to interactions between water molecules with the Na^+ and Cl^- ions. The TIP4P potential did not reproduce this layering, with only a single layer of water strongly bound to the surface with the rest acting as bulk water. Similar conclusion to that of Xantheas *et al* is drawn: the fitting of simple point potentials to reproduce the properties of bulk liquid water are unable to accurately reproduce the behaviour of water at surfaces and interfaces. In fact, the electrostatic interaction between the water molecule and the surface was found to be "*significantly smaller*" for the TIP4P model than for ASP-W4.

1.6.1.2. Hydrogen Bonding

Hydrogen bonds are amongst the most abundant and important non-covalent bond types, ubiquitous in both biochemistry [61-63] and materials chemistry [64-66]. Such interactions are not only strong, but are also observed to be highly directional, in many cases due to the geometry of the lone pairs on the acceptor atom [67-69]. Isotropic atomic point charges have been proven unable to accurately reproduce experimental bonding geometries for a range of molecules [70-76]. Efforts to model the directionality of hydrogen bonding within a point charge framework generally fall into two categories; the first being additional functions applied only to hydrogen bonding atoms, and the second being the addition of partial charges, typically in the positions of lone pairs. Although the two approaches have been successfully used to improve the description of hydrogen bonding in the simulation of polar molecules, both approaches are highly empirical and lack strong theoretical foundation. Allinger *et al*. implemented a directionality term into the hydrogen bonding potential of the MM3 force field and found that the interaction energies for 40 hydrogen bonded dimers was generally in better agreement with the *ab initio* MP2/6-31G** value [77, 78] than the standard MM3 force field. Kollman *et al.* developed a methodology for deriving additional lone pair point charges for use within a revised version of the AMBER force field [79]. The new potentials were tested on small dimers such as NMA…H_2O, and the results showed that the inclusion of the additional sites was able to reproduce much of the directionality observed in MP2 calculations. The additional point charges also led to improved molecular dipole moments, leading to more accurate thermodynamic properties upon molecular simulation.

Multipole moments have been shown capable of describing correctly both the directionality and strength of hydrogen bonding for many systems. Kong and Yan [80] found that multipole moments up to minimum interaction rank L=3 were required to reproduce the bent structures of the hydrides 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.* [35] where minimum interaction rank of L=5 was needed to reproduce the optimised *ab initio* structures for water clusters and the hydrated amino acids serine and tyrosine. Models including QCT multipole moment interactions of rank L=1-6 were compared with various point charge potentials, AMBER, CHARMM, MMFF, OPLS, TAFF and TIP4P. Again, in models where only point charges were included, equivalent to multipolar interaction rank L=1, pseudo-planar ring geometries were predicted that were too "flat" with the hydrogen atoms not enough sticking out of the plane of the oxygen atoms. It is, however, noted that as the number of water molecules in the cluster increases, models including only lower order moments such as monopole and dipole moments did recover to some extent. This is reasoned to be a result of two effects; the first being that for larger clusters there is an increase in the number of long range interactions, which are well described by low rank terms, and the second reason being that water molecules in larger clusters are locked into more rigid hydrogen bonding networks. If the torques generated from the interactions between higher moments are not strong enough to break the hydrogen bonds then the point charge should indeed perform comparably to higher order descriptions. This is in strong agreement with the observed success of many point charge potentials capable of describing 'bulk' properties such as the TIPxP potentials for water, despite their inability to provide reliable results when implicit water molecules are present. The superiority of multipolar electrostatics over point charges for describing hydrogen bonding is seen again in work by Ponder *et al*. [81]. The hydrogen bond association energy of the formaldehyde…water dimer O-H…O=C interaction with changing angle was calculated using both their own multipolar force field AMOEBA and also the point charge OPLS-AA force field. The results were compared to the MP2/aug-cc-pVTZ BSSE corrected values as an assessment of their ability to model the directionality. From **Figure 1.2** it is clear that the isotropic electrostatic potential of the OPLS-AA force field was incapable of reproducing the energy minima at ~100° and ~260° only showing the slightest of maximum at 180°. The AMOEBA results were more satisfactory, showing similar shape to the MP2 curve.

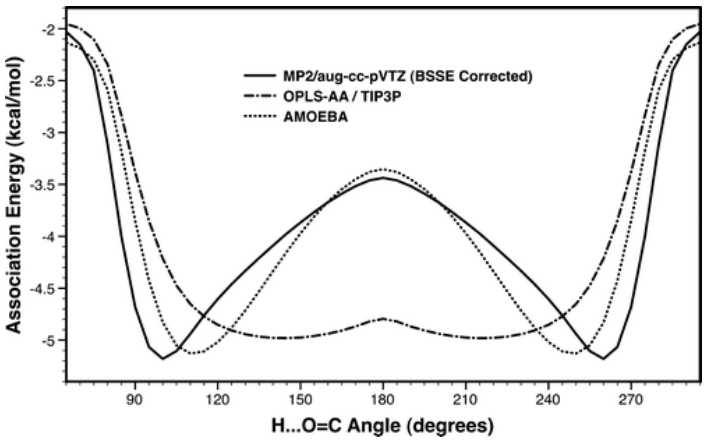


Figure 1.2: Change in the association energy of an O-H…O=C hydrogen bond interaction with changing bond angle for three different levels of theory. Reference [81].

In recent years, there has been a growing interest in what is termed the 'weak hydrogen bond' [63]. This is a hydrogen bond type interaction where the donor atom is not a strongly electronegative atom as found in conventional hydrogen bonding, for example C-H…N/O [82] or C-H…π [83, 84]. These interactions, although weaker, can be of significance for the chiral recognition of a substrate by proteins and also for stabilising the conformations adopted by important biomolecules [62]. Simulations utilising classical point charge force fields do to some extent pick up on such interactions, however the work of Westhof *et al*. showed that the cut-off distance for electrostatic interactions must be large for weak hydrogen bonds to be observed [85]. In simulations of a loop of $tRNA^{Asp}$ in water, upon increasing the cut-off distance from 8-16 Å the stabilising effects of two C-H...O interactions were observed to play a more important role in maintaining the structure of the loop. Solute-solvent interactions between non-polar C-H groups with water were observed, and were seen to play a small role in the solvation of $tRNA^{Asp}$ [86]. Despite interactions involving higher order multipole moments contributing little to the total energy of a system, their inclusion can be crucial where weak hydrogen bonding is present. DMA quadrupole and octopole moments were found to be 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 [87]. Obviously, the widely used point charge models such as AMBER, CHARMM and OPLS are currently unable to account for such interactions, and until high rank multipolar electrostatics are widely implemented, their subtle influence on many chemical systems and processes will remain unaccounted for.

1.6.1.3. Halogen Bonding

Halogen atoms are traditionally thought of as partially negatively charged nucleophilic atoms, interacting typically as hydrogen bond acceptors. Although this is often the case, 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 be both as strong as hydrogen bonding,

ranging from ~4-160 kJ mol^{-1}, and can also influence the structure of a system as a result of their directionality in similar fashion to hydrogen bonds. This pattern of bonding was first reported by Ramasubbu *et al* in 1986, where upon inspection of the adopted crystal structures of halogen atoms within the Cambridge Crystallographic Data Base, he wrote that *"the halogen X in a C-X bond is capable of significant interactions with electrophiles, nucleophiles, and other halogens. The electrophiles approach X of the C-X "side-on", nearly normal to C-X, and the nucleophiles nearly "head-on" and behind the C-X bond."* [88]. Since their discovery, the halogen bond has been the subject of many studies in order to elucidate their origin and nature [89-92]. Torii and Yoshida showed that the quadrupole moment Θ_{zz} of halogen atoms, where the z-axis is defined as the direction of the C-X bond, describe a positive region opposite the C-X bond. This region is commonly referred to as the σ-hole, and it is the position of this which accounts for the observed linear bonding to nucleophiles [90]. Halogen bonding was proven to be dictated primarily by electrostatic effects through the work of Tsuzuki *et al*. [89] through study of C_6F_6X and C_6H_6X interacting with pyridine, although induction and dispersion interactions were found to contribute. It was observed that the strength of a halogen bond is dependent upon the halogen atom involved, where I > Br > Cl, and F does not form halogen bonds. Due to the observed anisotropy in the electronic distribution, point charges fail to correctly model halogen bonding, and this has resulted in both the modification of existing potentials and the development of new potentials to reproduce these effects in molecular simulation. 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 [93]. The EP charge was placed opposite the C-X bond, and the partial atomic charges were recalculated via a restrained electrostatic potential (RESP) approach. MM interaction energies of the halogen containing molecules with a variety of Lewis bases were compared with DFT and MP2 energies, where the MM interaction energies 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 compared to when it was absent. In a medicinal chemistry application of the EP model [93], a simulations of 4,5,6,7-tetrachloro-, bromo-, and iodo-benzotriazoles in the active site of the enzyme phospho-CDK2/cyclin were performed. Two halogen bonds between the halogenated substrate and two carbonyl containing amino acids are known to be present from the x-ray crystal structure. The distributions of the halogen bond angles were in good agreement with the known order of strengths of the different halogens in their bonding, where the strongest bond, I…O was most linear and the weakest bond Cl…O was least linear. When the standard AMBER potentials were used without the EP charge sites, no halogen bonding was observed, with the X…O distances too large. This shows some success in accounting for halogen bonding within a point charge model, however when considering a multipolar force field, one would not have to deal with such "messy" extensions to the model, as the multipole moments, particularly the

quadrupole moments, would describe the electronic distribution sufficiently well. Until such force fields are readably available, QM/MM calculations have been suggested as an alternative to force fields [94], where halogen bonding between substrate and enzyme are described in the QM scheme.

1.6.1.4. Biomolecular

Biological macromolecules such as proteins and nucleic acids are highly polar molecules which engage in a wide range of interactions. Structural motifs such as the commonly observed α-helical and β-sheet structures of proteins are held together by networks of hydrogen bonds, and the double helix structure of DNA has been found to be stabilised largely by the stacking interactions between aromatic bases. Biological molecules frequently interact with and/or through their own aromatic π-electron density during such processes as recognition and catalysis. Point charges are known to provide a poor description of the electronic distribution of aromatic systems, and the XED force field of Chessari *et al*. was an early effort to capture the anisotropy by the addition of extra point charge sites [95]. The XED force field was able to correctly predict the edge to face stacking for a range of substituted polyphenyl species, where AMBER, OPLS, MM2 and MM3 were not. The work showed that "*a good electrostatic description is necessary in order to model non-covalent interactions due to electron anisotropy*." Hill *et al*. have also expressed the need for good description of electrostatics when considering aromatic systems. Their work demonstrates that electrostatics play an important role in determining the stabilisation of aromatic stacking interactions due to a degree of cancelling of the attractive correlation dispersion term by exchange repulsion and delocalisation effects [96]. Ghosh *et al*. used their 'Effective Fragment Potential' (EFP) method to investigate the interactions between nucleic acid bases, both the hydrogen bonding between base pairs and also the π-stacking interacting between the 'rungs' of bases[97]. The EFP method is described as a low cost alternative to *ab initio* calculations, and can be considered as a polarisable multipolar force field without empirically fitted parameters. A DMA is performed on atomic centres and bond midpoints up to octopole moment, with the interactions considered being charge-charge/dipole/quadrupole/octopole, dipole-dipole/quadrupole, and quadrupole-quadrupole. The EFP method was able to accurately reproduce the interaction energies between stacked dimers AA and TT, with deviation from MP2 energies within 1.5 and 3.5 kcal mol^{-1} respectively. They too found that the electrostatic interactions were key in describing the relative stabilities of different orientations of the stacked dimers, and the results can be seen in **Figure 1.3**. The coulombic contribution can be seen to have clear importance for the energies when rotating the dimers.

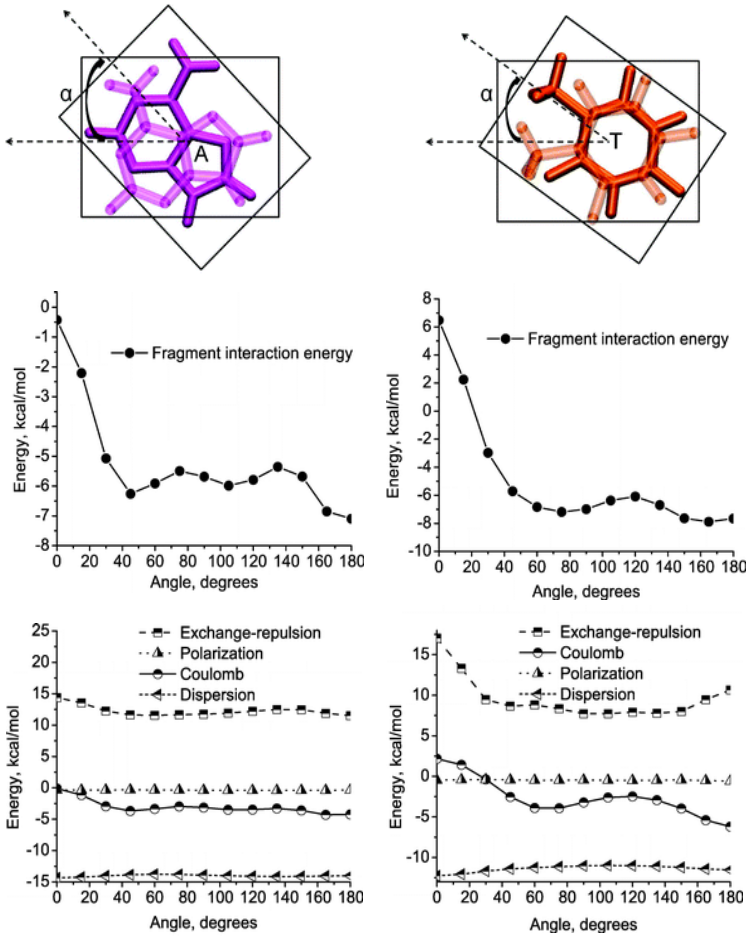


Figure 1.3: Effect of rotating the AA (*left*) and TT (*right*) base pairs around angle α on total interaction energy and on the individual contributions to the total energy. Reference [97].

An extension to the AMOEBA force field has been implemented by Tafipolski and Engles, which shows a much improved description of stacked aromatic systems [98]. This approach includes atomic multipole moments up to hexadecapole, with dipolar polarisabilities reparameterised from the existing AMOEBA values. The new model also includes 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 SAPT values. The new model was successful over a range of dimers of many poly aromatic hydrocarbons (PAHs), and the results can be seen in **Figure 1.4**.

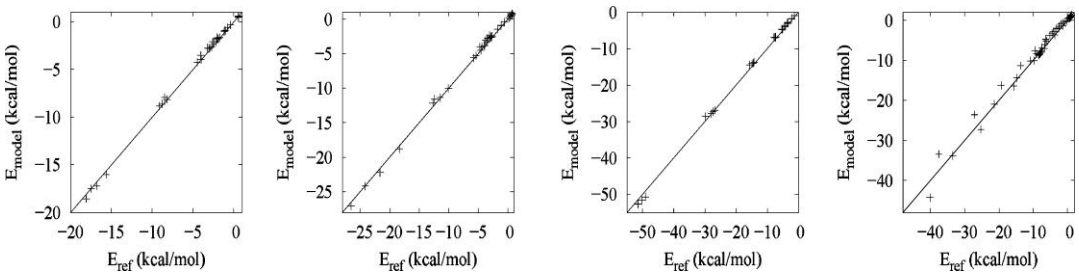


Figure 1.4: Scatter plots of electrostatic energies for selected dimers of PAHs (in kcal/mol). From left to right: naphthalene (35 configurations), anthracene (32 configurations), pyrene (44 configurations), and coronene (56 configurations). The reference data are taken from Podeszwa *et al.* [99, 100]

Cation-π interactions are important both in biomolecular recognition and in the structure adopted by biological macromolecules. Marshall *et al*. [101] ran simulations on β-hairpin structures of model polypeptides involving cation-π interactions between cationic $(Me)_N$-Lys^+ (MeNK) residues and two aromatic tryptophan side chains (where N = 0, 1, 2, 3). Simulations were run using the multipolar polarisable force field AMOEBA, and the point charge force fields OPLS-AA, CHARMM and AMBER. The results of the simulations were compared to experimental NOE values for distances between the lysine and tryptophan side chains. Only the AMOEBA force field was able to reproduce the experimental NOEs with any consistency, accurately predicting over 80% of the observed NOEs across the four systems (**Figure 1.5**). 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%. The relative success of AMOEBA over the point charge force fields is clear indication that detailed description of electrostatics is a requirement for modelling cation-π interactions.

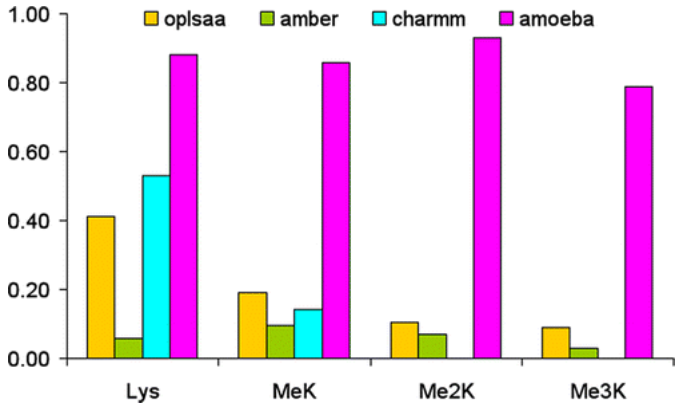


Figure 1.5: Fraction of successfully predicted experimental NOEs by the four force fields OPLSAA, AMBER, CHARM and AMOEBA for the four different residues Lys(K), MeK, Me_2K and Me_3K. Ref. [101]

1.6.1.5. Solvation

When designing novel drugs, catalysts, and other novel compounds, considerations must be made for their behaviour upon solution. Drug molecules may be protonated or deprotonated, leading to inactive and even toxic forms, and inorganic materials may remain in their crystalline forms. A key contribution to the solvation of a material in polar

solvents is its ability to form hydrogen bonds with solvent molecules. Abraham *et al* introduced a scale designed to describe the hydrogen bond donor/acceptor strength of a molecule, labelled α_2^H and β_2^H respectively[102]. An AIM based multipolar method for the prediction of α_2^H values for 39 hydrogen bond donor molecules with hydrogen cyanide was tested against a point charge only scheme [103]. The results showed that the correlation between α_2^H and donor hydrogen atomic charge was poor, R^2= 0.567, however this correlation improved to R^2=0.635 upon the use of atomic dipole moments, and improved further to R^2=0.725 upon inclusion of atomic quadrupole moments. The correlation of α_2^H with dipole and quadrupole only (no charge) was only slightly worse than when the donor hydrogens charge was included, with an R^2=0.721. It was found that a large charge on the donor hydrogen gives a larger α_2^H, whereas a large dipolarity on the donor hydrogen gives a smaller α_2^H. This is described qualitatively by there being more of the donor hydrogens electron density in weaker, less polarised X-H bonds.

Experimentally, measuring the thermodynamic properties relating to the solvation of a single ionic species is very difficult due to the presence of the counter ion(s). This is overcome by experimentalists through the application of extrathermodynamic assumptions. Computational methods, however, are able to model single ions in solution to obtain properties such as single ion enthalpies of solvation. One caveat to this is that ions are highly polarising and hence polarise their surrounding solvent molecules. This effect is not captured by force fields employing fixed isotropic point charges for their electrostatics, as polarization requires dipole moments, typically with their associated dipolar polarisabilities. Another dilemma concerning the use of traditional force fields such as AMBER, CHARMM and OPLS is that they commonly use ionic solvation free energies during their parameterisation. In an attempt to capture some degree of polarisation in the condensed phase, Kollman *et al*. [104] derived new atomic point charges for the AMBER force field using *ab initio* calculations performed with continuum solvent and dielectric constant of ε=4, chosen to mimic the hydrophobic interior of a protein. The new charges were reported to give "*encouraging results*", and were able to reproduce the experimental Ramachandran maps for two tripeptides.

The AMOEBA force field of Ponder has been designed to overcome the incapability of classical force fields such as AMBER and CHARMM of dealing with polarisation. 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 [105-107], including Cl^-, Na^+, K^+, Mg^{2+} and Ca^{2+}. Work by Grossfield *et al*. [106] showed that despite the parameters for AMOEBA being derived from calculations of gas phase molecules, inclusion of polarisation terms allows both accurate and transferrable single ion solvation free energies and also solvation

free energies of whole salts in both water and formamide. The whole salt free energies of solvation varied from experimental results on average by only 0.55 kcal mol^{-1}, whereas the OPLS-AA and CHARM27 force fields deviated from experiment by 9.8 and 6.6 kcal mol^{-1} respectively. The radial distribution of solvent molecules around the K^+ and Cl^- were plotted (**Figure 1.6**), and it is observed that the non-polarisable force fields show over structuring, which is a consequence of fixed point charges, favouring only a limited range of geometries.

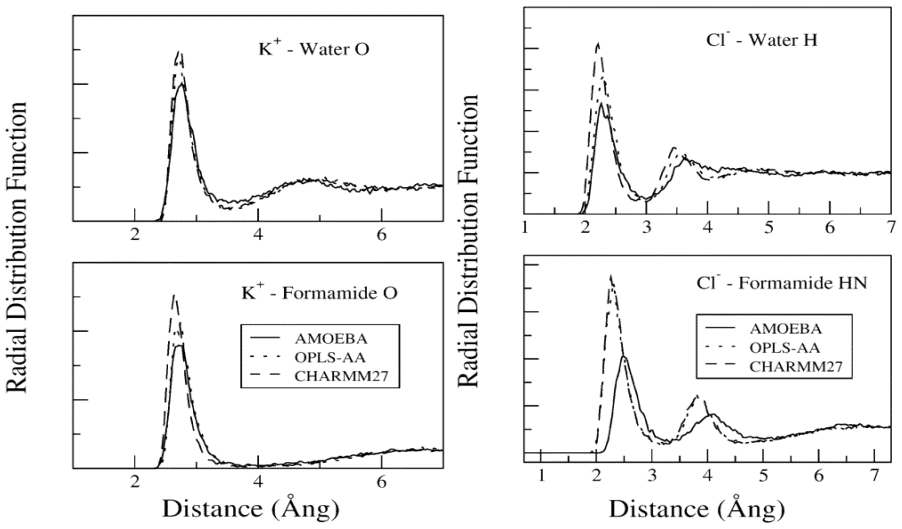


Figure 1.6: Radial distribution functions of *left*: the oxygen atoms of water (*top left*) and formamide (*bottom left*) around a K^+ ion and *right*: the hydrogen atoms of water (*top right*) and amide hydrogen of formamide (*bottom right)* around a chloride ion. Ref. [106]

1.6.2. Crystal Structure Prediction

Despite advances in the area over the last thirty years, there is still no reliable way to accurately predict the crystal structure adopted by a particular molecule. Reliable predictions will streamline many industrial processes such as pharmaceutical development, the screening of compounds for non-centrosymmetric lattices for use in non-linear-optics, development of novel metal organic frameworks where the pore size influences catalysis and potential for gas storage, and even in the synthesis of new explosives [108]. 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 potential energy surface. In the context of this review, the discussion will be restricted primarily to detailing how a more detailed 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 [108].

The most commonly used criteria for assessing the accuracy of a predicted crystal structure is by lattice energy calculation. It is suggested that a given molecule will adopt a crystal structure with the lowest possible lattice energy. This ranking criterion was used

by the Cambridge Crystallographic Data Centre (CCDC), who have been highly active in encouraging groups to participate in a series of four blind tests [109-112], where participants were invited to predict a range of unknown crystal structures as seen in **Figure 1.7**. In each test a range of computational methods were used by the participating groups including point charge, multipolar and statistical approaches. At first glance, the results of the early tests 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-PP [113, 114] method outperformed the multipolar DMAREL [115] method of Price in the CSP1999 test. It was found during post results analysis that DMAREL had found the experimental crystal structure during the search procedure, it would have predicted the experimental structure to have lower energy than the global minimum of earlier runs, indicating that the searching algorithm was to blame rather than the multipolar force field. This was the recurring message across all three early tests- small, rigid molecules containing only C, H, N and O were generally predicted correctly (with multipolar electrostatics providing a slight advantage over point charges), but molecules with a high degree of conformational flexibility were not being sampled thoroughly and as a result, the experimental structures were not identified. The results of the fourth blind test, CSP2007, showed that with the implementation of improved searching algorithms, the multipolar electrostatic methods of both Price *et al.* and of Ammon consistently outperformed methods with point charge electrostatics.

Within the assumption that the crystal structure adopted by an organic molecule will be that with the lowest lattice energy, it is important that the force field used to calculate the energy is the most accurate possible. Considering the electrostatic component only, multipole moments have been shown to provide more reliable contribution to the lattice energy than simple point charges. Work by Day *et al.* found that for 50 organic molecules, using atomic multipole moments up to hexadecupole increased the number of compounds for which only five or fewer crystal structures were predicted to have lower energy than the experimentally observed structure [116]. Of the 64 experimentally observed crystal structures for the 50 compounds, when using multipolar electrostatics 44 were predicted with fewer than five structures lower in energy than experiment, compared to only 36 when point charge electrostatics were used. Multipolar electrostatics also correctly predicted 32 of the compounds to have structures within 0.5 kJ mol^{-1} compared to only 23 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 [117]. 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.

The assumption that the structure with the lowest lattice energy will be the observed experimental structure is unfortunately an oversimplification of the problem. The presence of multiple crystal polymorphs and locally bound "metastable" structures can result in the experimental structure of a given molecule being higher in energy than the predicted global minimum [118]. Kinetically stable structures often arise due to strong intermolecular electrostatic forces such as hydrogen bonding, and as has been described elsewhere in this work, hydrogen bonds are strong and highly directional and require higher order moments for accurate description. It was observed by Day *et al.* that for 50 organic molecules with many polymorphic crystal structures, lattice energy minimisation using atomic point charges was considerably less accurate for molecules with hydrogen bond donor-acceptor groups than for those without [66] The primitive isotropic point charge descriptions within the FIT [119, 120], W99 [121-123], DREIDING [124], CVFF95 [125-127] and COMPASS [128] 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 potential energy surface, 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-amino-3-nitropyrimidine due to the crystal relaxing into the energy well of another polymorph. A similar result was seen by Price *et al.* [240] where an electrostatic potential containing DMA atomic multipoles moments up to hexadecupole failed to predict an experimental "buckled sheet" polymorph of 2-ammino-5-nitropyrimidine. Three polymorphs of 2-amino-5-nitropyrimidine are found experimentally: (i) Layered planar sheets with molecules linked by a network of hydrogen bonds, (ii) the previously mentioned "buckled" sheet structure which consists of the same hydrogen bonding motif as polymorph (i), and finally (iii) a highly symmetric non-layered structure. Polymorphs (i) and (iii) were correctly predicted by the multipolar methods, however the multipolar potential was too repulsive and polymorph (ii) always flattened during minimisation. This result was attributed to the isotropic repulsive terms in the force field used rather than inadequacies of a multipolar potential, and as such can be considered to support the use of anisotropic multipole moments over isotropic point charges.

There are sometimes cases in the literature where the use of multipole moments does 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 non-superiority of multipole moments. A novel electrostatic potential built for the MM3 force field was tested on the crystal structures of oligothiophenes [129] and it was found that atomic point charges outperformed multipole moments for all but one case, α-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. The RMSD between the reference and fitted electrostatic energies was also significantly higher for PFT4 than other test molecules, suggesting that fitted point charges were insufficient to model the electrostatic contribution to crystal structure. Other examples have been discussed where factors such as the flexibility of a molecule and the searching algorithms used in the lattice energy calculation were responsible for "hiding" the improved accuracy offered by atomic multipole moments. Brodersen *et al*. compared five electrostatic models including both ESP derived point charges and multipole moments were tested for the prediction of 48 crystal structures, again using the DREIDING force field [130]. Due to strong dependence on intramolecular terms in the force field, such as angle bends, bond stretches and torsion angles, the use of multipoles did not improve the accuracy of the predicted crystal structures for flexible molecules. They did however greatly improve the prediction for rigid molecule crystal structure, where the bonded terms are of less importance.

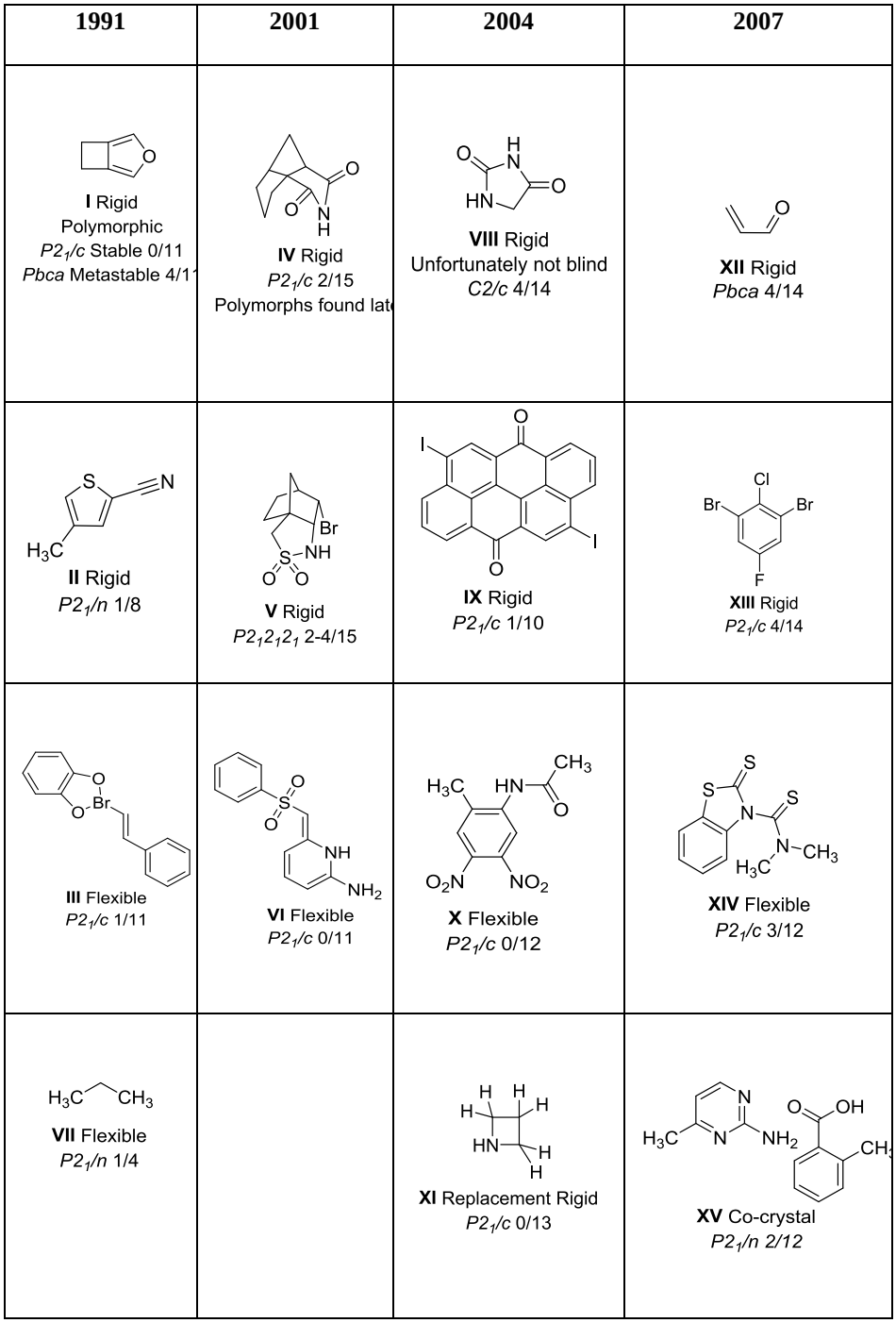


Figure 1.7: The 15 molecules used across the four blind studies of crystal structure prediction CSP1999, CSP2001, CSP2004 and CSP2007. Ref. [1]

1.7. The GAIA Protocol

An automated process has been developed for the streamlined generation of kriging models, named GAIA. In previous publications from the group, it has often been referred to by its older names "Pipeline" and "Autoline". GAIA is a Perl script written and maintained by a postdoctoral research assistant in our group, and so I claim no credit for its development, however the processes which it performs are fundamental to much of the

work in this thesis, in particular **Chapters 2 and 3** so a discussion and overview is required.

An overview of the GAIA protocol is provided in **Figure 1.8**. GAIA was developed in response to the high throughput nature of the work performed in the group. Here, a discussion of the default parameters is provided and unless specified in the corresponding results chapters apply to the work in all subsequent chapters.

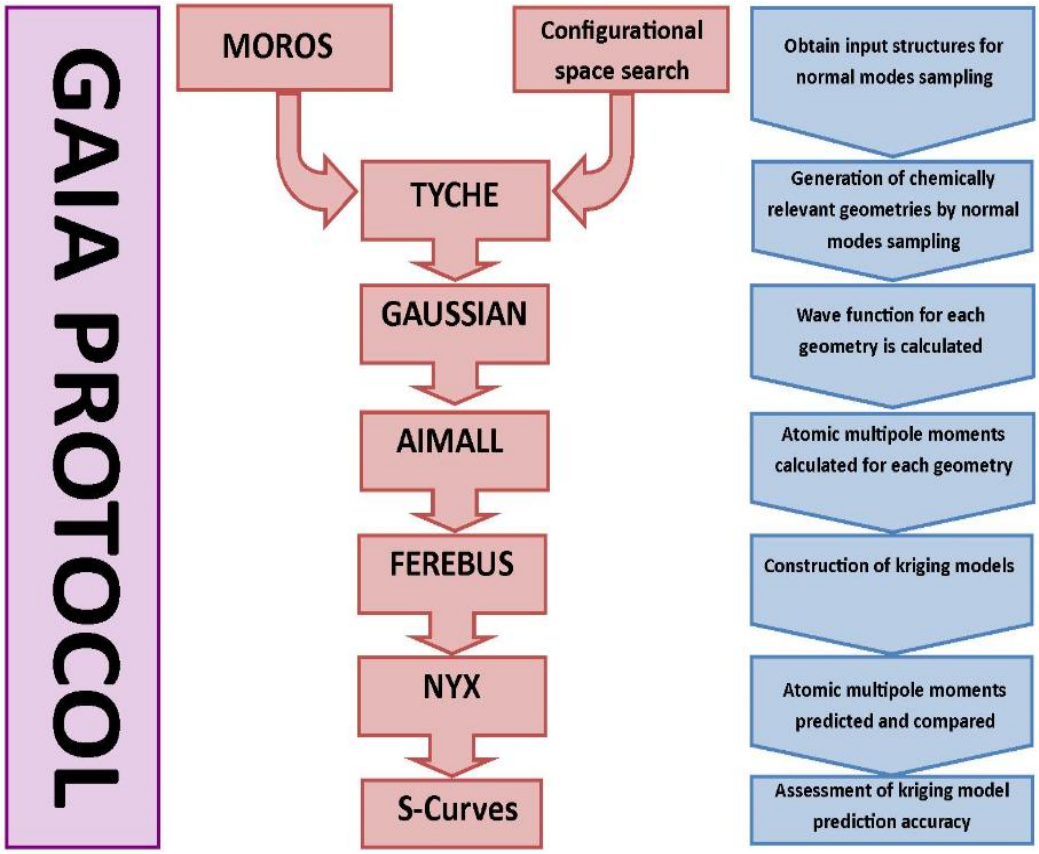


Figure 1.8: The fully automated GAIA protocol for building and testing QCTFF kriging models

Kriging requires two large sets of molecular geometries, one for training the models and the other for testing the models. Therefore, the first stage of the GAIA protocol is to sample a large number of relevant molecular geometries. A normal modes sampling approach is used, where pseudorandom quantities of energy are put into the normal vibrational modes of at least one "seed" geometry, and as the molecule is allowed to vibrate, "snapshots" of the molecular coordinates are taken. Each snapshot is a sampled molecular geometry that can then be used to build kriging models. The input seed geometries can be obtained, in theory, by any chemically justifiable approach. In the work discussed in this thesis three methods have been used. The first method is to take the seed structures direct from an external source, and this approach is followed in **Chapter 2** where molecular complexes are taken from the S22 database[131]. The second approach for obtaining seed geometries is to perform a search of the potential energy surface of the molecule, and to identify local energetic minima through a configurational space search. The local energy minima are

then used as seeds for the GAIA protocol. This is one of the two approaches used in **Chapter 3**, where the local energetic minima of the amino acids alanine and lysine are used. The details of the search used to obtain the energetic minima of alanine and lysine can be found in reference[132, 133]. The third method of obtaining seed geometries is to sample from X-ray crystal structures. I have developed a code named MOROS that enables the selective extraction of an amino acid from a number of protein crystal structures. An in depth discussion of sampling amino acids from both energetic minima and from protein crystal structures is provided in **Chapter 3**. A description of MOROS will now be provided, however a more technical description of the MOROS code can be found in **Appendix D**.

The first stage of sampling amino acids from crystal structures is to add hydrogen atoms to the .pdb files as these are not included in standard X-ray structures. Hydrogen atoms are added to all protein crystal structures using the HAAD code of Li *et al.*[134]. The HAAD algorithm was developed to add accurately hydrogen atoms by analysing the positions of nearby heavy atoms, following the basic rules of orbital hybridisation and through optimisation of steric and electrostatic parameters. HAAD was found to outperform the popular software CHARMM and REDUCE[135] 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 (that are able to locate the positions of hydrogen atoms unlike standard X-ray structures). MOROS then searches through the set of crystal structures for all examples of a given amino acid and uses the coordinates to output a Gaussian job file (.gjf) file for each sampled amino acid. Because we are interested in "capped" amino acids, i.e CH_3CO-(amino acid)-$NHCH_3$, the peptide bond and alpha carbon atoms of the residues either side of the extracted amino acids are extracted with the central amino acid and then hydrogen atoms are added. **Figure 1.9** shows the atoms extracted by MOROS including the amino acid of interest (blue box), and also atoms that make up the caps (red box).

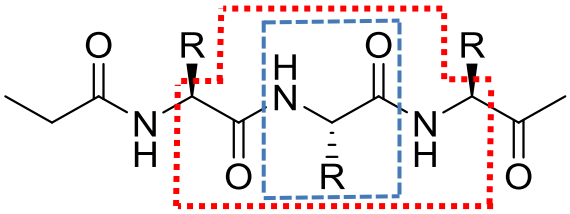


Figure 1.9: Diagrammatic representation of the atoms extracted by MOROS including the target amino acid (*blue box*) and also the full set of atoms including those used to make the peptide caps (*red box*).

Once a set of seed geometries has been obtained, the next step of the GAIA protocol is to use normal modes sampling. TYCHE is the program responsible for performing the normal modes sampling. A frequencies calculation using Gaussian is required for each seed geometry, and then TYCHE inserts energy into the normal modes of the seed molecule in a pseudo-random distribution. As the molecule is allowed to vibrate, snapshots of the distorted molecule are taken and these are then used to build the training sets for kriging.

A maximum bond-stretch parameter is defined to ensure that no molecule is over-distorted, and this is typically set to ±10%. This means that any sampled geometry with a bond length greater than 1.1 times the sum of the van der Waals radii of the bonded atoms is discarded. When using seed geometries that are not energetic minima, a non-stationary normal modes sampling approach must be used. This is because the first derivative term of the Taylor expansion used to calculate the vibrational modes is no longer zero and thus must be included in the calculation of the normal modes. The derivation of the non-stationary normal modes approach has been provided by Cardamone *et al* (in press) and is also provided in the supplementary information of Hughes *et al*[136].

The next stage in the GAIA protocol is to obtain molecular wave functions for each sampled geometry output from TYCHE. This is performed by Gaussian[137]. The level of theory used for the results in this work is stated in each chapter. The molecular wave functions are then used as input for the topological analysis, performed by AIMAll [138]. The integration grids used by AIMAll may be adjusted by changing the keywords "-breaq=" and "-boaq=". The parameters used as standard in the current work are "-boaq=high" and "-briaq=auto" as this gives a reliable number of low error results. Depending on the energetic components you want modelled by the kriging models, different levels of AIMAll calculation are available. The "-encomp=" keyword allows the user to determine what calculation AIMAll will perform, with the quickest calculation providing standard QTAIM metrics (BCP densities, values of the Hessian and energy densities for example), as well as the atomic multipole moments, and the most comprehensive calculation providing all IQA E_{inter}^{AB} terms explicitly. When only kriging the atomic multipole moments of a system "-encomp=1" was used. In **Chapter 2** where the IQA V_{self}^{A} and $V_{inter}^{A,A'}$ energies were used to build kriging models "-encomp=3" was used. In **Chapter 6** where the explicit AB pairwise terms such as V_{xc}^{AB} were needed, "-encomp=5" was used.

At this point, each molecular geometry undergoes "scrubbing", where the integration errors of AIMAll for a given molecular geometry are compared to a user defined threshold, and any geometries with an error above the cut-off are discarded. The standard value of the cut-off is 0.001 a.u. as this provides a compromise between high quality topological energy terms but not discarding too many geometries. Next, the molecular geometries are divided into two sets- the training set and the test set. The training sets are used to build the kriging models, and the test sets are used for testing of the kriging models once they have been built. The kriging is performed by the in-house code FEREBUS. There are many parameters that must be defined when kriging, and the following parameters listed apply to all calculations unless specified in the corresponding results chapter.

1. The full training set size was used to build models for all atomic multipole moments. It has been shown that some time can be saved at a negligible cost to the

accuracy if the higher order multipole moments have a reduced training set size. I did not do this, I used the full training set for all kriging models.

2. p was allowed to optimise during all kriging calculations (see **Equation 1.31**).
3. Particle swarm optimisation was used for all kriging models (see **Section 1.5**). An option to use an alternative differential evolution optimization algorithm has been recently added to FEREBUS, but this was not used in the current work.

FEREBUS then uses the models that it builds to make predictions for a set of untrained molecular geometries. Because GAIA includes the test set geometries in the Gaussian and AIMAll calculations, the true values of the given energy terms are known and so the prediction error can be calculated as the difference between the true value and that predicted by FEREBUS using the kriging model. If the models are describing an IQA energy term then the FEREBUS predictions can be plotted directly as S-curves. In **Chapter 2** of this work, kriging models are built the IQA atomic V_{self}^{a} and $V_{inter}^{a,a'}$ energies. In this case, the sum of all IQA energy prediction errors to give a total IQA energy prediction error:

$$Total\ IQA\ Energy\ Error = \sum_{A}^{N\ atoms} \left(\left(V_{self,true}^{A} - V_{self,predicted}^{A} \right) + \left(V_{inter,true}^{A,A'} - V_{inter,predicted}^{A,A'} \right) \right) \tag{1.35}$$

If the kriging models built have been for multipole moments, as is the case in both **Chapters 2 and 3**, the FEREBUS predictions are read by NYX and all 1,4 and higher electrostatic interactions between multipole moments up to interaction rank L=5 are interacted by the program NYX. The total error in this case is given by

$$\left|\Delta E_{system}\right| = \left|E_{system}^{true} - E_{system}^{predicted}\right| = \left|\sum_{AB} E_{AB}^{true} - \sum_{AB} E_{AB}^{predicted}\right| = \left|\sum_{AB} \Delta E_{AB}\right| \tag{1.36}$$

Chapter 2

Kriging the S22 Dataset

Summary

When applied to large biomolecular systems, QCTFF will be required to accurately model a wide range of intermolecular interactions. In the following chapter, kriging models have been built for molecules from the S22 dataset of small molecular dimers. For H-bonded dimers, such as the water dimer and the adenine-thymine base pair, the atomic multipole moments have been kriged as the interactions between such molecules are dominated by the electrostatic interaction term. Models were built at three levels of theory: HF/6-31G**, B3LYP/aug-cc-pVDZ and M06-2X/aug-cc-pVDZ. The quality of the kriging models was measured by their ability to predict the electrostatic interaction energy between atoms in external test examples for which the true energies are known. At all levels of theory, >90% of test cases for small van der Waals complexes were predicted within 1 kJ mol^{-1}, decreasing to 60-70% of test cases for larger base pair complexes. Models built on moments obtained at B3LYP and M06-2X level generally outperformed those at HF level. For all systems the individual interactions were predicted with a mean unsigned error of less than 1 kJ mol^{-1}. For a selection of dispersion bound complexes (benzene dimer, ammonia benzene dimer and water benzene dimer) where the electrostatic interaction is much weaker, the IQA self and interaction energies have been kriged instead. The IQA models were built using the M06-2X level of theory. The three systems had an average prediction error of less than 1.9 kJ mol^{-1} for the sum of the self and interaction energies, with 100% of the test systems predicted within 10.1 kJ mol^{-1}.

A note:

Much of the work in this chapter regarding hydrogen bonded complexes has been published in

> *"T.J. Hughes, S.M. Kandathil, P.L.A. Popelier, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 136, (**2015**), 32–41"*

*found in **Appendix F**. The work presented in this chapter contains only my own work, with all contributions from co-authors omitted.*

2.1 Introduction

Within a protein one may expect to find many different types of atomic interactions including hydrogen bonds, halogen bonds, π-π stacking interactions, ionic bonds, etc. It is important, therefore, that QCTFF is capable of accurately modelling all interaction types with ease. Building kriging models direct from *ab initio* data should in theory describe all interaction types without the need for specific interaction types. The one requirement is that the input training data includes molecular geometries where examples of the interaction types are included. Issues surrounding the sampling of molecular systems are discussed in a later **Chapter 3** and are not considered in detail here.

The following work is divided into two sections. The first section (**Section 2.2**) details the treatment of seven hydrogen bonded complexes present in the S22 dataset[139] and the second (**Section 2.3**) details the treatment of three weakly bound complexes that include a mix of π-π stacking and weak hydrogen bonding interactions. The S22 database was designed originally by Jurecka *et al*. [139] as a collection of highly accurate MP2 and CCSD(T)/CBS interaction energies as a means of validating the accuracy of new computational techniques. The database consists of 22 molecules (as the name suggests), that belong to one of three subgroups. There are seven hydrogen bonded complexes, eight dispersion bound complexes and seven complexes held together by a combination of both dispersion and hydrogen bonding.

2.2. Hydrogen Bonded Dimers

Kriging models were built for the atomic multipole moments of the seven hydrogen bonded complexes of the S22 dataset. As an interaction between atoms of partial charges, hydrogen bonds are highly electrostatic in nature. Atomic multipole moments are essential for accurately describing electrostatic interactions (see **Chapter 1.6**) so the first 25 atomic multipole moments of each atom were kriged for all systems. The hydrogen bonded dimers can be seen in **Figure 2.1**. **Figure 2.1** was built using the MORPHY software package, and topological features such as bond critical points, ring critical points, and bond paths can be observed. The surfaces are cut at 0.0005 au of electron density. The bond path between the two ammonia molecules is not what one may typically expect, as it appears to be connecting the two nitrogen atoms. This is not an error in the QCT analysis, but an interesting topological phenomenon that has been well documented for many small molecular dimers by Bone and Bader[140]. A notable example in their work is that of the CO_2 dimer in which bond paths between the oxygen atoms are observed for the side-on dimer.

2.2.1 Computational Methods

Kriging models were built using the GAIA protocol at three different levels of theory: HF/6-31G**, B3LYP/aug-cc-pVDZ, and M06-2X/aug-cc-pVDZ. This allows comparison of how well kriging performs at different levels of theory. Unpublished work has shown that

B3LYP/apc1 consistently outperforms HF/6-31G**, in that kriging models lead to interaction energies closer to the true energy. This is due to the higher levels of theory including electron correlation, which produces atomic monopole moments of smaller absolute value. It is possible to prove mathematically[141] why for the same kriging settings (e.g. number of data points in the training set) models built at a level of theory that included electron correlation will outperform Hartree-Fock.

The reason for including the Hartree-Fock level of theory in this work requires some justification in the light of its well-known limited accuracy. The first reason for its inclusion is that many currently used force fields, such as AMBER, include parameterisation from Hartree-Fock level data. Showing that our methods are able to produce accurate predictions relative to the "true" Hartree-Fock value proves we can compete with, and eventually supersede, the methodologies currently in place. The second justification for the use of Hartree-Fock is that the work presented here is intended to prove that intermolecular interactions, in particular hydrogen bonding, can be accurately described by a kriging model mapping *ab initio* values to nuclear coordinates. Assuming that any changes in the *ab initio* values for the multipole moments follow the correct patterns with respect to nuclear coordinates, the use of a lower level of theory is justified. It is stressed that all errors presented in this paper are relative to the correct value given at a specific level of theory, not relative to a true experimental or high level of theory *ab initio* value such as CCSD(T)/CBS.

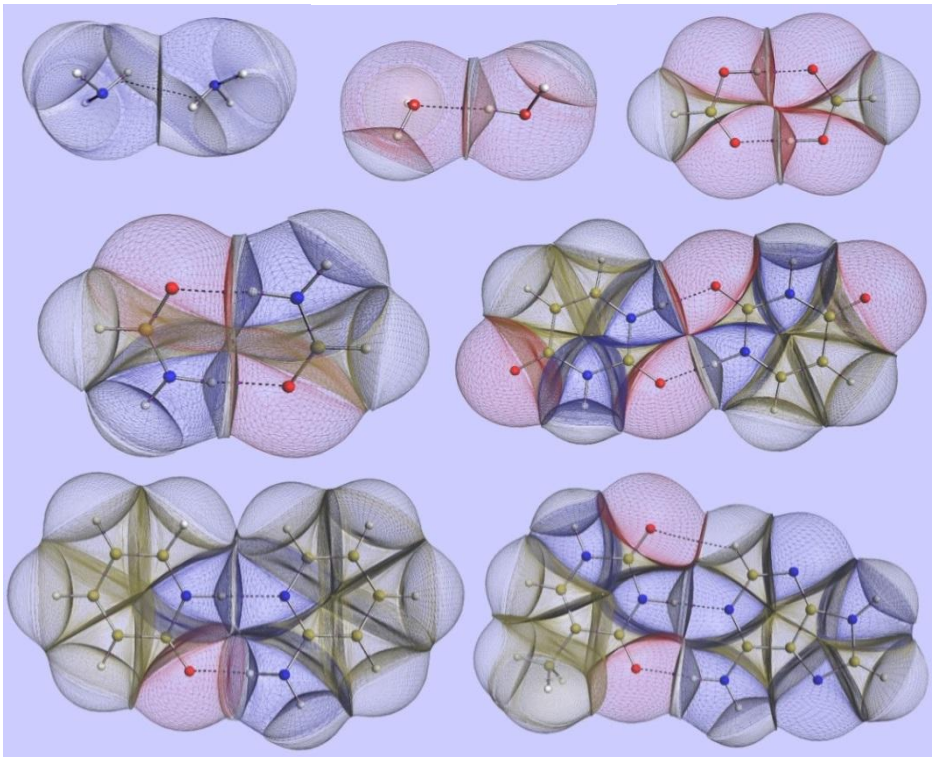


Figure 2.1: The hydrogen bonded dimers studied in this work. Topological atoms are capped at their 0.0001 a.u isosurface. Dimers are i) the ammonia dimer, ii) the water dimer, iii) the formic acid

dimer, iv) the formamide dimer, v) the uracil dimer, vi) the dimer of 2-pyridoxine and 2-aminopyridine, and vii) the adenine thymine base pair

A common criticism of the B3LYP density functional is that it is unable to provide description of long- range electron correlation effects which play a key role in the binding of many van der Waals complexes[142-145]. The M06-2X functional [146] has been specifically designed to provide accurate interaction energies for a range of intermolecular interaction types, in particular van der Waals dimers. In this work we use both the B3LYP and M06-2X functionals to see if improved modelling of the long-range electron correlation lowers the magnitude of the prediction errors of intermolecular interactions provided by our kriging models. In recent work [147] by Friesner *et al*., a database of highly accurate CCSD(T) noncovalent interaction energies was assembled. The database was then used to fit a correction term to be added to the B3LYP density functional to allow for accurate intermolecular interactions (see **Appendix B** for more information). This was tested using the aug-cc-pVDZ and LACVP* basis sets, and was compared with both the B3LYP-D3 method[148], and the M06-2X hybrid functional. In an effort to maintain some level of consistency with the work of Friesner *et al*., the aug-cc-pVDZ Dunning basis set was chosen in this work for building of the B3LYP and M06-2X kriging models.

2.2.2 Effect of the Level of Theory on the Training set

The training set geometries are sampled by putting energy into the normal modes of vibration of the system. These normal modes are calculated directly from the derivatives of the potential energy surface, and so are affected by the level of theory used to construct the Potential Energy Surface (PES). Therefore, one must keep in mind that true comparisons cannot be made between the performances of kriging models at different levels of theory. To generate the training set geometries at each level of theory, the maximum amount of energy is pumped into the sample without breaking any bonds. This maximum amount of energy changes when the PES is built at a different level of theory. For example, Hartree-Fock theory is known to predict bonds to be too polar. Subsequently, the force constants for these bonds are higher than those at B3LYP level, for example. This means that less vibrational motion may take place when pumping a large amount of energy into a HF PES compared to pumping a smaller amount of energy into a B3LYP PES. **Table 2.1** shows the amount of energy put into the systems at different levels of theory. Hartree-Fock does indeed show the greatest tendency to have the most energy pumped in, although is noted that this is not seen throughout. This is in part due to the random way in which energy is put into the vibrational modes.

Table 1.1: Energy (kJ mol^{-1}) pumped into the hydrogen bonded complexes. The highest values are in bold and the lowest values in italics. Numbers in brackets indicate the first lowest average prediction error across 600 external test examples, the second lowest and the highest.

System	**M062X**	**B3LYP**	**HF**
Ammonia Dimer	**150** (1)	120 (3)	*90* (2)
Water Dimer	**90** (1)	*40* (2)	69 (3)
Formic acid Dimer	50 (2)	*46* (1)	**90** (3)
Formamide Dimer	60 (2)	*50* (1)	**110** (3)
Uracil Dimer	180 (1)	180 (2)	**225** (3)
2-Pyridoxine…2-Aminopyridine	200 (3)	*190* (2)	**240** (1)
Adenine…Thymine	*150* (1)	**210 (2)**	180 (3)

As stated above, previous unpublished work of our group has shown that for *the same* training set geometries, B3LYP/apc-1 consistently outperforms HF/6-31G**, yielding kriging models that generate more accurate predictions of the electrostatic interaction between two topological atoms. To confirm that B3LYP/aug-cc-pVDZ also outperforms HF/6-31G**, kriging models were built at B3LYP/aug-cc-pVDZ level using the geometries sampled from the HF/6-31G** PES surface for the ammonia dimer and plotted in **Figure 2.2** as an S-curve. It can be seen for the red line of **Figure 2.2** that 50% of the test set of geometries had absolute prediction errors of less than 0.02 kJ mol^{-1}. Thus it follows that an S-curve that lies to the left is superior to one right of it. The results seen in **Figure 2.2** confirm that B3LYP outperforms HF methods when the same training geometries and test geometries are used. The results also show that the training set geometries obtained from a PES calculated at the B3LYP/aug-cc-pVDZ level (**Fig.2.2**, green line) lead to higher prediction errors for the two curves corresponding to the HF/6-31G** PES sampled training sets (**Fig.2.2,** red and blue lines). **Table 2.1** shows that more energy was put into the B3LYP PES than into the HF PES. Hence, one would expect the training set geometries to span a larger conformational space for kriging to capture in its models, and hence prediction errors will be slightly higher.

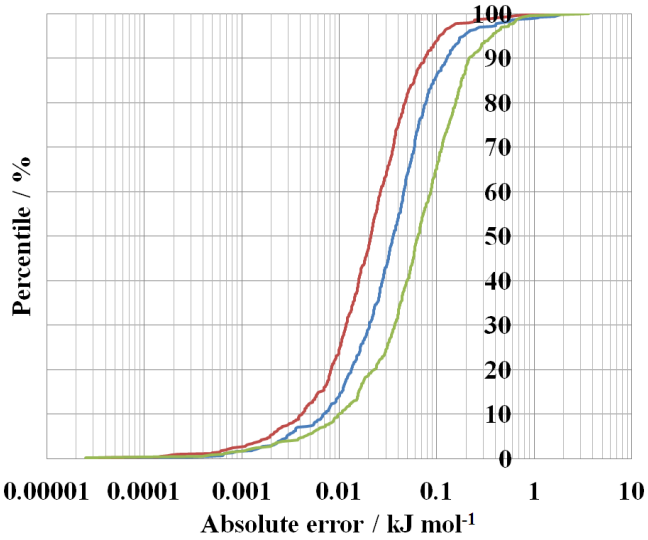


Figure 2.2: Comparison between the effect of the level of theory of the PES and the level of theory of the wave functions obtained to build kriging models for the ammonia dimer. *Blue*: training set geometries obtained from HF PES and training set wave functions obtained at HF; *Red*: training set geometries obtained from HF PES and training set wave functions obtained at B3LYP; *Green*: training set geometries from B3LYP PES and training set wave functions obtained at B3LYP.

2.2.3 Prediction of the Total Electrostatic Energy of the Hydrogen Bonded Complexes

Figure 2.3 shows the S-curves obtained, for all seven hydrogen bonded complexes of the S22 data set at all three levels of theory, and using 600 training examples. Looking at equation 1.49, we emphasize that the individual interaction errors (for each test geometry) are summed *before* the absolute value of this sum is taken. Hence, "cancellation of errors" is possible and indeed likely for each point on the S-curve. This cancellation is justified as the Coulomb law is itself additive. In other words, there is no summation of absolute values of atom-atom interactions when calculating a total electrostatic energy, but a summation of the actual values themselves (whether positive or negative). Analysis of the individual interactions is dealt with in **Section 2.1.5.**

Figure 2.3 shows that, for all three levels of theory used, the smaller systems lie furthest to the left, with a lower error, and the larger systems lie to the right. This is partially due to increased number of interactions present in the larger systems, and this is an almost linear relationship. Despite this increase in error with number of interactions, even the larger aromatic complexes are predicted within 1 kJ mol^{-1} for 70% of the test geometries, both at B3LYP and M06-2X level. For the ammonia dimer and the water dimer, almost 100% of test structures were predicted within 1 kJ mol^{-1}. None of the complexes have a single test geometry with an error greater than 9 kJ mol^{-1}. Almost all interactions are predicted within 1 kcal mol^{-1}, which is often referred to as "chemical accuracy".

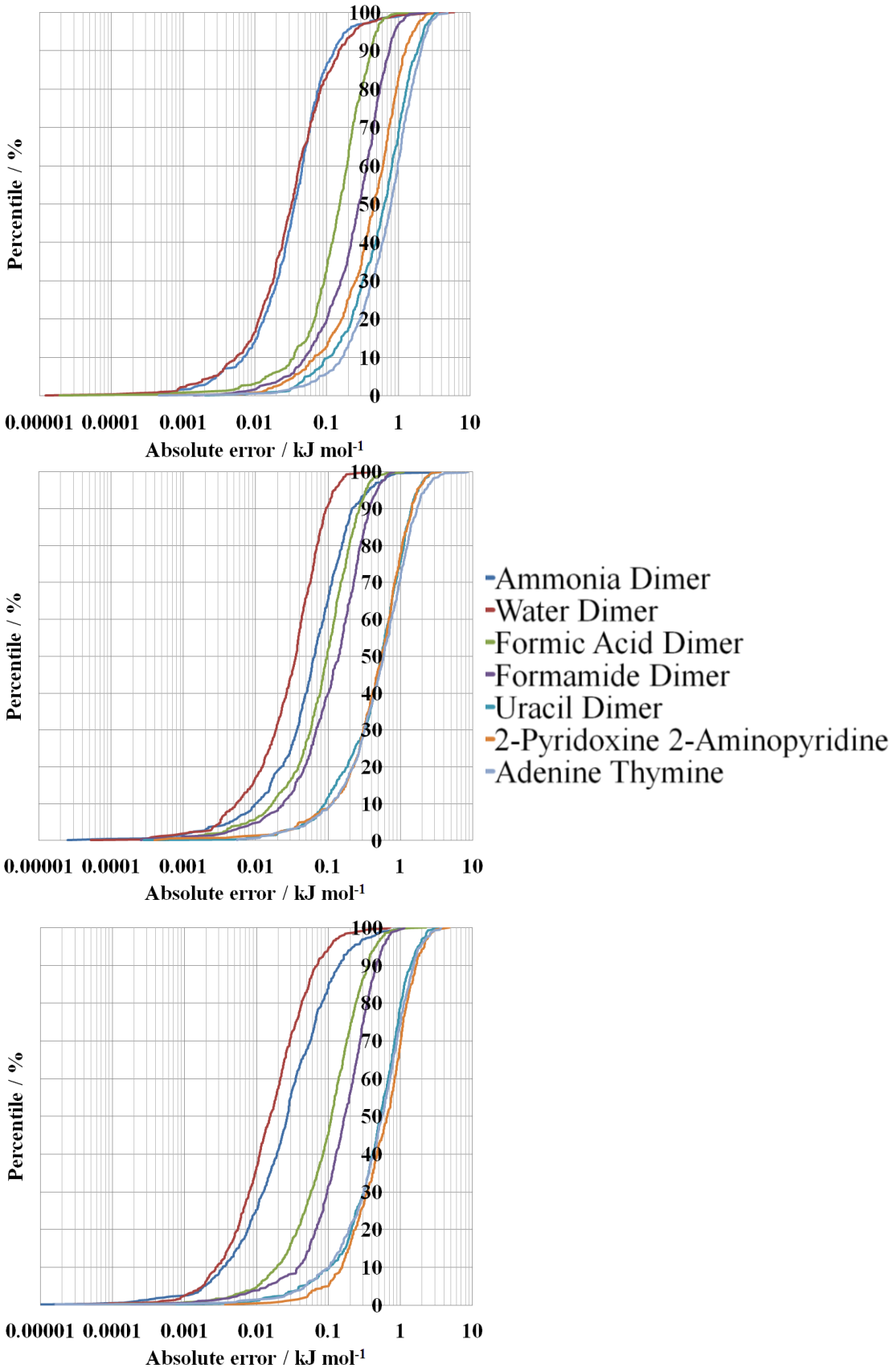


Figure 2.3: S-curves of the prediction error for the seven hydrogen bonded dimers in this work at the HF/6-31G** (*top left*), B3LYP/aug-cc-pVDZ (*top right*) and M06-2X/aug-cc-pVDZ (*bottom left*) levels of theory.

The errors for the Hartree-Fock complexes are on average higher than the error of the same complex at either B3LYP or M06-2X levels of theory, as expected. This is a consequence of the improved description of electron correlation as previously mentioned in **Section 2.1.1**. **Figure 2.4** shows the mean absolute prediction errors of the seven hydrogen bonded systems plotted against the number of intermolecular atomic interactions, for three levels of theory (wave functions and PES obtained at the same level). **Figure 2.4** demonstrates that neither of the two density functionals consistently outperforms the other. Plotting a trend line through the values of the average prediction error of each system against total number of interactions for B3LYP and M06-2X levels of theory yields overlapping lines (red and blue lines). Plotting a similar line for the HF level of theory (green line) shows that one can expect the average error to increase with a higher number of interactions at a faster rate. The R^2 value of 0.93 for the B3LYP data is higher than that of both HF and M06-2X ($R^2 = 0.88$ for both), suggesting that there is a stronger correlation between average error and number of interactions. However, due to the random sampling of the geometries this cannot be stated with certainty.

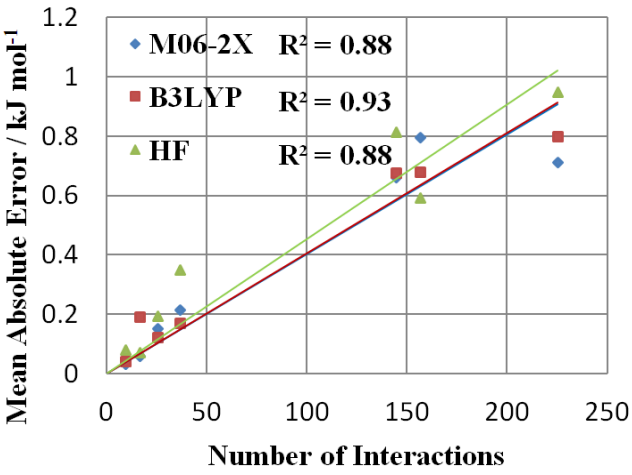


Figure 2.4: Mean absolute prediction errors of the seven hydrogen bonded systems plotted against the number of intermolecular atomic interactions.

2.2.4 Assessment of Individual Interactions Errors

For all seven systems, there is a general trend for the prediction errors for individual interactions to decrease with interaction distance. This is primarily due to the longer range interactions being smaller in magnitude, and hence any errors will be smaller in magnitude also. Because electrostatic interactions are dependent on $\frac{1}{r^L}$; the higher order interactions rapidly decrease in magnitude and hence the major interaction at longer distances is the monopole-monopole interaction. The decrease in interaction error with distance is good news when studying intramolecular interactions where only the 1,4 and higher interactions are taken into account as many such interactions are at a distance where individual interactions are typically predicted within ±1 kJ mol^{-1}. When studying intermolecular interactions (or indeed intramolecular interactions in which the molecule

folds back on itself) where interaction distances may be little of 1 Å it is still observed that the majority of interactions are predicted within ±2 kJ mol^{-1} of the true value.

Table 2.2 contains the average interaction error for each system at each level of theory, the standard deviation, and the average total error of all the points that make up the corresponding S-curve. It is seen that M06-2X performs best overall, with the lowest average interaction errors. The standard deviations for the M06-2X interaction errors is also the lowest, and this is can be observed by the smaller spread of interaction errors seen in F**igures 2.8-2.15**. Plots of the average interaction error of each system against the total S-curve energy, the number of atoms of each system against the average interaction error and the average interaction error of each system against the standard deviation of the interactions for each level of theory can be seen in **Figures 2.5, 2.6 and 2.7** respectively. **Figure 2.5** shows that as the S-curve moves to the right (increasing total error), the average interaction error increases also. However, most points in **Figure 2.5** lie below the line of y=x. indicating that the increase in the error of the individual interactions increases faster than the increase in the total error. Kriging models built at the Hartree-Fock level of theory are seen to have the highest interaction errors.

Figure 2.6 shows that the average interaction error is also seen to increase with the number of atoms in the system. Predictions by models built at the M06-2X level of theory appear to perform slightly better than B3LYP, with Hartree-Fock performing by far the worst, with the average interaction error for the larger systems significantly higher. **Figure 2.7** shows that as the average interaction error increases, the standard deviation, and hence the spread of the errors, increases. This is unsurprising.

Table 2.2: Average interaction errors, total errors, and standard deviations of the interaction errors for each of the seven hydrogen bonded dimers at all three levels of theory used. All errors in kJ mol^{-1}.

	No. Atoms	Int. error	Total Error	Standard Deviation
Ammonia Dimer	8			
HF		0.237	0.075	0.870
B3LYP		0.366	0.195	0.711
M06-2X		0.126	0.062	0.305
Water Dimer	6			
HF		0.186	0.084	0.534
B3LYP		0.181	0.045	0.274
M06-2X		0.060	0.037	0.138
Formic Acid Dimer	10			
HF		0.425	0.196	0.763
B3LYP		0.228	0.127	0.349
M06-2X		0.237	0.156	0.354
Formamide Dimer	12			
HF		0.529	0.354	1.096
B3LYP		0.316	0.175	0.531
M06-2X		0.327	0.219	0.543
Uracil Dimer	24			
HF		1.473	0.815	3.030
B3LYP		0.654	0.68	1.294
M06-2X		0.575	0.664	1.153
2-Pyridoxine 2-Aminopyridine	25			
HF		0.511	0.595	1.224
B3LYP		0.427	0.683	1.052
M06-2X		0.355	0.799	0.965
Adenine Thymine	30			
HF		1.321	0.951	4.224
B3LYP		0.883	0.802	2.013
M06-2X		0.912	0.715	2.031

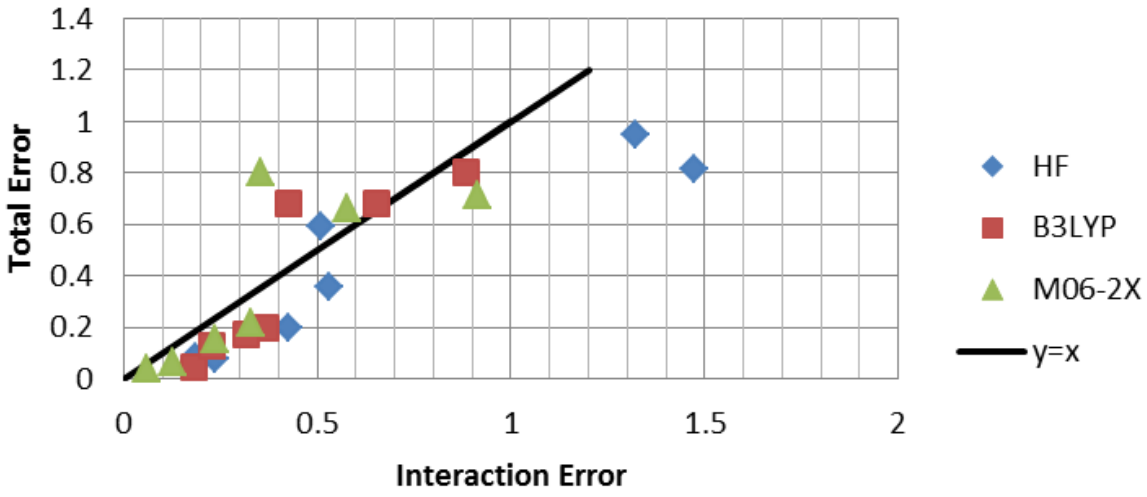


Figure 2.5: Average interaction error vs. average s-curve total error

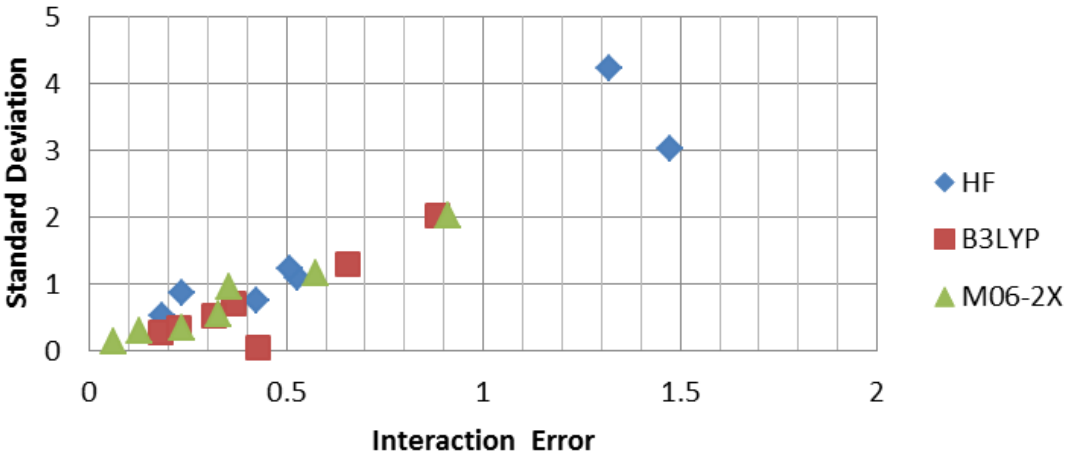


Figure 2.6: Average interaction error vs. standard deviation

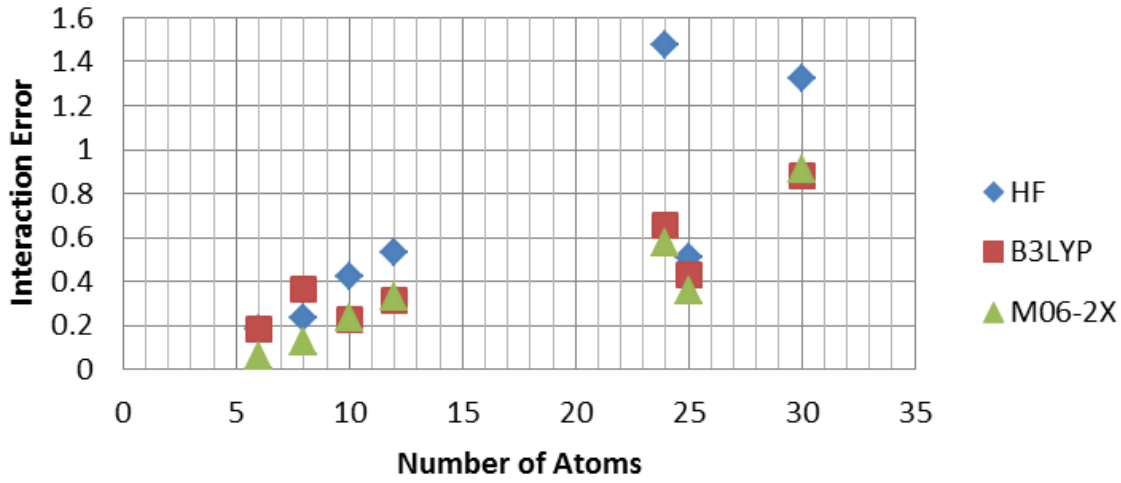


Figure 2.7: Number of atoms against average interaction error

It is observed in **Figure 2.8** that kriging models built at the M06-2X level of theory gave the most accurate predicted interaction energies for the ammonia dimer. The prediction errors for the individual interactions are observed to be clustered most tightly around 0 kJ mol^{-1} in the scatter plot. This is also seen in the histogram, where the peak around zero for M06-2X is highest and also narrowest. Hartree-Fock is seen in the scatter plot to produce the worst energy predictions out of any level of theory, with both highly negative and highly positive errors. The Hartree-Fock peak around zero is both higher and narrower than the B3LYP peak, indicating that despite producing the worst errors, it still predicts more interactions with a lower energy than B3LYP. The average interaction error for the B3LYP model of 0.366 kJ mol^{-1} is larger than the Hartree-Fock error of 0.237 kJ mol^{-1}. The similar standard deviations of the Hartree-Fock and B3LYP interactions reflects the fact that despite the histogram peak for Hartree-Fock being taller and narrower than the B3LYP peak, this is balanced by Hartree-Fock producing the worst individual interactions.

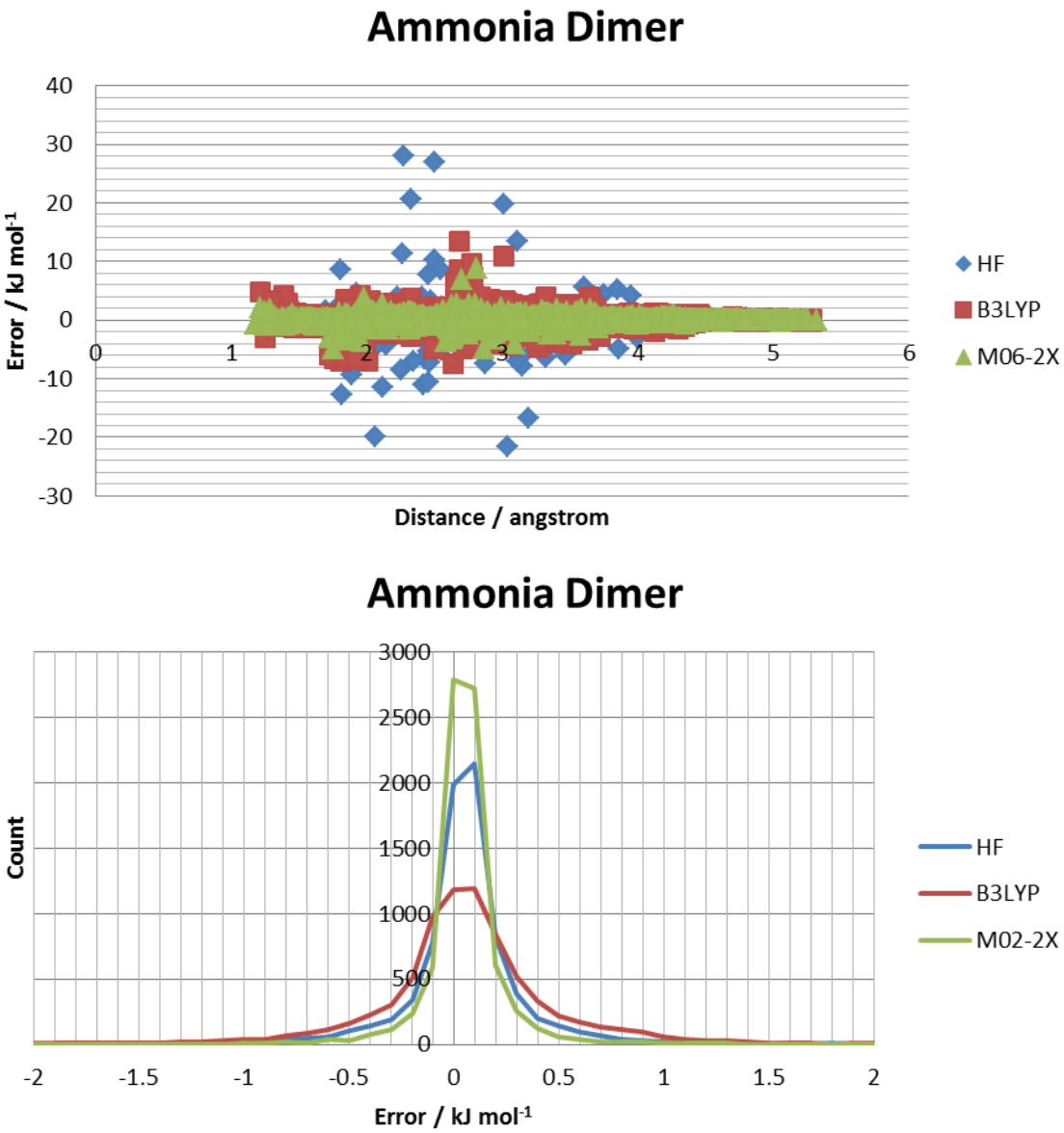


Figure 2.8: *Top* Scatter plot of the prediction errors for all interactions between atoms of the ammonia dimer against the interaction distance and *bottom* histogram depicting the number of interactions predicted at different errors.

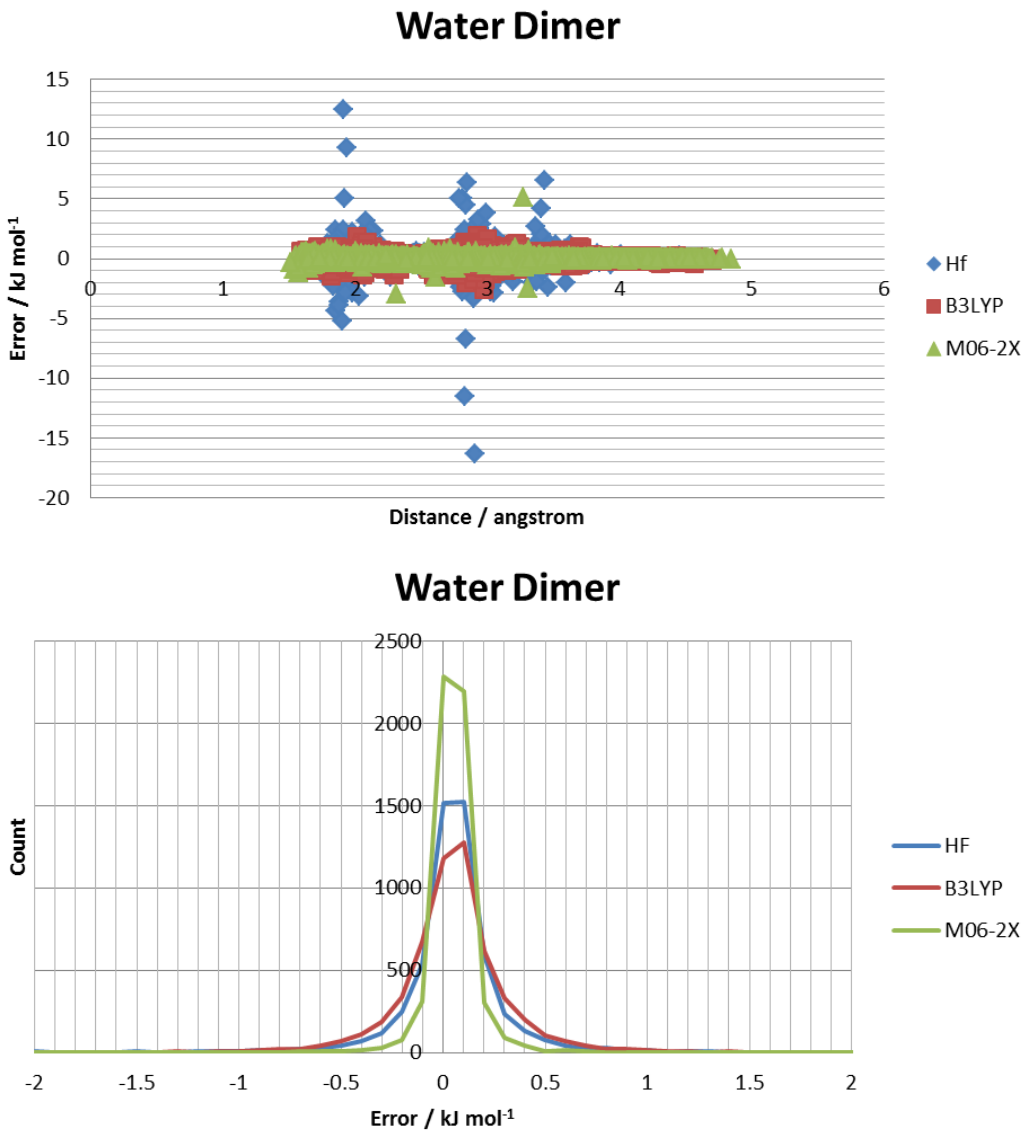


Figure 2.9: *Top* Scatter plot of the prediction errors for all interactions between atoms of the water dimer against the interaction distance and *bottom* histogram depicting the number of interactions predicted at different errors.

The same trends that were observed for the ammonia dimer are also observed for the water dimer (F**igure 2.9**). M06-2X performs best overall, with Hartree-Fock predicting the worst interactions out of any of the levels of theory. Unlike for the ammonia dimer, B3LYP has a lower average interaction energy than Hartree-Fock (0.181 and 0.186 kJ mol^{-1} respectively), however it can be seen from the histogram in **Figure 2.9** that B3LYP again has the lowest peak and the widest peak around 0 kJ mol^{-1}. Therefore B3LYP models are concluded to be more reliable than Hartree-Fock models because they do not predict large errors, however they do predict consistently less accurate interaction energies than M06-2X.

Kriging models for the ammonia dimer and the water dimer were both reconstructed using 1000 training examples rather than 600 at the B3LYP level of theory. The hypothesis is

that increasing the training set size should lower the error of the interactions. The results can be seen in **Figure 2.10**. Firstly, it is stated that increasing the training set size is computationally expensive due to the N^3 scaling of kriging, where N is the number of training examples. For both systems the mean unsigned error and standard deviation of the interaction prediction errors decreased (**Table 2.3**), and also the "worst offenders" are no longer present.

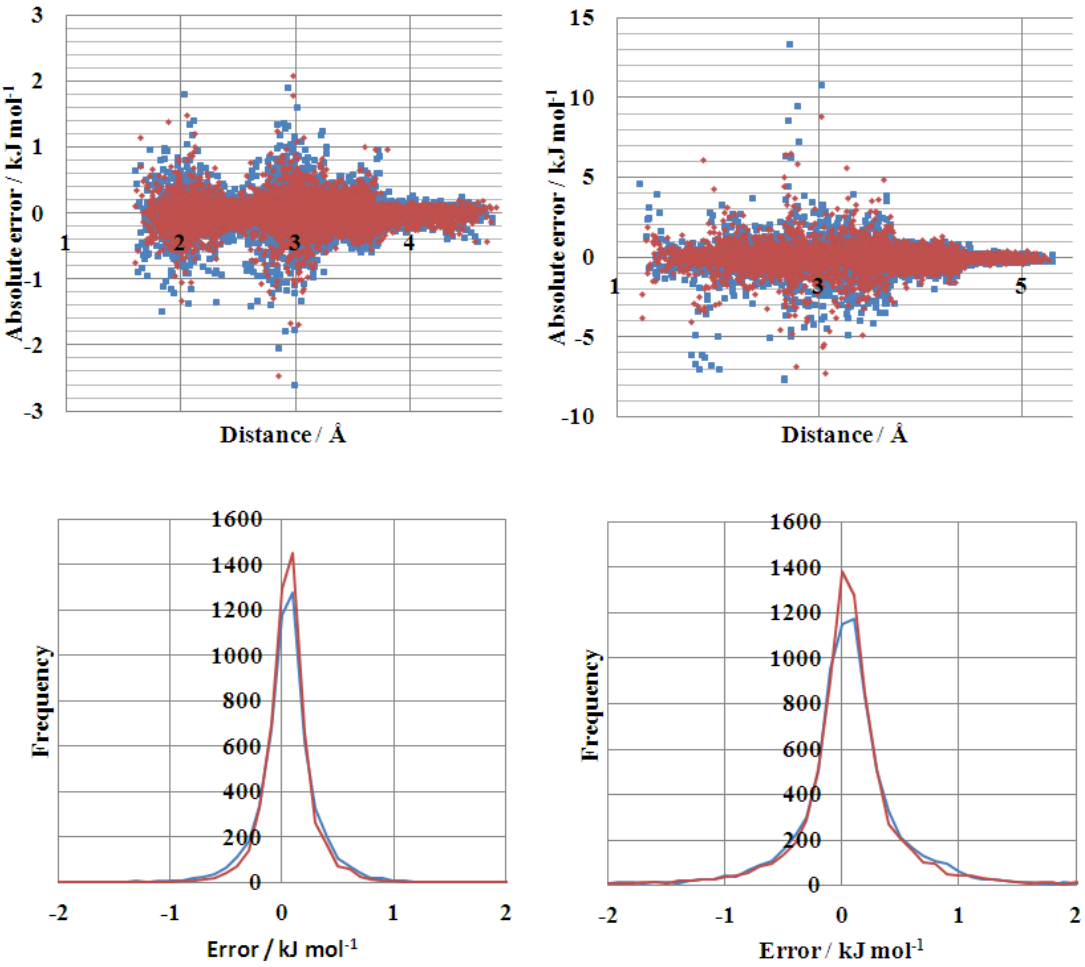


Figure 2.10: Scatter plots of and histograms for the individual interaction prediction errors for the B3LYP water dimer (*left*) and the B3LYP ammonia dimer (*right*) given by kriging models built with 600 examples (*blue*) and 1000 examples (*red*).

Table 2.3: Effect of increasing the training set size from 600 to 1000 examples for the ammonia dimer and the water dimer at the B3LYP level of theory.

System	Number of training examples	Standard deviation	MUE
Ammonia	600	0.71	0.37
Ammonia	1000	0.60	0.32
Water	600	0.27	0.18
Water	1000	0.23	0.15

The Hartree-Fock level of theory performs consistently worse than both M06-2X and B3LYP for the formic acid dimer (**Figure 2.10**). Again the worst predictions of any level of theory are made by the kriging models built at the Hartree-Fock level of theory. B3LYP outperforms the M06-2X level of theory in this instance, with both MUE and standard deviation lower for the interaction predictions at the B3LYP level than those from M06-2X. The scatter plot in **Figure 2.11** shows clearly four groups of interactions. The group around a distance of 6 Å corresponds to the interaction between the two C-H hydrogen atoms. This group is predicted most consistently well, which is to be expected because the interaction between the two hydrogen atoms is weak due to both the long range ($\frac{\mathbf{1}}{r^L}$ dependence) and also because H…H interactions are weak. Although this is to be expected, it is worth note as it shows that the kriging models have a level of stability- the rapidly increasing number of weak interactions within larger systems will not introduce large errors to the total energy of the system. It is also pleasing to observe that the interactions at short distances which correspond to the hydrogen bonds are predicted no worse than the mid-range interactions between, for example C…C.

Figures 2.12-2.15 show that the overall decrease in interaction error with interaction distance is also present for the formamide dimer, uracil dimer, the 2-pyridoxine 2-aminopyridine complex and the adenine thymine base pair. In all cases Hartree-Fock performs worse, with the largest spread of interaction errors, the smallest and widest peak around 0 kJ mol^{-1}, and the highest MUE. B3LYP and M06-2X perform similarly, with M06-2X performing overall best. For all levels of theory, the range of interaction energies increases with increasing system size, however the histograms show that despite larger errors becoming more common, the vast majority may still be found within ±2 kJ mol^{-1}.

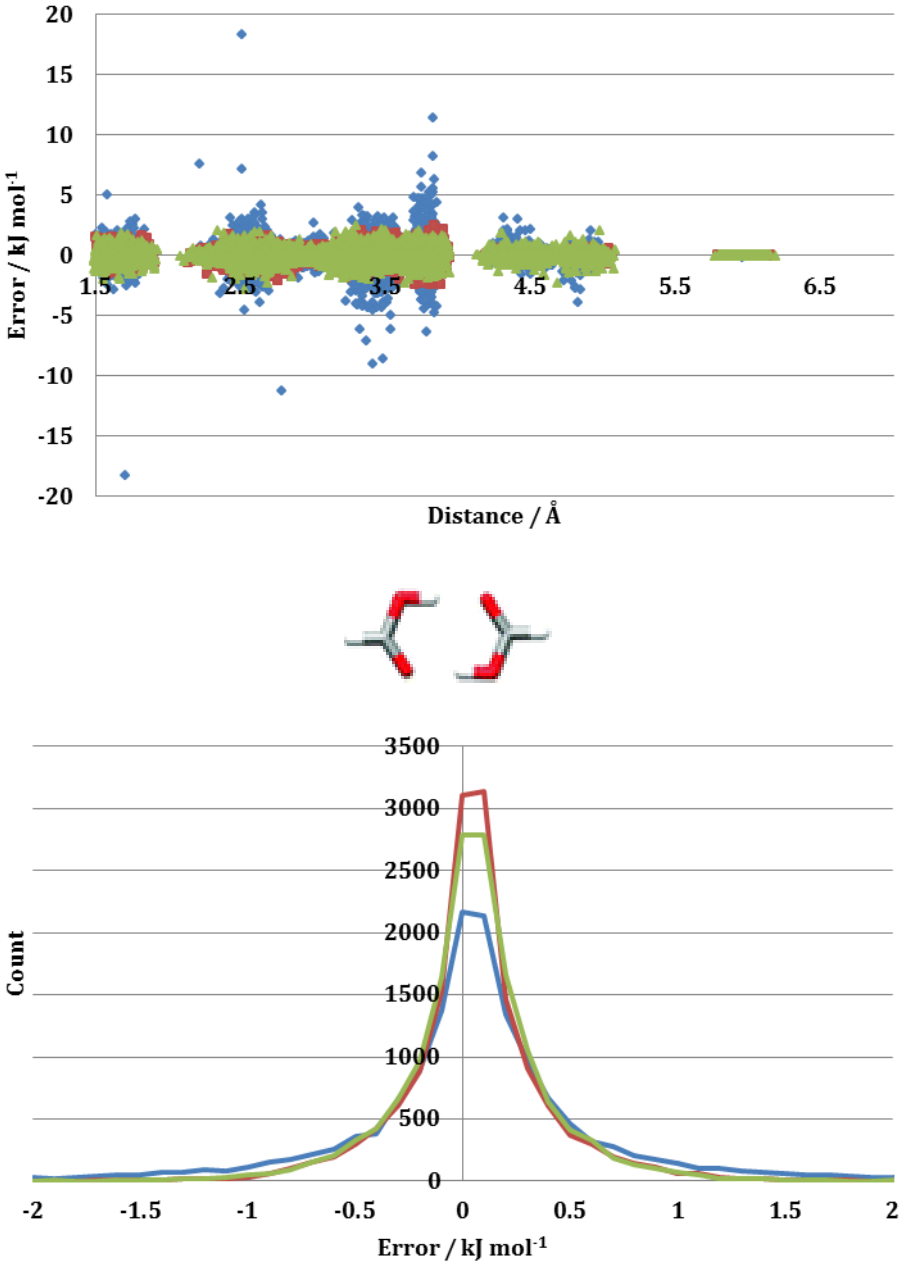


Figure 2.11: *Top* Scatter plot of the prediction errors for all interactions between atoms of the formic acid dimer against the interaction distance and *bottom* histogram depicting the number of interactions predicted at different errors. Blue: HF level of theory, Red: B3LYP level of theory, Green: M06-2X level of theory

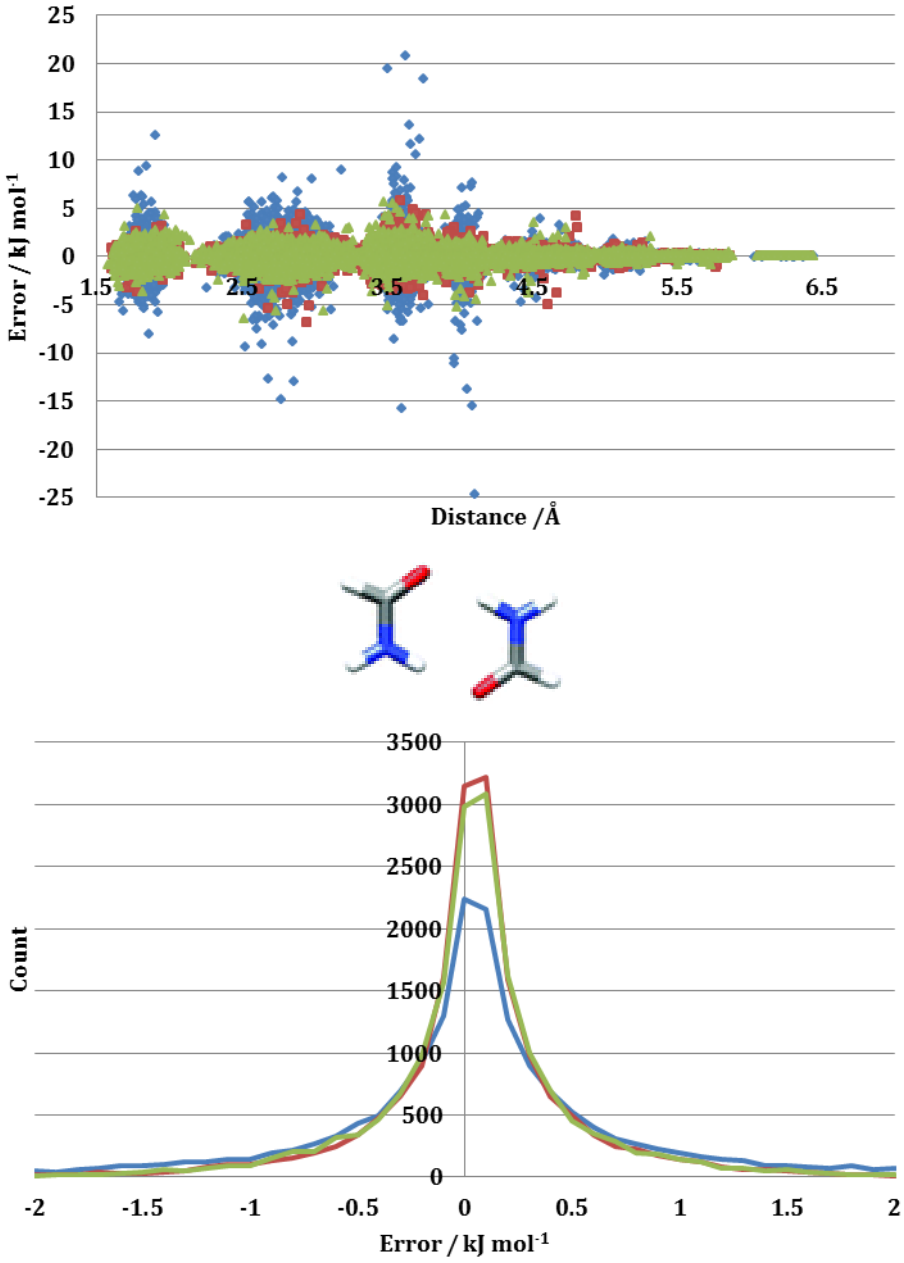


Figure 2.12: *Top* Scatter plot of the prediction errors for all interactions between atoms of the formamide dimer against the interaction distance and *bottom* histogram depicting the number of interactions predicted at different errors. Blue: HF level of theory, Red: B3LYP level of theory, Green: M06-2X level of theory

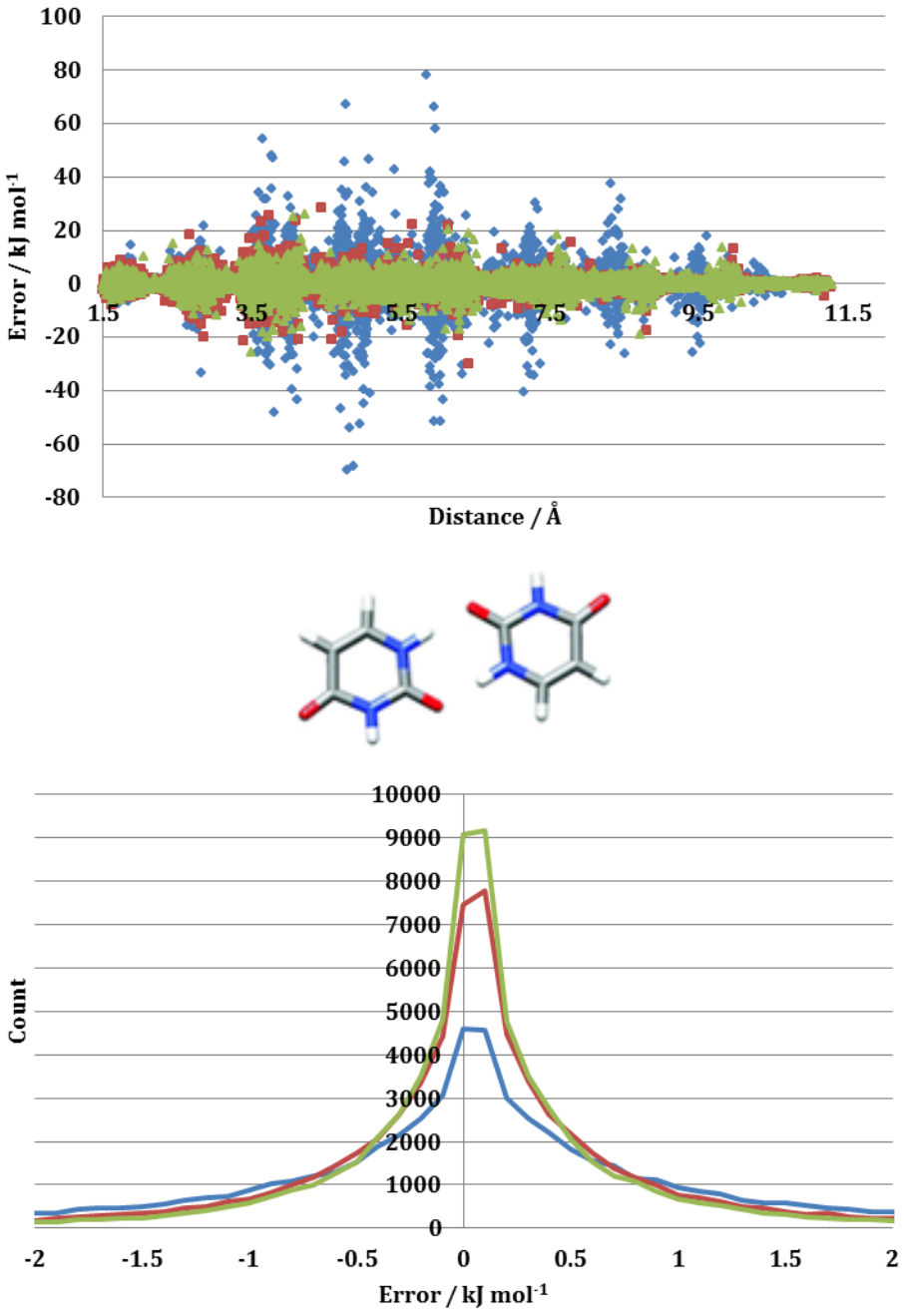


Figure 2.13: *Top* Scatter plot of the prediction errors for all interactions between atoms of the uracil dimer against the interaction distance and *bottom* histogram depicting the number of interactions predicted at different errors. Blue: HF level of theory, Red: B3LYP level of theory, Green: M06-2X level of theory

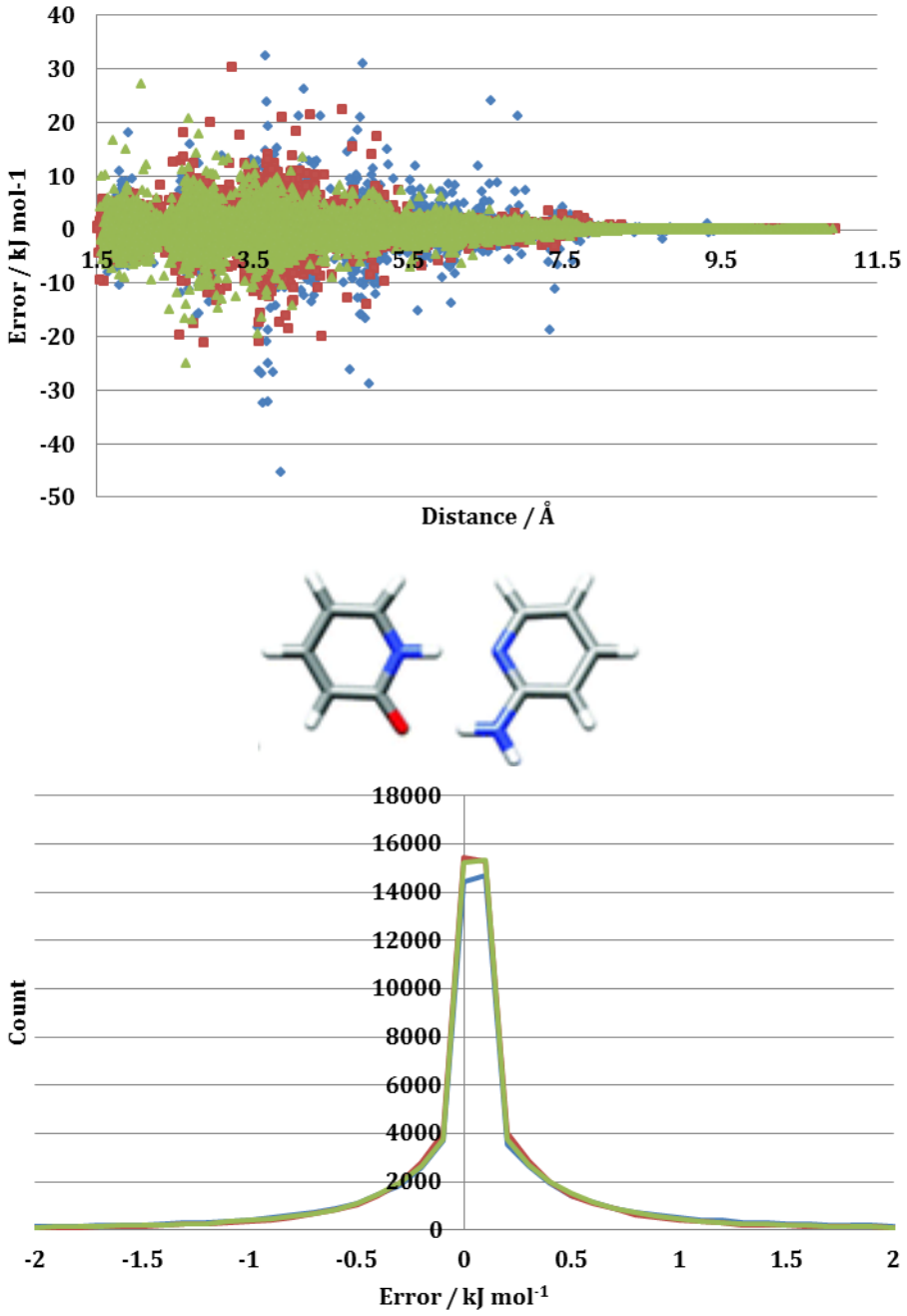


Figure 2.14: *Top* Scatter plot of the prediction errors for all interactions between atoms of the 2-pyridoxine 2-aminopyridine complex against the interaction distance and *bottom* histogram depicting the number of interactions predicted at different errors. Blue: HF level of theory, Red: B3LYP level of theory, Green: M06-2X level of theory

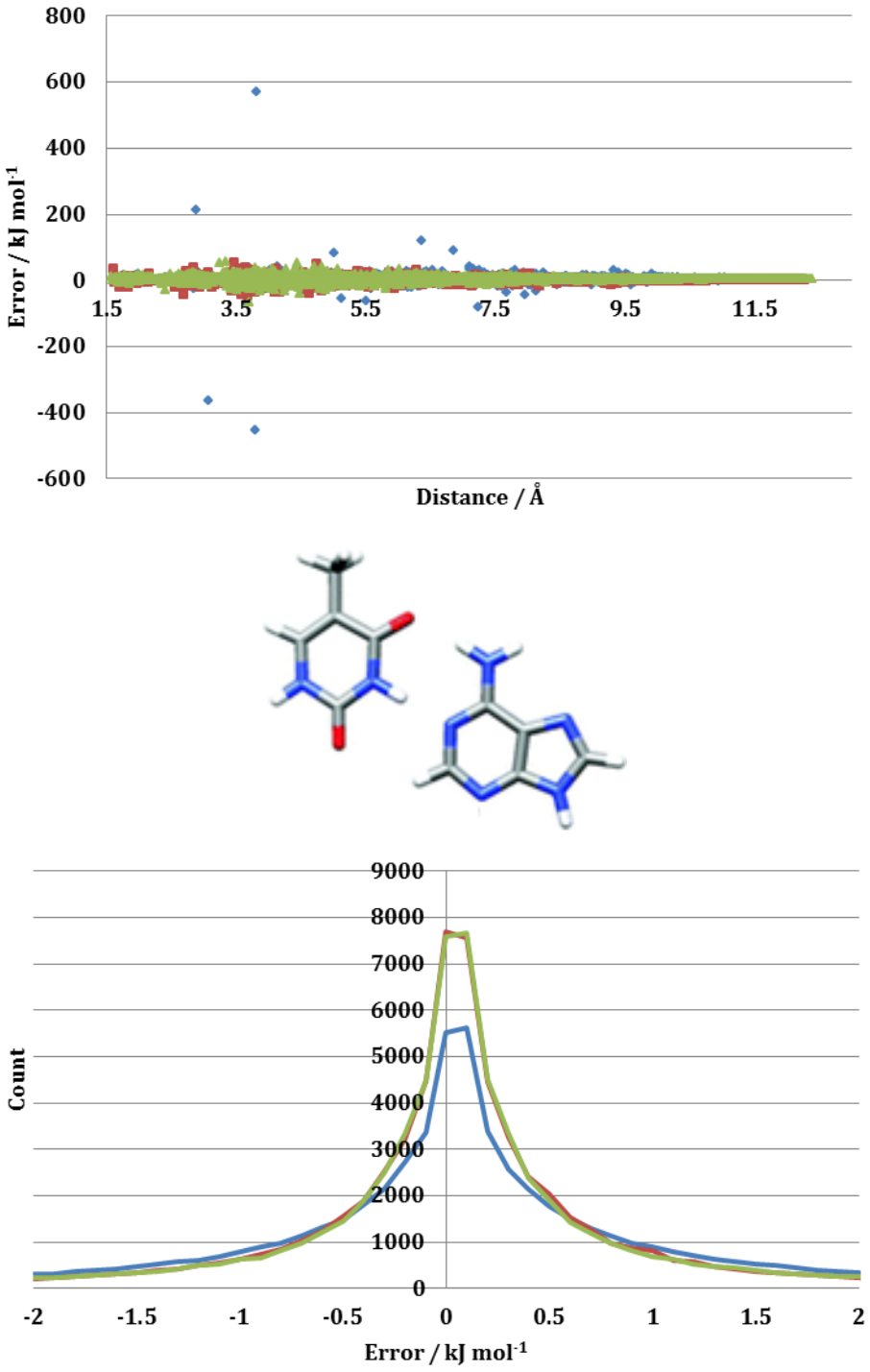


Figure 2.15: *Top* Scatter plot of the prediction errors for all interactions between atoms of the adenine-thymine complex against the interaction distance and *bottom* histogram depicting the number of interactions predicted at different errors. Blue: HF level of theory, Red: B3LYP level of theory, Green: M06-2X level of theory

2.3 Weakly Bound Complexes

The discussion now moves to three weakly bound complexes where hydrogen bonding is not the dominant interaction. The three systems studied are the ammonia-benzene complex, the water-benzene complex and the stacked benzene dimer. The ammonia-benzene and water-benzene complexes involve a weak hydrogen bond between the hydrogen atom of the donor ammonia or water molecule interacting with the delocalised π-system of the benzene ring. The benzene dimer involves a π-π stacking interaction. Topological pictures of the molecules can be seen in **Figure 2.16**. The electrostatics of the three systems discussed does not play as dominant a role as in H-bonded complexes. For this reason the choice was made to build kriging models for the IQA [25] self and interaction energies for the complexes instead (**see Chapter 1** for details of the IQA energy decomposition). It is noted that a rigorous, multipolar description of the electrostatic interaction is still important for a potential that aims to accurately model the energy profile of aromatic systems. For more details see **Chapter 1, Section 1.6**. The IQA self interaction energy, $V_{inter}^{a,a'}$refers to the total interaction energy that atom a has with all other atoms in the system. $V_{inter}^{a,a'}$ includes both Coulombic and non-classical components. It is the Coulombic component that has been expanded using spherical harmonics to give rise to the atomic multipole moments kriged for the hydrogen bonded systems. Thus, the treatment of the weakly bound complexes goes beyond that which was performed on the hydrogen bonded complexes.

In this work, dimers from the extended S22x5 dataset [149] were used in addition to the standard S22 dimer geometries as input for normal modes sampling. The former dataset includes the S22 molecules at 5 non-equilibrium geometries, where the molecules have been translated along the axis in the direction of the main intermolecular interaction. Further details of the sampling methods are provided below in **Section 2.3.2**.

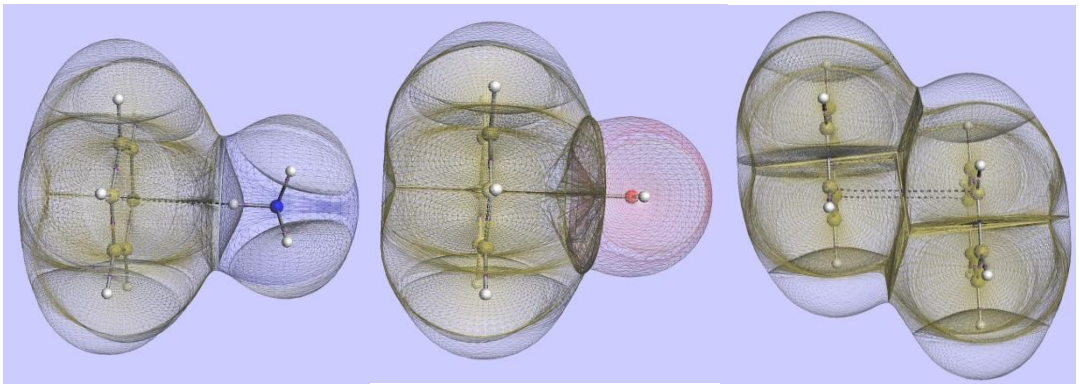


Figure 2.16: The three weakly bound complexes studied in this work: the ammonia benzene complex (*left*), the water benzene complex (*middle*) and the stacked benzene dimer (*right*)

Unlike the treatment of the H-bonded dimers, kriging models built for the non-H-bonded systems were only obtained at the M06-2X level of theory. This functional has been developed with the aim of improving the description of intermolecular energies, and has

been widely adopted by the DFT community due to its success [150-153]. As a consequence of the widespread use of M06-2X, an algorithm has been included in AIMAll version 14.11.23 [138] that allows the IQA decomposition to be performed on M06-2X wave functions. Both the HF and B3LYP levels of theory give poor interaction energies of weakly bound systems without the use of dispersion corrections [148]. Additionally, F**igure 2.4** in **Section 2.2.3** shows that training sets built using the M06-2X level of theory gave average errors lower than those obtained at HF and comparably to B3LYP. These reasons more than justify the exclusion of the HF and B3LYP levels of theory.

2.3.1 Computational Details

The GAIA protocol was followed in order to obtain the IQA kriging models of the three molecular complexes studied. As a consequence of using the S22x5 dataset, non-equilibrium geometries of the molecular complexes were present. This means that the standard normal modes sampling was not possible. Instead, Cardamone's non-equilibrium normal modes sampling algorithm (**Chapter 1 section 1.7**) implemented in TYCHE was used for the vibrational sampling of the hydrogen bonded dimers. All *ab initio* calculations were performed using the Gaussian09 software package at the M06-2X/aug-cc-pVDZ level of theory. The aug-cc-pVDZ basis set was chosen for its compromise between speed and accuracy. The IQA calculations were performed by AIMAll version 14.11.23 and the kriging models were built with the FEREBUS kriging engine. A training set size of 1000 was used for each of the three molecular complexes. Kriging IQA self and interaction energies means that NYX is not required as there is no need to interact atomic multipole moments. Instead, predictions made by FEREBUS were used to construct S-curves.

2.3.2 Sampling of the Molecular Complexes

Further details are now provided regarding the method by which the training set and test set geometries were obtained. For each molecular complex the S22x5 and S22 geometries were obtained directly from references [149] and [131], respectively, and then each of these 6 geometries had one molecule in each complex rotated by 90, 180 and 270 degrees to give a total of 24 molecular geometries (referred to from now as "*seeds*", or "*seed geometries*"). For example, in the case of the water-benzene complex, for each S22x5 and S22 geometry the water molecule was rotated around the axis defined as a line from the oxygen of water to the centre of the benzene ring. All 24 seed geometries were then input as minima for the non-equilibrium normal modes sampling routine within TYCHE. During the distortion, sampled geometries with angle bends and bond stretches in excess of 10% from the equilibrium distance were discarded. This ensured that the geometries were chemically reasonable. By including seed geometries from the S22x5 data set in addition to the geometries found in the S22 set, a greater sampling of conformational space is achieved. This gives rise to potentially more useful kriging models as they are able to predict energies for a greater number of systems- they describe a larger volume of configurational space. Although the example of weakly bound complexes is a relatively

trivial case, a sampling approach that covers a large area of conformational space is important when dealing with more complex systems such as amino acids. This is a topic of much discussion in **Chapter 3**. Wireframe images of 20 randomly selected sampled geometries for each of the three weakly bound complexes are shown in **Figure 2.17**.

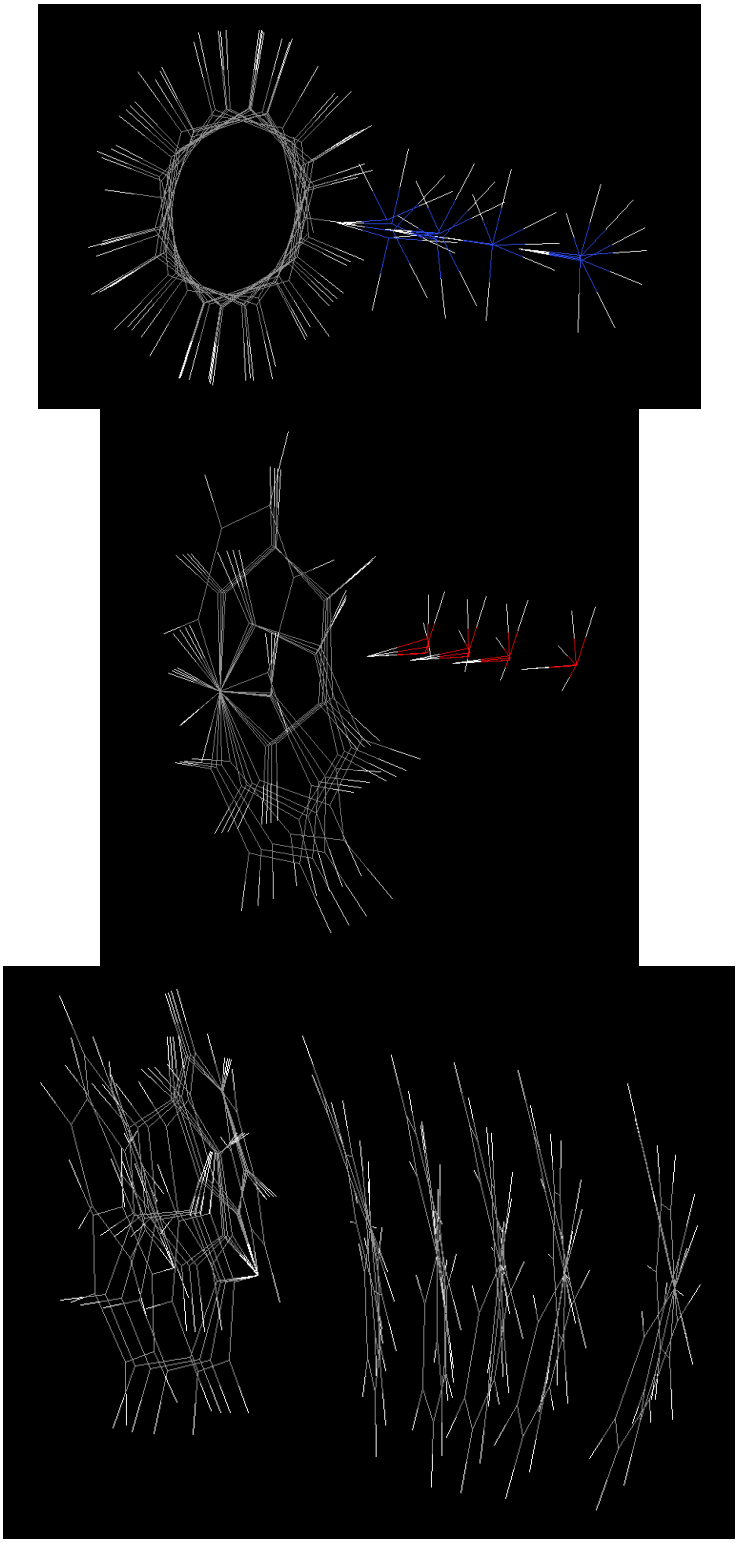


Figure 2.17: Wireframe images of 20 randomly selected geometries of the ammonia-benzene complex (*top*), water-benzene complex (*middle*) and bezene dimer (*bottom*).

2.3.3 Kriging Accuracy of Non-Hydrogen Bonded Complexes

The performance of the kriging models obtained from FEREBUS for the three complexes studied can be displayed using S-Curves, and these are provided in **figure 2.18**. The V_{self} and V_{inter} energies were predicted for 500 test geometries for both ammonia-benzene and the benzene dimer, and 400 test geometries for the water-benzene complex. The smaller test set for the water-benzene complex was required due to a greater number of geometries being filtered out due to high integration errors during the IQA analysis. Each point in the S-curve is equal to the sum of all atomic V_{inter} and V_{self} energy prediction errors. The exact formula is provided below:

$$Total\ IQA\ Energy\ Error = \sum_{a}^{N\ atoms} \left(\left(V^{a}_{self,true} - V^{a}_{self,predicted} \right) + \left(V^{a,a'}_{inter,true} - V^{a,a'}_{inter,predicted} \right) \right) \tag{2.1}$$

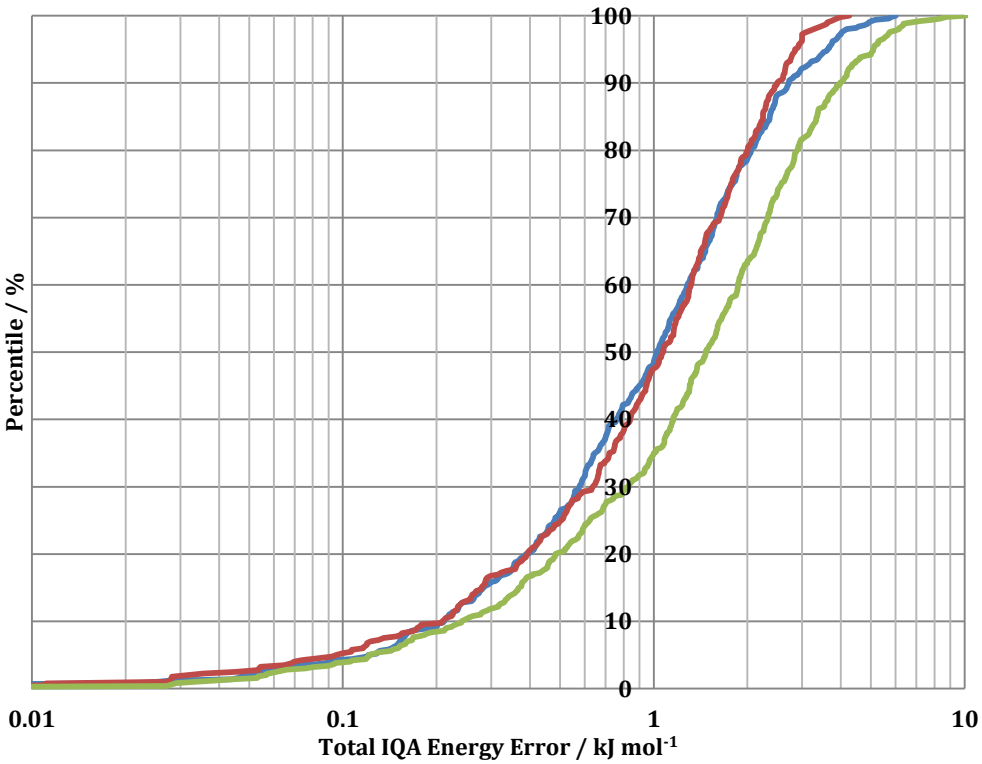


Figure 2.18: S-curve displaying the prediction error of the total IQA energy for the three weakly bound complexes: the ammonia-benzene complex (*blue*), the water-benzene complex (*red*) and the benzene dimer (*green*).

It is observed that the ammonia-benzene (blue line) and water-benzene (red line) complex kriging models perform comparably to one another and outperform the models obtained for the benzene dimer. In all instances, the total IQA energy is predicted within 10 kJ mol^{-1} accuracy. The ammonia-benzene and water-benzene complexes have 90% of the test structures predicted within 3 kJ mol^{-1}. **Table 2.4** contains the range in the total energy for each weakly bound complex as well as average prediction errors for the total energy.

Included is the average prediction error as a percentage of the range of the total IQA energy. The total IQA energy is predicted within 2% accuracy for all systems. The values in **table 2.4** show that as the range in the total energy increases, the average S-curve error also increases; however the increase in S-curve error is slower than that of the range, and therefore the average error is a smaller percentage of the range. This shows that the QCTFF protocol is capable of handling large ranges in molecular energies with only small cost to the accuracy of the kriging predictions.

The kriging performance of the separate V_{self}^{a} and $V_{inter}^{a,a'}$ energetic terms has also been analysed, where the two terms on the right hand side of **Equation 2.1** are each plotted as separate S-curves. Thus, each point on the V_{self}^{a} curve is given by:

$$Total\ Self\ Energy\ Error = \sum_{a}^{N\ atoms} \left(V_{self,true}^{a} - V_{self,predicted}^{a}\right) \tag{2.2}$$

and each point on the $V_{inter}^{a,a'}$ curve given by:

$$Total\ Interaction\ Energy\ Error = \sum_{a}^{N\ atoms} \left(V_{inter,true}^{a,a'} - V_{inter,predicted}^{a,a'}\right) \tag{2.3}$$

The two sets of S-curves can be seen in **Figure 2.19**. Both sets of curves perform similarly to the total energy S-curve with all kriging predictions within 10 kJ mol^{-1}. The average S-curve errors can be found in **Table 2.4** alongside the range in the self and interaction energies. The ranges in the two separate IQA energy terms are seen to be much larger than the range in the total IQA energy. For example, the ranges in the V_{self} and V_{inter} energies for the ammonia-benzene dimer are 235.5 kJ mol^{-1} and 244.4 kJ mol^{-1}, respectively, whereas the range in the total energy is only 69. kJ mol^{-1}. This is due to a cancellation between the energetic components. When the two molecules are close to one another, the self-energy is more positive than when they are at greater separation. This is because the atoms are deformed when brought close together. This always gives rise to a positive change in the self-energy, V_{self}. Conversely, the interaction energy, V_{inter}, is more negative the closer two molecules are because the intermolecular bonding is stronger. This effect has been previously documented by Pendas in his work on diatomics and hydrogen bonded dimers [39, 154]. **Table 2.4** shows that despite the large range in total V_{self} and V_{inter} values, the average S-curve error is relatively similar to the total IQA energy S-curve average error for all complexes. This means that the average S-curve error is much less than 1% of the range in the total self and interactions energies of all three weakly bound complexes.

Table 2.4: Summary of the kriging performance of the weakly bound complexes. Energies in kJ mol^{-1}.

	Ammonia-Benzene	Water-Benzene	Benzene Dimer
Total Energy			
Range	69.42	88.81	159.20
Standard deviation	8.58	12.44	17.43
Average S-curve error	1.30	1.22	1.86
Average Error as % of Range	1.87	1.38	1.17
Total Self-Energy			
Range	235.53	363.00	282.35
Standard deviation	35.92	39.82	48.88
Average S-curve error	1.42	1.57	2.16
Average Error as % of Range	0.60	0.43	0.77
Total Interaction Energy			
Range	244.36	387.35	250.28
Standard deviation	36.71	42.91	40.65
Average S-curve error	1.23	1.46	1.53
Average Error as % of Range	0.50	0.38	0.61

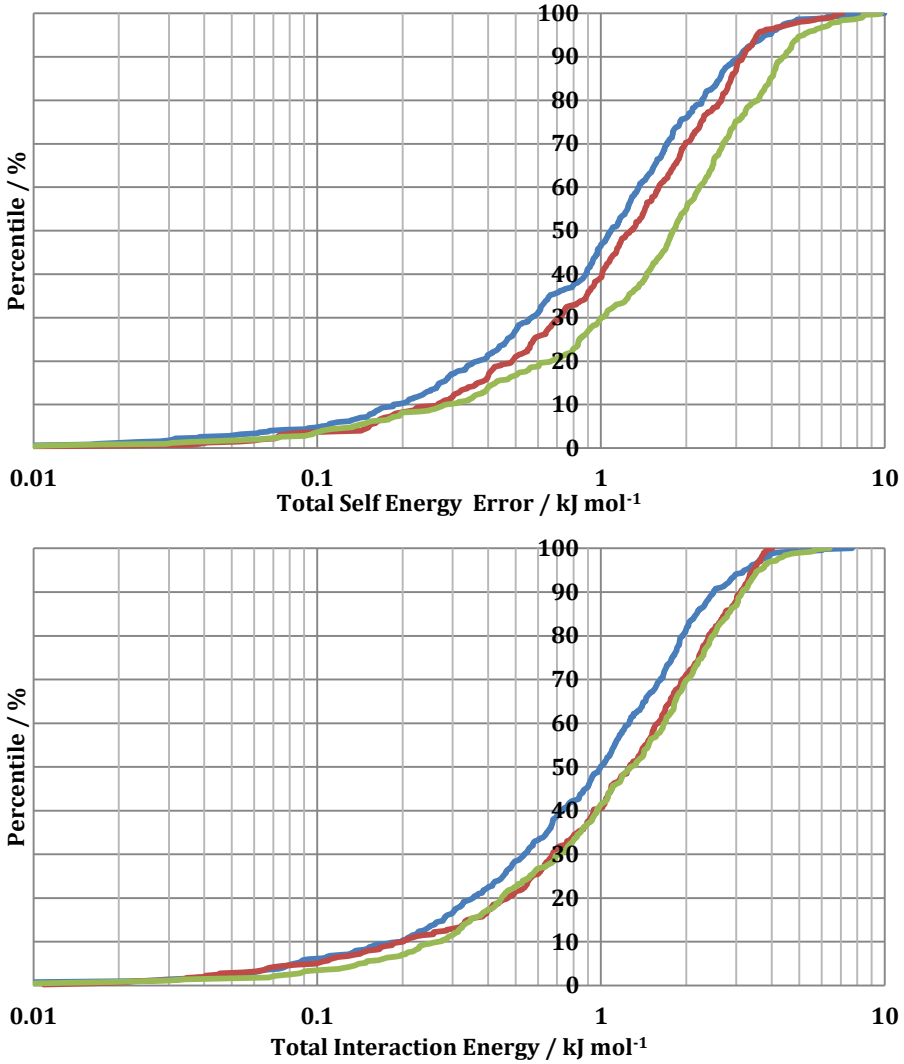


Figure 2.19: S-curve displaying the prediction error of the total self-energy (*top*) and total interaction energy (*bottom*) for the three weakly bound complexes: the ammonia-benzene complex (*blue*), the water-benzene complex (*red*) and the benzene dimer (*green*)

2.4. Conclusions and Further Work

The results of this chapter demonstrate that the high-rank multipole moments up to hexadecapole can be modelled by kriging as a function of nuclear coordinates to high accuracy for intermolecular hydrogen bonded systems. As these systems are ubiquitous within chemistry, the accurate modelling of intermolecular interactions is of great importance in the design of a next-generation force field such as QCTFF. Additionally, the work demonstrates that atomic energy components obtained from the IQA energy decomposition also may be described using kriging models. The models are built on *ab initio* values for the moments and IQA terms, and kriging allows for near-*ab initio* electrostatic interaction energies and atomic energies to be obtained in a fraction of the time. The models are able to model intermolecular interactions, including hydrogen bonding, mostly within ±2 kJ mol^{-1}, and the standard deviation and mean unsigned error of intermolecular interactions are shown to decrease with an increase in training set size. For the IQA kriging models, the predicted total energy of the test geometries for all three systems was within 10 kJ mol^{-1}.

The effect of the *ab initio* level of theory on the performance of the kriging was investigated for the hydrogen bonded dimers. In general, models built from moments obtained at the Hartree-Fock level of theory lead to larger errors in the prediction of electrostatic interactions than models built at B3LYP and M06-2X levels. There is no obvious difference between the accuracy of our results for the two density functionals, especially for larger systems.

The current work delivers proof-of-concept that machine learning can be used to accurately describe intermolecular interactions. This allows progress to be made on larger, more complex chemical systems. For example, knowledge that the hydrogen bond in the water dimer can be kriged to a high accuracy opens the door to working on larger water clusters as well as hydrated molecules. Recent work has been started by others in the group on such systems.

Chapter 3

PDB Sampling of Amino Acids

Summary

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. The traditional sampling method used within the group consists first of a search of the potential energy surface to find local minimum energy geometries. The minima are then used as "seeds" for a normal modes (NM) sampling approach where energy is pumped into the normal modes and snapshots of the vibrating molecule are taken. An alternative sampling of the "seed geometries" is presented in the current work, where single amino acid geometries were sampled 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. To address these issues, the hybrid PDB/non-stationary normal modes sampling approach was developed. I call this method "PDB/NM". This method is superior over standard normal modes sampling, which captures only geometries optimised from the stationary 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.

A couple of notes to the reader:

First, much of the work in this chapter may be found in the following publication:

"T.J. Hughes, S. Cardamone, P.L.A. Popelier, Journal of Computational Chemistry, ***2015****, 36, 1844-1857"*

found in ***Appendix F****. The work presented in this chapter contains only my own work, with all contributions from co-authors omitted.*

Second, the sampling of amino acids in the context of this work corresponds to the sampling of the amino acid with $H_3CC(O)$- *and* $-N(H)CH_3$ *caps to complete the peptide bonds. These*

structures are often referred to as "(amino acid) dipeptide" or "capped amino acids". Here they are simply referred to as "amino acids".

3.1. Introduction

To build a QCTFF kriging model, example molecular geometries must be obtained in order 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. In **Chapter 2**, the sampling approach used to obtain the S22 dimer molecules was a normal modes (NM) approach, as discussed in the description of the GAIA protocol in C**hapter 1**. In summary, this approach requires a small number of stationary points on the potential energy surface of a given molecule, and the normal modes at each stationary point (or local energy minimum) are calculated. Energy is then put randomly into the normal modes to distort the molecule, and "snapshots" are taken to obtain *distorted geometries*. The minimum energy conformations of all twenty naturally occurring amino acids have been reported in a comprehensive study[132], all obtained at the same level of theory. Kriging models built from normal modes 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[141, 155-160]. Recently, the kinetic energy of QCT atoms has been successfully incorporated into kriging models for methanol, NMA, glycine and triglycine[161]. The only other alternative sampling approach investigated draws snapshots from a molecular dynamics simulation, which has been done[162] 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 Φ and Ψ (**Figure 3.1**) describe the back bone and may be visualised using Ramachandran plots. These plots relate the values of Φ and Ψ 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 before[163, 164]. 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[165-170]. 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[171].

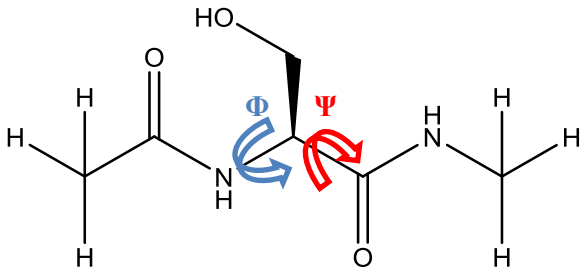


Figure 3.1: Definition of the Φ and Ψ dihedral angles of a peptide backbone.

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 geometries found in proteins. 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[172]. Ramos *et al.*[173] 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 intra-residue hydrogen bonding and steric considerations that have an effect on the amino acid conformation.

The work of Jha *et al.*[174] clearly shows the effect of the environment on the backbone angles Φ and Ψ. 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 geometries found in the β-sheet, PPII and α-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 geometries 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) [175, 176]. It has been shown, both experimentally (using NMR *J* couplings) and computationally, that disordered amino acid residues favour specific regions of the Ramachandran plot (typically β-sheet and PPII regions) in contrast to the conformational populations found in ordered protein secondary structures [177] [174, 178-180]. 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[181], and this relationship between environment and structure has been used successfully in rotamer libraries to predict side chain conformations[182]. 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 geometries.

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 geometries is impractical and inefficient. A number of efficient approaches have been presented in the literature including the use of molecular dynamics[183, 184], Monte Carlo[185], transition path sampling[186-188] and metadynamics[189]. Additionally, fragment based approaches may be used in order to improve a systematic approach by reducing the number of conformations searched though elimination processes. An example of such an approach is that of Luo *et al.*[190] 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. Firstly, only heavy atoms are detectable by X-ray crystallography and so the hydrogen atom coordinates are dependent upon the refinement process used. Secondly, 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-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.

Despite some challenges associated with the direct sampling of amino acids from protein crystal structures, the protein data bank (PDB sampling) remains an highly desirable source of amino acids for QCTFF development. **Sections 3.3.1-3.3.3** of this chapter are a study of the potential advantages that PDB sampling of amino acids has over normal modes sampling, and then in **Section 3.3.4** the chapter introduces a novel sampling approach that has been designed with the intention of overcoming the previously mentioned problems associated with direct PDB sampling. This methodology is named PDB/NM and a detailed technical description of this method is given in **Section 3.3.**

3.2 Computational Methods

Amino acids were sampled from protein crystal structures using the MOROS code (see **Chapter 1** and **Appendix D** for detailed discussion). The normal modes sampling was performed using TYCHE with a maximum bond stretch and angle bend of +/-10% of the equilibrium value. All *ab initio* calculations were performed by Gaussian09 at the B3LYP/aug-cc-pVDZ level of theory. Kriging models were built by following the GAIA protocol outlined in **Chapter 1**. A list of the protein crystal structures is provided in **Appendix C.**

3.3 Results

3.3.1 Normal Modes Sampling vs. PDB Sampling

As stated above, normal modes sampling require the generation of input "seed" geometries that are local minima on the potential energy surface of a molecule. The sampled geometries are all distortions of the seed structures, and the program TYCHE is very successful at sampling around the input minima. Clearly it is important that the collection of input seed geometries allows the sampling of all chemically important molecular geometries. Consider a hexane molecule, C_6H_{14}. The dihedral angle around each of the six C-C bonds has three minimum energy positions, $+60^o$, -60^o and 180^o. This means that there are 3^6=729 valid combinations of dihedral angles available that will give rise to a local minimum energy structure. Some of these geometries will have very high energy steric clashes and so could be omitted due to being inaccessible in standard conditions. Even if the possible number of local minimum geometries was halved due to these clashes, there would still be more than 350 possible local minima available to the hexane molecule. If one was to try to build a kriging model for hexane that comprehensively was able to describe all regions of conformational space using a normal modes approach, this means that roughly 350 minima would be needed.

Obviously, an exhaustive collection of minima is impractical as the molecules studied become larger and more complex. The minimum number of internal coordinates that describe a molecule is 3N-6 where N is the number of atoms. Assuming that each coordinate has three local minimum energy states, there are 3^{3N-6} potential input minima for each molecule. To limit the number of minima obtained for the 20 amino acids, a previous member of the Popelier group enforced a root mean squares approach (rms) to describe how similar two minima i and j are from one another. Sampled molecules that had an rms_{ij} value greater than a threshold of 40^o were determined as unique, whereas structures below that threshold were discarded. This criterion produced a manageable number of minima for each amino acid, however 40^o is a very high threshold. The result of the high threshold value is minima that are well-spaced throughout conformational space, however it means that some experimentally important regions of conformational space and certain combinations of internal coordinate "states" are omitted. It is for this reason

that alternative methods for sampling structures are of interest, for example the PDB sampling of amino acids presented here.

Figure 3.2a shows the obtained sampling of the dihedral angle around the C_α-C_β bond of glutamic acid (termed dihedral 1) via a normal modes approach using the minima obtained using the rms approach given above as seeds. It is observed that there are no structures sampled at a dihedral angle of 60° where one may expect a local minimum to exist (and therefore geometries to be present). **Figure 3.2b** shows the same data, only for geometries obtained from a collection of protein crystal structures (PDB sampling). The sampling from the PDB method sows that the population of dihedral 1 values sampled is highest around 60°. This is an example of the normal modes sampling approach failing to sample important experimentally observed molecular geometries due to a seed not being present with corresponding set of internal coordinates. In this example, one would be able to add an additional input seed geometry with a dihedral 1 value of 60° to correct this problem, but with larger more complex systems, an increasingly large number of additional seeds will require being added, and this is unfeasible in the long run.

To expand upon the previous point, now consider the next dihedral angle along the glutamic acid side chain around the C_β-C_γ bond (dihedral 2). **Figures 3.3a and 3.3b** show the values of dihedral 1 and 2 for glutamic acid structures sampled from normal modes and the PDB, respectively. One would expect a cluster of sampled geometries around all the combinations of the -60°, 60° and 180° minima for both dihedrals, but normal modes sampling is unable to sample geometries with a dihedral 1 value of 60°. PDB sampling, however, samples all the expected combinations of the local minima around each dihedral.

Figures 3.2 and 3.3 also provide an example another advantage that PDB sampling has over normal modes sampling, which will be referred to henceforth as "correlated dihedrals". This term relates to the sampled combinations of dihedral angles. Using a restricted set of seed geometries, as is the case when using the normal modes sampling approach, a complete spread of values for a given dihedral angle may be present, but important combinations of two dihedral angles may not be. In **Figure 3.2a** it appears that all values of dihedral 1 are sampled, with the exception of the previously discussed 60° local minimum. When the dihedral 1 values are plotted against the dihedral 2 values (**Figure 3.3a**), however, it is observed that there are no glutamic acid geometries sampled with dihedral 1 values of 180° when dihedral 2 has a value of 60°. This means that there is no seed geometry with this particular combination of dihedral angles. A more exhaustive discussion of the "correlated dihedrals" is provided in the results section.

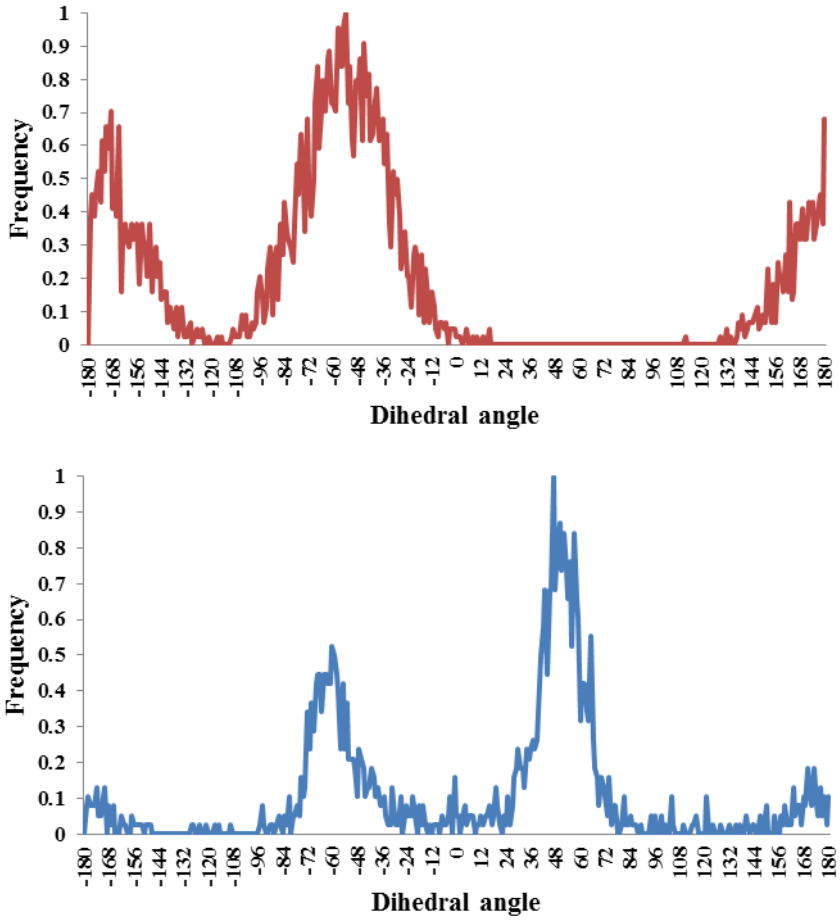


Figure 3.2a: *(top)* The frequency of occurrence of different side chain dihedral angles of glutamic acid molecules sampled using normal modes. **Figure 3.2b:** *(bottom)* The same as **3.2a** but for geometries sampled from the PDB

When sampling geometries from which kriging models will be built, it is important to consider the application that the kriging models will be applied to. For amino acids one would expect the kriging model to be applied to the simulation of a protein. Therefore, the amino acid geometries used to build the kriging models should cover regions of conformational space that one may expect the amino acid within a protein may exist during the simulation. PDB sampling satisfies this condition, as the amino acids are sampled direct from protein structures. It also allows one to be confident that the dihedrals that are not well sampled by the approach are unimportant because they are not present in the experimental crystal structures. The energetic minima used as seeds for the traditional normal modes approach obtained from gas phase calculations, and so there is no way to check the relevance of the obtained minima to the conformations of the amino acids in a proteins. Building training sets using geometries sampled from the PDB should also result in regions of conformational space that are common in nature, and therefore of interest in biomolecular simulation, being well represented within the training set. Equally, regions that are of less biological interest will be more poorly sampled. This will result in kriging models that are able to well describe regions of conformational space that are biochemically relevant.

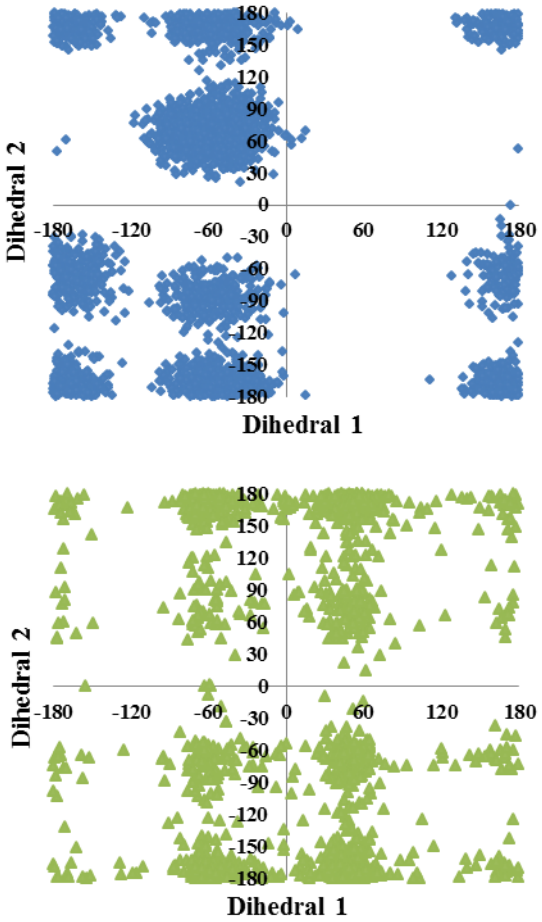
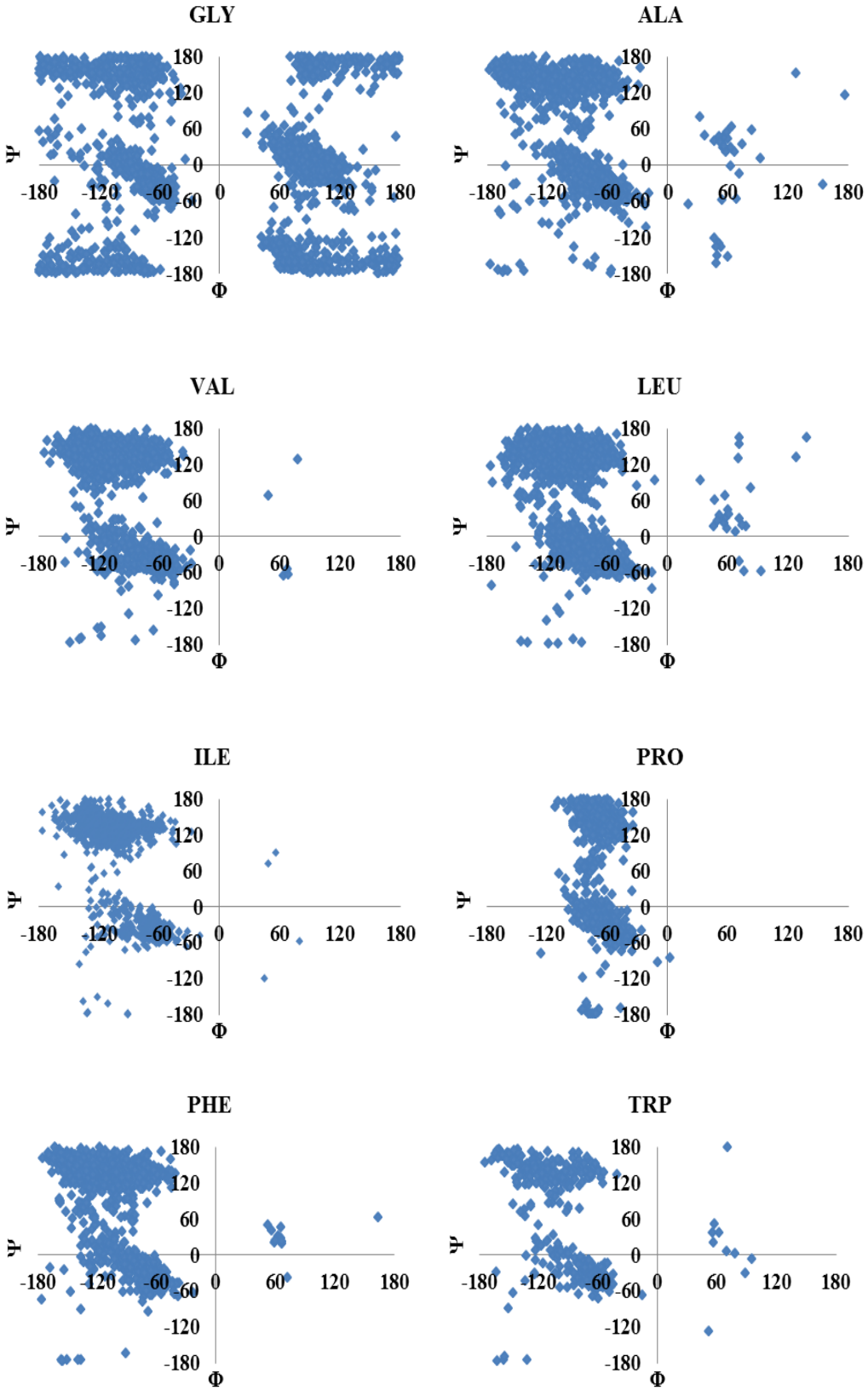
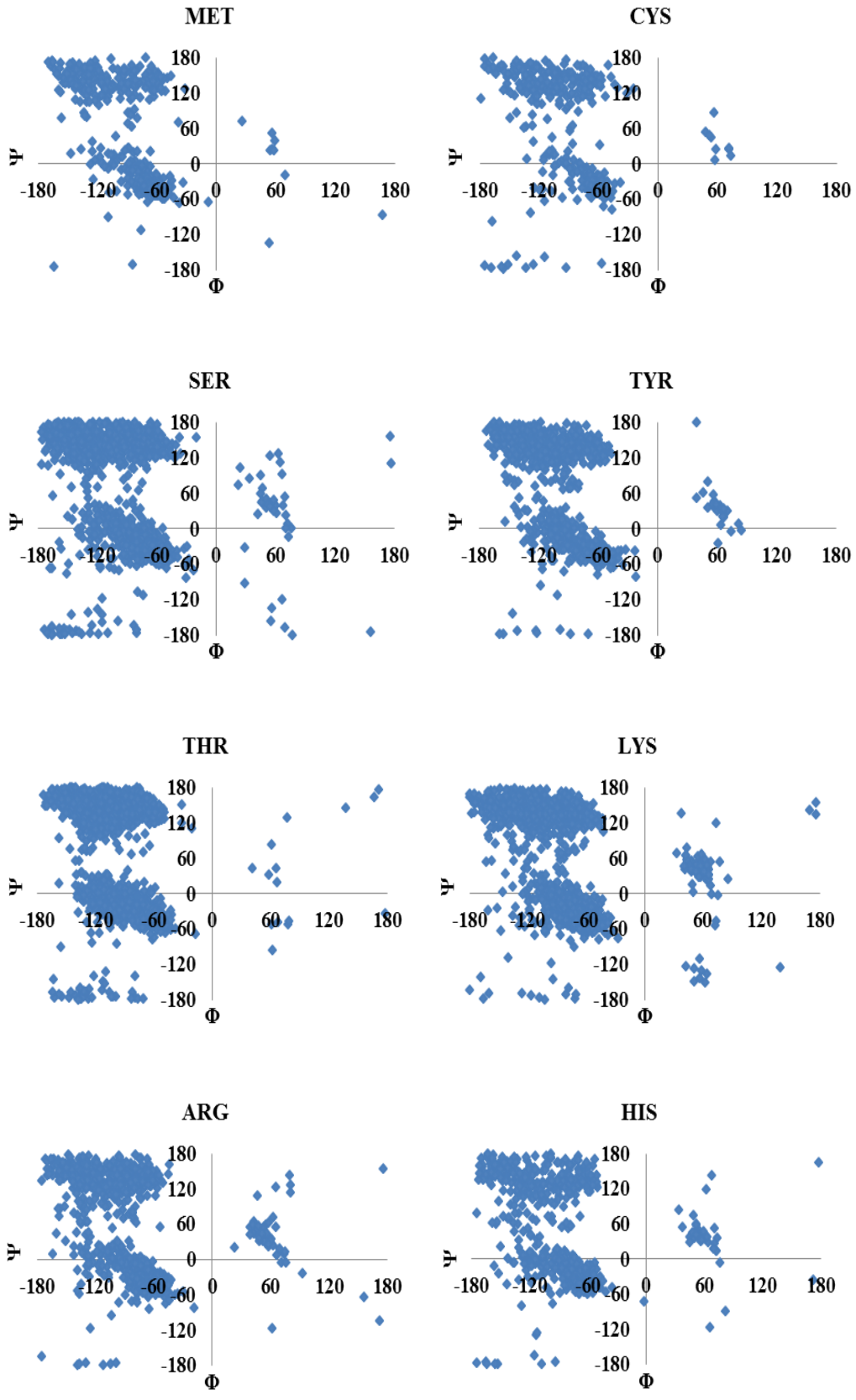


Figure 3.3a: *(Left)* plot of the two side chain dihedral angles of glutamic acid molecules sampled using normal modes. **Figure 3.3b:** *(Right)* the same but for geometries sampled from the PDB

3.3.2. Ramachandran Plots of the 20 Naturally Occurring Amino Acids

As stated earlier, sampling from the PDB should produce training set geometries that mimic the conformational preferences of peptides in nature. The Ramachandran plots for the 20 naturally occurring amino acids are well understood. In such a plot, the backbone ϕ and ψ angles (**Figure 3.1**) are plotted, and regions on the plot correspond to the different secondary structural motifs that are adopted by polypeptides. MOROS was used to sample amino acids from 80 protein crystal structures (listed in **Appendix C**). From the sampled geometries, Ramachandran plots have been drawn for all 20 naturally occurring amino acids. These are displayed in **Figure 3.4**.


GLY
ALA
VAL
LEU
ILE
PRO
PHE
TRP
Ψ
Φ
180
120
60
0
-60
-120
-180
-180
-120
-60
0
60
120
180


MET
CYS
SER
TYR
THR
LYS
ARG
HIS
180
120
60
0
-60
-120
-180
-180
-120
-60
0
60
120
180
Ψ
Φ

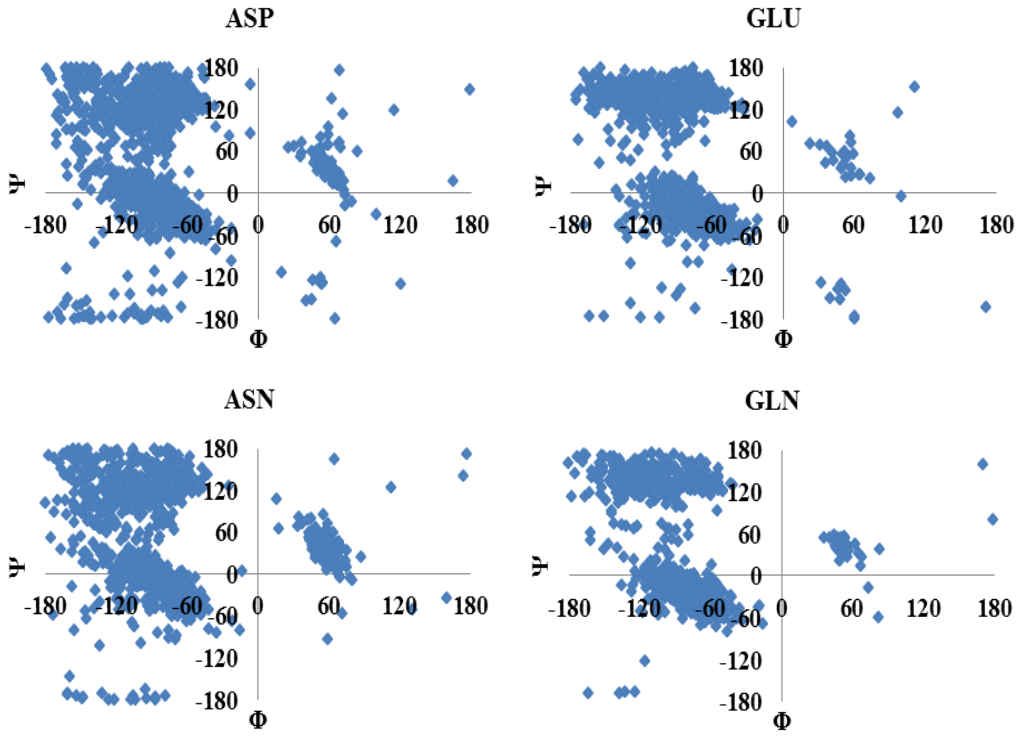


Figure 3.4: Ramachandran plots of the 20 naturally occurring amino acids. The geometries are sampled from the PDB.

The plots in **figure 3.4** illustrate the previously discussed point that PDB sampled geometries are found predominantly in the areas of conformational space associated with the secondary structural motifs of polypeptides. Negative Φ angles are typically well described, and in many cases (for example ASN, GLN, LEU and ASP) the two islands corresponding to α-helicies and β-sheets are bridged by a large number of intermediate geometries. One of the potential strengths of sampling geometries from the PDB is that the regions of conformational space between clusters should be better described than if normal modes were used as a sampling method. The Ramachandran plot of Gly shows a much greater range of Φ and Ψ angles than the other amino acids. This is due to the role that Gly plays within protein secondary structure. Due to the lack of side chain on Gly, it is significantly more flexible than other amino acids and therefore exists largely in loops and turns in the protein. All plots in **Figure 3.4** display a strong agreement with Ramachandran plots presented in work by Beck *et al.* where a much larger pool of crystal structures were sampled from[163].

The plots in **Figure 3.4** do not all contain the same number of points. Being sampled from a set number (260) of protein structures, more geometries were sampled for the more common amino acids such as Asp, Ala, Lys, and Leu, whereas the less frequently occurring such as Cys and Thr are sampled less frequently. It was discussed in section 3.2.3 that although PDB sampled geometries will show bias towards the regions of the Ramachandran corresponding to the α-helix and β-sheet secondary structures, as the number of sampled geometries increases, the other regions of the Ramachandran will

become more populated. This is seen in **Figure 3.4**, as highly sampled amino acids, such as Ala show a greater coverage of the Ramachandran plot than the less sampled residues, such as Cys.

To illustrate the above point further, geometries of Asn were obtained from an expanded pool of proteins (260 total crystal structures with codes provided in **Appendix C**), and the increased spread of geometries can be seen in **Figure 3.5a**. The expanded pool of proteins contained a total of 7476 Asn residues, while the original set of proteins contained 1183 residues. The right hand side of the Ramachandran plot that was previously poorly covered is now significantly more populated, particularly around the region corresponding to the left handed α-helical secondary motif. The left hand side of the Ramachandran is where the majority of the ~6000 new geometries lie (as can be observed in **Figure 3.6,** the shape of the Ramachandran does not change), giving a very complete description of regions of conformational space where Φ is negative. Obviously, if sampling by normal modes, increasing the number of sampled geometries will not add geometries to empty regions of conformational space as there are still no minima in that region to sample. **Figure 3.5b** shows a plot of the side chain dihedral against the back bone Φ angle for both the 1183 and the 7476 geometries. It is seen that an increase in the number of geometries sampled from the PDB results in the full range of possible side chain dihedral angles described for all sampled values of Φ.

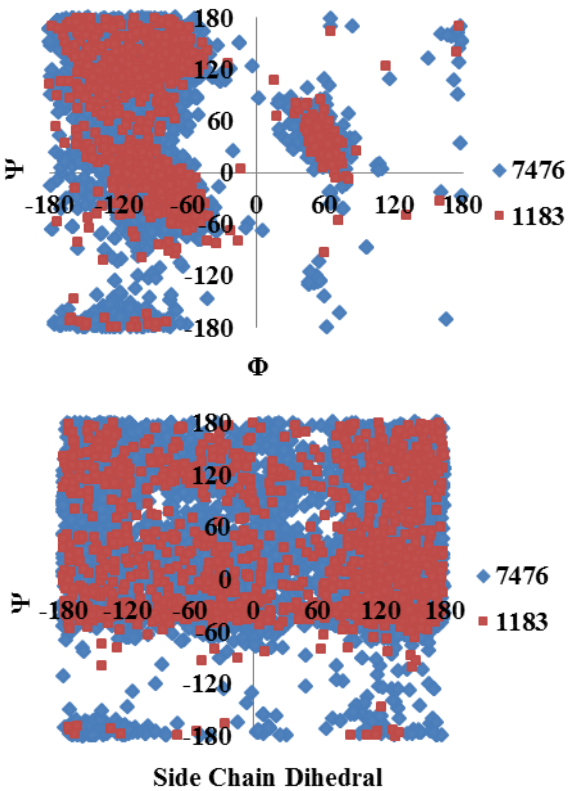


Figure 3.5a: (*Left*) Ramachandran plot of the Φ and Ψ angles of Asn residues sampled from a large pool of proteins (7476 sampled geometries) and a smaller pool of proteins (1183 sampled geometries). **3.5b:** (*Right*) Plot of Ψ against side chain dihedral for the same residues

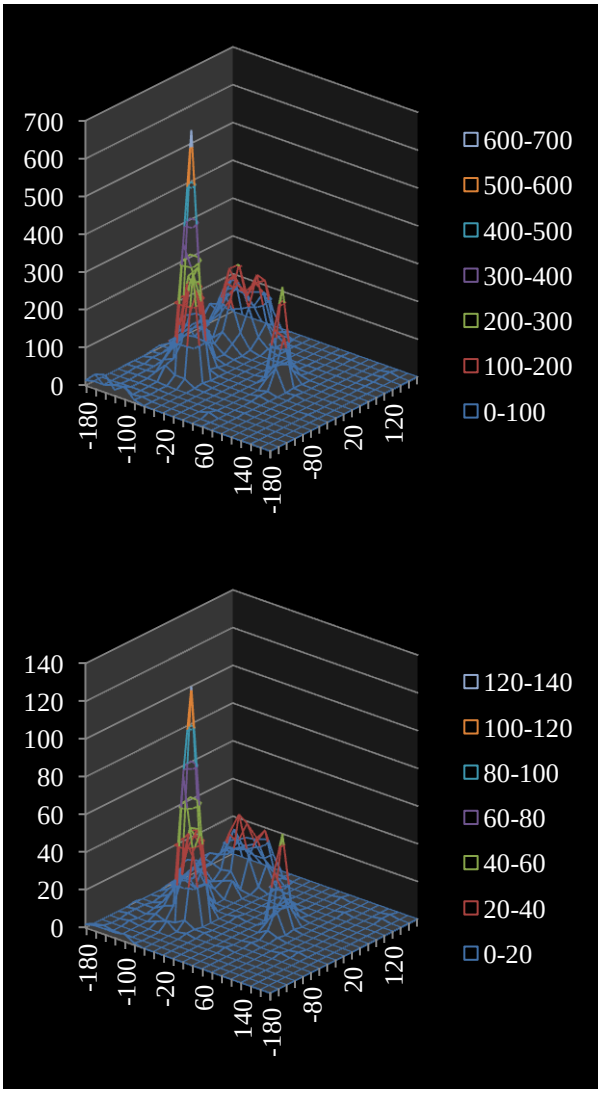


Figure 3.6: 3D Ramachandran plots of Asn. The *top* plot contains the 7476 residues sampled from the larger pool of proteins, and the *bottom* plot contains the 1183 residues from the smaller pool of proteins.

3.3.3. Correlated Dihedrals- A study of the Lysine Side Chain Rotamer Populations

Lysine is an amino acid with a particularly long and flexible side chain with four dihedral angles around C-C bonds (shown in **Figure 3.7**). For each of the dihedrals, there are three local minima- trans, gauche$^+$ and gauche$^-$. If using normal modes sampling, $3 \times 3 \times 3 \times 3 = 81$ minima would therefore be needed to ensure that geometries are sampled for all combinations of these dihedrals. Obviously, some of the possible rotamers are not energetically likely, due to factors such as steric clashing, so not all rotamers are equally likely. For example, the dynameomics rotamer library of Daggett and Scouras [165] shows that dihedrals 3 and 4 both strongly favour the trans conformation (66-75%), whereas dihedral 1 exists gauche$^-$ >75% of the time. Search of the potential energy surface of the gas phase lysine dipeptide molecule produced only 39 minima[133] to describe Lys, meaning that normal modes sampling is expected to struggle to cover all chemically relevant geometries. This is seen in **figure 3.8** where the number of sampled geometries

against dihedral angle is plotted for each of the four side chain dihedrals. It is seen that only dihedral 3 (green line) has geometries sampled for all three minima. Alarmingly, there is no example of dihedral 1 in the gauche$^-$ conformation, the most populous rotamer according to dynameomics.

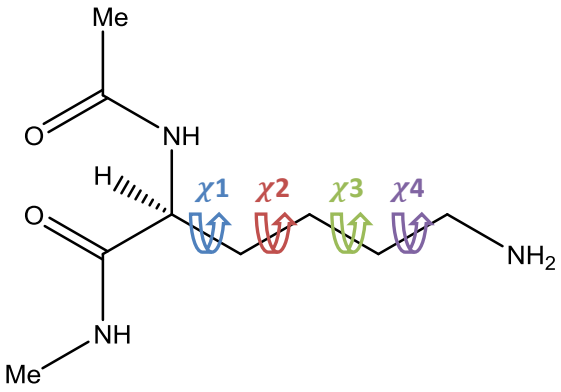


Figure 3.7: The four dihedral angles in the side chain of Lys, referred to as dihedral 1 (blue), dihedral 2 (red), dihedral 3 (green) and dihedral 4 (purple) in the text.

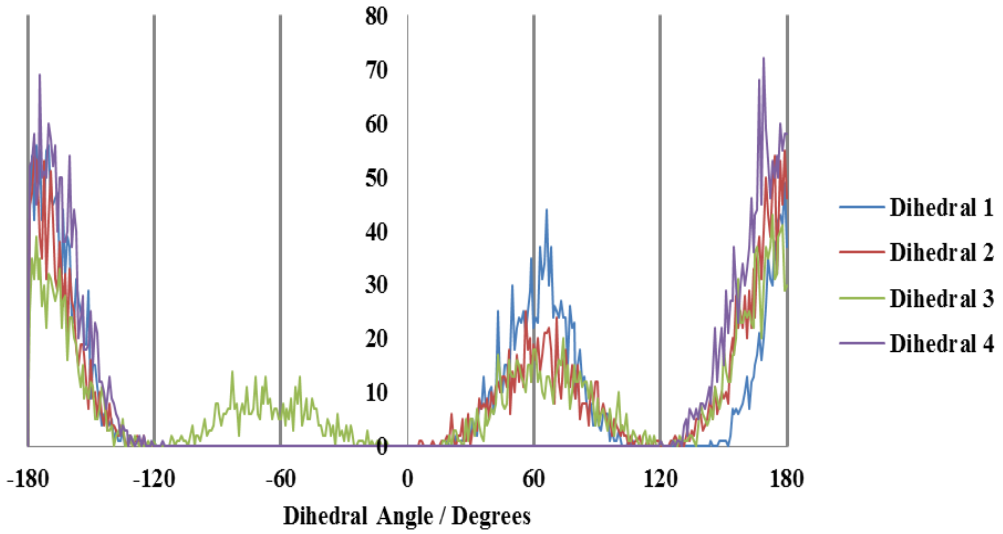


Figure 3.8: Plot of the number of geometries with a given dihedral angle obtained by normal modes sampling. 2964 geometries sampled in total.

Unlike the geometries sampled from normal modes, the geometries sampled from the PDB do include geometries for each of the three minimum energy conformations one would expect for the four side chain dihedrals (**Figure 3.9**). This reiterates one of the problems related to sampling using normal modes- due to the dependance on input minima, chemically relavent conformations of the molecule may be left unsampled. Because the PDB sampling produces these geometries direct from experimental crystal structures, one must accept that the geometries obtained by PDB sampling but ommited by normal modes are indeed chemically relevant. PDB sampling successfully reproduces the rotamer populations present in the dynameomics library, where dihedral 1 is primarilly gauche$^-$ with the majority of remaining geometries being trans, and dihedrals 2-4 favouring the trans conformation.

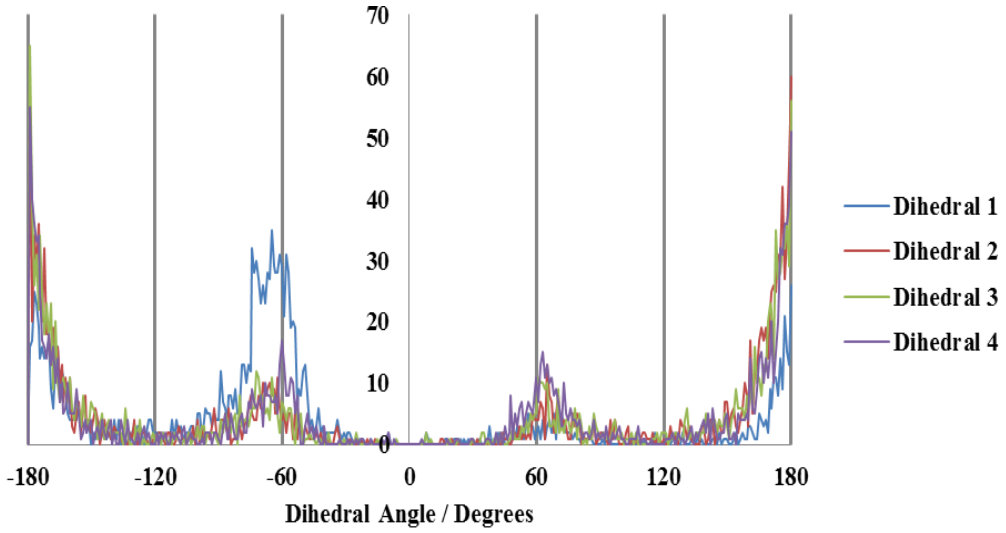


Figure 3.9: Plot of the number of geometries with a given dihedral angle obtained by PDB. 1556 geometries sampled in total.

Figure 3.10 is the same as **Figure 3.9**, however it is obtained with geometries sampled from the larger pool of proteins. There are 1556 geometries represented in **Figure 3.9** and 9425 in **Figure 3.10**, however both graphs show the same structure. This is in good aggreement with **Figure 3.6** where increasing the number of Asn residues sampled from the pdb had little effect on the relative peak heights of the Ramachandran plot. The peaks in **Figure 3.1** appear smoother than **Figure 3.9** due to the larger sample size. The relative peak heights for dihedrals 2-4 in **Figures 3.9** and **3.10** show the expected sizes- the peak at ±180 degrees (corresponding to the trans conformation) is larger than the peaks at ±60 degrees (the *gauche+* and *gauche-* conformations).

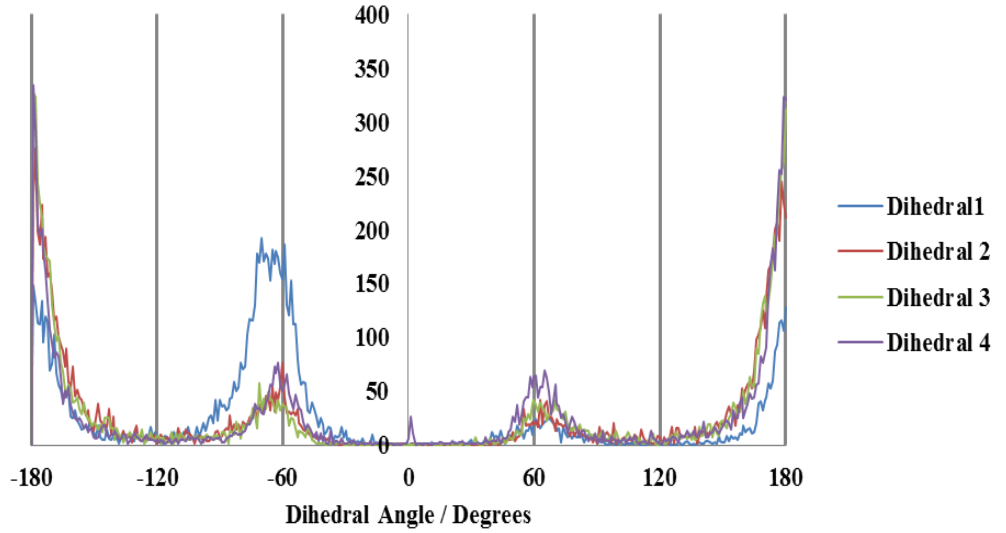


Figure 3.10: Plot of the number of geometries with a given dihedral angle obtained by PDB. 9425 geometries sampled in total.

The discussion now moves to address the previously introduced issue of "corellated dihedrals". A quick look at **Figure 3.8** would mislead one into thinking that using normal

modes sampling, all chemically relevant conformations of dihedral 3 are included in the training set (*trans, gauche+* and *gauche-*). This is seen not to be the case. **Figure 3.11** is analogous to the green line in **Figure 3.8** however only geometries where dihedral 2 lies within the range of 0-120^{o} are included. It can be seen that there are no sampled geometries where dihedral 3 and dihedral 2 lie in the ranges -120-0^{o} and 0-120^{o} respectively. This is shown in the boxed region of **Figure 3.12**.

For geometries sampled from the PDB (**Figures 3.13 and 3.14)**, it is observed that there are geometries with dihedral 3 and dihedral 2 in the ranges -120-0^{o} and 0-120^{o} respectively. Because the PDB sampled geometries come from real protein geometries, any areas of conformational space that are sampled are "chemically relevant". Because the strucures obtained by normal modes sampling miss regions of conformational space that are seen to be chemically relevant (as illustrated by the previous example), normal modes sampling had failed in this instance. As one would expect, the PDB sampled geometries do exhibit a clear preference for the less sterically hindered trans-trans geometry (clustering of points in the corners of **Figure 3.14**).

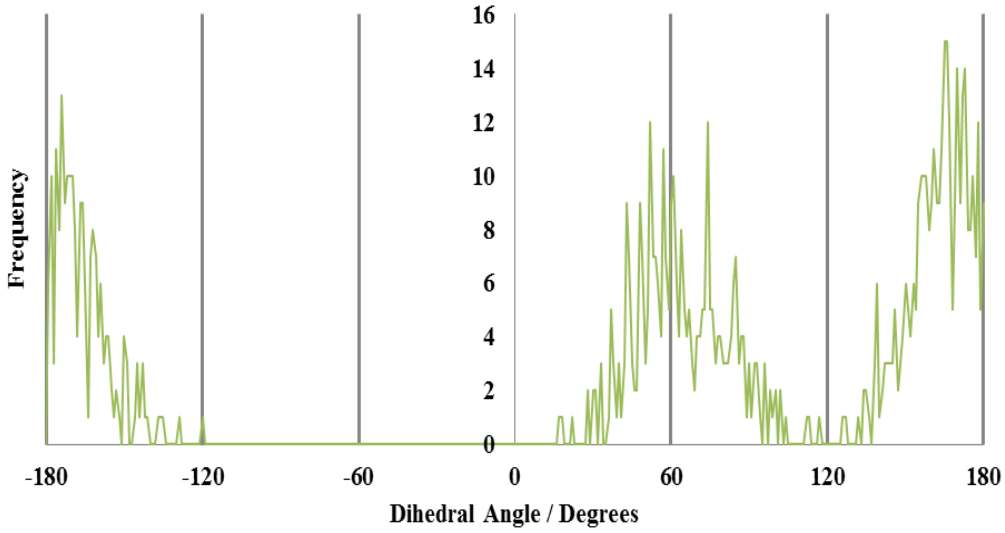


Figure 3.11: Plot of the number of geometries sampled for different values of dihedral 3 when dihedral 2 is between 0-120^{0}. Geometries sampled by normal modes.

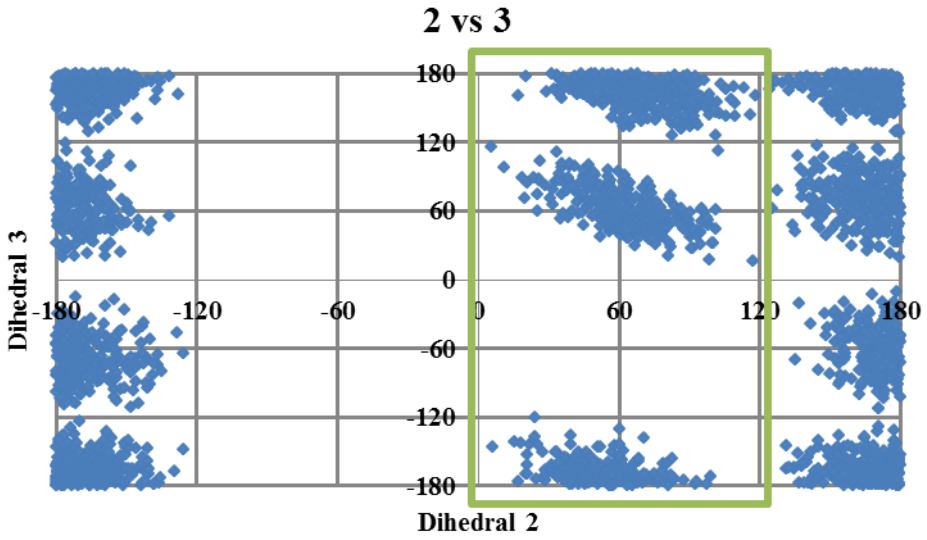


Figure 3.12: Scatter plot of dihedral 2 vs dihedral 3 for geometries sampled by normal modes. The boxed region corresponds to the geometries plotted in **figure 3.11**.

Spider plots can be drawn to display the dihedral sampling of a molecule efficiently. Each line on a spider plot corresponds to a sampled molecular geometry and each axis corresponds to a dihedral angle. Example spider plots for the four dihedral angles in the Lys side chain are provided for both PDB (blue) and normal modes (green) sampled geometries are given below in **Figure 3.15**. The spider plots reinforce the above arguments that normal modes sampling provides a set of locally clustered molecular geometries that poorly reproduce the distribution of molecular geometries sampled from real proteins. An important point must be made regarding the spider plot of the geometries sampled by normal modes as the geometries are not the same as those in the above histograms. The reason for this is that the geometries used in the spider plot were sampled for use in kriging models, and thus stricter requirements on quality of the geometries was required, such as lower allowed deviations in bond lengths. The means that the energy pumped into the normal modes is lower, leading to reduced dihedral sampling. In the histograms, the geometries were sampled purely for illustrative purposes with relaxed structural criteria, and hence the energy input to the minima was higher, and greater dihedral sampling was obtained. Due to the random input of energy to the normal modes by TYCHE, no formal quantitative comparison is available.

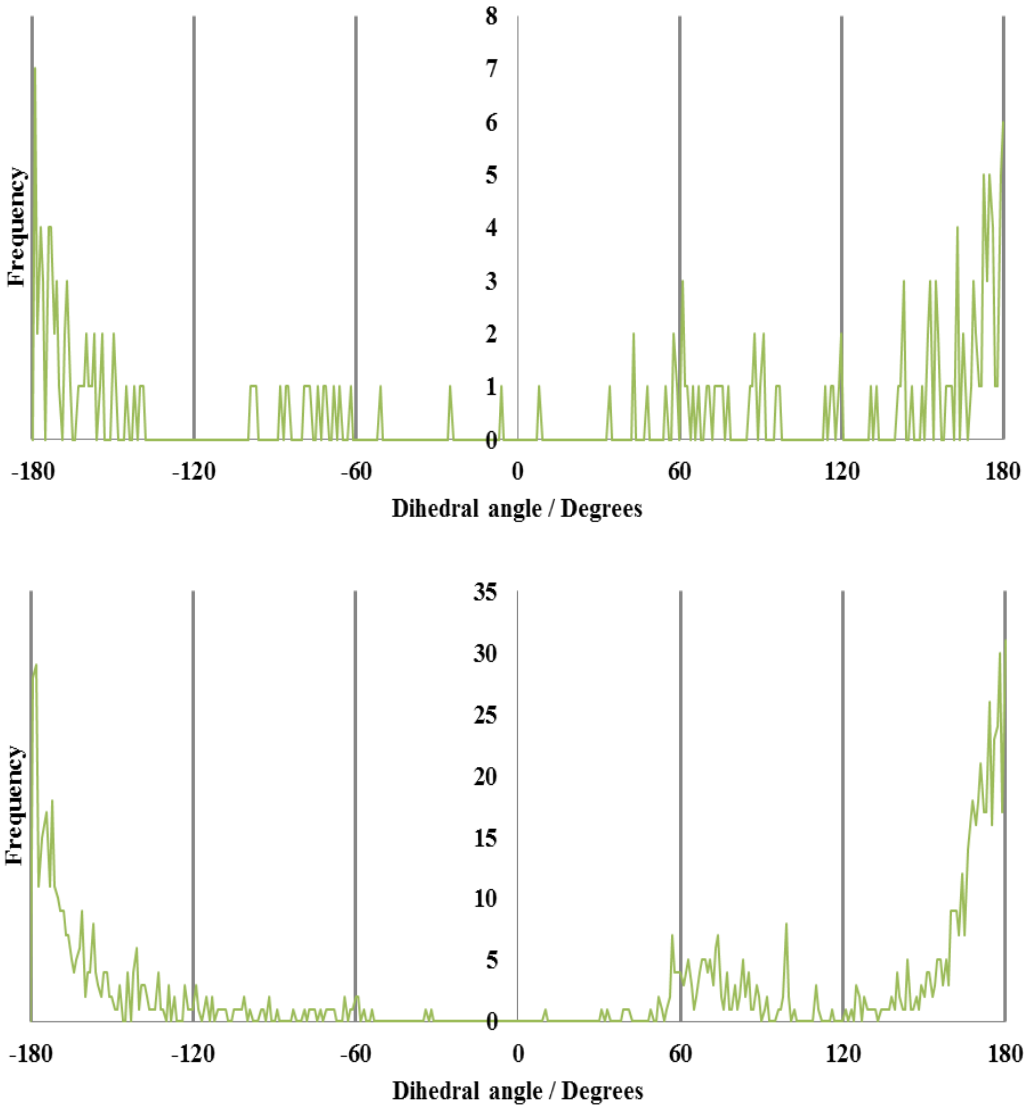


Figure 3.13: Plot of the number of geometries sampled for different values of dihedral 3 when dihedral 2 is between 0-120^0. Geometries sampled from the PDB, with the top graph sampled from a small pool of proteins, and the bottom graph sampled from a large pool of proteins.

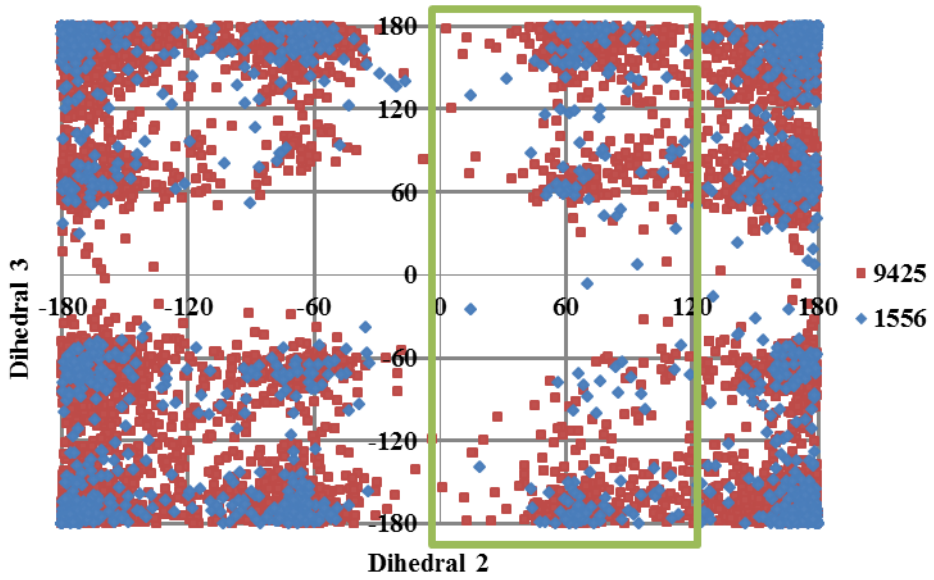


Figure 3.14: Scatter plot of dihedral 2 vs dihedral 3 for geometries sampled from the PDB (both from a large and a small pool of proteins containing 9425 and 1556 geometries respectively). The boxed region corresponds to the geometries plotted in **Figure 3.13**.

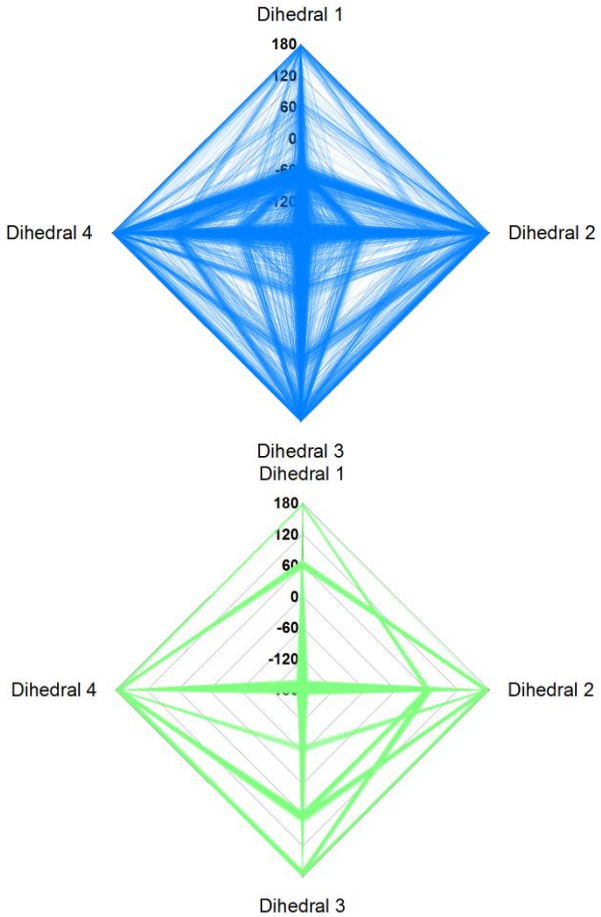


Figure 3.15: Spider diagrams of the four side chain dihedrals on Lys for geometries sampled from the PDB (*top, in blue*) and from normal modes (*bottom, in green*).

3.3.4 Development of the PDB/NM Hybrid Sampling Approach

The above discussion conveys to the reader that normal modes sampling of amino acid geometries is not able to sample the full conformational space occupied by a given amino acid when using gas phase energetic minima as input geometries. The simplest way to improve the performance of normal modes sampling is to increase the number of input minima in order to "fill the gaps" in conformational space. It has been discussed, however, that the number of possible geometries that a molecule may occupy increases rapidly with the size of the molecule, resulting in an increasingly large number of "gaps" that need filling. PDB sampling has been explored as an alternative method of sampling and shows promise. Dihedral sampling is better performed by PDB sampling than normal modes, and the amino acid geometries are inherently biologically relevant. The main problem associated with PDB sampling lies in the quality of the crystal structure of the protein and the possibility of unrealistic bond-lengths and angles included in the sampled geometries. The hybrid PDB/normal modes sampling approach has been developed, hereafter referred to as PDB/NM, with the aim of getting the positive features of both sampling approaches whilst avoiding their shortcomings.

An overview of the PDB/NM sampling approach is provided in **Figure 3.16**. The first stage of the sampling process is extraction of amino acids from protein crystal structures. The program MOROS performs this task. A large number (hundreds) of the PDB sampled amino acids are then randomly selected to be used as input minima for normal modes sampling. Before the normal modes sampling is performed, a partial optimisation is performed on each sampled structure where all dihedral angles are fixed but bond lengths and angles are allowed to relax. The relaxed amino acid geometries are then used as "seeds" for the normal modes sampling. The normal modes sampling of the partially optimised seeds requires the non-equilibrium normal modes of the molecule to be sampled (see **Chapter 1**). This is due to the first order term of the Taylor expansion to not be zero when not at an energetic minimum. A key difference to standard normal modes sampling of energetic minima at this stage in the sampling process is that a much smaller number of geometries are sampled per seed in the PDB/NM scheme due to the larger number of seeds. The molecular geometries output by TYCHE are then run though GAIA, starting at the Gaussian stage.

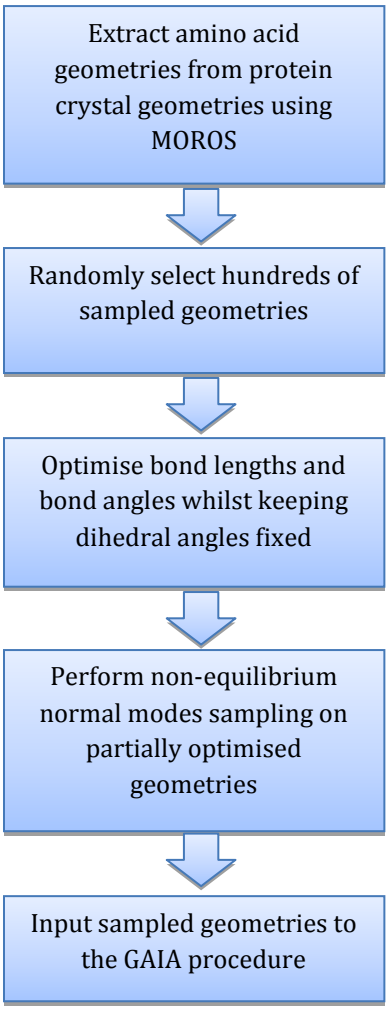


Figure 3.16: Flowchart outlining the stages of the PDB/NM sampling procedure.

The partial optimisation of the sampled geometries before input to TYCHE is of key importance in the PDB/NM sampling scheme for a number of reasons. By allowing the molecule to relax in this way poor quality information in the crystal structure in the form of poor bond lengths and angles is "tidied up". In addition, the program Haad[134] (which is used to add hydrogen atoms to the extracted amino acid geometries) adds all hydrogen atoms at fixed bond lengths. Relaxation gives more realistic R-H bond lengths. An early suggestion for the cleaning up of PDB sampled structures for use as a QCTFF training set was to sample all geometries from the PDB and to relax each structure in this way and input the relaxed structures into GAIA. Although this method is successful at "cleaning up" training set it ultimately leads to a poor sampling of conformational space as all bond lengths in the training set are at, or close to, their equilibrium value. PDB sampling in this manner offers improved dihedral sampling of an amino acid than normal modes sampling, but offers a poorer sampling of bond lengths and angles than normal modes. The normal modes component of PDB/NM is responsible for the sampling of the regions conformational space described by variation in bond lengths and angles.

In this work, kriging models have been built for the two amino acids, Ala and Lys, using four different sampling approaches (resulting in 8 sets of kriging models). The four approaches are detailed in **Table 3.1**. A comparative analysis of each sampling approach is given ranging from qualitative comparison of the dihedral sampling to quantitative analysis of the molecular energies and performance of the kriging models.

Table 3.1. An overview of the four sampling approaches.

PDB_OPT	Molecular geometries sampled directly from crystal structure coordinates and H atoms added by the HAAD program. GAUSSIAN fully optimises bond lengths and valence angles but all dihedral angles remain fixed.
PDB_NO_OPT	Molecular geometries taken directly from PDB coordinates and H atoms added by HAAD. Single-point GAUSSIAN calculations without any geometry relaxation.
NM	Standard normal modes 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.
PDB/NM	300 randomly selected PDB "seed geometries" sampled with PDB_OPT, each acquiring 7 geometries generated from the non-stationary normal modes. The "seed geometries" themselves are not included in either training or test sets.

All PDB sampling was performed using the larger pool of 260 protein crystal structures. Alanine was chosen because it is the smallest amino acid with a (non-trivial) side chain. Because there is only one side chain dihedral angle (χ_1), as opposed to the four dihedral angles (χ_1, χ_2, χ_3, χ_4) controlling the side chain of lysine, the ϕ and ψ 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 3.7** shows the four side chain dihedrals in lysine around C-C bonds, χ_1, χ_2, χ_3 and χ_4.

3.3.4.1 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 3.1**. Ramachandran plots for the sampled alanine geometries by each of the sampling methods are shown in **Figure 3.17**. The dihedral angles are fixed to the same values in both the PDB_OPT and PDB_NO_OPT approach, which is why **Figure 3.17** assigns the same colour (blue) to the distribution of ψ and ϕ angles of their

geometries. As expected, the PDB-sampled Ramachandran plots for both Ala and Lys display a sampling bias towards the α-helix and β-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, it is concluded here that normal modes 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 9 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 α-helix and β-sheet region. This is because the 300 "seed" geometries used as input for the normal modes sampling were randomly selected from the PDB_OPT sampled geometries and, statistically, they are most likely to be sampled from these well populated α-helix and β-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 3.18 shows 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 χ_n (n=1, 2, 3 or 4) can adopt, i.e. from -180 ^{o} to 180^o. Each sampled geometry then corresponds to a quadruplet of dihedral values (χ_1, χ_2, χ_3, χ_4), each marked by a point on each of the four corresponding axes. These four points are then linked by four coloured 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^o) conformation of χ_1 is the most sampled conformation in the protein crystal structures but this conformation is not at all present in NM. The preference of χ_1 to be in the *gauche*$^-$ conformation in proteins is a well-documented phenomenon[165] 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).

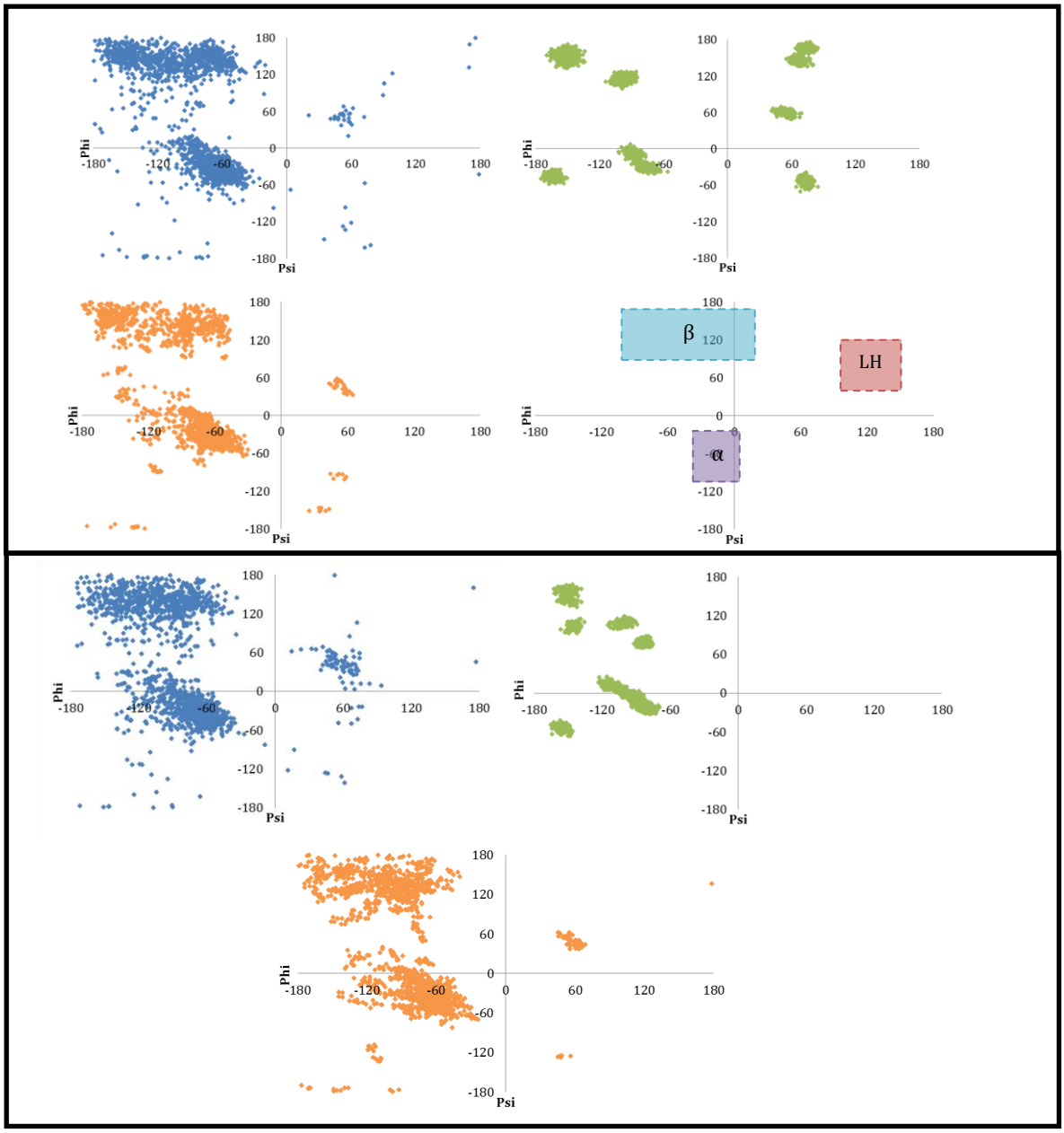


Figure 3.17: Ramachandran plots of Ala (*top box*) and Lys (*bottom box*) sampled using PDB_OPT and PDB_NO_OPT (*blue*), NM (*green*) and PDB/NM (*orange*). In the bottom right panel of the top box is a guide to the regions corresponding to the secondary structural motifs, β-sheet (labelled β), α-helix (labelled α), and left-handed alpha helix (labelled LH).

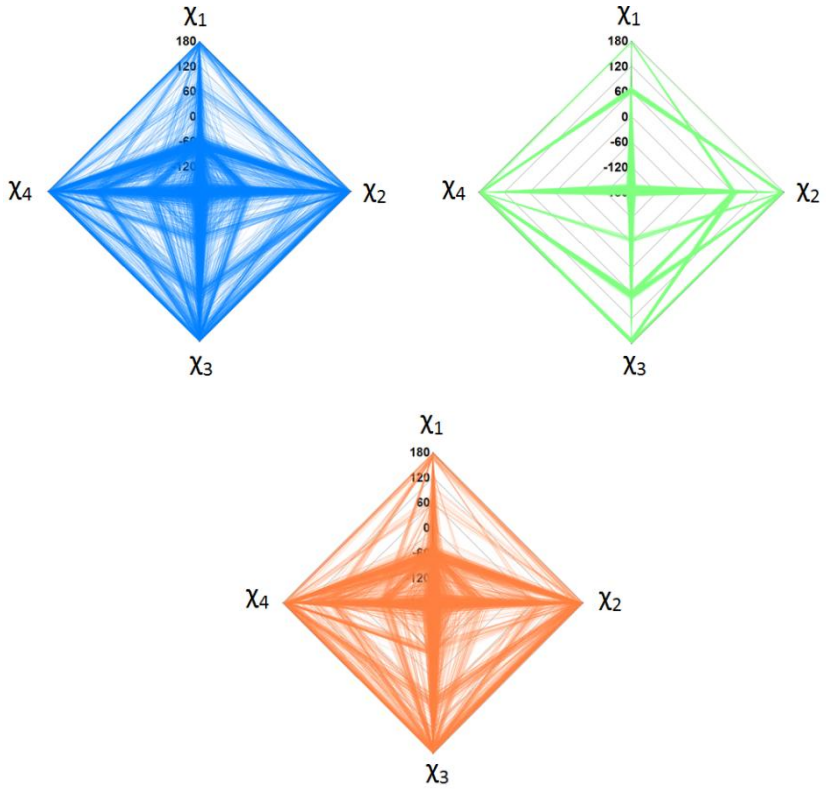


Figure 3.18: Spider plots displaying the Lys side chain conformations sampled by each of the four sampling approaches: PDB_OPT and PDB_NO_OPT (*blue*), NM (*green*) and PDB/NM (*orange*). Each axis ranges from -180 ° to 180°.

Table 3.2 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 3.2**. For both amino acids the NM sampled geometries show the smallest range in *ab initio* energy. 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 kJmol^{-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 non-stationary normal modes sampling.

Table 3.2 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 obviously no bond length variation is introduced by normal modes). The average bond length ranges of 0.07 Å and 0.08 Å for PDB_NO_OPT Ala and Lys, respectively, are the next lowest values. The reason for the low average bond

length range of the PDB_NO_OPT geometries is that the hydrogen addition software used, HAAD, adds hydrogens at a fixed length of 0.985 Å. Therefore the average range in bond length is reduced by all bonds containing a hydrogen atom. A more informative metric to describe the sampling of bond lengths by each method is to study the range of a single bond containing two heavy atoms. The bond between C_α and C_β was chosen for this purpose. Again, the PDB_OPT showed the lowest ranges of 0.03 Å and 0.05 Å, respectively, but the PDB_NO_OPT Ala geometries showed the highest range in C_α-C_β distance of 0.22 Å as expected. NM and PDB/NM showed the same range in C_α-C_β bond length of 0.14 Å. This highlights the similarity of both the stationary and non-stationary normal modes sampling algorithms in TYCHE.

Table 3.2. Statistical information detailing the sampling of Ala and Lys by the four sampling methods. All energies are in kJmol^{-1} and all distances in Å.

Alanine	PDB_OPT	PDB_NO_OPT	NM	PDB/NM
Range in *ab initio* Energy	132.5	281.0	84.4	111.0
Average Bond Length Range [a]	0.02	0.07	0.11	0.12
C_α-C_β Bond Length Range	0.03	0.22	0.14	0.14
Average $\left\|\Delta E_{system}\right\|$	0.7	1.8	4.0	3.4
Average $\left\|E_{AB}^{original} - E_{AB}^{predicted}\right\|$	0.1	0.2	0.4	0.4
Max $\left\|\Delta E_{system}\right\|$	6.8	25.8	18.4	17.2
Max $\left\|E_{AB}^{original} - E_{AB}^{predicted}\right\|$	10.0	9.4	13.7	9.4

Lysine	PDB_OPT	PDB_NO_OPT	NM	PDB/NM
Range in *ab initio* Energy	126.0	310.6	111.1	420.9
Average Bond Length Range [a]	0.02	0.08	0.13	0.14
C_α-C_β Bond Length Range	0.05	0.12	0.13	0.13
Average $\left\|\Delta E_{system}\right\|$	1.6	2.5	3.3	3.8
Average $\left\|E_{AB}^{original} - E_{AB}^{predicted}\right\|$	0.2	0.3	0.3	0.4
Max $\left\|\Delta E_{system}\right\|$	20.4	23.1	15.2	18.1
Max $\left\|E_{AB}^{original} - E_{AB}^{predicted}\right\|$	32.5	34.2	7.1	28.4

[a] The set of training geometries provides a range (i.e. maximum – minimum) for each bond length. The ranges of all bonds appearing in the system are then averaged (over these bonds).

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 3.19** shows the S-curves for all four sampling methods. As an reminder 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 $kJmol^{-1}$ (or 1 $kcalmol^{-1}$) (where the red curve intersects the purple dashed line).

In connection with the information shown in **Figure 3.19**, note that **Table 3.2** also reports the average absolute total error and the highest total error for each S-curve. The alanine models built using PDB_OPT geometries (blue curve) had the lowest average error of 0.7 kJ mol^{-1}. This is attributable to the lack of bond length and angle variation in the training and test sets and so the kriging problem is "less challenging" as there are fewer dimensions of conformational space being sampled. The second left-most S-curve corresponds to the predictions made using the models built using PDB_NO_OPT geometries (red curve). This is most likely a result of the lack of bond length variation of all hydrogen-containing bonds. However, the PDB_NO_OPT does have the highest maximum total error of all sampling approaches, amounting to 25.8 $kJmol^{-1}$, despite the low average error. This is attributable to an alanine residue extracted from a crystal structure with a significantly stretched C_α-C_β bond length and the H_α-C_α-C_β angle of 115°, which is significantly distorted from the stationary value of ~108°. This fact illustrates the unsuitability of sampling amino acid geometries directly from crystal structures for QCTFF development, and emphasises the need for a PDB/NM hybrid sampling approach. The kriging models obtained from the PDB/NM and NM sampled geometries perform worst overall, which is due to the large quantity of bond length sampling relative to the PDB_OPT and PDB_NO_OPT approaches. Despite being the S-curves furthest to the right, PDB/NM and NM have average S-curve errors of only 3.4 $kJmol^{-1}$ and 4.0 $kJmol^{-1}$, respectively. More than 60% of the test geometries of alanine were predicted by kriging models with an error of less than 1 $kcalmol^{-1}$, a value often described as "chemical accuracy".

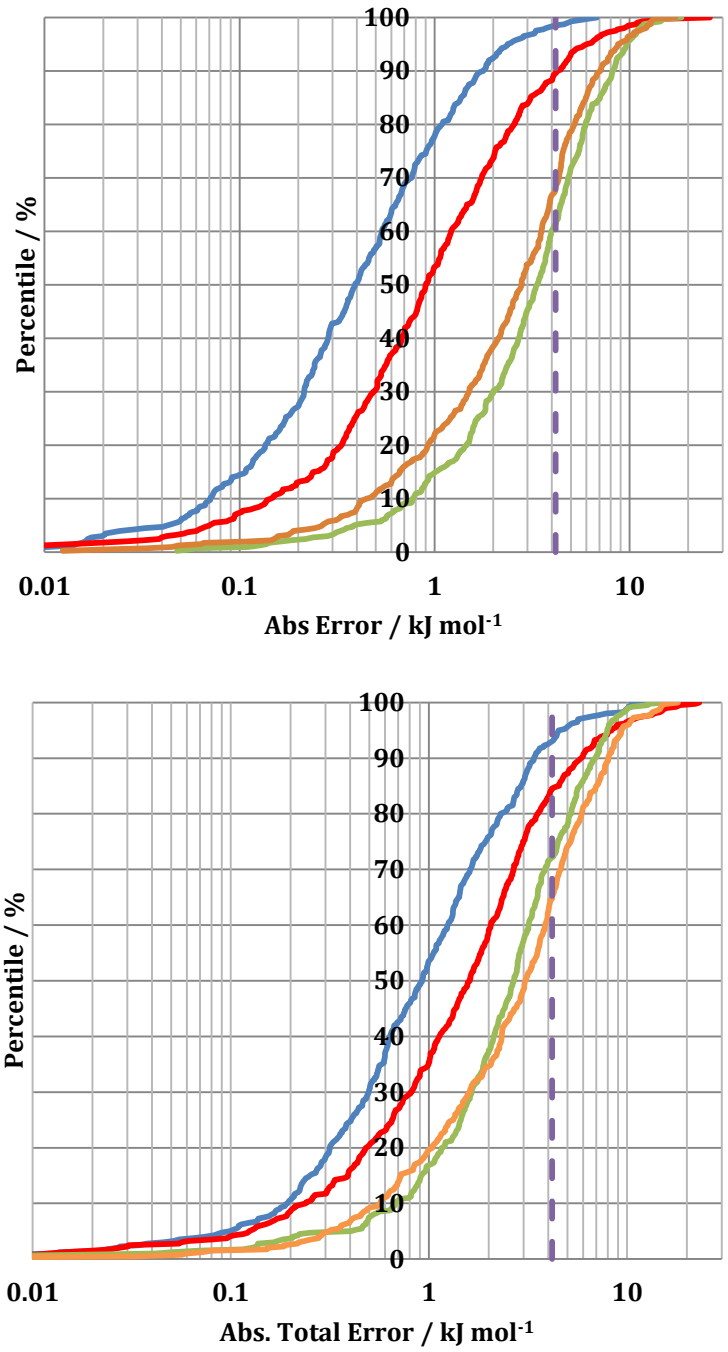


Figure 3.19: Errors in the predicted total electrostatic interaction energies (1-4 and higher) of alanine (*top*) and lysine (*bottom*) for kriging models trained with molecular geometries obtained by: PDB_OPT (*blue*), PDB_NO_OPT (*red*), NM (*green*) and PDB/NM (*orange*). The dashed purple lines mark the 1 kcal mol^{-1} threshold.

It is interesting to note that the dihedral sampling appears to have less effect on the difficulty of the kriging problem than well-sampled bond lengths. **Figure 3.20** plots the average bond length range against average total (S-curve) error for all four sampling approaches for Ala. The correlation between bond length and average S-curve error $\left(\frac{1}{N_{train}}\sum_{i=1}^{N_{train}}\left|\Delta E_{i,system}\right|\right)$ is fairly strong, with an R^2 value of 0.90 (see **Figure 3.20**). To illustrate this point further, the difference in average total error (S-curve error or

$|\Delta E_{system}|$) between PDB/NM and NM is 0.6 kJmol^{-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 sampling of dihedral space than NM but also the smallest average range of bond lengths, has an average total error 3.3 kJmol^{-1} 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 **Figure 3.21** where plots of C_α charge against both N-C_α bond length and backbone ψ 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_α charge shows correlation with the N-C_α bond length but not with the ψ angle. It is only in the plots obtained from the PDB_OPT geometries (where the N-C_α bond length range is significantly reduced as a result of partial geometry relaxation) that any correlation between C_α charge and ψ can be seen. In summary, the correlation patterns above prove the dominance of bond length variation over dihedral sampling in posing a challenge to kriging.

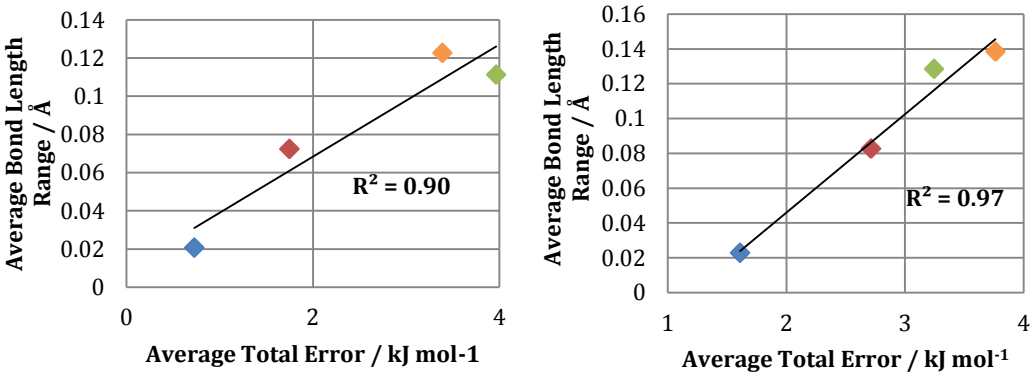


Figure 3.20: Average bond length deviation against average total (S-curve) error for the different sampling approaches of Ala (*left*) and Lys (*right*): PDB_OPT (*blue*), PDB_NO_OPT (*red*), NM (*green*) and PDB/NM (*orange*).

The same conclusions may be drawn from the Lys S-curves as from the Ala S-curves: average bond length deviation is the most import factor dictating the average S-curve error (**Figure 3.20**, right hand graph), and although larger dihedral sampling increases the average error, it does this to a lesser extent than a large average bond length deviation. PDB_OPT has the lowest average S-curve error (Lys: 1.6 kJmol^{-1} and Ala: 0.7 kJmol^{-1}) due to the optimised 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 to the less steep climb of this S-curve and its longer tail towards the 100% ceiling.

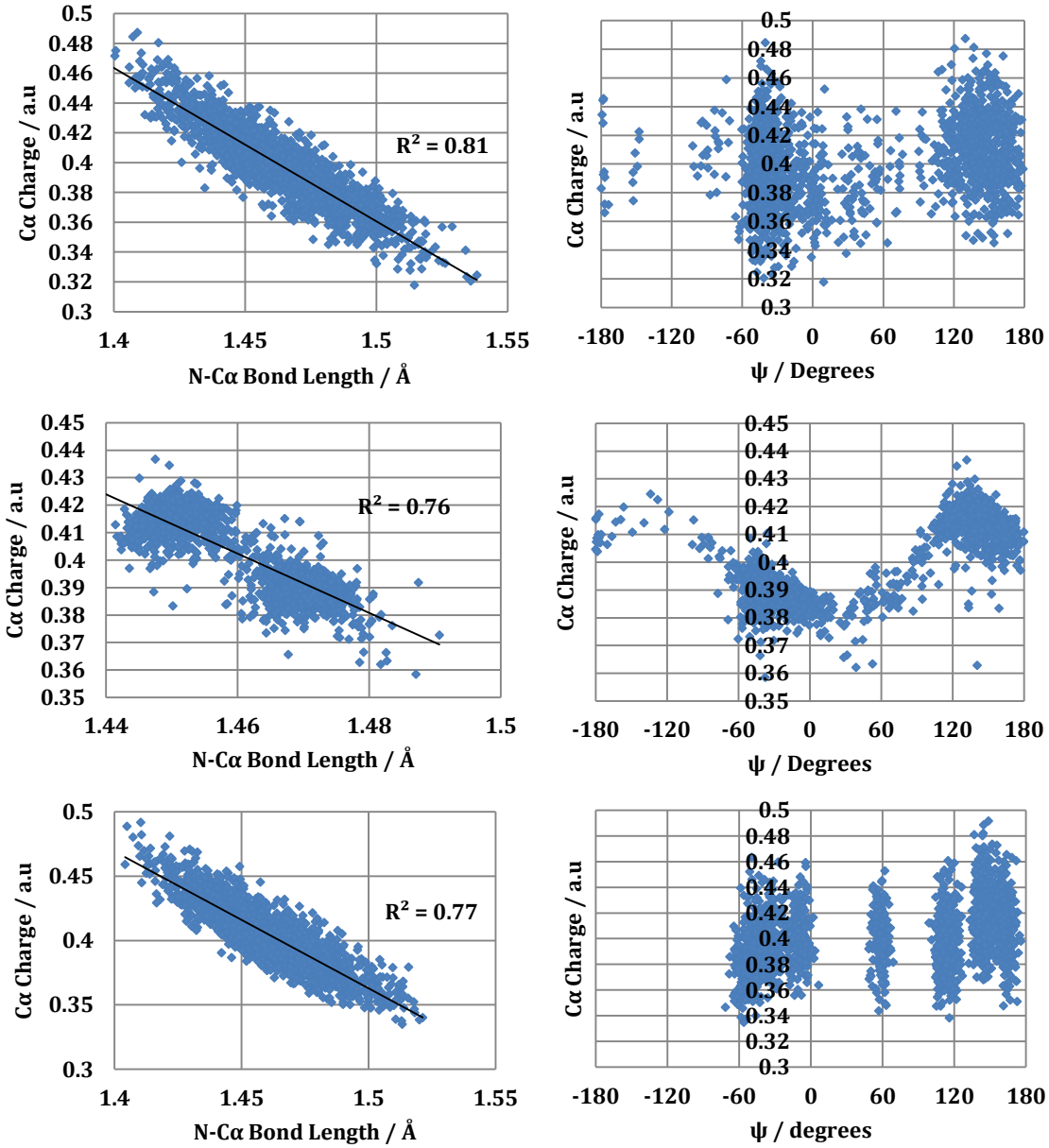


Figure 3.21: Dependence of Ala C_α charges (*left*) on N-C_α bond length and (*right*) on backbone ψ dihedral angle for PDB/NM sampled geometries (*top*), PDB_OPT sampled geometries (*middle*) and NM sampled geometries (*bottom*)

Each point on the S-curve is a sum of all 1,4 and higher intramolecular interaction prediction errors within a single test geometry ($\left|\sum_{AB}\left(E_{AB}^{true} - E_{AB}^{predicted}\right)\right|$ from **Equation 1.36** in **Chapter 1**). 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 I now focus on the construction of the S-curve. **Figure 3.22** shows all interaction errors for all Ala test geometries plotted against interaction distance for each sampling approach. The maximum absolute interaction error (max $\left|E_{AB}^{original} - E_{AB}^{predicted}\right|$) and average absolute interaction error (average $\left|E_{AB}^{original} - E_{AB}^{predicted}\right|$) for each approach is included in Table 2. **Figure 3.23** shows a plot analogous to **Figure 3.22** but for the sampled Lys geometries. The average absolute interaction errors follow the same trend as the total S-curve error (PDB/NM ≈ NM > PDB_NO_OPT > PDB_OPT). For all sampling approaches used, the largest average absolute interaction error was only 0.4 kJmol^{-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 **Figure 3.24**.

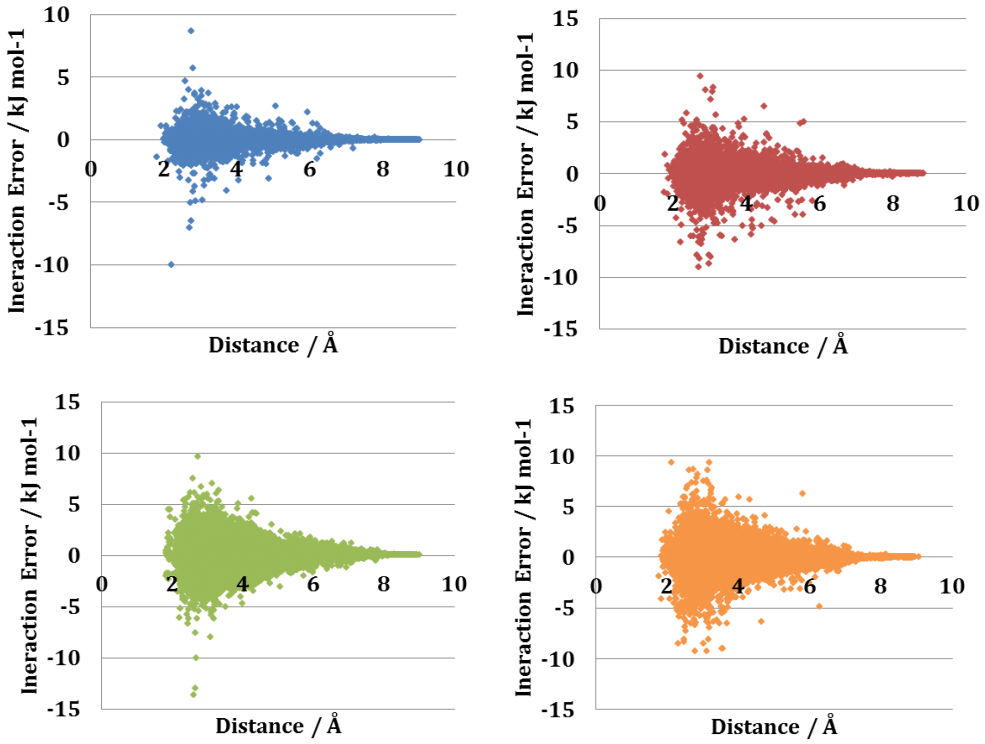


Figure 3.22: Individual intramolecular interaction prediction errors in Ala against interaction distance obtained for models built using the four sampling approaches: PDB_OPT (*blue*), PDB_NO_OPT (*red*), NM (*green*) and PDB/NM (*orange*).

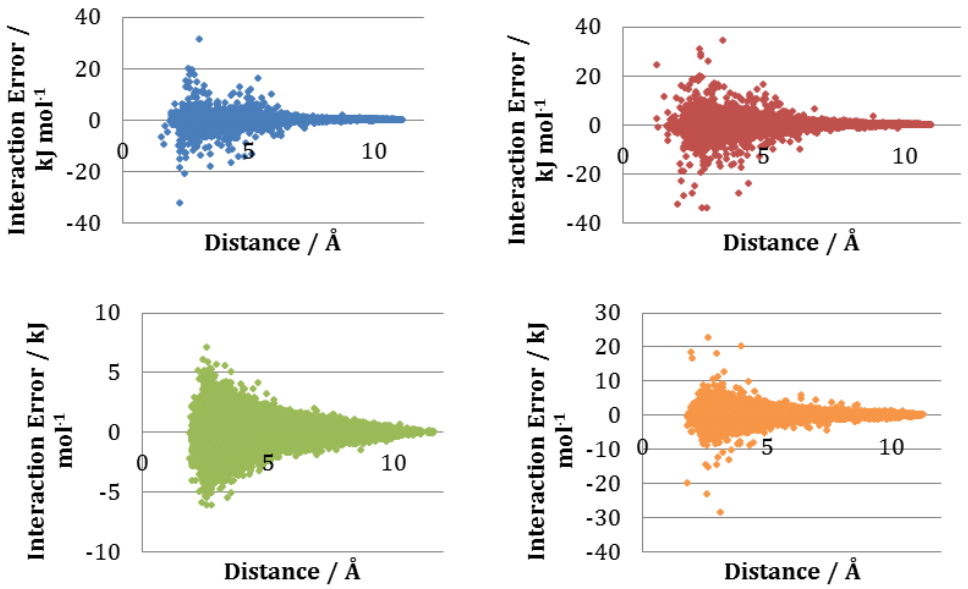


Figure 3.23: Individual intramolecular interaction prediction errors in Lys obtained for models built using the four sampling approaches: PDB_OPT (*blue*), PDB_NO_OPT (*red*), NM (*green*) and PDB/NM (*orange*).

The standard deviation of the interaction errors for each method is provided in Table 4 for both Ala and Lys. Both PDB_OPT and PDB_NO_OPT have significantly larger standard deviations for Lys (0.5 and 0.8 kJmol^{-1}, respectively) than for Ala (0.2 and 0.4 kJmol^{-1}, respectively) as is expected by comparison of the blue and green plots in Figures 8 and S4. The PDB/NM interactions in Lys also have a larger standard deviation (0.7 kJmol^{-1}) than the PDB/NM interactions in Ala (0.6 kJmol^{-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.

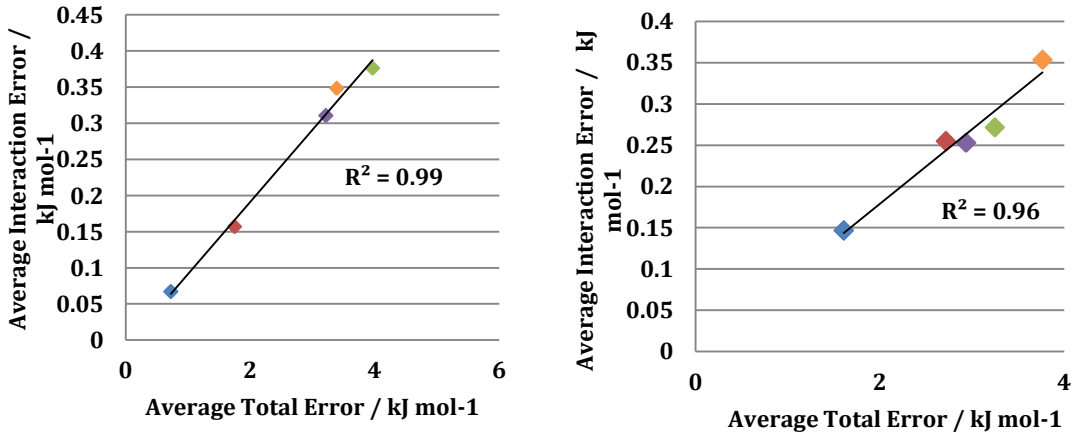


Figure 3.24: The average interaction energy prediction error versus average total (S-curve) error for Ala (*left*) and Lys (*right*) from kriging models trained with molecular geometries obtained by: PDB_OPT (*blue*), PDB_NO_OPT (*red*), NM (*green*) and PDB/NM (*orange*).

Table 3.3. Standard deviation of interaction prediction errors for both Ala and Lys from kriging models built from geometries sampled from the four sampling approaches (kJmol^{-1}).

Sampling	Ala	Lys
PDB_OPT	0.2	0.5
PDB_NO_OPT	0.4	0.8
NM	0.7	0.5
PDB/NM	0.6	0.7

3.3.4.2. 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 non-stationary 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 optimisations, 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.

Figure 3.25 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 3.26**. 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 kJmol^{-1}, between a 1:20 ratio of PDB-seed geometries to sampled geometries (average error of 3.8 kJmol^{-1}) and a 1:1 ratio (average error of 4.4 kJmol^{-1}).

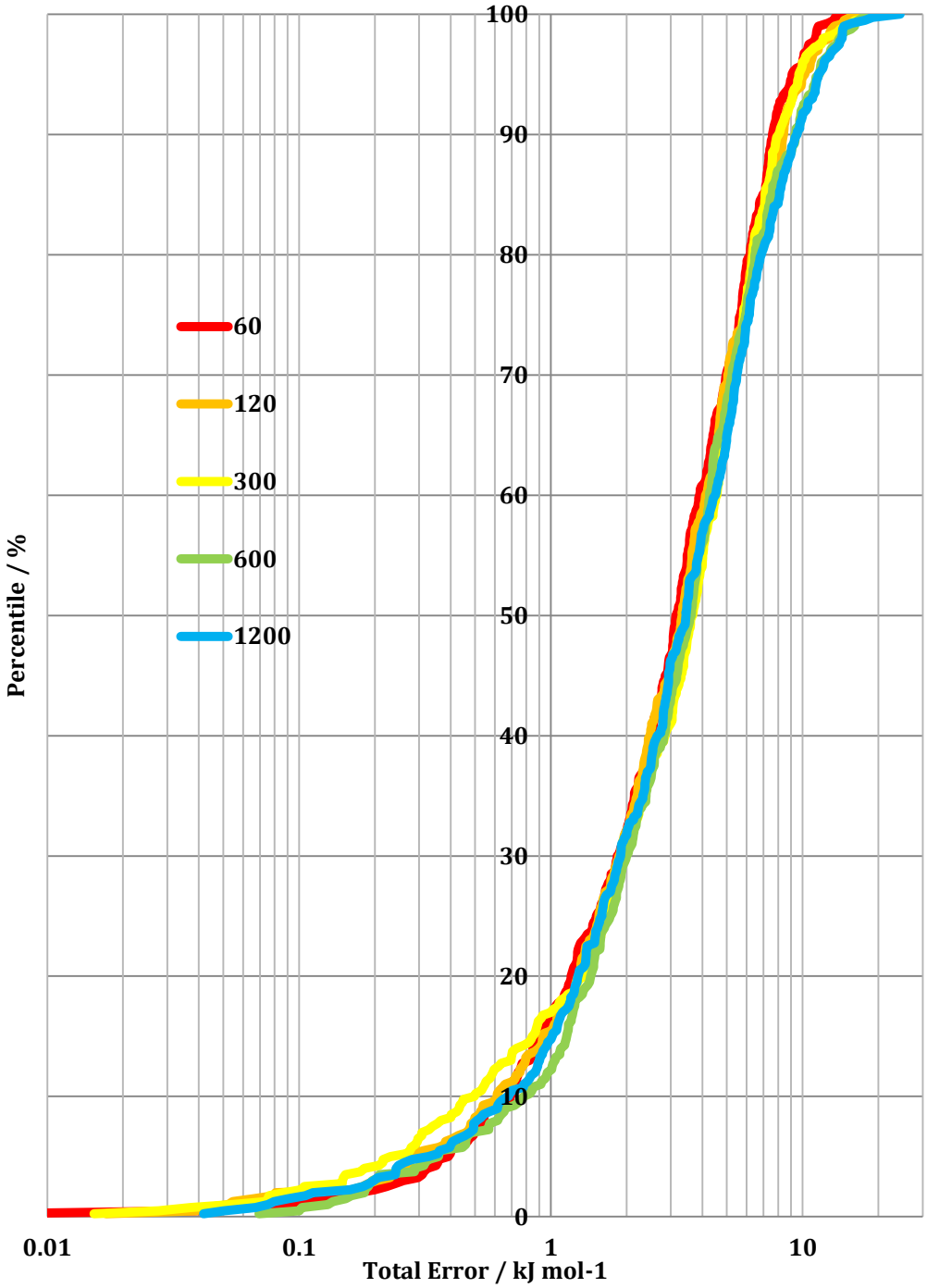


Figure 3.25: Errors in the predicted total 1-4 and higher electrostatic interaction energies of lysine by kriging models trained with molecular geometries obtained by the PDB/NM approach with different numbers of PDB-seed geometries (see key on graph, 1200 corresponds to the 1:1 ratio in the main text).

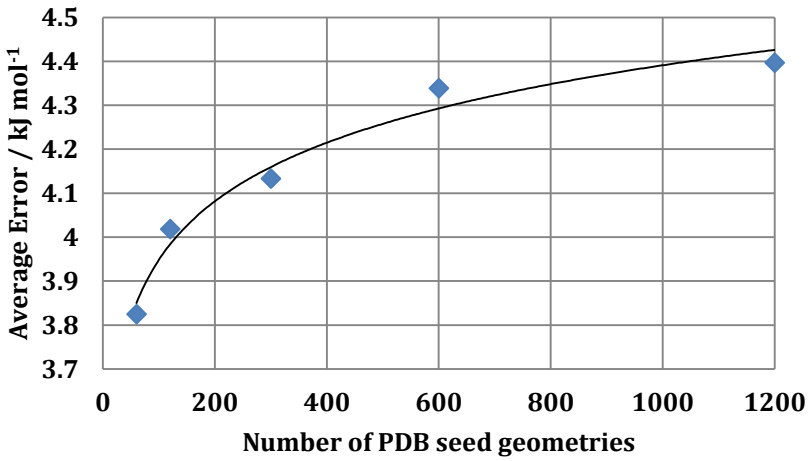


Figure 3.26: Average total error versus the number of PDB seed geometries for kriging models of lysine obtained from the PDB/NM sampling methodology.

3.4. Conclusions and Further Work

The topological force field QCTFF contains a machine learning component that handles polarisation and charge transfer (in a unified way). The machine learning method used, kriging, needs a data set of molecular geometries to train on. Here I focus on obtaining a more realistic and relevant training set for amino acids. Before the current study the training sets were sampled by distorting the local energy minima of (peptide-capped) amino acids (in the gas phase) according to normal modes obtained at those stationary points. Using the Protein Data Bank (PDB) I show here that these gas phase stationary points miss a number of important amino acid geometries that are present in a folded protein.

I have presented a new sampling approach that combines sampling of amino acid geometries from the Protein Data Bank (PDB) with non-stationary normal modes (NM) distortion. 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. The increased sampling in dihedral space by the PDB/NM approach does not cause a significant worsening of the quality of the kriging modes as 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 $kJmol^{-1}$. 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.

As further work I recommend that the next step should be to utilise rotamer libraries to guide the generation of seed geometries. Using rotamer libraries will allow the generation of seed geometries that allows for a bias towards the experimentally observed, biologically relevant amino acid geometries, and will allow a greater measure of control than is obtained by simply sampling the seeds directly from crystal structures. An important further test of any sampling method is to compare with other methodologies, and PDB/NM is no exception. Sampling approaches such as molecular dynamics and Monte-Carlo should be employed and their performance compared to the PDB sampling.

Chapter 4

How Are Protein Substrate Interactions Affected by the Surroundings? Application of the "Atomic Horizon Sphere" to Crambin and the tRNA-Guanine Transglycosylase-3,5-DAPH Complex

Summary

Molecular fragments of increasing size up to a radius of 10 Å have been built around selected atoms for three systems; the protein crambin, the cytochrome P450-camphor-carbon monoxide ternary complex and the tRNA-guanine transglycosylase-3,5-DAPH complex. Multipole moments have been obtained for both the central atom and a number of "probe" atoms, and the electrostatic interaction between the central atom and the probes have been calculated for each fragment to identify at what distance the environment no longer effects the electrostatic interaction between two atoms. This distance is given the name the "atomic horizon sphere".

4.1. Introduction

Although kriging models have been successfully built for amino acids, it is acknowledged that building kriging models from gas phase *ab initio* data will not produce kriging models immediately applicable to an amino acid within a protein in the condensed phase. It is well known that molecular properties obtained from gas phase *ab initio* calculations do not always show strong agreement with experimental condensed phase properties. An often used example is that of the molecular dipole moment of water. The dipole moment of gaseous water is calculated to be 1.855 D [191], whereas in the condensed phase, this increases to ~2.5 D [192]. This change is due to polarisation by the environment.

One of the core driving forces behind the development of QCTFF is to provide a rigorous treatment of the electrostatic interactions between atoms. This is achieved through the use of atomic multipole moments to describe the electronic distribution around an atom in place of atomic point charges (which is the first term of the multipolar expansion). Multipole moments provide an anisotropic description of the electron density around an atom, whereas point charges are spherically symmetric. This allows atomic multipole moments to describe non-spherical features of the electron density such as lone pairs and π-electron density, and has led to the development of a number of "next generation" force fields that include a multipolar description of the electrostatics, most notably AMOEBA [16, 193]and SIBFA[194]. It is the aim of the current work to determine at what distance the environment no longer has a significant polarizing effect on the value of the multipole moments of individual atoms. This is achieved by building multiple fragments of proteins

that include all atoms within a defined horizon radius, r_h, centred on an atom of interest. By calculating the multipole moments of the central atom in fragments of increasing r_h, one may identify the size of the fragment at which adding new atoms no longer has a polarising effect on the multipole moments. The polarisation of the multipole moments can be both directly observed, and can also be "probed" by interacting the atom with a number of "probe" atoms within the system, using fixed values for the probe atom multipole moments for all values of r_h. The minimum radius where both the atomic multipole moments and the central atom-probe atom interaction is no longer affected by increasing r_h is termed the "atomic horizon sphere".

Long range polarizing effects in the condensed phase is a problem when trying to derive generalised reference charges for atom types and so new approaches have been developed to tackle this issue. The most brute force approach is to perform *ab initio* molecular dynamics, where atom types are no longer required, and atomic charges may be extracted simply from a wave function. This is a computationally intensive approach, however studies have showed that charge transfer and polarization is well described by such a method[195]. A less computationally heavy approach is to obtain atomic charges from the system one wishes to study, then run the simulation, as usual, and use those charges. The DDEC/ONETEP approach of Lee *et al.* [241] is one such approach, and was shown to predict NMR coupling constants "at least as well as AMBER". DDEC/ONETEP was much better than AMBER when reproducing the electrostatic potential of protein crystal structures, but one would expect this due to the atomic charges being derived directly from the individual proteins. QM/MM approaches are another approach where polarization of an atom by the local environment is captured by both the QM part of the calculation and by the interactions between the MM and QM atoms. QM/MM approaches have been successfully applied to a number of biochemical systems[196-199].

Two test systems have been used to probe the atomic horizon sphere. The first system is the protein crambin (PDB code 2EYA), which has been extensively studied both by the Popelier group and also by others. Crambin has been isolated from the seeds of the cabbage *Crambe abyssinica* and is the smallest naturally occurring protein, consisting of only 46 amino acids. Previous horizon sphere studies on the protein crambin [157] found that the C_α of Ser_6 has an atomic horizon sphere of 10 Å. In this study, the horizon sphere is built around the carbonyl oxygen and the amide nitrogen of Phe_{13}. This is shown in **Figure 4.1**.

The second system that will be investigated in this work is the tRNA-guanine transglycosylase-3,5-DAPH complex (TGT-DAPH, PDB code 1F3E). TGT is involved in post-transcriptional modification of tRNA, catalysing the exchange of a guanine base with $preQ_1$. TGT is implicated in the pathogenicity of a number of bacteria, and so is a common drug target. 3,5-DAPH was identified as an inhibitor of TGT, and it is the inhibited enzyme that is

used in the studied. A horizon sphere will be built around one of the carbonyl oxygen atoms in 3,5-DAPH, as seen in **Figure 4.2**. This oxygen is involved in hydrogen bonding with multiple residues within the active site of the enzyme.

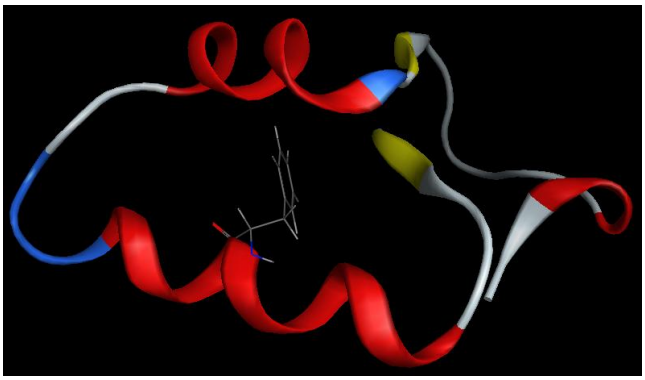


Figure 4.1: The protein crambin with Phe_{13} visible

Figure 4.2: TGT with 3,5-DAPH visible

4.2. Building the molecular fragments

A simple process was followed to obtain the horizon sphere fragments. Before building the molecular fragments, hydrogen atoms were added to the protein crystal structure using the program HADD[134]. Molecular fragments were then built by including all atoms within a specified radius, r from a central atom of interest. The fragments ranged in size from a radius of 1.5 Å up to a r =10 Å, with increasing radius of 0.5 Å. This resulted in 18 fragments of increasing size built around each atom probed. For each fragment, atoms

located on the edge of the fragment that had incomplete valency (due to bonded atoms lying beyond the fragment radius) had hydrogen atoms added to satisfy the valencey. Where possible, the valency was maintained (i.e sp^2 vs sp^3 carbon atoms), however when this was not possible (see **Figure 4.3**), the best possible alternative was chosen. Where two equally acceptable possibilities exist, a choice was made, and consistency over all fragments was maintained. The multipole moments of the probe atoms were obtained from the largest r calculation, and the same multipole moments are used to probe the central atom at all values of r.

4.3. Computational Details

Crystal structures for all three systems were downloaded from the protein data bank, and hydrogen atoms were added to the structures by HAAD[134]. The in house code MOROS extracted the protein fragments. All *ab initio* calculations were carried out using the Gaussian09 software package at the B3LYP/cc-pVDZ level of theory. Atomic multipole moments were calculated by AIMAll, and interaction energies were calculated by the in-house software NYX. Unless otherwise stated, all images were created using MOE.

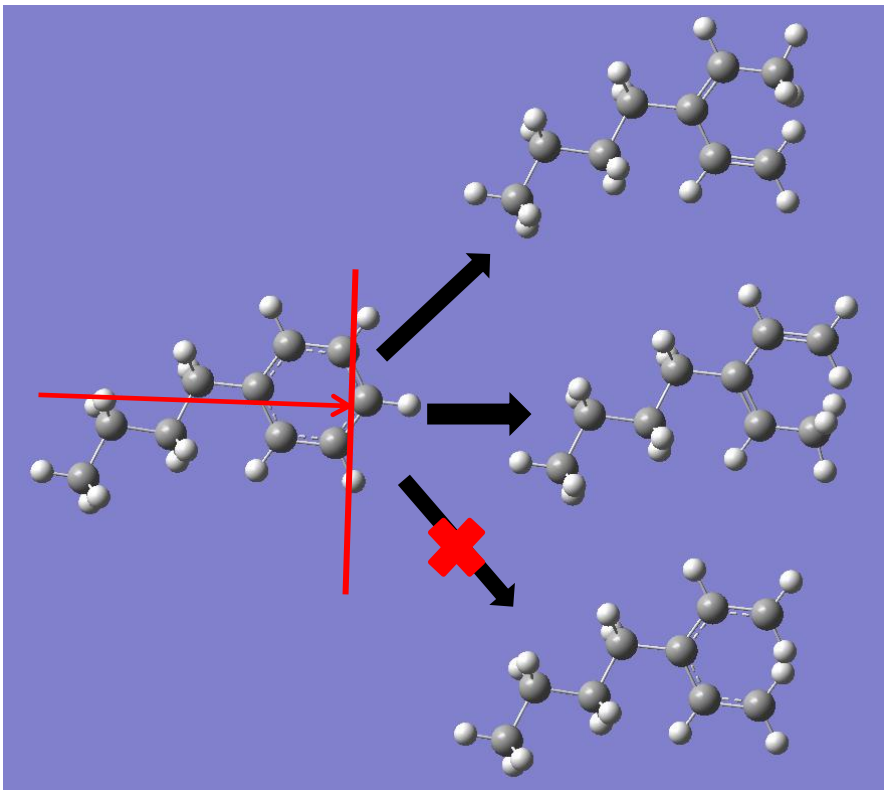


Figure 4.3: Illustration of where multiple possible capping alternatives exist. It can be seen that the rightmost carbon atom of the benzene ring lies outside of the fragment radius (shown as a *red line*). The top two capping possibilities on the right are both acceptable, despite one carbon in each case going from sp^2 to sp^3. Despite keeping hybridisation for both carbons constant, the bottom possibility is unfeasible, as the valence of at least one carbon is not fully satisfied due to the requirement of a double bond. Image drawn in GaussView.

4.4. Results

4.4.1 Crambin

Fragments of the protein crambin have been built around the amide nitrogen (N_{amide}) and amide oxygen (O_{amide}) of Phe_{13} (see **Figure 4.1**), and have been probed by three probe atoms, P, at the edge of the horizon sphere. The multipole moments and distance from the both N_{amide} and O_{amide} are given for each of the probes in **Table 4.1**. Plots of the electrostatic interaction energy between N_{amide} and the three probe atoms can be seen in **Figure 4.4**. Because it is only the moments of N_{amide} that are changing, all three graphs show the same overall shape as one another. The difference between the three graphs lies in the strength of the interaction between N_{amide} and the probe.

Table 4.1: The distance between the probe atoms and the amide nitrogen and oxygen atoms of Phe_{13} of crambin, and the values used for the multipole moments of the probe atoms

Probe, P	$r_{N\ldots P}$ /Å	$r_{O\ldots P}$ /Å	Q[0]	Q[1]	Q[2]	Q[3]	Q[4]
H279	10.71	13.5	-0.01E+00	-2.58E-02	-7.20E-02	-4.75E-02	7.75E-03
				-6.35E-02	-2.82E-03	5.43E-02	-1.79E-02
				1.27E-01	1.07E-02	1.25E-01	5.13E-02
					-5.30E-02	2.02E-02	1.69E-01
					-1.67E-01	-1.86E-03	-1.52E-01
						-1.79E-01	3.84E-02
						-8.86E-03	1.80E-03
							1.66E-01
							4.59E-02
N296	9.4	10.1	-1.11E+00	1.36E-01	6.68E-01	5.58E-01	-1.40E+00
				1.39E-02	-2.21E-02	3.06E-01	-8.58E-01
				-8.30E-02	-9.36E-02	-6.57E-01	7.64E-01
					-8.03E-01	1.84E-01	4.46E-01
					-1.12E-01	-2.65E-01	2.98E+00
						6.86E-01	-3.15E+00
						4.43E-01	-1.54E+00
							-1.08E+00
							5.18E-01
O166	9.8	7.5	-2.12E+00	-1.05E-02	-3.80E-01	-2.70E-01	1.15E-01
				-2.10E-01	1.87E-02	-2.79E-01	-2.38E-01
				1.01E-01	-8.36E-02	9.01E-02	5.50E-01
					2.08E-01	-5.95E-03	-4.55E-01
					3.25E-01	-2.80E-01	9.94E-01
						5.63E-02	-1.76E-01
						-3.06E-01	-2.57E-01
							4.55E-01
							-1.49E+00

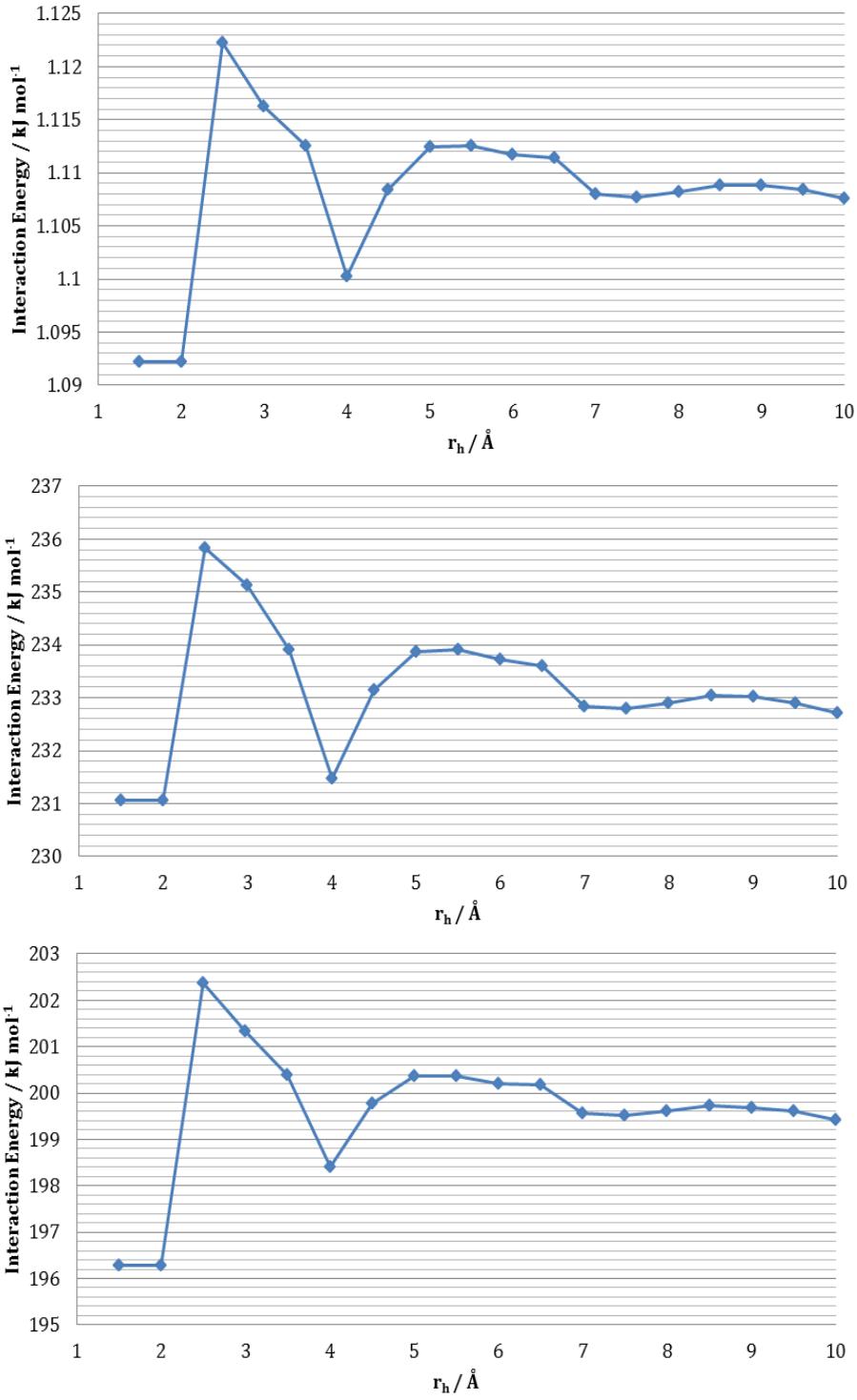


Figure 4.4: Plots of electrostatic interaction between the amide nitrogen of Phe_{13} and three probe atoms, H279 (*top*), N296 (*middle*) and O166 (*bottom*).

All three graphs appear to have reached a steady value around 7 Å. Thus, the horizon sphere of N_{amide} is defined as 7Å. The weakest probe interaction is that of N_{amide}…H279 and can be considered "converged" from the 1.5Å fragment onwards, as the interaction energy changes by less than 0.1 kJ mol^{-1} from the 1.5Å fragment for all other fragments of a greater size. The strongest probe interaction is that of N_{amide}…N296. N_{amide}…O166 is similar in magnitude to N_{amide}…N296. The graphs show a "jump" in the interaction energy between

values of r_h = 6.5 Å and 7 Å. Inspection of the fragments of crambin at these values of r_h, show that two sulphur atoms involved in a disulphide bridge are present on the edge of the 7 Å fragment but absent at 6.5 Å. These atoms have a significant polarising effect on N_{amide}. Although converged according to our set criterion, the r_h= 10 Å fragment shows another jump, although smaller than the jump between 6.5 Å and 7 Å..

Multipolar interactions between atoms A and B become increasingly short ranged as the expansion rank of the interacting moments, l, increases. A $1/r_{AB}^{L}$ dependence (where L is the interaction rank, defined by $L = l_A + l_B + 1$) is observed. The electrostatic interaction between N_{amide} and O166 at different values of maximum L can be seen in **Figure 4.5**. As expected for an interaction of a distance of 9.8 Å the charge-charge interaction ($L = 0 + 0 + 1 = 1$) dominates. Although the shape of the graph is dictated by the charge-charge term, the higher order interactions, in particular the charge-dipole interactions included when $L = 2$, act to scale the interaction. The difference between the $L = 5$ and $L = 1$ interactions are given in **Table 4.2**. After 8.5 Å the difference remains constant, around 1.05 kJ mol^{-1}. In the later example of TGT, the effect of L on interaction energies at different r_h is discussed in more detail.

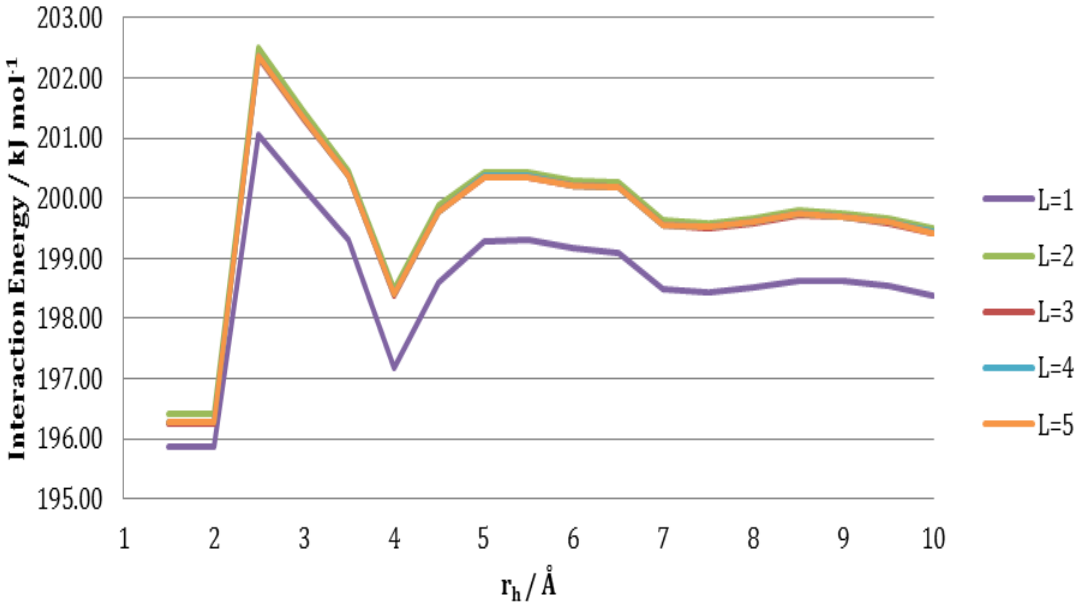


Figure 4.5: The electrostatic interaction energy between the amide nitrogen of Phe_{13} with the O166 at different interaction ranks for a number of r_h. The lines for $L = 3$ and $L = 4$ lie below the line for $L = 5$.

Table 4.2: The electrostatic interaction energy in kJ mol^{-1} at different rank L between the amide nitrogen of Phe$_{13}$ and the O166 probe atom for different values of r_h. The difference between the $L = 5$ and $L = 1$ energy is provided in the right hand column.

r_h	$L = 1$	$L = 2$	$L = 3$	$L = 4$	$L = 5$	$L = 5$ - $L = 1$
1.5	195.86	196.40	196.25	196.27	196.27	0.41
2	195.86	196.40	196.25	196.27	196.27	0.41
2.5	201.05	202.50	202.35	202.37	202.37	1.32
3	200.16	201.44	201.31	201.33	201.33	1.16
3.5	199.32	200.46	200.37	200.38	200.38	1.06
4	197.19	198.50	198.39	198.40	198.40	1.21
4.5	198.61	199.87	199.76	199.78	199.77	1.16
5	199.29	200.44	200.35	200.36	200.36	1.07
5.5	199.32	200.44	200.35	200.36	200.36	1.04
6	199.16	200.29	200.20	200.21	200.21	1.05
6.5	199.10	200.27	200.18	200.19	200.19	1.09
7	198.48	199.63	199.55	199.56	199.56	1.08
7.5	198.43	199.59	199.51	199.52	199.52	1.09
8	198.52	199.67	199.59	199.60	199.60	1.09
8.5	198.64	199.81	199.72	199.73	199.73	1.10
9	198.63	199.76	199.68	199.69	199.69	1.06
9.5	198.55	199.67	199.60	199.61	199.61	1.06
10	198.39	199.49	199.42	199.43	199.43	1.04

The effect of increasing r_h on the magnitude of the atomic moments of N_{amide} is now discussed. The magnitude of the atomic monopole, dipole, quadrupole, octopole and hexadecupole moments of N_{amide} plotted against r_h can be seen in **Figure 4.6**. The monopole moment curve reproduces the (inverse) shape of the probed interaction energies observed in **Figure 4.4**. It is the interactions involving the monopole moment that dominate the total probed interaction energy (due to the $1/r^L$ dependence of the interaction). It is, therefore, not surprising that both the interaction of N_{amide} with a probe and the monopole curves show similar shape. With the exception of the dipole moment, all other moments have converged to reache a stable value by ~7 Å. The dipole moment remains stable from r_h= 5 Å up to 8.5 Å, until a small increase between r_h=9 Å and 10 Å. The change corresponds to <3% of the total dipole moment so it will have a largely insignificant impact on the electrostatic interactions in which it takes place. The large jump in dipole moment between r_h= 4.5 Å and 5 Å corresponds to two nitrogen atoms being introduced to the system, one directly above and one directly below N_{amide}. These polarise the amide nitrogen, flattening it.

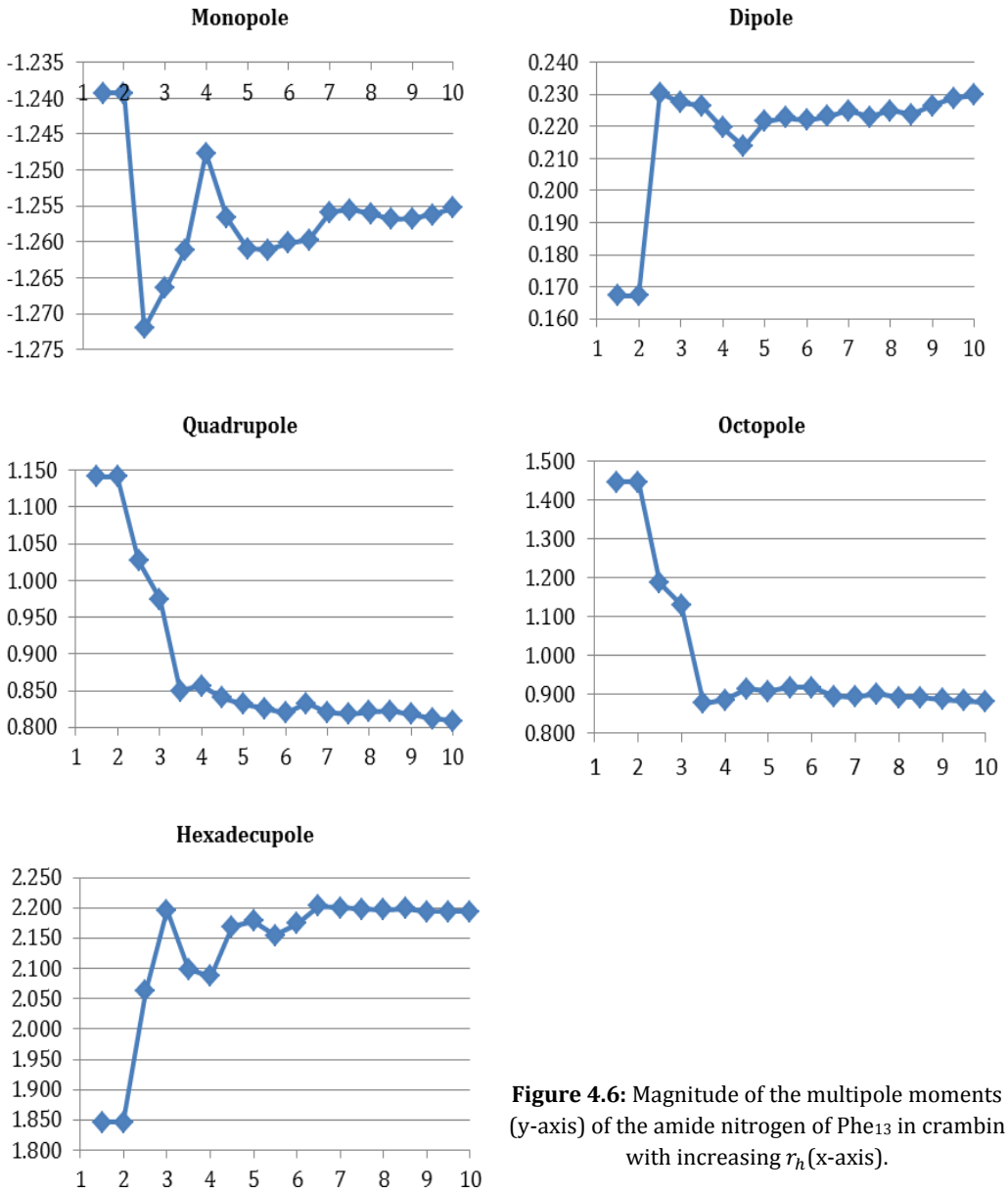


Figure 4.6: Magnitude of the multipole moments (y-axis) of the amide nitrogen of Phe_{13} in crambin with increasing r_h (x-axis).

A horizon sphere has also been built around the amide oxygen (O_{amide}) in Phe_{13}, and the probed interaction energies can be seen in **Figure 4.7**. The interactions exhibit a much smoother behaviour than the amide nitrogen interactions. The interactions between O_{amide} and the probes clearly have plateaued at r_h= 9 Å. The interaction with the O166 probe is particularly strong (nearly 250 kJ mol^{-1}) due to the short range (7.5 Å) of the interaction. Despite this, the difference in interaction energy between r_h= 9.5 Å and r_h= 10 Å is only 0.29 kJ mol^{-1}, which is less than 0.2% total interaction energy. There is a jump in the interaction energy between r_h= 6.5 Å and r_h= 7 Å, and this is due to the inclusion of a polarising sulphur atom when r_h= 7 Å. The jump is less significant for the amide oxygen that for the nitrogen due to the greater polarisability of the nitrogen atom. It is the "harder" nature of the oxygen atom that leads to the smooth shape of the curves in **Figure 4.8**. The effect of interaction rank on the interaction between O_{amide} and the O166 probe at

different r_h can be seen in **Figure 4.7**. As was observed for N_{amide}, increasing the interaction rank has no effect on the shape of the curve or the rate of convergence due to the dominance of the charge-charge interaction. Inclusion of the higher order interactions only shifts the interaction energy. The shift in interaction energy for O_{amide} ... O166 interaction is greater than for N_{amide} interaction (around 1.2% of the total interaction energy at r_h= 10 Å for O_{amide} and only 0.5% for N_{amide}). This is expected due to the shorter interaction distance meaning that the higher order interactions are more involved.

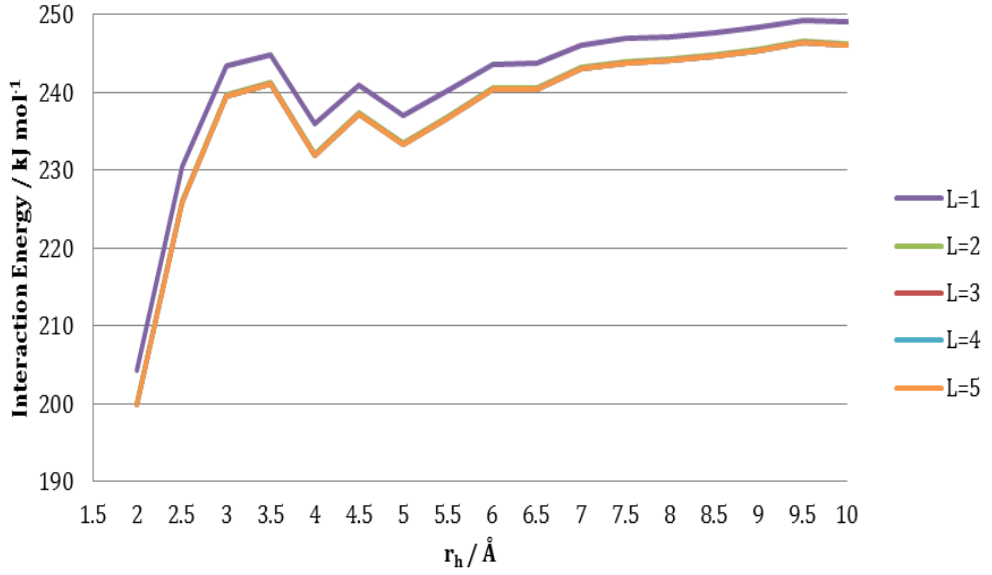


Figure 4.7: The electrostatic interaction energy between the amide oxygen of Phe_{13} with the O166 at different interaction ranks for a number of r_h. The lines for $L = 2, 3$ and 4 lie below the line for $L = 5$.

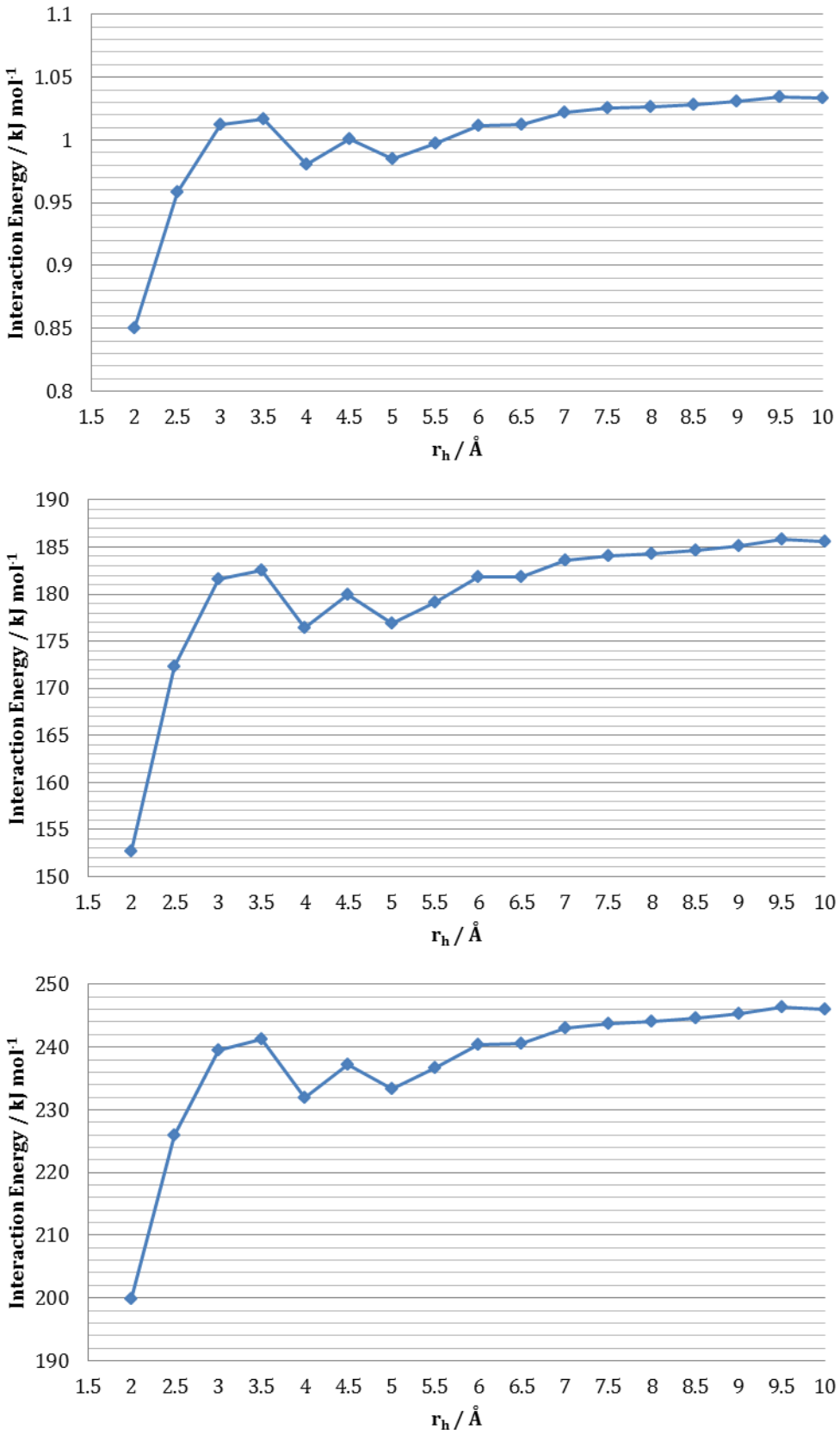


Figure 4.8: Plots of electrostatic interaction between the amide oxygen of Phe_{13} and three probe atoms, H279 (*top*), N296 (*middle*) and O166 (*bottom*).

The magnitude of the moments of O_{amide} have been plotted in the same way as for N_{amide}, and can be seen in **Figure 4.9**. The charge exhibits a gentle increase in magnitude from -1.15 a.u. to -1.21 a.u. between r_h=5-10 Å, however by r_h=6 Å the charge has already reached -1.18 a.u. This is the same shape the probe interactions for the same reasons as described for N_{amide}. Similarly the dipole moment changes by less than 0.20 Debye as r_h increases from 6 – 10 Å. Quadrupole, octopole and hexadecapole moments have all converged by 6 Å.

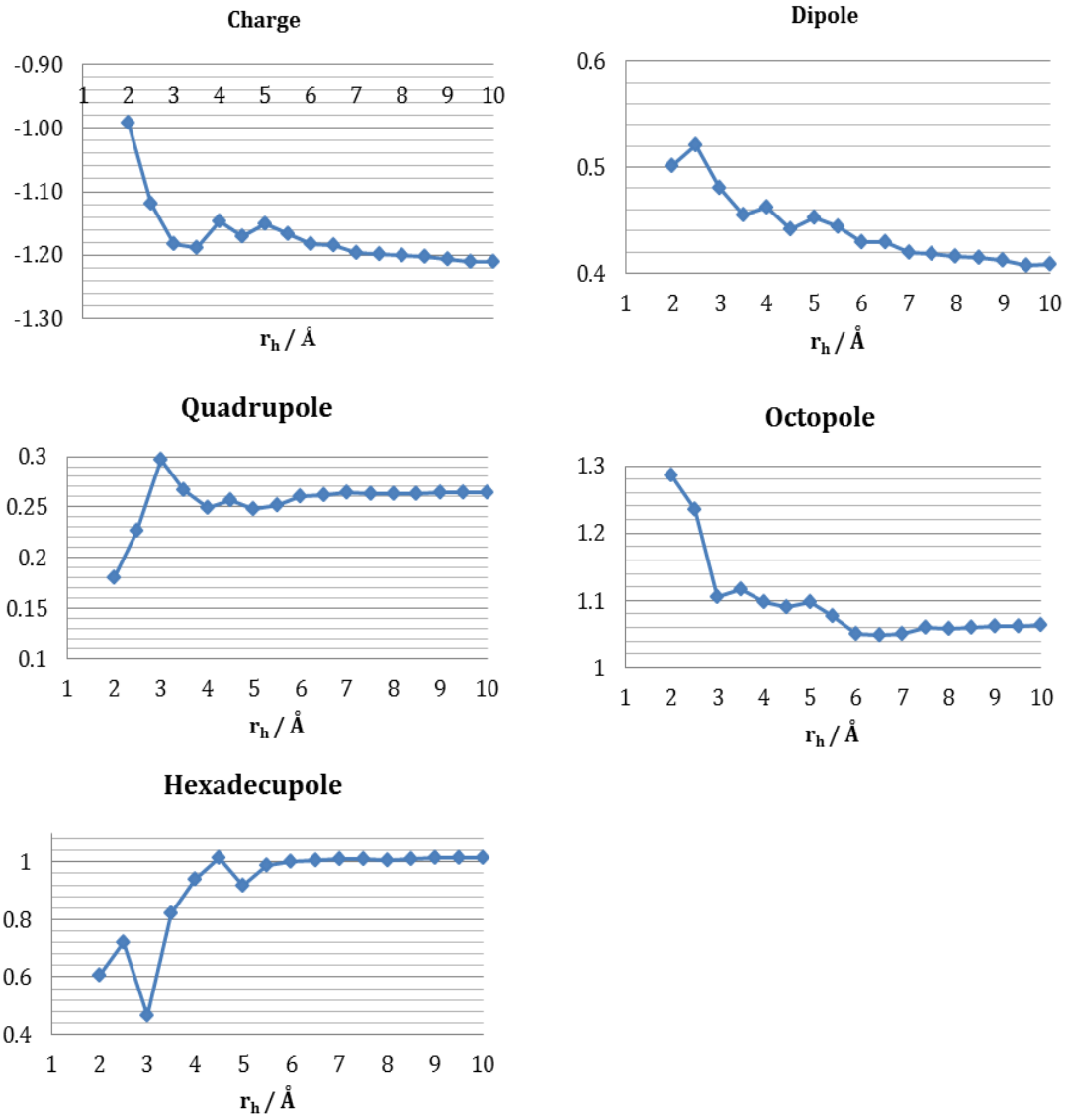


Figure 4.9: Magnitude of the multipole moments (y-axis) of the amide oxygen of Phe_{13} in crambin with increasing r_h (x-axis).

4.4.2 TGT-3,5-DAPH

A horizon sphere was built around a carbonyl oxygen of 3,5-DAPH involved in a hydrogen bond with the H-N of an active site glutamine side chain (H137). H137 was used as the probe atom, with the multipole moments calculated for H137 in the largest horizon sphere fragment used. **Figure 4.10** shows the interaction energy of the O…H137 interaction with

increasing values of interaction rank L. The separation between the lines in **Figure 4.10** is much more apparent than in **Figures 4.5 and 4.7**. This is because the interaction is much shorter and hence the higher order interactions play a greater role (due to the $1/r^L$ dependance). The separation between the lines is increasingly smaller as L increases. Despite the greater separation of the lines in **Figure 4.10**, the overall shape of each line is the same due to the dominance of the charge-charge interaction. The $L = 5$ interaction appears to have reached a steady value at $r_h = 6.5$ Å which is shorter than the crambin amide oxygen investigated above. It is at this value of r_h that the monopole moment of the carbonyl oxygen begins to converge (**Figure 4.11**).

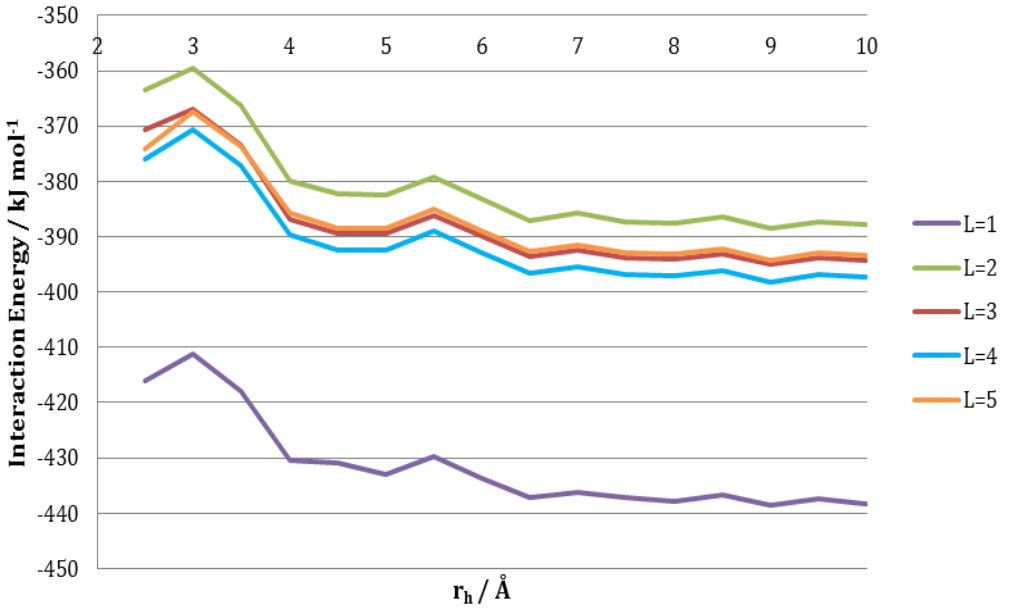


Figure 4.10: The electrostatic interaction energy of the O…H137 interaction at different interaction ranks for a number of r_h.

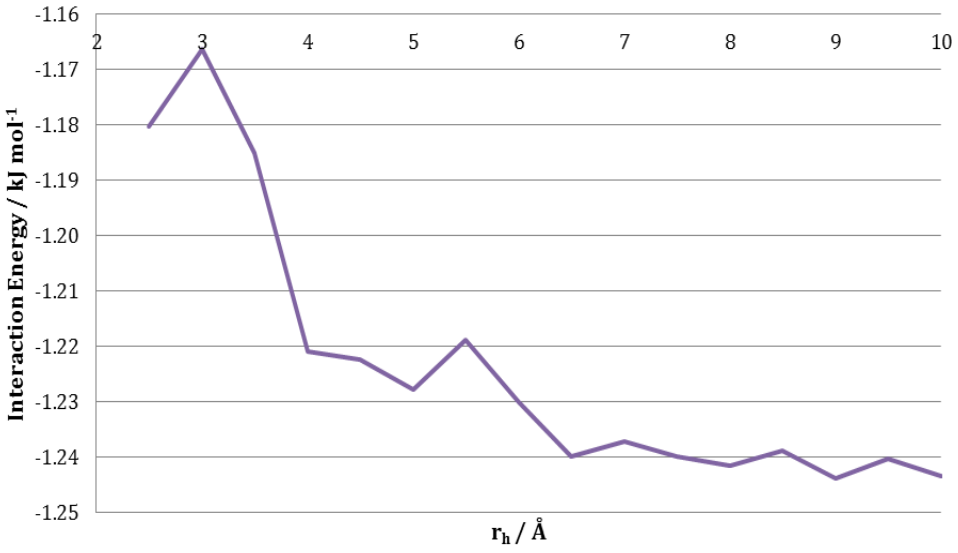


Figure 4.11: The monopole moment of the central carbonyl oxygen in the horizon sphere experiment for TGT at different values of r_h.

4.5 Conclusions and Further Work

Horizon sphere experiments have been performed on a number of atoms inside proteins and the central atoms have been probed. The interaction energies are dominated by the charge-charge interaction, and it is this interaction that dictates the convergence of an interaction between two atoms. Higher order multipole moments appear to converge quicker than the lower order moments, typically converged by r_h =5-6 Å. The charge takes longer to converge, with the smallest observed $r_h$value being 6.5 Å. Oxygen atoms appear to converge sooner than nitrogen atoms, however more tests are needed to confirm this. The inclusion of polarising atoms such as sulphur, oxygen and nitrogen atoms at the edge of a horizon sphere fragment can cause large changes in the multipole moments of the central atom relative to smaller fragments in which they are not present.

The above observations all indicate that kriging models built from gas phase calculations of capped amino acids are not suitable for direct implementation into an MM force field. Three lines of investigation must be pursued for the horizon sphere problem to be answered:

1. The building of a greater number of horizon sphere fragments to obtain a more solid value of r_h for different atom types.
2. The effect of strongly polarizing atoms such as nitrogen, oxygen, sulphur and metal ions on the multipole moments of the central atom.
3. A comparison of the multipole moments of a given atom type in a protein to those obtained from gas phase calculations of the amino acid dipeptide, and the moments obtained from an amino acid dipeptide under a number of different solvation models.

Other work in the group related to the horizon sphere is taking place regarding liquid water. Because proteins are typically solvated in aqueous media, the work on water may have implications towards this work.

Another interesting avenue of research that requires investigation is that of the horizon spheres of the IQA self and interaction energies. The computational cost of IQA calculations on large fragments is very high and so this project has not been possible in the current work.

Chapter 5

Where does charge lie in amino acids? The Effect of Side Chain Protonation State on the Atomic Charges of Asp, Glu, Lys, His and Arg

Summary

Quantum topological atomic charges have been calculated at the B3LYP/apc-1 level to identify where the charge is located on amino acid residues when the side-chain has been either protonated (Arg, Lys, His) or deprotonated (Glu, Asp). All local energy minima in the Ramachandran map of each (neutral) amino acid were populated with a number of distorted molecular geometries, summing up to a thousand geometries for each amino acid. The majority of the molecular charge is found on the side-chain (81-100%), with a large percentage of the charge located on the functional group undergoing protonation/deprotonation. Side-chain methylene groups were found to act as insulators for the amino acid backbone by accepting the majority of charge not located on the functional group. This results in no significant charge on backbone atoms relative to the neutral molecule. In the case of His^+ and Arg^+ where the charge is spread over a large number of atoms due to resonance, this reduces the influence of the positive charge on the backbone atoms.

A note:

Much of the work in this chapter has been published in

*"T.J. Hughes and P.L.A. Popelier, Computational and Theoretical Chemistry, 1053, (**2015**), 298-304"*

*found in **Appendix F**. The work presented in this chapter contains only my own work, with all contributions from co-authors omitted.*

5.1 Introduction

The complex mechanisms of enzymatic catalysis have been studied intensively for decades. A common feature in these mechanisms is the protonation and deprotonation of the active site amino acid side-chains involved in the catalysis. For example, the rate limiting step in the conversion of $CO_2 + H_2O \rightarrow HCO_3^- + H^+$ by the enzyme carbonic anhydrase is a proton transfer involving the residue His64 [200, 201]. Similarly, a proton transfer mechanism involving a Glu residue in the active site is found to be the rate-determining step in the mechanism of the enzyme glutaminylcyclase[202]. The subtle changes in the electronic charge of the active site atoms of glutaminylcyclase play a role in determining the path that the reaction follows. This effect arises through strengthening of hydrogen bonds within the active site upon proton transfer. The mechanism employed by enzyme horseradish peroxidase includes a nucleophilic attack by the hydroxyl oxygen of Ser195. However, this step requires activation through the deprotonation of the hydroxyl group[203]. Deprotonation results in the charge of the oxygen atom becoming more negative and hence more nucleophilic.

The above examples show that when developing a computational model to describe enzymatic reactions, any changes in electronic structure must be captured. Early potentials that enabled the modelling of reactions include the empirical valence bond approach[204] and the "ReaxFF" force field[205]. The popularity of QM/MM approaches is increasing in the study of such systems due to increases in computer power[206] . Currently under development in our lab is the quantum chemical topological force field (QCTFF). This is a novel approach to building a molecular mechanics force field, in which machine learning is used to map quantum mechanical properties (such as atomic multipole moments[141, 159, 207], kinetic energy[161] and exchange-repulsion) directly to the coordinates of the system. Preliminary work has shown that this methodology enables the modelling of changes in atomic charge as a reaction path is followed.

There is a perhaps surprising lack of literature detailing the changes in the atomic charges of amino acids upon a change of the side-chain protonation state, with studies [208-211] typically focusing on the zwitterionic states of amino acids. To address this gap in the literature, a thousand geometries for each of a total of five amino acids that most commonly undergo changes in protonation state (Asp, Glu, His, Lys, Arg) have been sampled for both the protonated and deprotonated state, and the changes in average atomic charges have been compared. In this work, charges have been obtained from the Quantum Theory of Atoms in Molecules (QTAIM) [22, 23]. The extensive QTAIM work[212-214] of Matta and Bader on all natural amino acids, provides a rich background to the current work but does not specifically address the question of where an excess or depletion of a formal unit charge resides compared to the neutral amino acid.

There are many methods of obtaining atomic charges but the question of which protocol produces the "best" atomic charges is contentious. Arguments for and against the different charge methods typically fall into one of two competing schools of thought. The first is a belief that the atomic charge should be capable of reproducing the electrostatic potential around an atom, and the second being that the charge should describe well the charge transfer in a molecule. The Hirshfeld charge[215] is an example of a charge that reproduces well the electrostatic potential around an atom, however it offers poor description of charge transfer effects, resulting in atoms with unrealistically low charges. The improved "iterative" Hirshfeld charge (Hirshfeld-I) method corrects for this to some extent[216]. QTAIM charges fall into the second category of charges- reproducing well the charge transfer in a molecule. This has led to QTAIM charges being been criticised for being unrealistically high[217]. The criticism that QTAIM charges do not reproduce the electrostatic potential is remedied by performing a multipolar expansion (of which the QTAIM charge is the first term of the expansion, the monopole moment) where it was shown[218] that reproduction of the *ab initio* electrostatic potential was achieved at interaction rank L=5. To quote from this work, *"This work makes clear that the atomic population (or rank zero multipole moment) is just one term of the expansion of a physically observable quantity, namely the electrostatic potential. Hence, QTAIM populations (and thus charges) cannot be judged on their reproduction of the electrostatic potential. Instead, they must be seen in the context of a multipolar expansion of the exact electrostatic potential of a topological atom."*.

5.2 Geometry Generation

Each amino acid was capped by a [$CH_3C=O$] group at the N-terminal, and by a [$NHCH_3$] group at the C-terminal to create the so-called "dipeptide". The minimum energy geometries for each neutral amino acid were obtained through a comprehensive search of the potential energy surface[219]. The number of energetic minima for each amino acid is given in **Table 5.1**.

Table 5.1: Number of local energy minima for each amino acid studied in this work.

Amino acid	No. Minima
Asp	36
Glu	36
His	24
Lys	39
Arg	61

A thousand geometries for each amino acid were obtained by normal modes sampling using TYCHE (see **Section 1.7**). All charged amino acid residues except Arg were obtained by direct addition or removal of a proton on the side-chain of the distorted geometries. For each of the thousand sampled neutral Asp and Glu residues, the acidic proton was removed

in order to obtain the geometries of the Asp^- and Glu^-, respectively. A similar approach was also taken in the case of Lys^+, where a proton was added to the primary amine to give the positively charged tetrahedral ammonium group. His^+ was similarly obtained by protonating the lone pair position of N29 to give the positively charged imidazolium group. Due to the more complex structural changes that take place in Arg upon protonation a different approach was taken in obtaining the Arg^+ geometries. In particular, the neutral Arg has a guanidine system with two pyramidal nitrogen atoms (N19 and N16) and one planar nitrogen (N34). However, in Arg^+ this group formally becomes a guanidinium group, which has three planar nitrogen atoms. The addition of a proton to N34 causes the geometrical change between guanidine and guanidinium. Therefore, an alternative approach was taken; a proton was added to each of the minimum energy geometries and then the guanidinium group alone ($[\text{-NH-C}(NH_2)_2]^+$) was allowed to relax by partial geometry optimisation. These new "minima" were then input to TYCHE to sample the thousand distorted Arg^+ geometries.

Appendix E contains a table that includes the average atomic charge, the range in atomic charge and the standard deviation in atomic charge of all atoms in both the neutral and charged amino acid systems studied in this work.

5.3 Computational methods

Normal modes sampling was performed by the in-house code TYCHE. The bond stretch parameters were set to ±10% max distortion from the equilibrium distance (see **Section 1.7**). All *ab initio* calculations were performed by GAUSSIAN09[137] at the B3LYP/apc-1[220] level of theory, taking advantage of a basis set with polarisation and diffuse functions optimised for use with density functionals. QTAIM charges for all atoms were calculated with the program AIMAll [221]. A topological representation of the protonated lysine can be seen in **Figure 5.1**, and this was obtained using MORPHY[222].

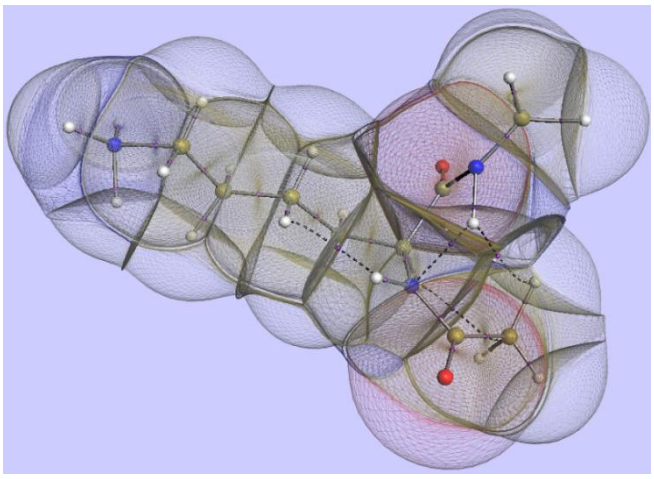


Figure 5.1: Finite-element representation of a molecular geometry of protonated lysine.

5.4 Results

Numbered geometries for all five protonated amino acids (Asp, Glu, Lys^+, His^+ and Arg^+) are provided in Fig. 5.2. For convenience, both the protonated and deprotonated amino acids share a common numbering system. The following discussion refers to the amino acids as consisting of both side-chain atoms and backbone atoms. The set of side-chain atoms consist of all atoms starting with C_β (including its methylene hydrogens), whereas the backbone corresponds to the C_α, the two peptide groups and the methyl caps, as well as all associated hydrogen atoms.

5.4.1 Acidic Amino Acids

The atomic charge (averaged over all thousand geometries) for all atoms of both Asp and Asp^- can be seen in **Figure 5.3.** The difference between the atomic charges in the neutral and in the charged amino acid is also plotted. Atom H25 is the acidic proton that is removed upon deprotonation. In the neutral molecule, the acidic proton has a charge of +0.56 a.u. (see **Figure 5.3**), which means that upon deprotonation a charge of (-1) + 0.56 = -0.44 a.u. is left over to be distributed over the remaining geometry. Of this 0.44 of an electron, 57% (-0.25 a.u.) moves onto the side-chain atoms. The remaining 43% (-0.19 a.u.) is found on the backbone atoms.

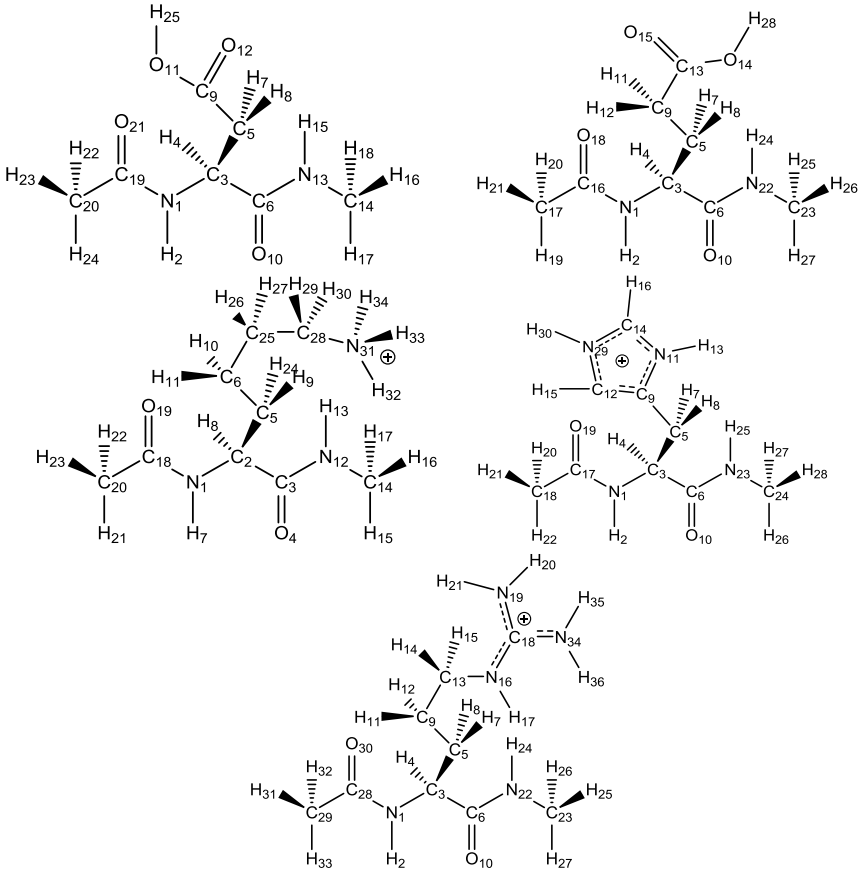


Figure 5.2: Numbered geometries for Asp (*top left*), Glu (*top right*), Lys^+ (*bottom left*), His^+ (*bottom middle*) and Arg^+ (*bottom right*). The numerical labels of the atoms ("atom number") of the deprotonated geometries are the same. In all five cases the proton removed upon deprotonation is the highest numbered proton.

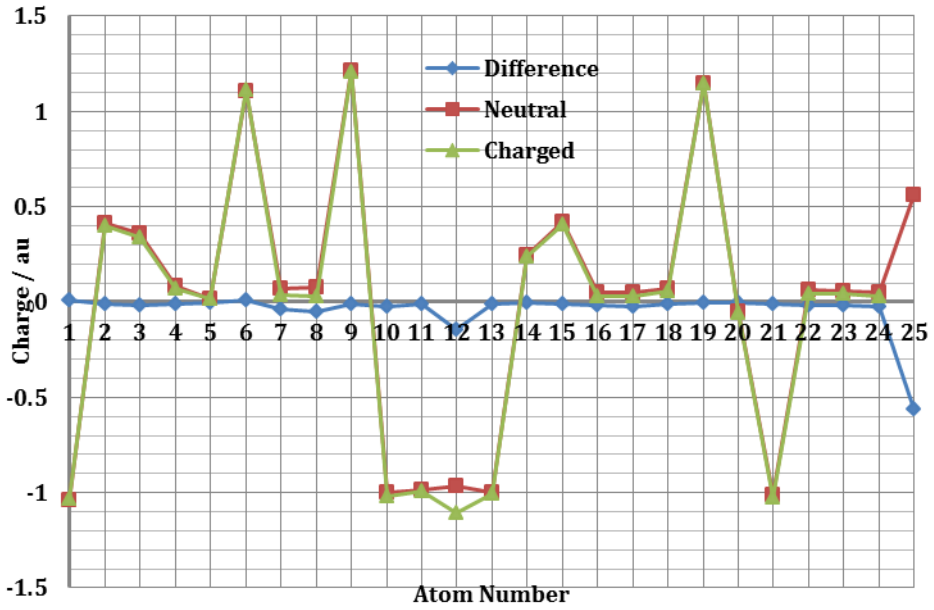


Figure 5.3: The averaged atomic charges of both Asp (*green*) and Asp$^-$ (*red*) and the difference (*blue*) between the neutral and charged atomic charges.

Despite the even spread of H25's charge over the whole molecule upon deprotonation, the total molecular charge is highly concentrated on the side-chain of the molecule of Asp$^-$. Upon deprotonation the sum of all side-chain atomic charges (including H25 for the neutral side-chain) decreases from -0.01 a.u. to -0.81 a.u. meaning that 81% of the total molecular charge is found on the side-chain atoms. The carboxylate group of Asp$^-$ has a summed charge of -0.89 a.u. (89% of the molecular charge). The methylene group of the side-chain increases in charge from Asp to Asp$^-$, with a summed (group) atomic charge of 0.08 a.u. (= |-0.89 – (-0.81)|). There are no chemically significant changes in atomic charges of the backbone atoms. Curiously, one of the most significant changes in backbone charge is that the hydrogen atoms on the methyl capping groups undergo a difference in summed charge of -0.10 a.u. when going from Asp to Asp$^-$.

Similar results are found for the deprotonation of Glu to Glu$^-$. The differences in average atomic charge over a thousand conformations are shown in **Figure 5.4**. Atom H28 corresponds to the acidic proton that is removed when going from Glu to Glu$^-$. The charge of H28 in Glu is 0.55 a.u. meaning that in Glu$^-$ only -0.45 a.u. of additional negative charge is available to the molecule for redistribution. A value of -0.33 a.u. of the additional charge (73%) remains on the side-chain atoms, and the remaining -0.12 a.u. is shared by the backbone atoms.

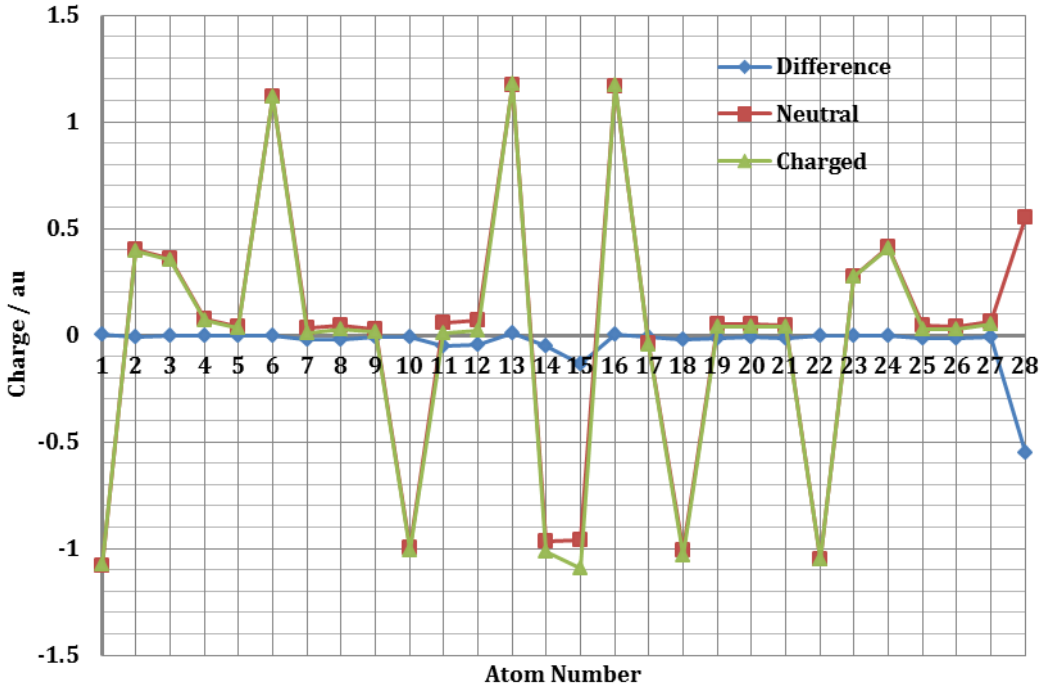


Figure 5.4: The averaged atomic charges of both Glu (*green*) and Glu^- (*red*) and the difference (*blue*) between the neutral and charged atomic charges.

Similar to Asp^-, it is apparent that the majority of the negative molecular charge of Glu^- is found on the side-chain atoms (-0.88 a.u., 88% of total molecular charge). A similar situation to that of Asp^- arises where the majority of the side-chain charge of Glu^- is concentrated on the carboxylate group. In Glu^- the carboxylate atoms have a summed charge of -0.93 a.u., which is an increase in charge of -0.73 a.u. relative to the summed charge of the neutral carboxylic acid group. There is no significant change in backbone atom charges. The methyl hydrogens increase in summed charge by -0.7 a.u., which is less than in the case of Asp^-.

There are differences between the changes seen in atomic charges for the two systems Asp^- and Glu^-. Eight percent more charge is located on the side-chain of Glu^- than on the side-chain of Asp^-. Also, less of the additional charge available upon deprotonation is found on the backbone atoms for Glu^- (26%) compared to Asp^- (43%). This observation has led to the idea of a "buffering" methylene group. Methylene groups are neutral fragments in the side-chain that act to separate the polar carboxylic acid/carboxylate group from the rest of the amino acid. The additional methylene group in the side-chain of Glu^- creates a more insulating buffer between the charged carboxylate group and the amino acid backbone. This buffering is responsible for the increased localisation of the charge on the side-chain in Glu^- than in Asp^-.

In summary, the deprotonation of the acidic hydrogen in Asp and Glu, causes the newly available negative charge to predominantly reside on the side-chain atoms (81% and 88% for Asp^- and Glu^-, respectively). In particular, the charge is localised on the three

carboxylate atoms (COO^-). Changes in the charge of backbone atoms, when going from the protonated to the deprotonated state, are insignificant due largely to "buffering" methylene groups. The buffering effect is greater in the case of Glu^- where there are two methylene groups.

5.4.2 Basic Amino Acids

Figure 5.5 shows the atomic charges of Lys and Lys^+. The acidic proton in Lys^+ (H33) has a charge of 0.48 a.u. This means that the atoms present in Lys undergo a sum increase in positive charge of 0.52 a.u. when going from neutral Lys to protonated Lys^+ (because 0.52 of an electron has moved onto H33). A positive charge of 0.44 a.u. (85%) is generated on side-chain atoms. As one would expect, the backbone atoms of Lys^+ remain relatively unaffected by the protonation of the amine group due to the four methylene groups "buffering" the ammonium group from the backbone. This explains the summed charge of the backbone atoms increasing by only (0.52 - 0.44 =) 0.08 a.u. upon protonation.

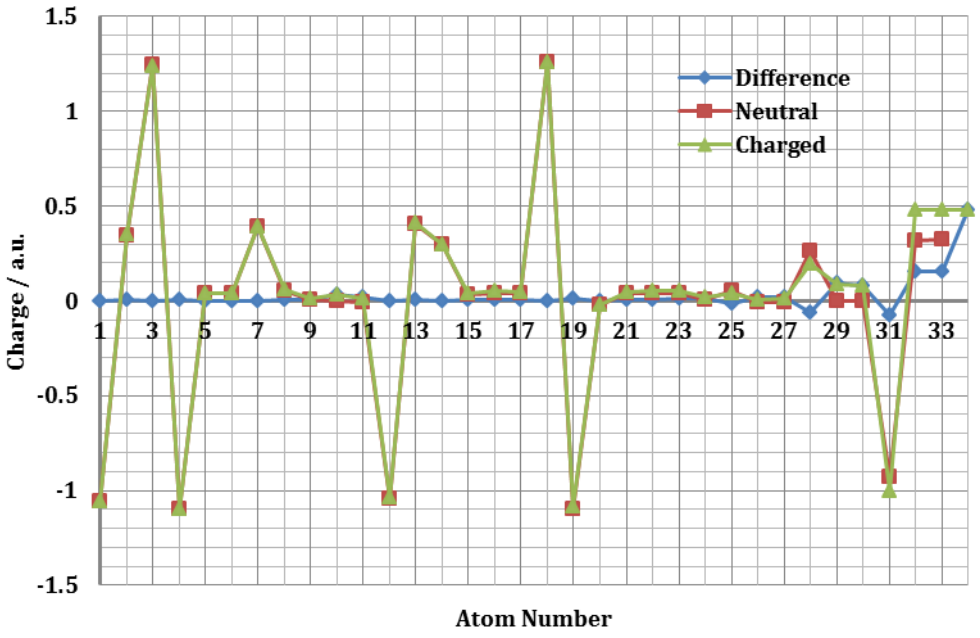


Figure 5.5: The averaged atomic charges of both Lys (*red*) and Lys^+ (*green*) and the difference (*blue*) between the neutral and charged atomic charges.

Fragmenting the molecule into side-chain atoms and backbone atoms and summing the atomic charges gives a clear illustration of the buffering effect. The summed charge of all side-chain atoms in Lys is 0.09 a.u., whereas in Lys^+ the summed charge is 1.01 a.u. (an increase of 0.92 a.u.), whereas the backbone atoms have a summed charge of -0.01 a.u. This shows that all of the positive molecular charge is found on the side-chain. The ammonium atoms ($[-NH_3]^+$) of Lys^+ have a summed charge of 0.43 a.u. , which is the largest contribution to the molecular charge. The remaining charge resides on the methylene groups. The summed charge of each methylene group is plotted in **Figure 5.6** against the number of covalent bonds between the carbon atom and the ammonium nitrogen. The

summed charge of the methylene atoms decreases as the number of covalent bonds away from the ammonium nitrogen increases. The summed charge of the methylene groups in the neutral Lys molecule are also plotted in F**igure 5.6**. From left to right, the gap between the neutral and charged values narrows, and by the fourth methyl carbon the difference between the charged and neutral methylene groups is only a summed charge of 0.02 a.u. This illustrates clearly the "buffering" effect of the methylene groups; the backbone atoms are almost unaware of the protonation of the amine group.

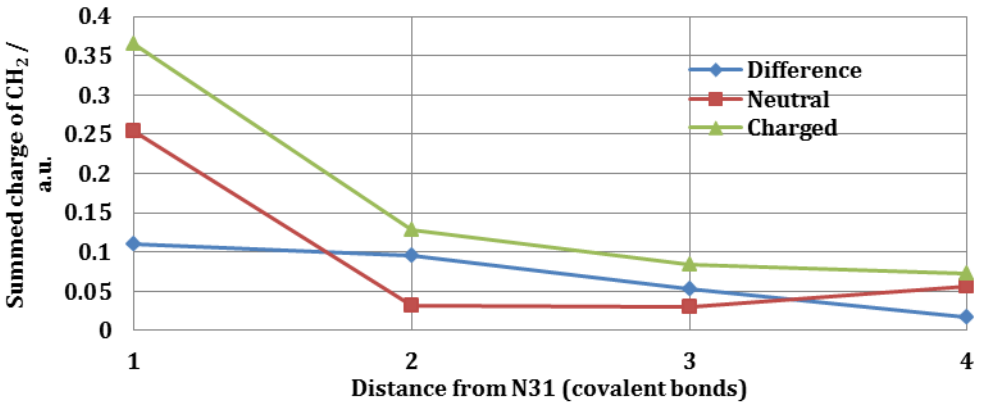


Figure 5.6: Summed charges of the methylene groups of Lys (*red*), Lys^+ (*green*) and their difference (*blue*) against the number of covalent bonds from the side-chain nitrogen atom (N31). (1=C_ε, 2=C_δ, 3=C_γ and 4=C_β).

The atomic charges of His and His^+ can be seen in **Figure 5.7**. The acidic proton of His^+ (H30) has a charge of 0.51 a.u. meaning that 0.49 a.u. of positive charge much be built up on the atoms present only in His (0.49 of an electron has moved onto H30). Of this charge, 76% (0.37 a.u.) lies on the side-chain atoms. The summed charge of the side-chain atoms is 0.93 a.u, which is 0.88 a.u more positive than the neutral side-chain. This again shows that the molecular charge is predominantly located on the side-chain, with the backbone atoms of His^+ undergoing a change in summed charge of 0.12 a.u. The only other amino acid that only has a single methylene group to protect the side-chain from the effects of side-chain protonation is Asp/Asp^-. The backbone atoms of Asp^- experience a greater change in summed charge (-1.9 a.u.). An incorrect assumption would be that the methylene (C5H7H8, Fig.2) in Asp^- is a worse "buffer" than the methylene (C5H7H8) in His^+. This is not true. Instead, in His^+ the positive charge is delocalised over the imidazolium and therefore its methylene group is no longer directly bonded to a charged atom but rather a group of atoms charged to a lesser extent. Thus, the methylene group in His^+ is only 0.07 a.u. more positive than the methylene in the neutral His, compared to a difference of -0.16 a.u. for the methylene of Asp and Asp^-.

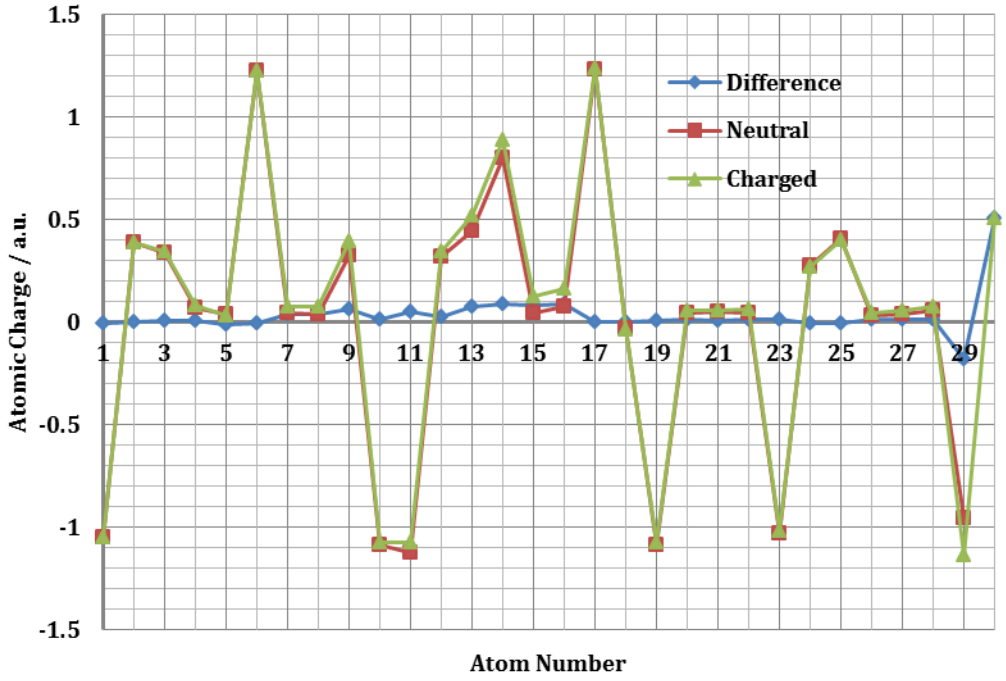


Figure 5.7: The averaged atomic charges of both His (*red*) and His$^+$ (*green*) and the difference (*blue*) between the neutral and charged atomic charges.

The atomic charges of Arg and Arg$^+$ can be seen in **Figure 5.8**. The acidic proton of Arg$^+$ (H36) has a charge of 0.48 a.u. meaning that 0.52 a.u. of positive charge is built up on the atoms present in the neutral Arg molecule. The side-chain atoms of Arg increase by a total summed charge of 0.44 a.u. when the proton is added, which accounts for 86% of the charge build up. The small contribution to this charge by the backbone atoms is due to a combination of the factors previously discussed. Firstly, there are three buffering methylene groups to separate the protonated guanidinium group from the backbone. The summed charge of the methylene groups can be seen in **Figure 5.9**. By the second methyl group (C_γ) the difference between the charged and neutral methylene groups is less than 0.05 a.u. The second reason for the low increase in backbone charge is that the positive charge is stabilised by the delocalised π-system of the guanidinium group. The eight guanidine atoms present in Arg account for 81% of the total side-chain increase in summed charge of Arg$^+$.

The side-chain atoms of Arg$^+$ have a summed charge of 1.01 a.u., accounting for all of the positive charge of the molecule. The backbone atoms have a summed charge of -0.01 a.u. due to the three "buffering" methylene groups and the spread of the charge over the guanidinium group. The guanidinium group has a summed charge of 0.45 a.u, which is the largest contribution to the molecular charge. The next largest contributor to the molecular charge is the methylene group adjacent to the guanidinium group, with summed charge of 0.41 a.u.

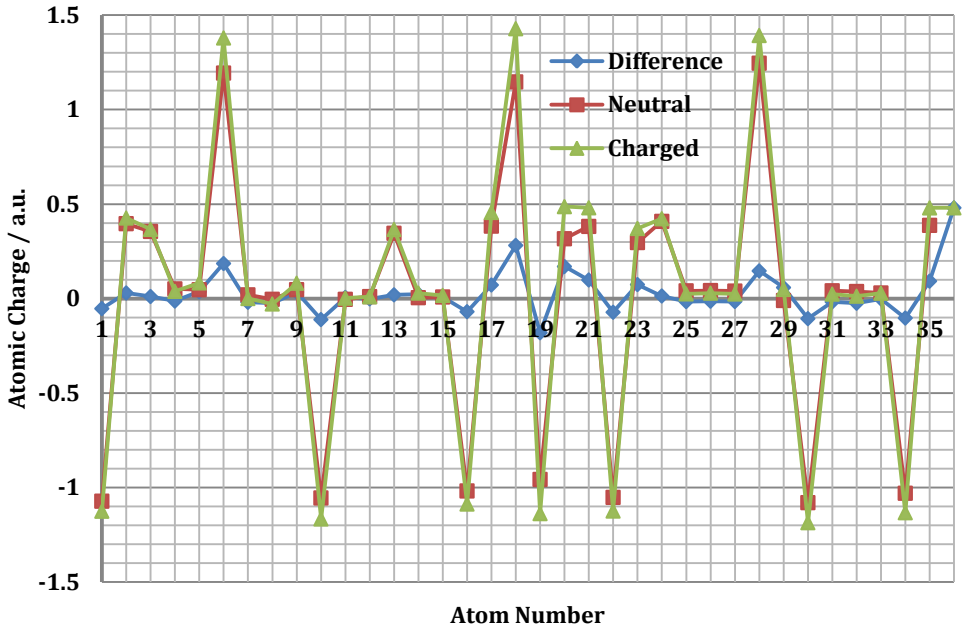


Figure 5.8: The averaged atomic charges of both Arg (*red*) and Arg^+ (*green*) and the difference (*blue*) between the neutral and charged atomic charges.

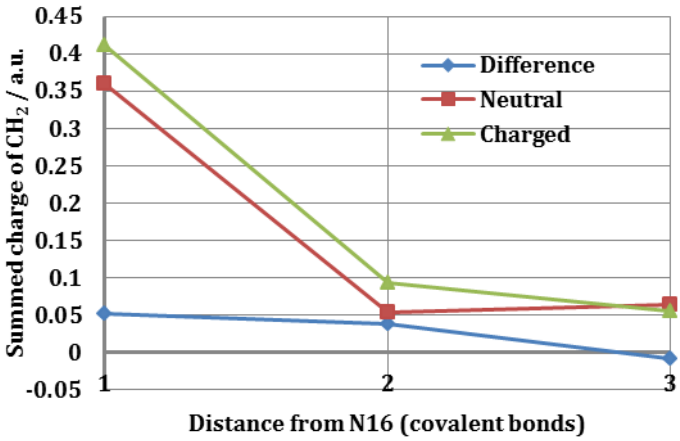


Figure 5.9: Summed charges of the methylene groups of Arg (*red*), Arg^+ (*green*) and their difference (blue) against the number of covalent bonds counting from the side-chain nitrogen atom (N16) (1=C_δ, 2=C_γ and 3=C_β).

5.5 Conclusions

The atomic charges of five amino acids that undergo protonation (Lys, His and Arg) and deprotonation (Asp and Glu) have been studied. The QCT atomic charges of all atoms, averaged over a thousand conformations, for both charged and neutral amino acids have been compared. For Asp and Glu, which are deprotonated to form Asp^- and Glu^-, the majority of the negative charge is located on the side-chain atoms (81% and 88% respectively). Less charge is found on the backbone of Glu^- than Asp^- due to the additional side-chain methylene group "buffering" the charge. The buffering effect of methylene groups is more apparent in the positively charged amino acids Lys^+, His^+ and Arg^+ due to the large number of methylene groups in Lys^+ and Arg^+. By the third methylene group counting from the site of protonation, the summed charge of the CH_2 group is comparable

to that of the neutral molecule. Spread of the charge over multiple side-chain atoms (such as in the imidazolium ring of His^+ and the guanidinium group of Arg^+) also reduces the effect of the charge on backbone atoms.

Chapter 6

Relating the IQA Exchange-Repulsion Energy to Classic Repulsive Potentials

Summary

The short-range repulsion between atoms and molecular fragments is considered within the context of QCTFF. Following a brief recap on the IQA energy decomposition, the IQA interpretation of the exchange-repulsion (XR) energy is calculated for a selection of small molecular systems (R-OH...HOH and R-NH_2...HOH where R=H, Me, Et, as well as serine...HOH and lysine...HOH) and it is then compared to a number of classical force fields. The concept of a transferable atom type is central to the ideology of classical force fields, and therefore transferability in the IQA XR energy is investigated. The XR energy consists of both a sum of pairwise v_{xc} interactions and also the deformation energy of the atoms/fragments, and it is the latter energy that prevents one from obtaining pairwise potentials between IQA atom types due to the definition of the deformation energy as a self-term, rather than an interaction. Despite this, the deformation energy is found to be within 6.4% of the R=H systems when R=Me and Et and so some transferability is observed.

6.1. Introduction

Short range repulsion between atoms is responsible for a number of chemical effects that are taken for granted by most chemists, most notably preventing all matter from falling into infinitely attractive potential energy wells. It also plays a major role in describing steric clashes between atoms, which has consequences ranging from reaction pathways to crystal structure formation. The quantum mechanical origin of short range repulsion arises as a consequence of the Pauli principle. This states the requirement that the wave function must be anti-symmetric with respect to exchanging electrons between molecular orbitals. As a consequence, no two electrons may occupy the same spin orbital. This effect gives rise to repulsive “exchange holes” around each electron that becomes increasingly repulsive as two electron approach one another.

As stated above, an important consequence of the atomic short range repulsion is that of the steric interaction. Being of such great importance, many models have been developed to describe the steric interaction between atoms in a system. Often, people include the electron-electron and nucleus-nucleus Coulombic repulsion in addition to the short range repulsion when discussing sterics, and so the reader is reminded that the short range repulsion and steric interaction are not always equivalent terms, however due to short range repulsion being a large contributor to the effect it is still considered relevant to warrant discussion in the present work.

Because the steric energy is a concept defined by chemists to explain the effects of short range repulsive interactions, there have been very few attempts to derive a "steric energy" from *ab initio* methods. One exception is the 2007 work of Liu, where a DFT derived definition of the steric repulsion between atoms was provided[223]. By assuming that the total DFT energy of a chemical system, $E[\rho]$, is a sum of electrostatic, quantum and steric contributions ($E_e[\rho]$, $E_q[\rho]$ and $E_s[\rho]$, respectively), the steric energy of a system is simply defined as the total DFT energy minus the electrostatic and quantum contributions to the total energy

$$E_s[\rho] = E[\rho] - E_e[\rho] - E_q[\rho] \tag{6.1}$$

The electrostatic term includes all nucleus-nucleus repulsion, all electron-electron repulsion and all electron-nucleus attraction, and so includes all the electrostatic interactions within the system. The quantum mechanical term, $E_q[\rho]$, is equal to the sum of the exchange correlation energy (defined by the exchange-correlation functional used), $E_{xc}[\rho]$, and the Pauli energy, $E_P[\rho]$ defined as the difference between the non-interacting kinetic energy and the Weizsäcker kinetic energy[224]:

$$E_P[\rho] = T_s[\rho] - T_w[\rho] \tag{6.2}$$

Which when substituting all the terms into **Equation 6.1** leads to the steric energy being equal to the Weizsäcker kinetic energy.

$$E_s[\rho] = T_W[\rho] = \frac{1}{8}\int \frac{|\nabla\rho(\boldsymbol{r})|^2}{\rho(\boldsymbol{r})} d(\boldsymbol{r}) \tag{6.3}$$

This partition successfully excludes all of the electrostatic energy from the steric energy, therefore providing a purely quantum mechanical value. The steric energy was then calculated for a number of small chemical systems to provide analysis of the steric contributions. For example the ethane rotation barrier was shown to be consequence of the steric energy being highly positive. This approach has since been applied to a number of model chemical problems where the steric interaction has been considered a key contributor, for example the anomeric effect in sugars[225], and the origin of more rotation barriers[226]. Despite the success of the method in providing quantitative results that agree with chemical intuition, the authors describe the steric effect as a "noumenon", stating "*there is no physically observable value associated with the steric effect and thus, it is an object , though chemically significant and conceptually relevant in understanding the behaviour of molecules, of purely rational apprehension and intellectual intuition*"[223].

As stated previously, the short range repulsion energy is a quantum mechanical effect and is a function of the molecular wave function. By this reasoning, the assumption that the repulsion between atoms is a pairwise interaction is an oversimplification. However, when building quick chemical potentials, such as MM force fields, the use of pairwise potentials has proven itself a successful method for the estimation of atomic repulsion. These pair-wise potentials typically take one of two forms, the first being $E = R^{-n}$ where n is an integer. The Lennard-Jones potential is one such potential and it is applied to chemical systems to model both short range repulsion and the attractive dispersion interaction. The full Lennard-Jones potential is provided in **Equation 6.4**, however the R^{-6} term may be overlooked as this relates to the attractive dispersion interaction.

$$E_{LJ} = 4\varepsilon_{ab}\left(\frac{\sigma_{ab}^{12}}{R_{ab}^{12}} - \frac{\sigma_{ab}^{6}}{R_{ab}^{6}}\right)$$

(6.4)

where ε_{ab} is the depth of the potential well, and σ_{ab} is the value of R_{ab} where $E_{LJ} = 0$. Many popular force fields use a Lennard-Jones potential to describe the short range repulsion between atoms including CHARMM[227], AMBER[12] and GROMOS[15]. The pairwise constants used in the Lennard-Jones potential are usually obtained through the use of mixing rules. These are functions used to calculate pairwise constants from monatomic constants corresponding to the atom-types being interacted. There are a number of mixing rules that may be used, however the most common are the Lorentz and Bethelot rules,

$$\varepsilon_{ab} = (\varepsilon_a \varepsilon_b)^{\frac{1}{2}} \quad , \quad \sigma_{ab} = \frac{1}{2}(\sigma_a + \sigma_b)$$

(6.5)

The theoretical reasoning underlying the Lorentz and Berthelot rules is weak, however they do yield reasonable pairwise constants at an insignificant computational cost. There are instances where the coefficients given by mixing rules yield poor results, particularly in simulations involving ionic species, however this is often "overcome" by modifying the monatomic coefficients [228, 229]. Despite the existence of more involved combination rules [230-232], little improvement is gained through their use, although some perform better for specific tasks[233]. The use of $n = 12$ in R^{-n} potentials is an arbitrary choice made only to produce a steeply repulsive potential that mimics repulsion. Other values for n may be used, for example, both Halgren's MMFF94[234] and the AMOEBA force field of Ponder et al[16, 193] use an R^{-14} potential with pair-wise constants obtained from mixing rules given by Halgren in 1991[18].

The second form of potential commonly used to describe the short-range repulsion between two atoms is an exponential function $E = \exp(-R_{ab})$ where R is the interatomic

separation. The Born-Mayer potential was an early implementation of such a potential where the interactions between two atoms a and b is given by

$$E_{rep} = A_{ab} e^{-B_{ab}\left(\frac{R_{ab}}{R_{ab_o}}\right)} \tag{6.6}$$

where A_{ab}, B_{ab} and R_{ab_0} are constants and R_{ab} is the internuclear distance. The pairwise constants are often obtained by combining monatomic constants through the use of mixing rules. The MM2[235], MM3[236] and MM4 [237] force fields of Allinger et al describe repulsion in this way.

The use of the functional forms $\exp(-R)$ and R^{-12} to model repulsion has no theoretical grounding (whereas, for example, the R^{-6} attractive term in the Lennard-Jones potential does roughly describe the induced dipole-induced dipole interaction between atoms). They are chosen only because they form a steep repulsive potential. The accuracy and wide ranging applicability of a molecular potential is determined by both its theoretical grounding and also the accuracy of the fitting procedures used to obtain constants. It is, therefore, desirable to find a potential for short range repulsion that has a theoretical foundation when developing new molecular potentials. The theory of interacting quantum atoms (IQA) [25], is a theoretically rigorous energy decomposition that breaks down the energy of a chemical system into atomic self and pairwise contributions, and therefore fits the above criterion as a basis for developing a novel potential. IQA is described in detail in the following section. A comparison of the IQA exchange repulsion (XR) energy with classical MM potentials is provided and then the concept of transferable IQA atom-types is investigated, whereby a discussion of potential mixing rules will be included.

6.2. DFT IQA Calculations

Until recently, the IQA decomposition has been restricted to the HF level of theory, however new releases of the AIMAll software package (versions 14.11.23 onwards) [138] have included algorithms capable of performing full IQA analysis on both the B3LYP and M06-2X density functionals. In the current work both HF and M06-2X wave functions are used. M06-2X has been developed by Truhlar *et al* [146] with the aim of improved description of intermolecular interaction by including some dispersion effects. The energy that the M06-2X functional aims to account for, which both HF and other density functionals fail to describe, is the electronic correlation energy. Thus V_x should be written in full as V_{xc} when considering the M06-2X functional. Unfortunately the correlation energy is a two-electron property whereas density functionals, including M06-2X, are one electron functionals and so approximations must be made in the IQA decomposition. It is for this reason that there is work both inside the group and by others to extend IQA to MP2 wave functions that include the dynamic two electron correlation effects. Until this is readily

available, the following assumptions concerning the IQA decomposition of density functionals have been made:

1. The inter-atomic exchange-correlation energy between two atoms is treated in an identical fashion for both HF and density functional theory (DFT) wave functions, $V_{xc}^{a,b}(DFT) = V_{x}^{a,b}(HF)$.
2. The DFT intra-atomic exchange correlation is the total exchange-correlation energy of atom a minus the HF exchange between atom a and all other atoms a', $V_{xc}^{a,a}(DFT) = V_{xc}^{a}(DFT) - V_{x}^{a,a'}(HF)$.

Although not wholly satisfactory due to the neglect of inter-atomic correlation energy, the above assumptions reproduce the total energy of the system (when used as part of a whole IQA decomposition) therefore accounting for the improved description of intermolecular interactions offered by the M06-2X functional.

6.3. Computational Details

All *ab initio* calculations were performed using Gaussian09[137] at either HF/6-31++G(d,p) or M06-2X/6-31++G(d,p) levels of theory. The topological analysis including the IQA analysis was performed by AIMAll version 14.11.23[138]. The default integration grids used by AIMAll in the presence of the "briaq=auto" and "boaq=auto" keywords gave erratic results, and so finer grids were employed using the "briaq=skyhigh" and "boaq=skyhigh" keywords. Although, the results presented in **figures 6 and 7** in **section 4.2** do not show smooth curves, they represent a marked improvement over the original results obtained using default parameters. The non-linear least squares fit used to obtain **equation 14** was obtained using a simple Perl script.

6.4. Results and Discussion

6.4.1. Fitting XR to Classical Potentials

First, we shall compare the overall shape of the IQA XR energy profile to the repulsive potentials included in the classical potentials AMBER, MM2, OPLSAA, MMFF94 and AMOEBA. **Figure 6.1** shows the total IQA XR energy of the water dimer as the two water molecules are brought towards one another at both the HF and M06-2X levels of theory. Also plotted on **Figure 6.1** are the repulsive components of the classic potentials listed above. It is immediately apparent that the IQA decomposition partitions energy in a much different way to classical force fields and so a quantitative comparison is not possible. The overall shape of the XR energy curves are much softer than the classical potentials, with the energy rising at a relatively more gradual rate over a longer distance. The classical potentials do not begin to introduce significant intramolecular repulsion until a H-bond distance of 2 Å. As previously stated, the overall shape of the two IQA XR curves is similar to the classical descriptions of repulsion. When fitted to Born-Mayer type potential by a ,

an average error of only 9 kJ mol^{-1} from the HF XR energy is obtained. The optimum potential is given below.

$$XR = 3401e^{-3.1\left(\frac{R_{ab}}{R_{ab_o}}\right)} \tag{6.7}$$

where R_{ab} is the OH…O separation and R_{ab_o} is the equilibrium separation (2.04 Å).

The M06-2X functional is parameterised to account for some electron correlation effects and this leads to a lower XR energy than that which is obtained at the HF level. The total interaction energy of the water dimer at the two levels of theory can be seen in **Figure 6.2**. The XR curve in **Figure 6.1** at the M06-2X level of theory is lower in energy than the HF XR curve and this is in agreement with inclusion of correlation effects. The difference in energy is particularly apparent at short H-bond distances, with a difference of 94 kJ mol^{-1} at a distance of 1.25 Å. The difference in total energy in **Figure 6.2** is much smaller than the observed difference in XR energy in **Figure 6.1** but because the XR energy is not the only IQA component to be affected by a change in the level of theory, the disagreement in energy difference gives no undue concern (for example, V_{cl}^{ab} is affected as the HF level of theory gives much more polar atoms than M06-2X giving a larger V_{cl} interaction for the HF results).

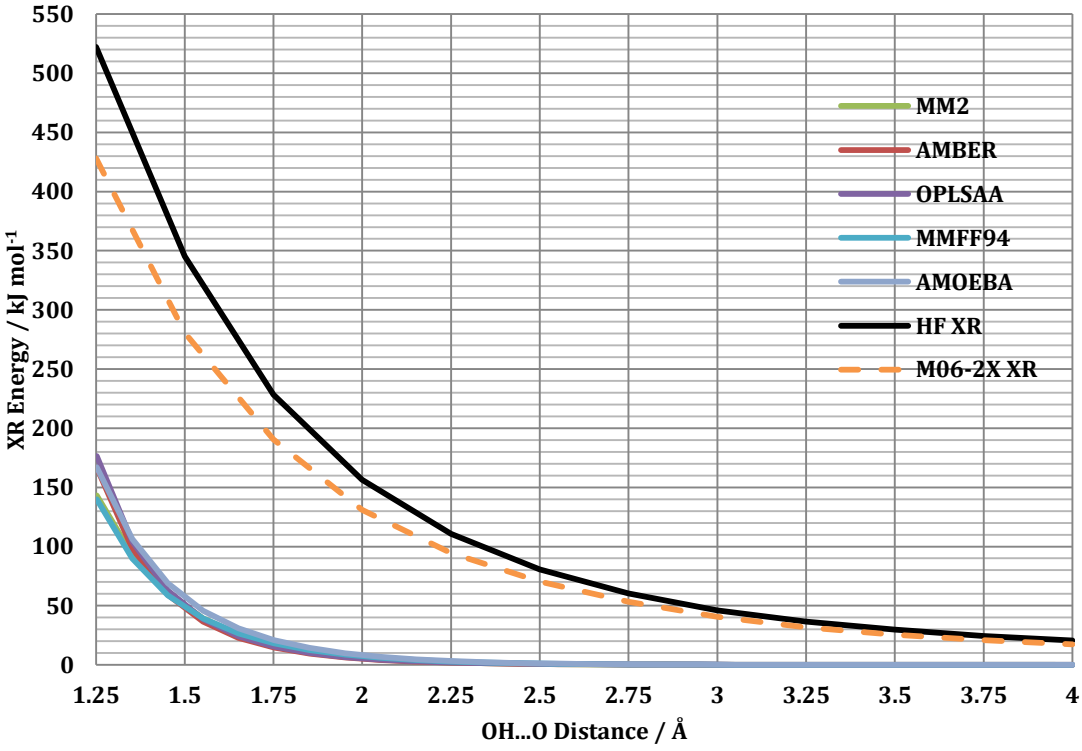


Figure 6.1: Comparison of the different repulsive potentials in popular force fields with the IQA XR energy at both the HF and M06-2X levels of theory.

Figure 6.3 breaks down the XR energy into its individual contributions and shows the difference in energies obtained from the M06-2X and HF levels of theory. Note that the V_{xc} value given in **Figure 6.3** corresponds to the sum of all nine intermolecular V_{xc} interactions in the water dimer. At all distances, >85% of the difference in the XR energy is accounted for by a change in the V_{def}^{a} energy. This initially gives rise to concern, as the inclusion of correlation effects would be expected to express itself the exchange-correlation term (V_{xc}). As stated at the end of the **Section 6.2**, however, only the calculation of the atomic self-exchange-correlation energy is different in the DFT IQA calculations. Because the self-exchange-correlation energy is a component of V_{def}^{a} and not V_{xc}^{ab}, the V_{def}^{a} should account for a large amount of the difference between the HF and M06-2X XR energies. The small difference in the V_{xc}^{ab} between the HF and M06-2X levels of theory seen in **Figure 6.3** is attributable to the different molecular electron density obtained at different levels of theory.

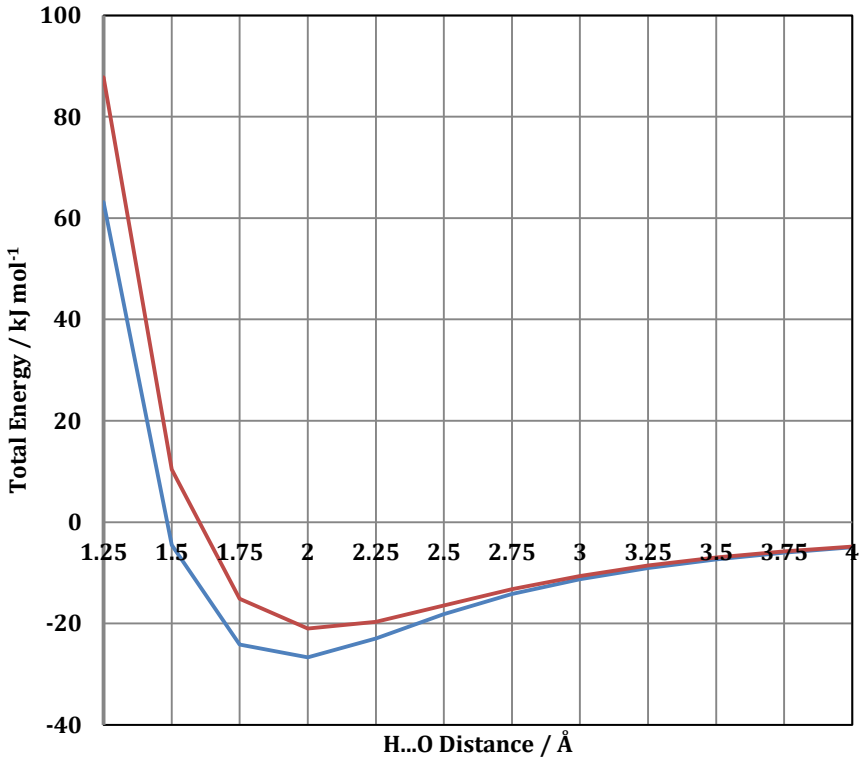


Figure 6.2: The total interaction energy of the water dimer against the H-bond separation relative to the energy at a 30 Å separation. Blue line obtained at the M06-2X level of theory and the red line at the HF level.

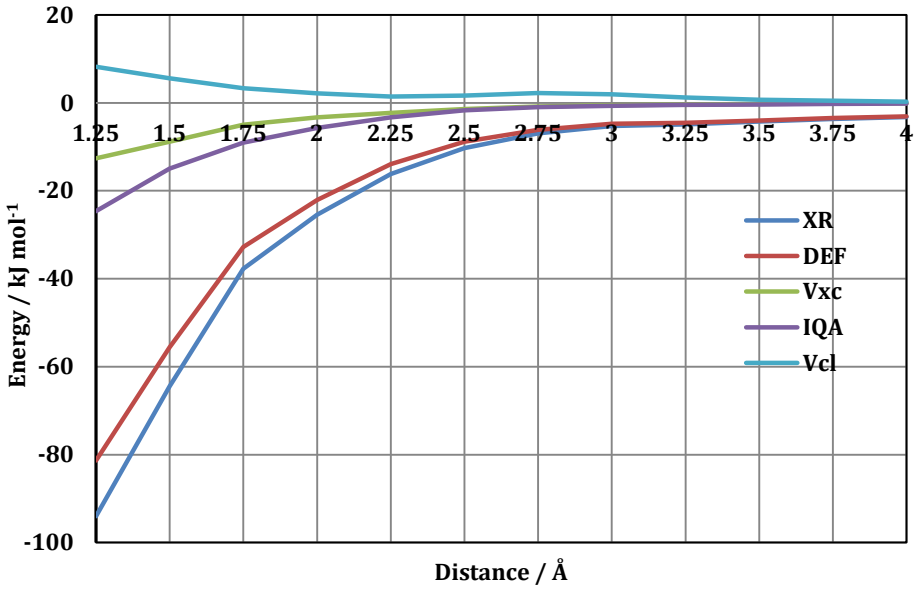


Figure 6.3: Plot of the difference between the M06-2X and HF values of a number of IQA energy terms and also the total IQA energy.

In conclusion, it is not possible to directly compare the IQA XR energy with the repulsive components of classical potentials, despite the roughly exponential form of the XR vs distance profile. The HF level of theory gives rise to a more repulsive XR energy than that obtained using the M06-2X functional as one would expect from the capturing of electron correlation effects in the M06-2X parameterisation. Misleadingly, the deformation energy obtained from M06-2X analysis is more affected than the V_{xc} term due to the DFT IQA algorithm, and so the HF level will be used exclusively from this point. The overall shape of the HF IQA is similar to that of M06-2X.

6.4.2. Does the XR Energy Produce Transferable Atom Types?

Classic repulsive potentials have been developed as a fast means of obtaining reasonable estimates of the energy at the expense of some chemical accuracy. In order to achieve this, the concept of a transferable atom type has been successfully developed over many years. In **Section 6.4.1** it was shown that the XR energy can be fit to a Born type exponential with an average error of 9 kJ mol^{-1}. Hence, the next question asked is "do transferable IQA atoms types allowing accurate reproduction of the XR energy exist?" In an attempt to answer this question, transferability of the atoms within two model systems has been studied. The first system is that of an amine nitrogen (R-NH_2) acting as a hydrogen bond donor to a water molecule, with a range of R groups. The second model system is that of a hydroxyl oxygen (R-OH) also acting as a hydrogen bond donor to a water, again with a range of R groups. Topological representations of all complexes studied in this work may be seen in **Figure 6.4**, and a numbered schematic of the two functional groups is provided in **Figure 6.5**. The changes in selected V_{def} and V_{xc}^{ab} energies at H-bond distances between 1.25 – 4 Å are analysed.

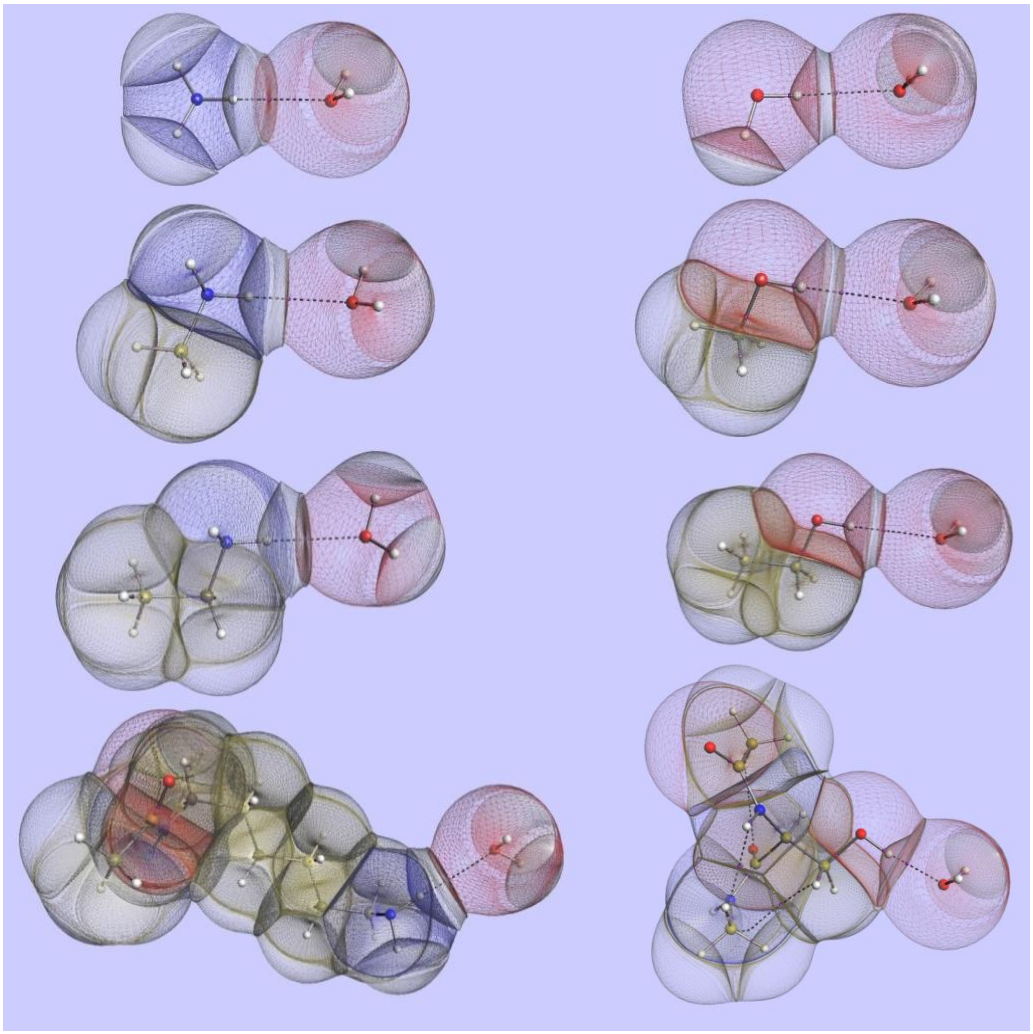


Figure 6.4: The eight dimers studied, clockwise from the top right: the water dimer, methanol-water, ethanol-water, serine-water, lysine-water, ethylamine-water, methylamine-water, ammonia-water.

As stated above, a number of R groups were used to test transferability. For both the amine nitrogen and hydroxyl oxygen tests, the R groups were H, methyl (Me) and ethyl (Et). In addition, the full amino acids lysine (Lys) and serine (Ser) were also used. The fully extended side rotamer of Lys (all side chain dihedrals *trans*) was used in order to minimise any perturbation of the amine and water atoms by the polar amino acid back bone atoms. Serine has a much shorter side chain, with only one methylene group separating the hydroxyl group from the amino acid backbone and as a result it is not possible to achieve the same separation as achieved in the case of Lys. Because the hydroxyl oxygen can act as a H-bond acceptor to the peptide hydrogen atoms, care was taken to select a structure where this was not present in order to prevent polarisation and delocalisation of the O-H…O hydrogen bond. In order to assess the transferability of individual atoms in the presence of different R groups, the difference in the value of the IQA terms relative to those when R=H has been plotted.

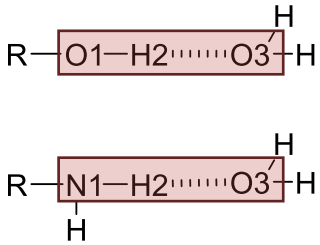


Figure 6.5: The atoms studied in this work.

Because the deformation energy of an atom is sensitive to small changes in molecular geometry, an effort was made to ensure that all systems with different R groups were as similar as possible. For example, in the hydroxyl case, the following steps were taken in order to obtain the molecular geometries:

1. The water dimer (R=H system) was optimised at the HF/6-31++G(d,p) level of theory. This geometry was then used as a "template" for all systems with different R groups.
2. Next, the methanol, ethanol and serine monomers were optimised at the HF/6-31++G(d,p) level.
3. The H-atom of the hydrogen-bond-donating water molecule that is not involved in the hydrogen bond was then substituted for each of the optimised monomers (see **Figure 6.4**).
4. Each system then had the R-H…OH_2 H-bond artificially stretched 1.5 Å to 4 Å with the coordinates taken at intervals of 0.25 Å.
5. A single point calculation of each intermediate structure was taken and the molecular wave function was input to AIMAll for IQA analysis.

For the amine case, a similar strategy was employed with the following amendments. In step 1 the ammonia-water complex replaced the water dimer, and in step two the methylamine, ethylamine and lysine monomers were optimised.

All of the atoms studied excluding the amine nitrogen display a degree of transferability in V_{def} as the R group is changed. The mean absolute deviation (MAD) and mean percentage deviation from R=H for all systems can be found in **Table 6.1**. In all cases other than that of the amine nitrogen, the deformation energy stays within 5 kJ mol^{-1} of the value obtained when R=H. Plots of the difference between the R=H and R=methyl, ethyl and amino acid V_{def} values for both amine and hydroxyl systems are provided in **Figures 6.6 and 6.7**.

Table 6.1: Mean absolute deviations (MADs) and mean % deviations from the R=H values for the atoms studied in this work. All energies in kJ mol^{-1}.

	MAD from R=H			Mean % Difference from R=H		
	Methanol	Ethanol	Serine	Methanol	Ethanol	Serine
O1 v_{def}	2.4	2.4	3.1	4.1	2.0	7.0
H2 v_{def}	0.8	0.9	0.1	1.2	2.0	0.7
O3 v_{def}	2.1	2.3	2.3	2.4	5.5	6.8
H2_O3 v_{xc}	0.1	0.1	0.1	1.2	1.0	3.6
	MAD from R=H			**Mean % Difference from R=H**		
	Methylamine	Ethylamine	Lysine	Methylamine	Ethylamine	Lysine
N1 v_{def}	11.0	12.4	8.9	42.2	48.9	25.7
H2 v_{def}	0.7	0.6	0.6	2.9	2.1	2.7
O3 v_{def}	0.7	0.6	0.6	3.3	6.3	7.4
H2_O3 v_{xc}	0.1	0.1	<0.1	1.5	1.0	1.8

In the amine systems, **Figure 6.6** shows that the H2 and O3 atoms show clear transferability across the different R groups, with very little deviation from the R=H value. The MAD values in **Table 6.1** for the larger R groups are all lower than 0.8 kJ mol^{-1} with low % differences. The % differences are misleadingly large, as at short range where the V_{def} values are large, the % differences are very small. The % differences are larger when the magnitude of the V_{def} energies are very small. This explains the low MAD values. The N1 atom in the amine systems does not present transferability relative to the R=H system, with MAD values of 42%,49% and 25% for R=Me, Et and Lys, respectively. The volume of the N1 atomic basin is much larger (~15 au) in the R=H system than when R=Me, Et and Lys, and it is therefore reasonable to suggest that it more polarisable, leading to much larger deformation as the H-bond distance is reduced. This hypothesis is supported by the observation that the N1 deformation energy when R=H is between 30-40 kJ mol^{-1} greater than the R=Me, Et and Lys cases. To support the hypothesis of greater polarizability of the R=H system being responsible for the irregular behaviour of the N1 atom one would expect the magnitude of the atomic dipole moment to be larger in the R=H case at short H-Bond separation. Unfortunately this is not observed, with a negligable difference in dipole moments between the N1 atoms when different R groups are present.

Similar to the amine systems, the atoms in the alcohol systems present some elements of transferability, with MADs from the R=H system of less than 3.1 kJ mol^{-1} for all atoms. The H2 atom is particularly stable with MAD values of 0.77, 0.90 and 0.14 kJ mol^{-1} for R=Me, Et and Ser, respectively. It is unsurprising that the H2 atom V_{def} values in both the alcohol and amine systems remains similar to the R=H cases for the following reasons: the H atom is small and not easily polarisable (having a partial positive charge due to being bonded to

an electronegative nitrogen or oxygen atom), and also because it is being deformed by a nearly identical system in all cases (a water molecule). The same cannot be said about the O1, N1 and O3 atoms. In the cases of O1 and N1 the functional group directly attached is changing, and in O3 there is an increasingly bulky molecule moving towards it as the R groups are substituted from H to Me to Et and finally Ser/Lys. It is also noted that with the exception of the ethanol system, that the O1/N1 atoms are have greater mean % differences than the O3 water oxygen. This shows that changing the R group directly attached to an atom has a greater effect on the deformation pattern of that atom than on the H-bond acceptor.

Studying the top two plots in **Figure 6.7** (corresponding to the O1 and H2 deformation energies as a function of H-bond distance) it is clear that the Ser system behaves differently to the methanol and ethanol systems. The close proximity of the polarizing backbone atoms present in Ser clearly have an effect on the deformation profile. It is difficult to tell if this effect is as pronounced in the case of Lys due to the "volume problem" discussed previously.

The previous point leads to the greatest conceptual problem when discussing IQA deformation energies in a pseudo pair-wise fashion: the deformation energy is defined as an atomic self-energy, and therefore, unlike for V_{xc}, no explicit atomic contribution to the deformation of an atom may be obtained. Unfortunately, the deformation is still a function of the atoms around it but we have no way of determining the individual contribution of each atom to a given atomic deformation. It is for this reason that IQA atom types for use in a pair wise potential to describe the XR energy do not naturally arise. In the examples of the alcohol-water and amine-water dimers described above, it means that although one may observe similarities in an energy, one is unable to take this further and describe the specific contributions of each atom on the deformations. One is equally unable to derive pair-wise coefficients. Of course classical potentials are not claiming to be theoretically sound in their use of pair-wise potentials. These potentials are used due to convenience, simple parameterisation, and ability to provide some chemical insight. The concept that atomic repulsion is more complex than a pair-wise term is not new, for example the work of Badenhoop and Weinhold in 1997 studying steric interactions, showed that the repulsion was a complex function of the total system [238].

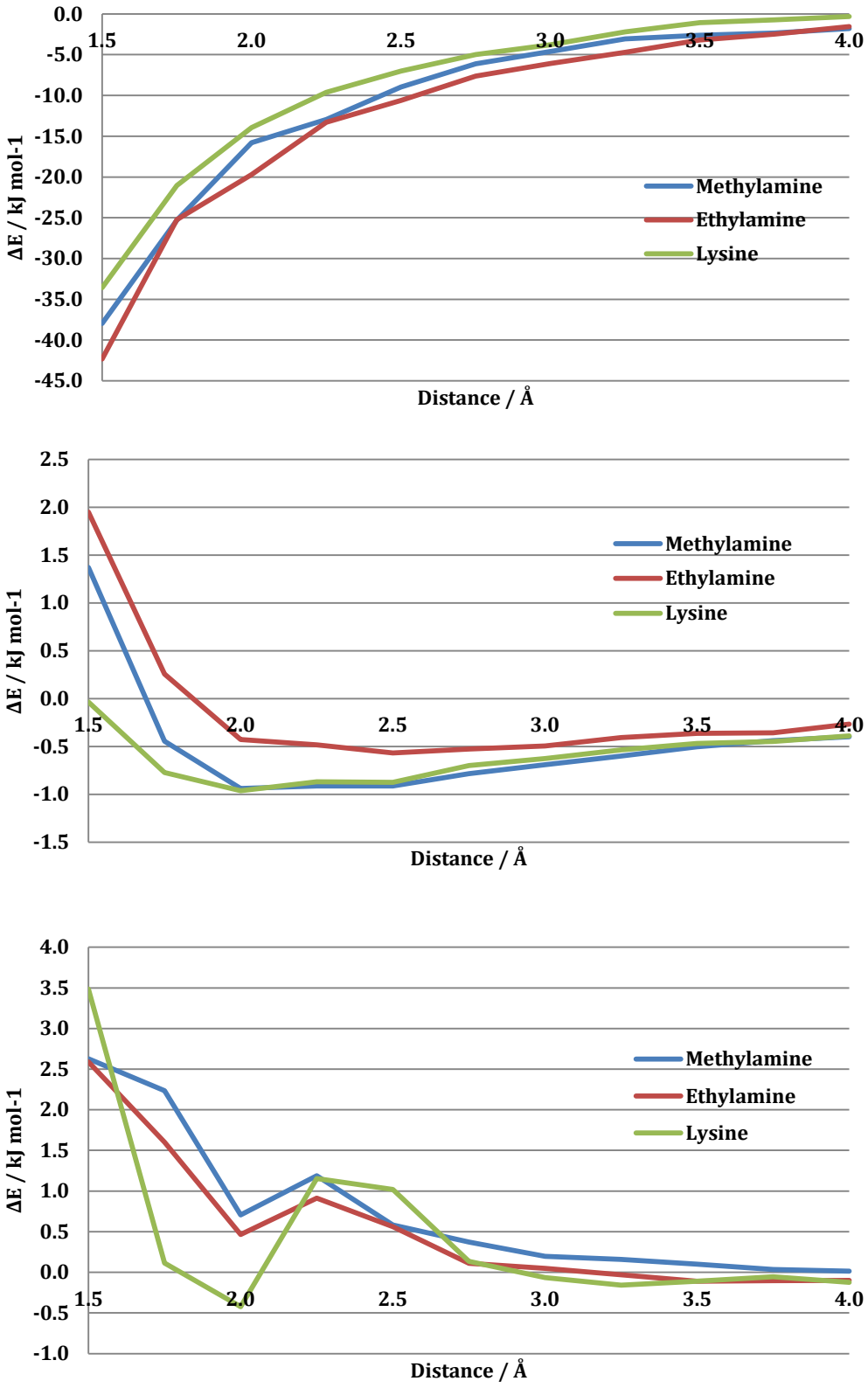


Figure 6.6: Plot of the V_{def} values of the amine N1 (*top*), H2 (*middle*) and O3 (*bottom*) obtained from R=Me, Et and Lys relative to the value for R=H against NH…O hydrogen-bond distance.

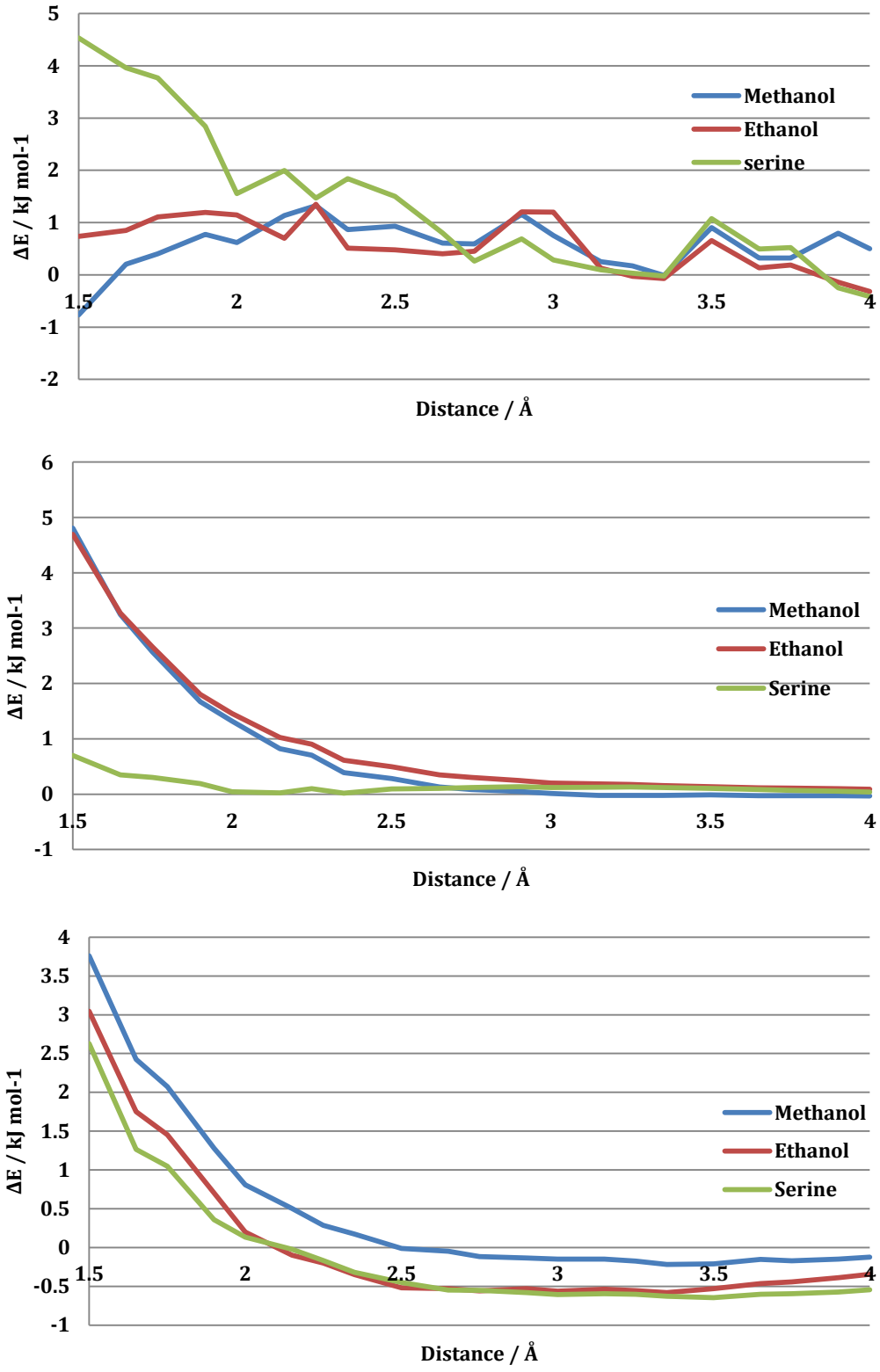


Figure 6.7: Plot of the V_{def} values of the hydroxyl O1 (*top*), H2 (*middle*) and O3 (*bottom*) obtained from R=Me, Et and Lys relative to the value for R=H against OH…O hydrogen-bond distance.

The H2…O3 V_{xc} interactions for both the hydroxyl and amine complexes remain highly transferable between the different R groups, showing on average less than a 2% deviation from the R=H values for all systems except the Ser-water dimer. Both the R=Me and Et as

well as the Ser and Lys systems all have more negative V_{xc} interactions than R=H at short range indicating a more covalent interaction. The results obtained in the present work suggest that the methanol-water dimer is a stronger interaction than the ethanol-water dimer, however Suhm *et al.* [195] have shown that higher level calculations are needed to reproduce the correct order of ethanol-water being stronger than methanol-water. Despite this, it is believed that the high degree of transferability shown across all the hydroxyl-water V_{xc} interactions remains a favourable property in the search for IQA atom types for a repulsive potential.

6.5. Conclusions and Further Work

In **Section 6.4.1** it was shown that the IQA XR energy is not directly comparable to the repulsive potentials present in a number of popular MD force fields due to a different partitioning of the total energy, however the XR energy does exhibit a similar shape to the classic potentials. In **Section 6.4.2** it was shown that there is a high level of transferability in the behaviour of the IQA components that make up the XR energy. Thus one would hope to find a mixing rule that will allow the generation of pairwise constants that can be input into a Born-type potential. Such a methodology will require pairwise terms for both the V_{xc}^{ab} and $V_{\mathrm{def}}^{\mathrm{ab}}$.

In principle, V_{xc}^{a} atomic terms could be parameterised from known V_{xc}^{ab} interactions using a list of atom pairs, however such an approach is not feasible for V_{def}^{ab}. The deformation energy is a self-term, and so V_{def}^{ab} is undefined. It is impossible to determine the deformation that one atom, *a*, inflicts upon another atom, *b*. To obtain such a value, one would be required to calculate the deformation energy of atom *b* in the system where *a* is present, and also in a system that is chemically identical except without the presence of atom *a*. The second system is impossible to obtain because atom *a* influences all other atoms in the system. Thus, removing *a* will have an effect on the deformation of *b* by all atoms, not just the deformation caused by *a*.

The quantum chemical topological force field (QCTFF) removes the need for artificial pairwise potentials. Instead of the bonded and non-bonded potentials used in classic force fields, machine learning is used to build models capable of mapping changes in an atomic property to the coordinates of the system. By following such an approach, mapping IQA energetic components to a system's coordinates, the XR energy may be extracted naturally from the v_{def} and v_{xc} models without the need for pairwise potentials. Because of this, it is not necessary to obtain pair-wise XR atom types for QCTFF. The present work was simply an opportunity to compare the IQA description of chemical systems with the classical world. The fact that IQA is unable to fit into the pair-wise scheme is in no way a flaw in the methodology as the IQA derivation is rigorous. Rather it instead invites a fresh perspective on the analysis of chemical systems, where one is no longer able to think in the simplistic

pair-wise sense. When interpreting atomic repulsion, one has to think of an atom being put into an environment it is "uncomfortable" in, rather than a repulsive interaction pushing the atoms away. The IQA interaction terms may well be highly attractive in repulsive regions on the total PE surface. This has been seen by other work in the group studying the biphenyl rotation barrier, where there is an attractive interacting between "clashing" hydrogen atoms despite the repulsive total energy.

Chapter 7

Conclusions and Further Work

Complex biological systems require accurate potentials that are not too computationally taxing for molecular simulation. QCTFF takes a radical approach to this problem by using the machine learning method kriging to build models that model changes in QCT atomic properties with respect to the system's coordinates. The work in this thesis provides examples of this methodology, where, for example, in **Chapter 2** kriging models have been built for the atomic multipole moments of a number of hydrogen bonded complexes. The kriging models have then been used successfully to predict accurate electrostatic interaction energies for a number of test molecular geometries that were not present in the training set used to build the kriging models. Also in **Chapter 2** kriging models have successfully been built to describe the atomic self- and interaction energies as defined by IQA for three weakly bound complexes. The systems that are described by both IQA self- and interaction energies have the entire energy of the system described by kriging models, therefore providing a complete molecular potential that is, in theory, ready for application in a molecular simulation. Currently, QCTFF is being implemented into the DLPOLY molecular dynamics package, and geometry optimisations have been performed for systems using IQA self- and interaction energy kriging models. The next stage must be to perform a molecular dynamics simulation using QCTFF kriging molecules.

The potential applicability of a QCTFF kriging model is defined in part by the collection of molecular geometries that are used to train the kriging models. In **Chapter 3**, a new sampling approach has been provided that enables kriging models to be built for amino acids that use chemically important molecular geometries. The new approach, PDB/NM, has been compared to the traditional normal modes sampling approach used within the group, and shows that kriging models are fully capable of describing the changes in atomic multipole moments across a broad range of molecular geometries. In future, it is advised that other sampling approaches are tested, in particular sampling by molecular dynamics (MD). An argument against sampling using MD is the introduction of sampling bias due to the choice of potential used in the dynamics. MD sampling of water clusters and hydrated amino acids has now begun by others in the group, however the results are still in a preliminary stage and so I am unable to comment. A further method of sampling that I recommend is to use existing rotamer libraries to guide the construction of seed geometries for input to normal modes sampling.

Transferability is essential for molecular potentials, and QCTFF is no exception. The "horizon sphere" experiments in **Chapter 4** begin to address this issue by investigating the

maximum size of a protein fragment around an atom that is required for atomic properties to converge. Unfortunately, the fragment size is observed to be large, however further work is required to obtain a firm value for the "horizon sphere" radius. In my opinion, this must be performed as a matter of urgency. If the horizon sphere is indeed large, then additional polarization terms may be required. Work by others in the group shows that the horizon sphere of liquid water is also large, however there is evidence that the IQA self-energy has a much smaller horizon sphere. It has been shown by others in the group that a single model can be used to describe a particular atomic property for multipole atoms along a repeating polymer chain. This demonstrates that the kriging models can indeed be transferable. Further experiments of this type must be performed on more complex systems and an automated procedure must be put in place to develop these transferable models.

The future for QCTFF is bright, and I am optimistic that QCTFF will become a powerful tool for biomolecular simulation. The issue of transferability is the last major challenge, and this has begun to yield (see above). The decision to incorporate IQA energy terms in place of classic potentials is a bold choice due to the computational cost involved in running the calculations. Despite this cost, it is seen that as more groups are using the theory the algorithms used to perform the calculations are becoming faster. Using the IQA terms in place of the classical potentials will provide a greater level of interpretation when analysing the results when QCTFF simulations are performed.

Bibliography

1. Bardwell, D.A., C.S. Adjiman, Y.A. Arnautova, E. Bartashevich, S.X.M. Boerrigter, D.E. Braun, A.J. Cruz-Cabeza, G.M. Day, R.G. Della Valle, G.R. Desiraju, B.P. van Eijck, J.C. Facelli, M.B. Ferraro, D. Grillo, M. Habgood, D.W.M. Hofmann, F. Hofmann, K.V.J. Jose, P.G. Karamertzanis, A.V. Kazantsev, J. Kendrick, L.N. Kuleshova, F.J.J. Leusen, A.V. Maleev, A.J. Misquitta, S. Mohamed, R.J. Needs, M.A. Neumann, D. Nikylov, A.M. Orendt, R. Pal, C.C. Pantelides, C.J. Pickard, L.S. Price, S.L. Price, H.A. Scheraga, J. van de Streek, T.S. Thakur, S. Tiwari, E. Venuti, and I.K. Zhitkov, Acta Crystalography B, 2011. **67**: p. 535-551.
2. Jensen, F., *Introduction of Computational Chemistry. Second edition*. 2007, Chichester, Great Britain: Wiley.
3. Nazeeruddin, M.K., F. De Angelis, S. Fantacci, A. Selloni, G. Viscardi, P. Liska, S. Ito, B. Takeru, and M. Grätzel, Journal of the American Chemical Society, 2005. **127**: p. 16835-16847.
4. Pastore, M., E. Mosconi, F. De Angelis, and M. Grätzel, The Journal of Physical Chemistry C, 2010. **114**: p. 7205-7212.
5. Tantanak, D., M. A. Vincent, and I. H. Hillier, Chemical Communications, 1998: p. 1031-1032.
6. Xu, L., Q. Zhu, G. Huang, B. Cheng, and Y. Xia, The Journal of Organic Chemistry, 2012. **77**: p. 3017-3024.
7. Yang, Y.-F., G.-J. Cheng, P. Liu, D. Leow, T.-Y. Sun, P. Chen, X. Zhang, J.-Q. Yu, Y.-D. Wu, and K.N. Houk, Journal of the American Chemical Society, 2014. **136**: p. 344-355.
8. Grädler, U., H.-D. Gerber, D.M. Goodenough-Lashua, G.A. Garcia, R. Ficner, K. Reuter, M.T. Stubbs, and G. Klebe, Journal of Molecular Biology, 2001. **306**: p. 455-467.
9. Borhani, D. and D. Shaw, Journal of Computer-Aided Molecular Design, 2012. **26**: p. 15-26.
10. Cole, D.J., J. Tirado-Rives, and W.L. Jorgensen, Journal of Chemical Theory and Computation, 2014. **10**: p. 565-571.
11. Takatani, T., E.G. Hohenstein, and C.D. Sherrill, The Journal of Chemical Physics, 2008. **128**: p. 124111-7.
12. Case, D., V. Babin, J. Berryman, R. Betz, Q. Cai, D. Cerutti, T. Cheatham Iii, T. Darden, R. Duke, and H. Gohlke, 2014.
13. Huang, J. and A.D. MacKerell, Journal of Computational Chemistry, 2013. **34**: p. 2135-2145.
14. Vanommeslaeghe, K., A. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, I. Lopes, I. Vorobyov, and A.D.J. McKerell, Journal of Computational Chemistry, 2010. **31**: p. 671–690.
15. Oostenbrink, C., A. Villa, A.E. Mark, and W.F. van Gunsteren, Journal of Computional Chemistry, 2004. **25**: p. 1656-1676.
16. Ponder, J.W., C. Wu, V.S. Pande, J.D. Chodera, M.J. Schnieders, I. Haque, D.L. Mobley, D.S. Lambrecht, R.A.J. DiStasio, M. Head-Gordon, G.N.I. Clark, M.E. Johnson, and T. Head-Gordon, Joutnal of Physical Chemistry B, 2010. **114**: p. 2549–2564.
17. Allinger, N.L., Y.H. Yuh, and J.H. Lii, Journal of the American Chemical Society, 1989. **111**: p. 8551-8566.
18. Halgren, T.A., Journal of the American Chemical Society., 1992. **114**: p. 7827–7843.
19. Popelier, P.L.A., International Journal of Quantum Chemistry, 2015. **Early View , DOI: 10.1002/qua.24900**.
20. Popelier, P.L.A. and É.A.G. Brémond, International Journal of Quantum Chemistry, 2009. **109**: p. 2542-2553.
21. Popelier, P.L.A., *The Quantum Theory of Atoms in Molecules*, in *The Nature of the Chemical Bond Revisited*, G. Frenking and S. Shaik, Editors. 2014, Wiley-VCH, Chapter 8. p. 271-308.
22. Bader, R.F.W., *Atoms in Molecules. A Quantum Theory.* 1990, Oxford, Great Britain: Oxford Univ. Press.

23. Popelier, P.L.A., *Atoms in Molecules. An Introduction.* 2000, London, Great Britain: Pearson Education.
24. Matta, C.F. and R.J. Boyd, *The quantum theory of atoms in molecules*. 2007: John Wiley & Sons.
25. Blanco, M.A., A.M. Pendas, and E. Francisco, Journal of Chemical Therory and Computation, 2005. **1**: p. 1096-1109.
26. Arnold, P.L., A. Prescimone, J.H. Farnaby, S.M. Mansell, S. Parsons, and N. Kaltsoyannis, Angewandte Chemie International Edition, 2015. **54**: p. 6735-6739.
27. Mountain, A.R.E. and N. Kaltsoyannis, Dalton Transactions, 2013. **42**: p. 13477-13486.
28. Farrugia, L.J., C. Evans, D. Lentz, and M. Roemer, Journal of the American Chemical Society, 2009. **131**: p. 1251-1268.
29. Mandado, M., M.J. González-Moa, and R.A. Mosquera, Journal of Computational Chemistry, 2007. **28**: p. 127-136.
30. Grabowski, S., Journal of Molecular Modeling, 2013. **19**: p. 4713-4721.
31. Popelier, P.L.A., Journal of Physical Chemistry A, 1998. **102**: p. 1873-1878.
32. Popelier, P.L.A. and D.S. Kosov, Journal of Chemical Physics, 2001. **114**: p. 6539-6547.
33. Popelier, P.L.A., L. Joubert, and D.S. Kosov, Journal of Physical Chemistry A, 2001. **105**: p. 8254-8261.
34. Stone, A.J., *The Theory of Intermolecular Forces*. 1 ed. The International Series of Monographs on Chemistry, ed. J.S. Rowlinson. Vol. 32. 1996, Oxford: Clarendon Press. 264.
35. Shaik, M.S., M. Devereux, and P.L.A. Popelier, Molecular Physics, 2008. **106**: p. 1495-1510.
36. Liem, S.Y., P.L.A. Popelier, and M. Leslie,International Journal of Quantum Chemistry, 2004. **99**: p. 685-694.
37. Eskandari, K. and C. Van Alsenoy, Journal of Computational Chemistry, 2014. **35**: p. 1883-1889.
38. Jarzembska, K.N. and P.M. Dominiak, Acta Crystallographica Section A, 2012. **68**: p. 139-147.
39. Pendas, A.M., E. Francisco, and M.A. Blanco, Journal of Physical Chemistry A, 2006. **110**: p. 12864-12869.
40. Matheron, G., Economic Geology, 1963. **58**: p. 21.
41. Krige, D.G., Journal of Chemical Metals and Mining Society, 1951. **52**: p. 119–139.
42. Jones, D.R., M. Schonlau, and W.J. Welch, Journal of Global Optimisation, 1998. **13**: p. 455-492.
43. Rasmussen, C.E. and C.K.I. Williams, *Gaussian Processes for Machine Learning.* 2006, Cambridge, USA: The MIT Press.
44. Kennedy, J. and R.C. Eberhart, Proceedings of the IEEE Int. Conf. on Neural Networks 1995. **4**: p. 1942-1948.
45. Guillot, B., Journal of Molecular Liquids, 2002. **101**: p. 219-260.
46. Jorgensen, W.L. and J. Tirado-Rives, Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**: p. 6665-6670.
47. Jorgensen, W.L., J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein, The Journal of Chemical Physics, 1983. **79**: p. 926-935.
48. Mahoney, M.W. and W.L. Jorgensen, The Journal of Chemical Physics, 2000. **112**: p. 8910-8922.
49. Stillinger, F.H. and A. Rahman, The Journal of Chemical Physics, 1974. **60**: p. 1545-1557.
50. Walsh, T.R. and T. Liang, Journal of Computational Chemistry, 2009. **30**: p. 893-899.
51. Batista, E.R., S.S. Xantheas, and H. Jonsson, J. Chem. Phys., 2000. **112**: p. 3285-3292.
52. Hawe, G.I. and P.L.A. Popelier, Canadian Journal of Chemistry, 2010. **88**: p. 1104-1111.
53. Millot, C. and A.J. Stone, Molecular Physics, 1992. **77**: p. 439-462.
54. Cieplak, P., P. Kollman, and T. Lybrand, The Journal of Chemical Physics, 1990. **92**: p. 6755-6760.

55. Niesar, U., G. Corongiu, E. Clementi, G.R. Kneller, and D.K. Bhattacharya, The Journal of Physical Chemistry, 1990. **94**: p. 7949-7956.
56. Barnes, P., J. Finney, J. Nicholas, and J. Quinn, Nature, 1979. **282**: p. 459-464.
57. Millot, C., J.-C. Soetens, M.T.C. Martins Costa, M.P. Hodges, and A.J. Stone, The Journal of Physical Chemistry A, 1998. **102**: p. 754-770.
58. Hodges, M.P., A.J. Stone, and S.S. Xantheas, Journal of Physical Chemistry A, 1997. **101**: p. 9163-9168.
59. Engkvist, O. and A.J. Stone, The Journal of Chemical Physics, 2000. **112**: p. 6827-6833.
60. Foster, M.C. and G.E. Ewing, The Journal of Chemical Physics, 2000. **112**: p. 6817-6826.
61. Bissantz, C., B. Kuhn, and M. Stahl, Journal of Medicinal Chemistry, 2010. **53**: p. 5061.
62. Melandri, S., Physical Chemistry Chemical Physics, 2011. **13**: p. 13901-13911.
63. Panigrahi, S.K. and G.R. Desiraju, Proteins: Structure, Function, and Bioinformatics, 2007. **67**: p. 128-141.
64. Kato, T. and J.M.J. Fréchet, Macromolecular Symposia, 1995. **98**: p. 311-326.
65. Kato, T., N. Mizoshita, and K. Kanie, Macromolecular Rapid Communications, 2001. **22**: p. 797-814.
66. Day, G.M., J. Chisholm, N. Sham, W.D.S. Motherwell, and W. Jones, Crystal Growth Descriptions, 2004. **4**: p. 1327-1340.
67. Taylor, R., O. Kennard, and W. Versichel, Journal of the American Chemical Society, 1983. **105**: p. 5761-5766.
68. Lommerse, J.P.M., S.L. Price, and R. Taylor, Journal of Computational Chemistry, 1997. **18**: p. 757-774.
69. Nobeli, I., S.L. Price, J.P.M. Lommerse, and R. Taylor, J. Comp. Chem., 1997. **18**: p. 2060-2074.
70. Umeyama, H. and K. Morokuma, J. Am. Chem. Soc., 1977. **99**: p. 1316-1332.
71. Kollman, P.A., Acc. Chem. Res., 1977. **10**: p. 365-371.
72. Hurst, G.J.B., P.W. Fowler, A.J. Stone, and A.D. Buckingham, International Journal of Quantum Chemistry, 1986. **29**: p. 1223-1239.
73. Rendel, A.P.L., G.B. Bacskay, and N.S. Hush, Chemical Physical Letters, 1985. **117**: p. 400-413.
74. Rowlands, T.W. and K. Somasundram, Chemical Physical Letters, 1987. **135**: p. 549-552.
75. Hobza, P. and C. Sandorfy, J.Am.Chem.Soc., 1987. **109**: p. 1302-1307.
76. Alagona, G. and A. Tani, J. Chem. Phys., 1981. **74**: p. 3980-3988.
77. Lii, J.-H. and N.L. Allinger, J. Phys. Org. Chem., 1994. **7**: p. 591-609.
78. Lii, J.-H. and N.L. Allinger, J. Comp. Chem., 1998. **19**: p. 1001-1016.
79. Cieplak, P., J. Caldwell, and P. Kollman, Journal of Computational Chemistry, 2001. **22**: p. 1048-1057.
80. Kong, J. and J.-M. Yan, Int.J.Quant.Chem., 1993. **46**: p. 239-255.
81. Ren, P., C. Wu, and J.W. Ponder, Journal of Chemical Theory and Computation, 2011. **7**: p. 3143-3161.
82. Desiraju, G.G.R. and T. Steiner, *The weak hydrogen bond: in structural chemistry and biology*. Vol. 9. 2001: Oxford University Press on Demand.
83. Steiner, T. and G. Koellner, Journal of Molecular Biology, 2001. **305**: p. 535-557.
84. Levitt, M. and M.F. Perutz, Journal of Molecular Biology, 1988. **201**: p. 751-754.
85. Auffinger, P., S. Louise-May, and E. Westhof, Journal of the American Chemical Society, 1996. **118**: p. 1181-1189.
86. Auffinger, P., S. Louise-May, and E. Westhof, Faraday Discussions, 1996. **103**: p. 151-173.
87. Maris, A., S. Melandri, M. Miazzi, and F. Zerbetto, ChemPhysChem, 2008. **9**: p. 1303-1308.
88. Ramasubbu, N., R. Parthasarathy, and P. Murray-Rust, Journal of the American Chemical Society, 1986. **108**: p. 4308-4314.
89. Tsuzuki, S., A. Wakisaka, T. Ono, and T. Sonoda, Chemistry – A European Journal, 2012. **18**: p. 951-960.

90. Torii, H. and M. Yoshida, Journal of Computational Chemistry, 2010. **31**: p. 107-116.
91. Alkorta, I., F. Blanco, M. Solimannejad, and J. Elguero, The Journal of Physical Chemistry A, 2008. **112**: p. 10856-10863.
92. Politzer, P., J.S. Murray, and T. Clark, Physical Chemistry Chemical Physics, 2010. **12**: p. 7748-7757.
93. Ibrahim, M.A.A., Journal of Computational Chemistry, 2011. **32**: p. 2564-2574.
94. Lu, Y., T. Shi, Y. Wang, H. Yang, X. Yan, X. Luo, H. Jiang, and W. Zhu, Journal of Medicinal Chemistry, 2009. **52**: p. 2854-2862.
95. Chessari, G., C.A. Hunter, C.M.R. Low, M.J. Packer, J.G. Vinter, and C. Zonta, Chemistry – A European Journal, 2002. **8**: p. 2860-2867.
96. Hill, G., G. Forde, N. Hill, W.A. Lester Jr, W. Andrzej Sokalski, and J. Leszczynski, Chemical Physics Letters, 2003. **381**: p. 729-732.
97. Ghosh, D., D. Kosenkov, V. Vanovschi, C.F. Williams, J.M. Herbert, M.S. Gordon, M.W. Schmidt, L.V. Slipchenko, and A.I. Krylov, The Journal of Physical Chemistry A, 2010. **114**: p. 12739-12754.
98. Tafipolsky, M. and B. Engels, Journal of Chemical Theory and Computation, 2011. **7**: p. 1791-1803.
99. Podeszwa, R. and K. Szalewicz, Physical Chemistry Chemical Physics, 2008. **10**: p. 2735-2746.
100. Podeszwa, R., The Journal of Chemical Physics, 2010. **132**: p. 044704-8.
101. Zheng, X., C. Wu, J.W. Ponder, and G.R. Marshall, Journal of the American Chemical Society, 2012. **134**: p. 15970-15978.
102. Abraham, M.H., P.L. Grellier, D.V. Prior, J.J. Morris, and P.J. Taylor, J.Chem.Soc.-Perkin Trans. 2, 1990: p. 521-529.
103. Platts, J.A., Physical Chemistry Chemical Physics, 2000. **2**: p. 973-980.
104. Duan, Y., C. Wu, S. Chowdhury, M.C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman, Journal of Computational Chemistry, 2003. **24**: p. 1999-2012.
105. Jiao, D., C. King, A. Grossfield, T.A. Darden, and P. Ren, The Journal of Physical Chemistry B, 2006. **110**: p. 18553-18559.
106. Grossfield, A., P. Ren, and J.W. Ponder, Journal of the American Chemical Society, 2003. **125**: p. 15671-15682.
107. Grossfield, A., The Journal of Chemical Physics, 2005. **122**: p. 024506-10.
108. Price, S.L., International Reviews in Physical Chemistry, 2008. **27**: p. 541-568.
109. Lommerse, J.P.M., W.D.S. Motherwell, H.L. Ammon, J.D. Dunitz, A. Gavezzotti, D.W.M. Hofmann, F.J.J. Leusen, W.T.M. Mooij, S.L. Price, B. Schweizer, M.U. Schmidt, B.P. van Eijck, P. Verwer, and D.E. Williams, Acta Crystallographica Section B, 2000. **56**: p. 697-714.
110. Motherwell, W.D.S., H.L. Ammon, J.D. Dunitz, A. Dzyanchenko, P. Erk, A. Gavezzotti, and e. al., Acta Crystallographica Section B, 2002. **B58**: p. 647-661.
111. Day, G.M., W.D.S. Motherwell, H.L. Ammon, S.X.M. Boerrigter, R.G. Della Valle, E. Venuti, A. Dzyabchenko, J.D. Dunitz, B. Schweizer, B.P. van Eijck, P. Erk, J.C. Facelli, V.E. Bazterra, M.B. Ferraro, D.W.M. Hofmann, F.J.J. Leusen, C. Liang, C.C. Pantelides, P.G. Karamertzanis, S.L. Price, T.C. Lewis, H. Nowell, A. Torrisi, H.A. Scheraga, Y.A. Arnautova, M.U. Schmidt, and P. Verwer, Acta Crystallographica Section B, 2005. **61**: p. 511-527.
112. Day, G.M., T.G. Cooper, A.J. Cruz-Cabeza, K.E. Hejczyk, H.L. Ammon, S.X.M. Boerrigter, J.S. Tan, R.G. Della Valle, E. Venuti, J. Jose, S.R. Gadre, G.R. Desiraju, T.S. Thakur, B.P. van Eijck, J.C. Facelli, V.E. Bazterra, M.B. Ferraro, D.W.M. Hofmann, M.A. Neumann, F.J.J. Leusen, J. Kendrick, S.L. Price, A.J. Misquitta, P.G. Karamertzanis, G.W.A. Welch, H.A. Scheraga, Y.A. Arnautova, M.U. Schmidt, J. van de Streek, A.K. Wolf, and B. Schweizer, Acta Crystallographica Section B, 2009. **65**: p. 107-125.
113. P. Verwer, F.J.J.L., *Reviews in Computational Chemistry*, D.B.B. K. B. Lipkowitz, Editor. 1998, Wiley-VCH: New York. p. 327-365.
114. Karfunkel, H., F.J. Leusen, and R. Gdanitz, Journal of Computer-Aided Materials Design, 1994. **1**: p. 177-185.

115. Willock, D.J., S.L. Price, M. Leslie, and C.R. Catlow, American Journal of Computional Chemistry, 1995. **16**: p. 628.
116. Day, G.M., W.D.S. Motherwell, and W. Jones, Crystal Growth & Design, 2005. **5**: p. 1023-1033.
117. Mooij, W.T.M. and F.J.J. Leusen, Physical Chemistry Chemical Physics, 2001. **3**: p. 5063-5066.
118. Price, S.L., Physical Chemistry Chemical Physics, 2008. **10**: p. 1996-2009.
119. Cox, S.R., L.-Y. Hsu, and D.E. Williams, Acta Crystallographica Section A, 1981. **37**: p. 293-301.
120. Williams, D.E. and S.R. Cox, Acta Crystallographica Section B, 1984. **40**: p. 404-417.
121. Williams, D.E., Journal of Molecular Structure, 1999. **485–486**: p. 321-347.
122. Williams, D.E., Journal of Computational Chemistry, 2001. **22**: p. 1-20.
123. Williams, D.E., Journal of Computational Chemistry, 2001. **22**: p. 1154-1166.
124. Mayo, S.L., B.D. Olafson, and W.A. Goddard, Journal of Physical Chemistry, 1990. **94**: p. 8897-8909.
125. Hwang, M.J., T.P. Stockfisch, and A.T. Hagler, Journal of the American Chemical Society, 1994. **116**: p. 2515.
126. Maple, J.R., M.J. Hwang, T.P. Stockfisch, U. Dinur, M. Waldman, C.S. Ewig, and A.T. Hagler, Journal of Computational Chemistry, 1994. **15**: p. 162-182.
127. Peng, Z.W., C.S. Ewig, M.J. Hwang, M. Waldman, and A.T. Hagler, Journal of Physical Chemistry A, 1997. **101**: p. 7243-7252.
128. Sun, H., The Journal of Physical Chemistry B, 1998. **102**: p. 7338-7364.
129. Marcon, V. and G. Raos, The Journal of Physical Chemistry B, 2004. **108**: p. 18053-18064.
130. Brodersen, S., S. Wilke, F.J.J. Leusen, and G. Engel, Physical Chemistry Chemical Physics, 2003. **5**: p. 4923-4931.
131. Jurecka, P., J. Sponer, J. Cerny, and P. Hobza, Physical Chemistry Chemical Physics, 2006. **8**: p. 1985-1993.
132. Yuan, Y., M.J.L. Mills, P.L.A. Popelier, and F. Jensen, Journal of Physical Chemistry A, 2014. **118**: p. 7876−7891.
133. Yuan, Y., *A polarisable multipolar force field for pepides based on kriging: towards application in protein crystallography and enzymatic reactions*, in *School of Chemistry*. 2012, PhD thesis, School of Chemistry, University of Manchester: Manchester.
134. Li, Y., A. Roy, and Y. Zhang, PloS one, 2009. **4**: p. e6701.
135. Word, J.M., S.C. Lovell, J.S. Richardson, and D.C. Richardson, Journal of Molecular Biology , 1999. **285** p. 1735-1747.
136. Hughes, T.J., S. Cardamone, and P.L.A. Popelier, Journal of Computational Chemistry, 2015. **36**: p. 1844-1857.
137. Frisch, M.J., G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. Montgomery, J. A., J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, N.J. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Ö. Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, and D.J. Fox, *Gaussian 09*. 2009, Gaussian, Inc.: Wallingford CT.
138. Keith, T.A., *AIMAll*. 2014, TK Gristmill Software: Overland Park KS, USA.
139. Jurecka, P., J. Sponer, J. Cerny, and P. Hobza, Physical Chemistry Chemical Physics, 2006. **8**: p. 1985-1993.
140. Bone, R.G.A. and R.F.W. Bader,Journal of Physical Chemistry, 1996. **100**: p. 10892-10911.
141. Mills, M.J.L. and P.L.A. Popelier, Theoretical Chemistry Accounts, 2012. **131**: p. 1137-1153

142. Johnson, E.R., R.A. Wolkow, and G.A. DiLabio,Chemical Physics Letters, 2004. **394**: p. 334-338.
143. Kannemann, F.O. and A.D. Becke, Journal of Chemical Theory & Computation, 2009. **5**: p. 719-727.
144. Rappé, A.K. and E.R. Bernstein, Journal of Physical Chemistry A, 2000. **104**: p. 6117-6128.
145. Tsuzuki, S. and H.P. Luthi, Journal of Chemical Physics, 2001. **114**: p. 3949-3957.
146. Zhao, Y. and D. Truhlar, Theoretical Chemical Accounts, 2008. **120**: p. 215-241.
147. Schneebeli, S.T., A.D. Bochevarov, and R.A. Friesner, Journal of Chemical Theory & Computation, 2011. **7**: p. 658-668.
148. Grimme, S., J. Antony, S. Ehrlich, and H. Krieg, Journal of Chemical Physics, 2010. **132**: p. 154104-19.
149. Gráfová, L., M. Pitoňák, J. Řezáč, and P. Hobza, Journal of Chemical Theory and Computation, 2010. **6**: p. 2365-2376.
150. Gu, J., J. Wang, and J. Leszczynski, Chemical Physics Letters, 2011. **512**: p. 108-112.
151. Walker, M., A.J.A. Harvey, A. Sen, and C.E.H. Dessent, The Journal of Physical Chemistry A, 2013. **117**: p. 12590-12600.
152. Tiwary, A.S., K. Datta, and A.K. Mukherjee, Computational and Theoretical Chemistry, 2015. **1068**: p. 123-127.
153. Tiwary, A.S. and A.K. Mukherjee, Chemical Physics Letters, 2014. **610–611**: p. 19-22.
154. Pendás, A.M., M.A. Blanco, and E. Francisco, Journal of Chemical Physics., 2006. **125**: p. 184112-1-20.
155. Mills, M.J.L. and P.L.A. Popelier, Comput.Theor.Chem. , 2011. **975**: p. 42-51.
156. Mills, M.J.L., *A multipolar polarisable force field method from Quantum Chemical Topology and machine learning*, in *School of Chemistry*. 2011, PhD Thesis, School of Chemistry, University of Manchester: Manchester.
157. Yuan, Y., M.J.L. Mills, and P.L.A. Popelier, J.Mol.Model., 2014. **20**: p. 2172-2186.
158. Fletcher, T., S.J. Davie, and P.L.A. Popelier, Journal of Chemical Theory & Computation, 2014. **10**: p. 3708-3719.
159. Kandathil, S.M., T.L. Fletcher, Y. Yuan, J. Knowles, and P.L.A. Popelier, Journal of Computational Chemistry, 2013. **34**: p. 1850-1861.
160. Hughes, T.J., S.M. Kandathil, and P.L.A. Popelier, Spectrochimica Acta A, 2015. **136** p. 32-41.
161. Fletcher, T.L., S.M. Kandathil, and P.L.A. Popelier, Theoretical Chemical Accounts, 2014. **133**: p. 1499:1-10
162. Handley, C.M., G.I. Hawe, D.B. Kell, and P.L.A. Popelier, Physical Chemistry Chemical Physics, 2009. **11**: p. 6365–6376.
163. Beck, D.A.C., D.O.V. Alonso, D. Inoyama, and V. Daggett, Proceedings of the National Academy of Sciences, 2008. **105**: p. 12259-12264.
164. Muñoz, V. and L. Serrano, Proteins: Structure, Function, and Bioinformatics, 1994. **20**: p. 301-311.
165. Scouras, A.D. and V. Daggett, Protein Science, 2011. **20**: p. 341-352.
166. Francis-Lyon, P. and P. Koehl, Proteins: Structure, Function, and Bioinformatics, 2014. **82**: p. 2000-2017.
167. Shapovalov, Maxim V. and Roland L. Dunbrack Jr, Structure, 2011. **19**: p. 844-858.
168. Shandler, S.J., M.V. Shapovalov, J.R.L. Dunbrack, and W.F. DeGrado, Journal of the American Chemical Society, 2010. **132**: p. 7312-7320.
169. Lovell, S.C., J.M. Word, J.S. Richardson, and D.C. Richardson, Proteins: Structure, Function, and Bioinformatics, 2000. **40**: p. 389-408.
170. Dunbrack, R.L. and F.E. Cohen, Protein Science, 1997. **6**: p. 1661-1681.
171. Dunbrack Jr, R.L., Current Opinion in Structural Biology, 2002. **12**: p. 431-440.
172. Schweitzer-Stenner, R., Molecular BioSystems, 2012. **8**: p. 122-133.
173. Sousa, S.F., P.A. Fernandes, and M.J. Ramos, The Journal of Physical Chemistry A, 2009. **113**: p. 14231-14236.
174. Jha, A.K., A. Colubri, M.H. Zaman, S. Koide, T.R. Sosnick, and K.F. Freed, Biochemistry, 2005: p. 9691-9702
175. Hagarman, A., D. Mathieu, S. Toal, T.J. Measey, H. Schwalbe, and R. Schweitzer-Stenner, Chemistry: A European Journal, 2011. **17**: p. 6789-6797.

176. Hagarman, A., T.J. Measey, D. Mathieu, H. Schwalbe, and R. Schweitzer-Stenner, Journal of the American Chemical Society, 2010. **132**: p. 540–551.
177. Schweitzer-Stenner, R., Molecular Biological Systems, 2012. **8**: p. 122-133.
178. Lindorff-Larsen, K., N. Trbovic, P. Maragakis, S. Piana, and D.E. Shaw, Journal of the American Chemical Society, 2012. **134**: p. 3787-3791.
179. Pizzanelli, S., C. Forte, S. Monti, G. Zandomeneghi, A. Hagarman, T.J. Measey, and R. Schweitzer-Stenner, Journal of Physical Chemistry B, 2010. **114**: p. 3965-3978.
180. Cruz, V.L., J. Ramos, and J. Martinez-Salazar, Journal of Physical Chemistry B, 2011. **116**: p. 469-475.
181. Rost, B., Journal of Structural Biology, 2001. **134**: p. 204-218.
182. Subramaniam, S. and A. Senes, Proteins: Structure, Function, and Bioinformatics, 2014. **82**: p. 3177-3187.
183. Chipot, C. and A. Pohorille, *Free energy calculations: theory and applications in chemistry and biology*. Vol. 86. 2007: Springer.
184. Higo, J., J. Ikebe, N. Kamiya, and H. Nakamura, Biophysical Reviews, 2012. **4**: p. 27-44.
185. Okamoto, Y., Journal of Molecular Graphics and Modelling, 2004. **22**: p. 425-439.
186. Bolhuis, P.G., C. Dellago, and D. Chandler, Farad. Discuss., 1998. **110**: p. 421-436.
187. Dellago, C., P.G. Bolhuis, F.S. Csajka, and D. Chandler, Journal of Chemical Physics, 1998. **108**: p. 1964-1977.
188. Dellago, C. and P.G. Bolhuis, *Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events*, in *Advanced Computer Simulation Approaches for Soft Matter Sciences Iii*, C. Holm and K. Kremer, Editors. 2009. p. 167-233.
189. Barducci, A., M. Bonomi, and M. Parrinello, Wiley Interdisciplinary Reviews: Computational Molecular Science, 2011. **1**: p. 826-843.
190. Li, H., Z. Lin, and Y. Luo, Chemical Physics Letters, 2014. **610–611**: p. 303-309.
191. Lovas, F.J., Journal of Physical and Chemical Reference Data, 1978. **7**: p. 1445-1750.
192. Coulson, C.A. and D. Eisenberg, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 1966. **291**: p. 454-459.
193. Ren, P.Y., C.J. Wu, and J.W. Ponder, Journal of Chemical Theory & Computation, 2011. **7**: p. 3143-3161.
194. Gresh, N., G.A. Cisneros, T.A. Darden, and J.-P. Piquemal, Journal of Chemical Theory & Computation, 2007. **3**: p. 1960-1986.
195. Ufimtsev, I.S., N. Luehr, and T.J. Martinez, The Journal of Physical Chemistry Letters, 2011. **2**: p. 1789-1793.
196. Senn, H.M. and W. Thiel, Angewandte Chemie International Edition, 2009. **48**: p. 1198-1229.
197. Menikarachchi, L.C. and J.A. Gascón, Current topics in medicinal chemistry, 2010. **10**: p. 46-54.
198. Senthilkumar, K., J.I. Mujika, K.E. Ranaghan, F.R. Manby, A.J. Mulholland, and J.N. Harvey, Journal of The Royal Society Interface, 2008. **5**: p. 207-216.
199. Nam, K., J. Gao, and D.M. York, Journal of Chemical Theory and Computation, 2004. **1**: p. 2-13.
200. Maupin, C.M. and G.A. Voth, Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2010. **1804**: p. 332-341.
201. Mikulski, R.L. and D.N. Silverman, Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, 2010. **1804**: p. 422-426.
202. Calvaresi, M., M. Garavelli, and A. Bottoni, Proteins: Structure, Function, and Bioinformatics, 2008. **73**: p. 527-538.
203. Dunford, H.B., Progress in Reaction Kinetics and Mechanism, 2013. **38**: p. 119-129.
204. Warshel, A. and R.M. Weiss, Journal of the American Chemical Society, 1980. **102**: p. 6218-6226.
205. van Duin, A.C.T., S. Dasgupta, F. Lorant, and W.A. Goddard, Journal of Physical Chemistry A, 2001. **105**: p. 9396-9409.
206. van der Kamp, M.W. and A.J. Mulholland, Biochemistry, 2013. **52**: p. 2708-2728.
207. Mills, M.J.L., G.I. Hawe, C.M. Handley, and P.L.A. Popelier, Physical Chemistry Chemical Physics, 2013. **15**: p. 18249-18261.

208. Nagy, P.I. and B. Noszal, Journal of Physical Chemistry A, 2000. **104**: p. 6834-6843.
209. Remko, M., D. Fitz, and B.M. Rode, Amino Acids, 2010. **39**: p. 1309-1319.
210. Deplazes, E., W. van Bronswijk, F. Zhu, L.D. Barron, S. Ma, L.A. Nafie, and K.J. Jalkanen, Theoretical Chemical Accounts., 2008. **119**: p. 155-176.
211. Weixin, F., K.R. Amareshwar, K. Rai, Z. Lu, and Z. Lin, Journal of Molecular Structure, 2009. **895**: p. 65-71.
212. Matta, C.F. and R.F.W. Bader, Proteins: Structure, Function and Genetics, 2003. **52**: p. 360-399.
213. Matta, C.F. and R.F.W. Bader, Proteins:Structure, Function and Genetics, 2002. **48**: p. 519-538.
214. Matta, C.F. and R.F.W. Bader, Proteins: Structure, Function and Genetics, 2000. **40**: p. 310-329.
215. Hirshfeld, F., Theoretica Chimica Acta, 1977. **44**: p. 129-138.
216. Bultinck, P., C. Van Alsenoy, P.W. Ayers, and R. Carbó-Dorca, Journal of Chemical Physics, 2007. **126**: p. 144111.
217. Fonseca Guerra, C., J.W. Handgraaf, E.J. Baerends, and F.M. Bickelhaupt, Journal of Computational Chemistry, 2004. **25**: p. 189-210.
218. Kosov, D.S. and P.L.A. Popelier, Journal of Physical Chemistry A, 2000. **104**: p. 7339-7345.
219. Yuan, Y., M.J.L. Mills, P.L.A. Popelier, and F. Jensen, Journal of Physical Chemistry A, 2014. **in press**.
220. Jensen, F., J.Chem.Phys., 2002. **117**: p. 9234-9240.
221. Keith, T.A., *AIMAll (Version 10.07.25), aim.tkgristmill.com*. 2010, AIMAll (Version 10.07.25), aim.tkgristmill.com.
222. Popelier, P.L.A., R.G.A. Bone, and D.S. Kosov, *MORPHY01*. 2001, UMIST: Manchester, England.
223. Liu, S., The Journal of Chemical Physics, 2007. **126**: p. 191107.
224. Weizsäcker, C.F.v., Zeitschrift für Physik, 1935. **96**: p. 431-458.
225. Huang, Y., A.-G. Zhong, Q. Yang, and S. Liu, The Journal of Chemical Physics, 2011. **134**: p. 084103.
226. Liu, S., The Journal of Physical Chemistry A, 2013. **117**: p. 962-965.
227. MacKerell, A.D.J., D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evansek, M.J. Field, S. Fischer, J. Gao, H. Guo, and S. Ha, Journal of Physical Chemistry B., 1998. **102**: p. 3586-3616.
228. Chen, A.A. and R.V. Pappu, The Journal of Physical Chemistry B, 2007. **111**: p. 11884-11887.
229. Jorgensen, W.L., D.S. Maxwell, and J.J. Tirado-Rives, Journal of the Americal Chemistry Society, 1996. **118**: p. 11225-11236.
230. Fender, B.E.F. and G.D. Halsey, The Journal of Chemical Physics, 1962. **36**: p. 1881-1888.
231. Kong, C.L., The Journal of Chemical Physics, 1973. **59**: p. 2464-2467.
232. Waldman, M. and A.T. Hagler, Journal of Computational Chemistry, 1993. **14**: p. 1077-1084.
233. Delhommelle, J. and P. MilliÉ, Molecular Physics, 2001. **99**: p. 619-625.
234. Halgren, T.A., Journal of Computational Chemistry, 1996. **17**: p. 490-519.
235. Allinger, N.L., Journal of the American Chemical Society, 1977. **99**: p. 8127.
236. Allinger, N.L., Y.H. Yuh, and J.-H. Lii, J.Am.Chem.Soc., 1989. **111**: p. 8551-8556.
237. Allinger, N.L., K. Chen, and J.H. Lii, Journal of Computational Chemistry, 1996. **17**: p. 642-668.
238. Badenhoop, J.K. and F. Weinhold, The Journal of Chemical Physics, 1997. **107**: p. 5406-5421.
239. Smith, B. J., Swanton, D. J., Pople, J. A., Schaefer, H. F., III, and Radom, L., Journal of Chemical Physics, 1990, ***92***, p.1240-1251
240. Aakeroy, B.B., Nieuwenhuyzen, M. and Price, S.L, Journal of the American Chemical Society, 1998, **120**: p,8986-8993
241. Lee, L.P., Cole, D.J., Skylaris, C-K., Jorgensen, W.L. and Payne, M.C., Journal of Chemical Theory and Computation, 2013, **9:**p. 2981–2991

Appendices

1. S. Cardamone, T. J. Hughes, P. L. Popelier, Multipolar Electrostatics, *Phys. Chem. Chem. Phys.*, 2014, **16**, p. 10367-10387
2. T. J. Hughes, S. M. Kandathil, P. L. Popelier, Accurate prediction of polarised high order electrostatic interactions for hydrogen bonded complexes using the machine learning method kriging, *Spectrochemica Acta A*, 2015, **136**, p. 32-41
3. T. J. Hughes, P. L. Popelier, Where does charge reside in amino acids? The effect of side-chain protonation state on the atomic charges of Asp, Glu, Lys, His and Arg, *Comp. Theor. Chem.*, 2015, **1053**, p. 298-304
4. T. J. Hughes, S. Cardamone, P. L. Popelier, Realistic sampling of amino acid geometries for a multipolar polarizable force field, *J. Comp. Chem*, 2015, **36**, p. 1844-1857

Appendix A

Literature review of Potential Applications of QCTFF in Medicinal Chemistry- Docking

A note:

Being sponsored by AstraZeneca, it was originally planned that during my PhD I would be the first group member to apply QCTFF to medicinal chemistry problems, namely that of small molecule docking. Instead, the direction that my research took led down related avenues of interest, however the docking application was never performed. For example, my work in chapter 2 looking at interactions between S22 dimers is highly relevant to how QCTFF will handle the intermolecular interactions identified during a docking experiment. Similarly, the work of chapter 4 describing the horizon sphere is important for deciding how big an active site of a protein may need to be to get reliable results of a docking experiment.

Here I include the literature review of docking that I wrote in my first year of my PhD, as it represents a significant amount of my time and effort.

Medicinal Chemistry

Development of a novel force field, such as QCTFF, is a challenging and time consuming task. The motivation for undertaking such a project is that the new force field should be able to provide more accurate and reliable solutions to real world problems than those provided by existing methods. Therefore the discussion now moves from description of theoretical and methodological considerations of force field design to the possible applications of QCTFF. In particular, the discussion shall focus on the role of MM force fields in the drug discovery process which is one of the key aims of QCTFF.

Rational Drug Design

There are over 20,000 proteins encoded by genes within the human genome. Upon translation many proteins are then subject to a range of different post-translational modifications such as methylation, addition of both prosthetic groups and polysaccharides, and the formation of multiple subunit protein complexes. It has therefore been stated that it would be irrational to blindly purify and experimentally assay thousands of proteins with hundreds of thousands of possible drug candidates. This led to the development of new techniques and the application of protocol for the efficient screening large databases of drug candidates to provide shortlists of '*lead molecules*', followed by more computationally expensive simulation and experimental studies. Such a process is given

the general title '*rational drug design*'. Docking and MD simulation are the two stages of rational drug design where MM force fields are most widely used with the former topic discussed here in greatest detail.

Before discussing the above topics it is useful to provide brief definition of some of the language used by medicinal chemists. *Receptor* or *drug target* refers to a binding site, typically of a protein. This may be either the active site of the protein or any other region where a *guest* molecule may bind. A *guest* molecule is synonymous with *ligand* and simply refers to a small molecule. Such a molecule can also be thought of as a *drug candidate*. Drug candidates which have passed through a screening process are then referred to as *leads*, or *lead compounds*. The structure adopted by the ligand within the receptor is called its *pose*, *binding mode*, or *conformation*. Multiple binding modes may be possible for a given molecule.

Docking

Once a receptor such as a protein active site has been identified and an accurate 3D structure has been obtained (details of the methods used in the acquisition of such structures are not discussed here, typical sources are NMR, x-ray crystallography and homology modelling), the next step in the rational drug design process is to screen databases of small molecules for possible candidates. This task is performed typically by a piece of docking software such as GOLD[1], AutoDock[2], DOCK[3] and GLIDE[4]. The aim of docking is to rapidly insert a large number of small molecules into the binding site of a receptor and to score each molecule based on criteria such as the strength of interaction between ligand and receptor, the number of hydrogen bonds formed, and the internal energy calculated by a MM force field. Multiple orientations (*binding modes*) of each ligand may be docked, and the highest scoring ligands become lead compounds. Different pieces of docking software operate at differing levels of complexity. For example the flexibility of the receptor molecule may range from being fixed, to allowing the side chains to rotate based upon databases of commonly found rotamers, and even to combining the docking process with full MD simulation. The scoring function also ranges in complexity, from simple empirical scores to full MM energy calculation.

The popular GOLD program of Jones *et al.*[1] uses a genetic algorithm (GA) to explore both the conformational space of flexible ligands and to simultaneously sample a range of different binding modes of a ligand. The active site of the protein is also partially flexible. The use of the genetic algorithm allows 'good' solutions to a docking problem to be found rapidly. The scoring function used by GOLD consists of three parts; an empirical hydrogen bonding term, a pairwise dispersion term to account for hydrophobicity, and an MM calculation of the ligand's internal energy. In a test of the GOLD program, NADPH was

docked into the active site of dihydrofolate reductase (DHFR), with the aim being to reproduce the experimental crystal structure. This was a challenging test, with the flexibility of NADPH described by 17 rotatable bonds and partial cyclic flexibility. Despite this the predicted structure of the complex from docking had a root mean square deviation of the heavy atoms from the experimental structure of only 1.2 Å. The ability of GOLD to predict 100 protein-ligand complexes was tested in order to analyse its strengths and identify any possible weaknesses. The systems were chosen due to the structures of the ligands being "drug like". Each system was docked 20 times to allow the genetic algorithm to find its best solution, and the optimal docking solutions were then ranked based upon the scoring function as belonging to one of four categories; *good*, *close*, *with errors*, or *wrong*. In 71% of the test cases GOLD predicted either *good* or *close* docking solutions. As stated, each system was docked 20 times to allow GOLD a chance of finding the correctly docked structure, however 49 complexes were correctly predicted within only two docking runs, and 65 complexes predicted after just 10 docking runs. Thus it was concluded that generally GOLD does not need to be run 20 times to obtain accurate docking solutions. A failing of GOLD is that the scoring function used is largely dependent upon hydrogen bonding, resulting in difficulty in correctly predicting the experimentally observed binding mode of hydrophobic ligands.

The AutoDock software package of Morris *et al.*[2] implements a hybrid Lamarckian genetic algorithm (LGA) in which a standard GA is used for global screening of ligand conformations with the addition of a local search method to perform an energy minimisation. The name Lamarckian comes from the now discredited theory of Jean Batiste de Lamarck that phenotypic characteristics acquired throughout an organism's life may be inherited by offspring. This relates to the 'phenotypic traits' picked up by a parent during the local search minimisation being 'inherited' by its offspring. The ability of the LGA to correctly predict the experimental crystal structure of seven protein-ligand test systems was assessed by comparison with both standard GA and simulated annealing (SA) search algorithms. For each search algorithm, 10 docking runs were performed, with the resulting optimised docked structures being grouped into clusters and then ranking the clusters based on the lowest energy docking solution in each cluster. LGA performed best overall, with 78% of the docked structures found in the lowest energy (rank 1) cluster and an average root mean squared deviation of heavy atoms from the experimental structure of only 0.88 Å. LGA also gave the fewest number of clusters across the seven test systems with an average of just 2.29 meaning that it was most consistent at finding the optimum docking solution across all 10 docking runs without becoming trapped in local minima. The average difference between the effective global minimum energy (the lowest docked energy for each complex as found by any of the three searching methods used) and the best docked solution for each complex found by the LGA search method was the lowest of

all search methods, 0.40 kcal mol^{-1}, meaning that the effective global minimum was most often that found by LGA rather than GA or SA. The standard GA was the second most efficient search method with 40% of the docked structures in the cluster of rank 1 and an average RMSD of 3.06 Å. The SA approach suffered being trapped in local minima which resulted in a mean difference between the docked energy and the effective global minimum energy was 2.62×10^5 kcal mol^{-1}, and a mean RMSD of 3.63 Å. Due to becoming trapped in local minima, SA was the worst performing search method used.

DOCK, of Ewing and Kuntz [3], uses a more simple docking algorithm than the GAs of Gold and AutoDock where instead ligands are superimposed onto predetermined site points within the cavity of the receptor in multipole orientations. Although initially ligands were treated rigidly DOCK has been developed to include a flexible description of ligands[5]. The flexible docking algorithm proceeds via an incremental construction of the docked ligand within the receptor active site. Before docking the ligand is cut at each rotatable bond to give a number of rigid fragments. One or more of the fragments is defined as being an anchor, meaning that it is the initially docked fragment that will be built upon in sequential steps. Typically the anchor is the section of the ligand that forms strong intermolecular interactions with the receptor. The anchor is docked to find valid docking orientations and then the rigid groups are incrementally added to build up the ligand. At each stage of the incremental building of the ligand local optimisations are performed including the rotation of the flexible torsional angles, and the different structures are scored and ranked. The basic scoring function used is a sum of intermolecular van der Waals's and Coulombic terms obtained from the AMBER force field and intramolecular terms to prevent steric clashes consisting of Coulombic, van der Waals and simplified torsional potentials.

The flexible docking algorithm was tested in two ways: first in its ability to correctly predict the experimental crystal structure of a docked ligand in a receptor molecule, and secondly in the ability of the algorithm for the efficient screening of a molecule database. It was noted that the simple scoring function used lowers the accuracy of the docking process (for example the neglect of solvation effects) and it is stated that the aim of the test is to show that the flexible docking algorithm is capable of finding "reasonable binding modes" even with a "minimal scoring function". For the former test, 15 test cases were docked and each was performed five times to allow DOCK to find the optimal docked structure. The RMSD of the top scoring binding mode for each test case from the crystal structure ranged from 0.9 Å to 6.8 Å. Seven of the docked complexes had an RMSD of less than 2 Å. For six out of the eight test systems where the best docked solution had an RMSD greater than 2 Å the score was more favourable than the score of the experimental structure. This indicated that the scoring function and not the flexible search algorithm was largely responsible for the large RMSDs. The two remaining test cases had less

favourable scores than the experimental structure and so in these cases the search algorithm was to fault. Despite this, the flexible algorithm was found to perform generally well. In the second test of the flexible search algorithm, 49 randomly selected molecules from the Current Medicinal Chemicals (CMC) database were selected and docked into two receptors, streptavidin (1stp) and dihydrofolate reductase (3dfr). To ensure that a strongly binding ligand was present for each of the receptors 1stp and 3dfr, the naturally occurring ligands biotin (complexes with 1stp) and methotrexate (complexes with 3dfr) were added to the database to give 51 test molecules in total. The performance of the incremental flexible algorithm was compared with the performance of both flexible and rigid algorithms that dock random ligand orientations into the receptor. The accuracy of a given method was determined by the number of times the optimal solution scored the most favourable score out of all the three different methods tested. For 1stp the incremental flexible algorithm gave the most favourable docking score for 63% of the ligands when 15-30 s sampling time per molecule was allowed. At shorter sampling times all three methods were equivalent in accuracy. For 3dfr the incremental method was again the most successful technique, with a success rate of 82% at longer sampling times. For both test systems the naturally occurring ligands were given favourable scores.

The final piece of docking software presented in this work is Glide of Friesner *et al.*[4]. Glide differs from the previously discussed software packages in that the search algorithm utilised by Glide approximates a complete systematic search of conformational, orientational and positional space of the docked ligand. This is achieved by a series of hierarchical filters that search for possible binding modes with increasingly more accurate scoring functions. Ligands are divided into core and rotamer groups and the core is then assigned a number of conformers dependent upon the number of degrees of freedom within the core. Methyl groups and primary amines are not classed as rotatable however all other rotatable bonds are classed as rotatable and so are defined as being part of a rotamer group. High energy conformations are removed though the screening of all of the combinations of rotamer conformations using a simplified version of the torsion angle term in the OPLS-AA force field. The remaining conformations are then docked into the receptor by superimposing the ligand onto defined site points within the binding site, and then screening for both steric clashes and also any hydrogen bonds. Successful structures are scored and the highest scoring proceed to energy minimisation and then Monte-Carlo simulation to probe remaining torsional conformations. The highest scoring conformations after this stage are then assigned a final score. This is a sum of an intermolecular OPLS-AA score, an internal ligand strain score, and a score given by a modified scoring function named GlideScore 2.5 that included solvation, metal ion-ligand and hydrogen bonded terms. The combination of GlideScore 2.5 and an MM score was found to be more reliable than either scoring method on their own.

The accuracy of Glide was evaluated by its ability to reproduce the experimentally observed structure of 282 co crystallised protein-ligand complexes. The list included the complexes that were used to test the GOLD and FlexX docking programs, which allowed for direct comparison of the performance of Glide relative to these programs. For the 93 complexes that were used to test GOLD, Glide had an average RMSD from the crystal structures of only 1.85 Å compared to 3.06 Å by GOLD. For the 189 complexes that were used to test FlexX, Glide performed even more favourably with an average RMSD of 1.95 Å compared to 3.72 Å by FlexX. Overall, nearly half of the 282 test complexes predicted by Glide had an RMSD of less than 1 Å, with only one third of cases predicted with an RMSD of greater than 2 Å.

Because the purpose of docking is to reduce large databases of compounds to a smaller subset of leads quickly and efficiently, a balance between accurate scoring functions, ligand flexibility and receptor flexibility must be struck. Kolb and Irwin outline two criteria that must be met by docking software for it to be deigned as successful. Docking software must be both able to correctly predict the ligand pose to allow a meaningful analysis of the observed binding interactions, and also to accurately score docked molecules from a database in order to identify ligands to proceed to later stages in the drug design process. Often the scoring function comprises of a sum of both empirical terms such as the number of hydrogen bonds between ligand and receptor and MM energies. Both terms are quick to evaluate however lack accuracy. In particular MM force fields typically use atomic point charges and so any directional bonding such through lone pairs in hydrogen bonding will not be accurately reproduced. This can lead to experimentally more stable ligands being scored lower than experimentally less stable ligands where, for example, the hydrogen bonding atoms are closer in distance despite the hydrogen bond not being at an angle where a lone pair would lie. This is an area in which QCTFF will offer improvement as the directionality of intermolecular interactions are only reproduced when using higher order multipole moments to describe electrostatics[6].

It has been found that the first of the criteria, that docking must be able to correctly predict the experimental pose of a ligand in a receptor, is not always met, even when it appears that docking has been successful. For example, DOCK was used to screen a database of 55 000 molecules docked into a binding site of thymidylate synthase (TS) [7]. Despite a number of phenolphthalein analogues found to inhibit TS, the pose predicted by DOCK was found to be considerably different to the crystal structure. Although docking was successful at correctly identifying the ligands that bind strongly to TS, the pose of binding was not the same as that observed experimentally. Hence the success of docking was somewhat fortuitous. In this case Kolb and Irwin state that they “feel that it is hard to argue that docking worked for the right reasons in this case”. Docking can, however, make

predictions very close to the crystallographic structure. The docking software LUDI [8] was used to screen 120 000 molecules for possible tRNA-guanine transglycosylase (TGT) inhibitors[9]. The screening narrowed the database down to three ligands, with the highest scoring compound being 4-aminophthalhydrazide. An x-ray structure of the docked compound was taken and it was seen that the docked pose had an RMSD of only 0.24 Å from experiment. This was described by Kolb and Irwin as "a clear success of docking". Unfortunately, there are very few examples in the literature of where docking solutions have been directly compared to a crystal structure of the ligand-receptor complex. Therefore the question of whether the high scoring molecules in a given docking experiment are predicted in the "true" pose is difficult to answer. For example, GOLD has been used to screen a library of 58 855 compounds for ligands that bind to a protease of the SARS coronavirus[10]. Although docking was successful in producing two hits, and a crystal structure was obtained, no direct comparison was made and so whether the correct pose was predicted by docking is not clear.

Beyond Simple Docking

Due to the limited power of computers, early docking experiments made the assumptions of rigid ligands docked into rigid receptors, with scoring based solely on sterics. Peishoff *et al.* referred to early docking experiments as "a gentleman's pursuit" and "at worst, a fool's errand"[11]. Docking algorithms have since incorporated ligand flexibility and improved scoring functions that have made docking a viable technique in the rational drug design process. The next challenge facing docking methodology is the incorporation of flexible receptor molecules. This is done in a number of ways. Again, key considerations for the incorporation of receptor flexibility into docking algorithms are speed and efficiency. There have been many attempts to incorporate receptor flexibility ranging from statistical approaches to full MD simulation.

The simplest approach to incorporate receptor flexibility into a docking experiment is simply to reduce the size of the van der Waals's radii of the receptor atoms. This allows the ligand to "penetrate" the active site in an attempt to reproduce small amounts of receptor readjustment and flexibility, and has been termed "soft docking"[12]. This approach is included in the Glide software[4], where non-polar ligand and receptor atoms have scaled van der Waals's radii. To achieve this practically, it is the repulsive term of the Lennard-Jones potential that is scaled. It was found by Ferrari *et al.* [13] that a scoring function including soft docking performs better when only one configuration of a ligand is docked into a receptor, however as more configurations are allowed, "hard docking" recovers to such an extent that it is the superior model. Hence more involved treatment of ligand flexibility is required. Two methods are discussed here; the use of rotamer libraries to

explore the conformational space of specific active site residues, and combining docking with molecular dynamics simulation.

One of the first to use rotamer libraries as a means to introduce molecular flexibility into docking experiments was Leach [14]. In this work, benzamide was docked into trypsin, where 61 side chains were allowed to vary. The allowed rotamers for each side chain were obtained from the library of Ponder *et al*. [15] and the possible ligand conformations were obtained before docking all combinations of receptor and ligand. The complexes were ranked in order of their AMBER MM energy. The advantage of such a model is that the optimum docking solution is guaranteed to be found, however calculating the energies of all receptor conformations is costly. An improvement is the SLIDE algorithm of Schneck and Kuhn [16] in which an "anchor" fragment of the ligand is docked into the receptor initially and then the ligand is built up by the addition of flexible ligand fragments. Early versions of SLIDE used a rotamer library in an attempt to remove any steric clashes between receptor and ligand. It was found, however, that using rotamer conformations directly from database values often led to new steric clashes. It is in fact found that side chains close to a ligand often adopt atypical conformations due to effects such as intermolecular interactions. Heringa *et al*. found in a study of 112 tertiary protein-ligand structures that Asp, Glu, His, Met and Asn residues within 9 Å of a ligand inside a protein active site are the most likely to adopt atypical rotamer conformations [17]. The use of rotamer libraries to incorporate side chain flexibility has since been dropped within SLIDE in favour of an approach based on mean-field theory. In this, each rotatable bond that has the ability to resolve a steric clash between two atoms is awarded a probability weighted by the minimum energy angle and the number of non-hydrogen atoms displaced upon rotation. The details of mean-field theory are not included here as it is beyond the scope of this review. The minimum amount of rotation is then performed in order to relieve the maximum amount of steric clash over ten iterations of the mean-field theory algorithm.

In the above example of SLIDE, rotamer libraries were used to incorporate receptor flexibility after the ligand had already been docked into the receptor molecule. Kallblad and Dean [18] designed an approach in which an ensemble of receptor structures are generated using a rotamer library with a statistically representative sub set extracted. Docking experiments are then performed on each of the representative receptor conformations individually. This model was tested on the docking of an inhibitor (RS-104966) of human collegenase-1 (MMP-1). It is known that the RS-104966 induces large conformational change within the active site of MMP-1. Due to this induced change in receptor conformation it was not possible to dock RS-104966 directly into the crystal structure of MMP-1. It was found that RS-104966 was able to fit into some members of the rotamer conformation ensemble highlighting the importance of molecular flexibility in docking experiments.

Receptor flexibility can be incorporated into a docking experiment using MD simulation in one of two ways. The first approach involves running a simulation of the protein before docking and then extracting a number of structures of the protein from its trajectory. It is sometimes possible to have multiple experimental structures of a protein for example both NMR and x-ray structures, however this is not often the case and so MD simulation presents an alternative. The multiple receptor structures can then be either docked individually or combined to generate a "docking grid". Carlson *et al*. used the latter method to generate a "dynamic pharmacophore" for the HIV-1 integrase protein by combining 11 conserved binding site structures from a 500 ps MD simulation. The dynamic model was more successful than the "static" model, correctly predicting 15 out of 20 very active HIV-1 integrase inhibitors docked, 12 out of 23 active HIV-1 integrase inhibitors docked, and 62% of the ineffective inhibitors docked. The static model was unable to correctly predict any inhibitors.

The second way in which MD simulation may be used to incorporate ligand flexibility is in the refinement of docked complexes by MD simulation. In such an approach, rigid docking is performed to screen a database quickly, and then the highest scoring complexes are then simulated. Although this approach still involves a rigid description during the screening, effects such as induced fit and complete flexibility of both ligand and receptor are included during the simulation.

References

1. Jones, G., P. Willett, R.C. Glen, A.R. Leach, and R. Taylor, Journal of Molecular Biology, 1997. **267**: p. 727-748.
2. Morris, G.M., D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson, Journal of Computational Chemistry, 1998. **19**: p. 1639-1662.
3. Ewing, T.J.A. and I.D. Kuntz, J.Comp.Chem, 1997. **18**: p. 1175-1189.
4. Friesner, R.A., J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, and P.S. Shenkin, Journal of medicinal chemistry, 2004. **47**: p. 1739-1749.
5. Ewing, T.A., S. Makino, A.G. Skillman, and I. Kuntz, Journal of Computer-Aided Molecular Design, 2001. **15**: p. 411-428.
6. Shaik, M.S., M. Devereux, and P.L.A. Popelier, Molec.Phys., 2008. **106**: p. 1495-1510.
7. Shoichet, B.K., R.M. Stroud, D.V. Santi, I.D. Kuntz, and K.M. Perry, Science, 1993. **259**: p. 1445-1450.
8. Böhm, H.-J., Journal of Computer-Aided Molecular Design, 1994. **8**: p. 623-632.
9. Grädler, U., H.-D. Gerber, D.M. Goodenough-Lashua, G.A. Garcia, R. Ficner, K. Reuter, M.T. Stubbs, and G. Klebe, Journal of Molecular Biology, 2001. **306**: p. 455-467.
10. Lu, I.L., N. Mahindroo, P.-H. Liang, Y.-H. Peng, C.-J. Kuo, K.-C. Tsai, H.-P. Hsieh, Y.-S. Chao, and S.-Y. Wu, Journal of medicinal chemistry, 2006. **49**: p. 5154-5161.
11. Leach, A.R., B.K. Shoichet, and C.E. Peishoff, Journal of medicinal chemistry, 2006. **49**: p. 5851-5855.
12. Alonso, H., A.A. Bliznyuk, and J.E. Gready, Medicinal Research Reviews, 2006. **26**: p. 531-568.

13. Ferrari, A.M., B.Q. Wei, L. Costantino, and B.K. Shoichet, Journal of medicinal chemistry, 2004. **47**: p. 5076-5084.
14. Leach, A.R., Journal of Molecular Biology, 1994. **235**: p. 345-356.
15. Ponder, J.W. and F.M. Richards, Journal of Molecular Biology, 1987. **193**: p. 775-791.
16. Schnecke, V. and L. Kuhn, Perspectives in Drug Discovery and Design, 2000. **20**: p. 171-190.
17. Heringa, J. and P. Argos, Proteins: Structure, Function, and Bioinformatics, 1999. **37**: p. 44-55.
18. Källblad, P. and P.M. Dean, Journal of Molecular Biology, 2003. **326**: p. 1651-1665.

Appendix B

Literature review of Empirical Dispersion Corrections for Density Functional Theory

A note:

During the first year of my PhD, before the uptake of IQA methods, it was thought that QCTFF may account for the dispersion energy of a system using an empirical dispersion correction term, similar to those present in the literature. I was assigned the project and I wrote the following review of DFT dispersion corrections for my first year report.

As the group moved towards IQA, where the dispersion energy of a system will be accounted for by use of kriging models describing the exchange-correlation energy component (assuming a correlated ab initio *method is used), the project was never undertaken as it did not follow the direction in which QCTFF was headed, and I instead began to look at the PDB sampling covered in chapter three of this thesis.*

This appendix contains the original review of dispersion corrections. I am loath to omit it from the thesis entirely as it represents a significant amount of time and effort familiarising myself with the topic, however I am of the opinion that it does not fit into the introduction and therefore may not sit in the main text.

Dispersion Corrections

Density functional theory (DFT) has been successful in providing accurate *ab initio* calculation at relatively low computational cost. Due to the aforementioned reasons, DFT methods have been applied to aid in the understanding of a multitude of different chemical systems. A major drawback of DFT methods is that they are unable to model dispersion interactions. This is due to approximations in the treatment of the exchange and correlation of electrons which are responsible for dispersion interactions. Subsequently intermolecular interaction energies calculated by DFT methods are typically under bound, and require correction[1]. It is important that when using DFT to obtain intermolecular interaction energies to correct for the lack of dispersion.

There are a number of methods of correcting DFT to incorporate dispersion corrections. The first approach is to parameterise the density functional using data that includes the accurate binding energies of dispersion bound complexes. This is the approach that is taken by Truhlar *et al*. in the development of their "Minnesota" functionals [2]. The dataset used to parameterise a new functional, M05-2X, includes hydrogen-bonded complexes,

charge transfer complexes, dipole interaction complexes, π-π stacked complexes, and weakly bound complexes. The M05 and M05-2X functionals of Truhlar *et al*. were tested against 13 common density functionals including B3LYP, LSDA and PBE for their ability to reproduce the accurate binding energies of the 22 complexes of the S22 database (discussed in more detail in **chapter 3**). M05-2X outperformed all other functionals when predicting interaction energies for complexes bound only by dispersion. M05 also performed well, being the fourth best functional for dispersion bound complexes. For hydrogen bonded complexes, M05-2X and M05 performed third and eighth best, respectively, and for mixed dispersion bound and hydrogen bonded complexes M05-2X and M05 performed third and sixth best. More recently, Truhlar *et al*. have introduced the M06 family of density functionals. These are parameterised in a similar way to the M05 family. M06-2X had a mean unsigned error (MUE) of only 0.47 kcal mol^{-1} for the interaction energies of the S22 complexes, whereas an MUE of 0.75 kcal mol^{-1} was calculated by M05-2X for the same data. M06-2X had an MUE of only 0.36 kcal mol^{-1} for the dispersion only bound complexes.

The second approach to correcting DFT interaction energies for dispersion interactions is the use of an empirical correction that is performed after the DFT energy has been obtained from standard functionals such that

$$E_{DFT,corrected} = E_{DFT} + E_{dispersion} \tag{B.1}$$

There are many such corrections, $E_{dispersion}$, however most take a form similar to

$$E_{dispersion} = -s\frac{C_6^{AB}}{r_{AB}^6} f_{damp}(r) \tag{B.2}$$

where C_6^{AB} is a constant dependent on the atom types A and B, r_{AB} is the interatomic separation, and $f_{damp}(r)$ is a damping function that prevents the correction tending to infinity at small internuclear distances. Finally, s is a scaling factor that can be included to tailor the correction to different functionals.

The most widely used dispersion correction of this type is DFT-D of Grimme[3-5]. There are three generations of DFT-D; DFT-D1,-D2 and –D3. DFT-D1 and DFT-D2 both take the form of equation B.1 with the main difference being the manner in which the C_6^{AB} coefficients are obtained. In DFT-D1, for interacting atoms *A* and *B*,

$$C_6^{AB} = 2\frac{C_6^A C_6^B}{C_6^A + C_6^B} \tag{B.3}$$

where C_6^A and C_6^B are atomic coefficients that have been averaged across the possible hybridisation states of atoms A and B. Grimme acknowledged that the use of averaged C_6^A

coefficients rather than hybridization dependent coefficients lowers the accuracy of the correction, however inclusion would lead to problems arising in situations when hybridisation state is poorly defined. It is estimated that the omission of hybridisation dependent coefficients may lead to errors 10-20% of the order of the binding energy. The C_6^A parameters are included in **table B.1**.

Table B.1: C_6 coefficients for 6 atoms used in the DFT-D1 dispersion correction

	H	C	N	O	F	Ne
C_6 / J nm^6 mol^{-1}	0.16	1.65	1.11	0.70	0.57	0.45
R_0 / pm	111	161	155	149	143	138

For interacting atoms A and B, DFT-D2 calculates the C_6^{AB} coefficients by

$$C_6^{AB} = \sqrt{C_6^A C_6^B} \tag{B.4}$$

where atomic C_6^A coefficients are no longer averaged constants, but are derived from the London formula for dispersion. Thus,

$$C_6^A = 0.05 N I_p^A \alpha^A \tag{B.5}$$

where N is a constant that has values of 2, 10, 18, 36, and 54 for atoms of rows 1-5 on the periodic table respectively, I_p^A is the ionisation potential of atom A calculated at the DFT/PBE0 level of theory, and α^A is the static dipolar polarizability of A. Using such a method allows for C_6 coefficients for almost all elements of the periodic table to be obtained.

The damping function of both DFT-D1 and DFT-D2 is of the form

$$f_{damp}(r_{AB}) = \frac{1}{1+e^{-d(R_{AB}/R_r - 1)}} \tag{B.6}$$

where d is either 23 or 20 for DFT-D1 and DFT-D2 respectively, R_{AB} is the separation of atoms A and B, and R_r is the sum of the van der Waals radii of atoms A and B. At values of R_{AB} greater than R_r, $f_{damp}(r_{AB})$ has a value close to one, however as R_{AB} falls below R_r, $f_{damp}(r_{AB})$ rapidly tends to zero. An example of the role played by the damping function is present in the discussion of DFT-D3. The performance of DFT-D2 using the B97 functional (B97-D2) was tested by the prediction of 40 noncovalent complex interaction energies. The test set included hydrogen bonded complexes, non-aromatic complexes, aromatic

complexes, rare gas dimers, complexes with third row elements, and DNA base pairs. The mean absolute error of interaction energy for all 40 complexes was only 0.39 kcal mol^{-1}.

Hillier *et al*. [6] tested DFT-D2//BLYP/TZV(2d,2p) on a large database of 142 complexes including the S22 dataset, hydrogen bonded DNA base pairs, interstrand base pairs, stacked base pairs, and amino acid pairs. The mean unsigned error (MUE) relative to the accurate database values for the prediction of the S22 complexes' interaction energies was 0.72 kcal mol^{-1}. For the stacked base pairs, for which the interaction is dominated by dispersion, the MUE was only 0.53 kcal mol^{-1}. For all 142 complexes considered, the MUE was 0.76 kcal mol^{-1}. It was concluded that DFT-D2 is worthy of consideration when performing DFT calculation of biologically relevant molecules.

The most recent generation of DFT-D departed from the "elegant simplicity" of the previous generations. Three body terms are now considered, as well as pairwise C_8 coefficients. Both the pairwise C_6 coefficients and the van der Waals radii are also now obtained from first principles. The coordination of atoms A and B is accounted for through use of fractional coordination numbers. The corrected DFT energy, E_{DFT-D3}, is given by

$$E_{DFT-D3} = E_{DFT} + E_{disp} \tag{B.7}$$

where E_{disp} is a sum of two body and three body terms,

$$E_{disp} = E^{(2)} + E^{(3)} \tag{B.8}$$

The two body term can be generalised as

$$E^{(2)} = \sum_{AB} \sum_{n=6,8,10\ldots} s_n \frac{C_n^{AB}}{r_{AB}^n} f_{damp,n}(r_{AB}) \tag{B.9}$$

The first sum runs over all atom pairs A and B, and the second sum is over all n^{th} order dispersion terms C_n^{AB} included (6^{th}, 8^{th}, 10^{th} … etc). To maintain stability it is recommended to truncate the sum at n = 8. As before, s_n is a scaling factor dependent upon the density functional used in conjunction with the correction. The damping function $f_{damp,n}(r_{AB})$ used in DFT-D3 differs to DFT-D1/2, although it is noted that the importance in choosing a particular form of damping function is often overcomplicated, as when averaged over a number of density functionals, DFT-D is only weakly dependent on the choice. The form of the damping function used in DFT-D3 is based on the work of Chi and Head-Gordon due to its stability over multiple dispersion orders. It takes the form

$$f_{disp,n}(r_{AB}) = \frac{1}{1+6\left(r_{AB}/\left(s_{r,n}R_0^{AB}\right)\right)^{-\alpha_n}} \tag{B.10}$$

in which R_o^{AB} is the cut off radius; this is calculated for all possible atom pairs, resulting in possible 4465 values. $s_{r,n}$ is a scaling parameter dependent upon the density functional used, and α_n is a steepness parameter, set to 14. The effect of the damping function on the dispersion correction can be seen in **figure B.1. Figure B.1** (top left) shows the undamped values of $E^{(2)}$ where the energy can clearly be seen to become increasingly negative as r_{AB} falls below the equilibrium distance R_o^{AB} = 2.9 Å. **Figure B.1** (top right) shows the damping function switching from ~1 to ~0 as r_{AB} falls below R_o^{AB}, and the product of the two curves can be seen in **figure B.1** (bottom right).

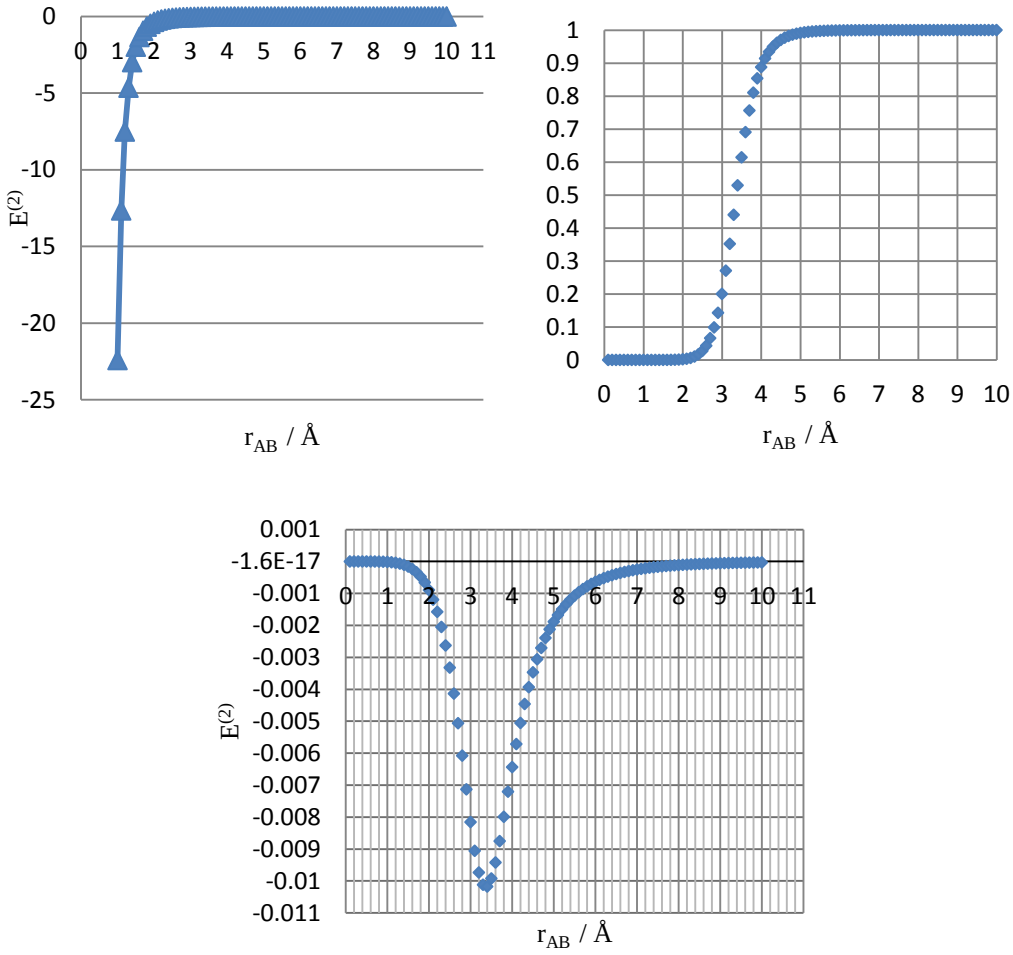


Figure B.1: The role played by the damping function in the DFT-D3 correction of two sp^3 hybridised C atoms. *Top left* the undamped interaction, *top right* the damping function, *bottom* the damped interaction

In DFT-D2 the dispersion coefficients were derived empirically from values such as the ionisation potential of an atom. In DFT-D3 the C_6^{AB} coefficients are derived using time dependent DFT starting from the Casimir-Polder formula

$$C_6^{AB} = \frac{3}{\pi}\int_0^{\infty} \alpha^A(i\omega)\alpha^B(i\omega)d\omega \qquad \text{(B.11)}$$

where $\alpha^A(i\omega)$ is the averaged dipole polarisability at imaginary frequency ω. The polarisability of free atoms is generally higher than that of an atom involved in chemical bonding, due to valence electrons being held tightly in covalent bonds. Therefore equation B.11 requires modification. Bonded C_6^{AB} coefficients were obtained by using the hydride of each atom, rather than the free atoms to give

$$C_6^{AB} = \frac{3}{\pi}\int_0^\infty d\omega \frac{1}{m}\left[\alpha^{A_mH_n}(i\omega) - \frac{n}{2}\alpha^{H_2}(i\omega)\right] \times \frac{1}{k}\left[\alpha^{B_kH_l}(i\omega) - \frac{l}{2}\alpha^{H_2}(i\omega)\right] \quad \text{(B.12)}$$

where m, n, k and l are stoichiometric factors, $\alpha^{A_mH_n}(i\omega)$, $\alpha^{B_kH_l}(i\omega)$ and $\alpha^{H_2}(i\omega)$ correspond to the average dipolar polarisabilities of molecules A_mH_n, B_kH_l and H_2 at imaginary frequencies ω. The hydrides of each atom are used simply because every atom other than the noble gas atoms forms a stable hydride. In equation B.12 for each atom the contribution of the hydrogen atoms to the molecular polarisability is removed by use of the H_2 dipolar polarisability. Although using reference molecules may be seen as a disadvantage, it allows the introduction of coordination number (CN) dependent coefficients $C_6^{AB}(CN^A, CN^B)$. This is achieved by replacing the reference hydride for common atoms such as carbon with a list of reference molecules with a range of coordination numbers for an atom. For example, a simple list of ethane, ethene, ethyne, C-H and C could be used to cover the different coordination environments in which carbon is found. When calculating the $C_6^{AB}(CN^A, CN^B)$ for the interaction between two atoms during DFT-D3, initially a coordination number is assigned to each atom *A* and *B* using the formula

$$CN^A = \sum_{B\neq A}^{N_{at}} \frac{1}{1+e^{k_1(K_2(R_{A,cov}+R_{B,cov})/r_{AB}^{-1})}} \quad \text{(B.13)}$$

Equation b.13 is a sum that runs over all other atoms in the system. k_1 and k_2 are constants with values 16 and 4/3 respectively. The new $C_6^{AB}(CN^A, CN^B)$coefficient is obtained from a two dimensional interpolation based on the values of the reference compounds $C_{6,ref}^{AB}(CN^A, CN^B)$. A larger number of reference compounds for a given element will mean a greater number of reference points during the interpolation and will produce more accurate C_6^{AB} coefficients.

As stated earlier, DFT-D3 introduces 8th order dispersion coefficients. These are obtained in a manner derived from the work of Starkschall and Gordon [7]:

$$C_8^{AB} = 3C_6^{AB}\sqrt{Q^AQ^B} \quad \text{(B.14)}$$

where

$$Q^A = s_{42}\sqrt{Z^A}\frac{\langle r^4\rangle^A}{\langle r^2\rangle^A} \quad \text{(B.15)}$$

$\langle r^4 \rangle^A$ and $\langle r^2 \rangle^A$ are multipole expectation values that have been obtained from atomic charge densities. The nuclear charge stabilising factor, Z^A is he nuclear charge, and s_{42} is a scaling factor that is redundant due to the scaling factor in equation B.9.

Returning to equation B.8, the three body term is given by

$$E^{(3)} = \sum_{ABC} f_{d,(3)}(\bar{r}_{ABC}) E^{ABC} \tag{B.16}$$

where $f_{d,(3)}(\bar{r}_{ABC})$ is a damping function and E^{ABC} is given by

$$E^{ABC} = \frac{C_9^{ABC}(3cos\theta_a cos\theta_b cos\theta_c + 1)}{(r_{AB} r_{BC} r_{CA})^3} \tag{B.17}$$

in which θ_a, θ_b and θ_c are the internal angles of the triangle formed by r_{AB}, r_{BC} and r_{CA}. C_9^{ABC} is approximated to be the geometric average of the C_6 coefficients,

$$C_9^{ABC} = -\sqrt{C_6^{AB} C_6^{AC} C_6^{BC}} \tag{B.18}$$

Despite the inclusion of three body terms, Grimme states that their contribution to the total energy is very small, especially for small to medium sized systems. For that reason the default setting for DFT-D3 is for three body terms to be switched off. DFT-D3 was tested on a number of data bases of accurate interaction energies including the S22 database, the large S22+ database, and others including PCONF, SCONF, ACONF and RG6, using a range of density functionals. In all cases DFT-D3 gave interaction energies with a mean absolute deviation lower than both the uncorrected density functional and the DFT-D2 interaction energies. The improvement of DFT-D3 over uncorrected DFT is significant, and therefore recommended for consideration when performing DFT calculation on such systems. The improvement upon DFT-D2 is much lower, of the order of only a few tenths of a kcal mol^{-1}. The increase in complexity from DFT-D1/2 to DFT-D3 is large, and although still insignificant with respect to the DFT calculation, DFT-D2 remains a simpler alternative to DFT-D3. DFT-D2 has the additional advantage of being incorporated into the popular GAUSSIAN *ab initio* software.

More recently, Friesner *et al*. have amassed a large database of highly accurate CCSD(T)/CBS level interaction energies[8]. The database includes 2027 CCSD(T) energies, which includes almost all the published data available. The database has been used to parameterise a DFT correction for B3LYP consisting of three parts- a Lennard-Jones type dispersion correction E_{LDC}, a hydrogen bonded term E_{HBC}, and a cation-pi interaction correction $E_{\pi+}$. The correction is named B3LYP-MM, and is described by

$$E_{B3LYP-MM} = E_{B3LYP} + E_{LDC} + E_{HBC} + E_{\pi+} \tag{B.19}$$

E_{LDC} takes the form of a Lennard-Jones 12-6 potential with parameters $r_{AB,equilibrium}$ and ε_{AB} where

$$r_{AB}^{equilibrium} = q(R_A^{VDW} + R_B^{VDW}) \tag{B.20}$$

and

$$\varepsilon_{AB} = \varepsilon_A \varepsilon_B \tag{B.21}$$

q in equation B.20 is a global scaling factor and R_A^{VDW} is the experimental van der Waals radius of A. Therefore the E_{LDC} correction term has only three parameters per atom that require parameterisation.

E_{HBC} is given by

$$E_{HBC} = \sum_{A<B} -\epsilon^{hb}(r_{AB} - r_o^{hb}) \tag{B.22}$$

and $E_{\pi+}$ is given by

$$E_{\pi+} = \sum_{A<B} -\epsilon^{\pi+}(r_{AB} - r_o^{\pi+}) \tag{B.23}$$

where ϵ^{hb}, r_o^{hb} and $\epsilon^{\pi+}$ are parameters that must be fit, and $r_o^{\pi+}$ was set to a value of 5.0 Å. Therefore only 5 parameters (q, ε_A, ϵ^{hb}, r_o^{hb} and $\epsilon^{\pi+}$) need be fitted for the complete B3LYP-MM correction. The performance of the correction is very impressive. A comparison of B3LYP-MM, B3LYP-D3 and M06-2X can be seen in **figure B.2**. Overall it is seen that B3LYP-MM performs better than both M06-2X and B3LYP-D3. All three methods were used (with the aug-cc-pVDZ basis set and counterpoise corrections) to predict the interaction energy for 1715 test systems and the results were compared with the CCSD(T) values. B3LYP-MM had a mean unsigned error (MUE) of 0.32 kcal mol^{-1}, M06-2X had an MUE of 0.67 kcal mol^{-1}, and B3LYP-D3 had an MUE of 0.87 kcal mol^{-1}.

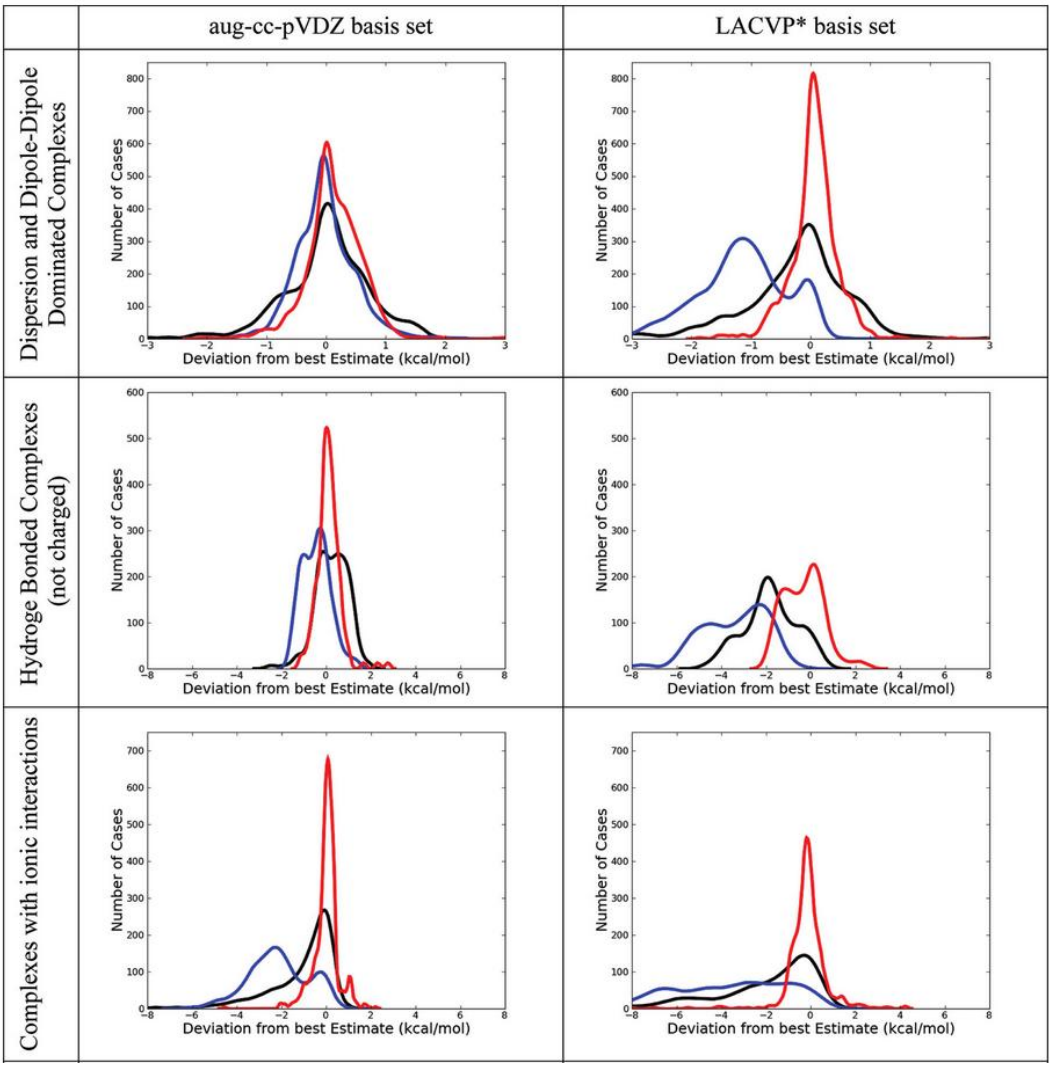


Figure B.2: Comparison of B3LYP-MM (red line), B3LYP-D3 (blue line) and M06-2X (black line) for the prediction of the interaction energies of intermolecular complexes. Ref. 111

References

1. Klimes, J. and A. Michaelides, The Journal of Chemical Physics, 2012. **137**: p. 120901-12.
2. Zhao, Y. and D.G. Truhlar, Journal of Chemical Theory and Computation, 2006. **3**: p. 289-300.
3. Grimme, S., Journal of Computational Chemistry, 2004. **25**: p. 1463-1473.
4. Grimme, S., Journal of Computational Chemistry, 2006. **27**: p. 1787-1799.
5. Grimme, S., J. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys., 2010. **132**: p. 154104-19.
6. Morgado, C., M.A. Vincent, I.H. Hillier, and X. Shan, Physical Chemistry Chemical Physics, 2007. **9**: p. 448-451.
7. Starkschall, G. and R.G. Gordon, The Journal of Chemical Physics, 1972. **56**: p. 2801-2806.
8. Schneebeli, S.T., A.D. Bochevarov, and R.A. Friesner, J. Chem. Theo. & Comp, 2011. **7**: p. 658-668.

Appendix C

List of Protein Crystal Structure Codes

The 80 protein crystal structure codes that amino acids were sampled from in **Chapter 3** when referring to the "small" pool of crystal structures.

1DP4	1HCL	1DMH	1CCD	1JW9	1REQ	1F6Y
1KC1	1O70	1PT7	1GLJ	1B3Q	1Q8R	1JMO
1A53	1CJC	1A6Q	1DOV	1JLY	1EWF	1N8P
1FL7	1FHE	1GOH	1AZO	2ETA	1CD0	1GH7
1KW8	1ILV	1JH5	1DT6	1QB3	1FC4	1JB6
2MSB	1GQW	1GCO	1SZJ	1ZXQ	1I72	1QKI
1B7V	1HBH	1DNC	1GMO	1AZY	1NUL	1QFJ
1GRE	1QF7	1A39	1K8T	1DLM	1HLG	1DIR
2BVW	1DZJ	1F8R	1AHP	1GSE	1QAG	
1EAE	1B2P	1FG3	2FHI	1TRB	1AY9	
1FUJ	1L5X	1M32	1TGJ	1FCJ	1C8B	
1M3K	1A8O	1HMD	1L5Z	1QMV	1M6B	

The 260 protein crystal structure codes that amino acids were sampled from in **Chapter 3** when referring to the "large" pool of crystal structures.

1A70	1PRG	1F9G	1K32	1GUZ	1JZ0	1FUJ	1QF9	1M5U	1HV8
1DP4	1RA4	1YTI	1RWR	1G2V	1MVX	1F6Y	1KIU	1DM3	1ILV
1KC1	1DIR	1A22	1N2M	1FAE	1AHP	1HXY	1E9N	1F8R	1SWA
1A53	1F5W	1PYP	1CQK	1L5X	2FHI	1KEX	1VHH	1MHN	1GQW
1FL7	1GN9	1GYV	2RAP	1BF2	1TGJ	3PFL	1FC5	1LNS	1TN3
1MRU	1A48	1E0J	1B3Q	1HJ6	1K1O	1CR7	1QUS	1ATZ	1GR3
1A99	1IO1	1IQA	1FD4	1BA1	1IJB	1JMO	1Q5Z	1FG3	1H4R
1G0T	1KJN	1FC3	1JLY	1FEC	1L5Z	1N8P	1AZO	1SXB	1HBH
1MA3	1DUV	1PT7	2ETA	1GMG	1A8O	1A7Q	1XWL	1GQN	1KMM
1FT9	1VAL	1A6Q	1HYQ	1QQ2	1KAO	1GH7	1DT6	1M32	1QF7
1KW8	1FDR	1E6J	1QB3	1H0O	1MB0	1MA1	1SZJ	1QAG	1BEE
1H6V	1QMV	1FTX	1ZXQ	1GSE	1M3K	1JB6	1IHN	1A0D	1AYB
1ZOO	1M6B	1KKE	1AZY	1IK4	1B7D	1QKI	1HQ0	1LM7	1EVQ
1QOL	1HMD	1B6C	1SBP	1PCF	1VCP	1RBC	1K44	1AY9	1PCZ
2MSB	1CCD	1CSM	1F6B	1TRB	1FEB	1QB4	1IHO	1I6I	1M0Z
1B7V	1K0E	1D1P	1GL0	1A04	1MMI	1MVB	1G88	1C8B	1DZJ
1DFN	1HF2	1GOH	1DLM	1I4W	1EZX	1JUQ	1REQ	2MHR	1YDV
1GRE	1A7N	1FVR	1HQN	1GMJ	1GMO	1KZQ	1CJV	1QML	1VJS
1TYF	1CJ1	1JH5	1HCL	1A0Z	1DMH	1K8T	1Q8R	1E2E	1DNC
1CQP	1MOL	1QD9	1QQC	1KXG	1NYL	1QFJ	1EWF	1O70	1CJQ
1MIZ	1K2F	1GCO	1O8Q	1IUG	1JW9	1O88	1QIM	1CJC	1A39
1A5W	1IN5	1I01	1GMI	1FCJ	3LYN	1AGN	1CD0	1FHE	1RG7
1IK3	1GOS	6PTD	1QC7	1LNH	1AYX	1B2P	1FC4	1BOI	1JV3
1A2Z	1PM4	2BVW	1MXE	1AD1	1EAE	1JIZ	1J54	2DBV	1B24
1I72	1QO4	1JA3	1TXX	1ML1	1A00	1LFK	1EH9	1MTZ	1GLJ
1RJ1	1NUL	1CLL	1CII	1HLG	1GEG	1DFQ	1IK6	1LJP	1DOV

Appendix D

MOROS.pl

Latest stable version: MOROS2.1.pl

Introduction

MOROS is a Perl script that fulfils two roles. Firstly, it extracts all amino acid residues of a certain type from .pdb files and writes .gjf files for each extracted residue. It is intended for use as an alternative to normal modes as a means of sampling structures that will then be used to build kriging models from. The second role of MOROS is to perform hydrogen sphere experiments.

This document aims to present the technical details of MOROS, and provide a discussion of the reasons why it works in the manner described. The structure of MOROS is based around seven core subroutines that extract perform the residue extraction experiments. A separate routine performs the horizon sphere experiment. This document will be split up according to the subroutine being discussed, with the following order:

A. Amino acid Extraction:
    1) Setting up for MOROS
    2) Initial input and new directories
    3) Subroutine: ExtractResidue
    4) Subroutine: runHAAD
    5) Subroutine: ChangeAndRemove
    6) Subroutine: makeXYZ
    7) Subroutine: makeGJF
    8) Screening .gjf files, subroutines: checkAtoms and checkBonds
    9) Post subroutine file handling
B. Horizon Sphere:

An overview of each subroutine is given at the beginning of each section, followed by a technical discussion of key operations.

Finally, a list of the future aims of MOROS will be presented. When writing MOROS, much care was taken to comment the code, almost line by line. If the reader finds that a section of this document is not clear, looking at the comments in code may make my description in this document clearer.

Before discussing the code itself, a brief comment must be made on the choice of writing MOROS in Perl rather than a scientific language such as Fortran. The reason is simple- most of the actions performed by MOROS are file handling tasks, and so a scripting language was the obvious choice. There are very few mathematical operations performed within MOROS, so Fortran would not have provided any advantage.

By looking at the structure of the code in more detail, one may initially think that it is doing more than it needs to just to write .gjf files (a .pdb and .xyz file is written before writing the .gjf file for each sampled geometry). This structure is a relic of how MOROS was originally written, but has been left as having a number of file formats for each sampled geometry is not a bad thing. Also, the most time consuming task of MOROS is searching the large .pdb

files for the desired residues, with the rewriting of the sampled structures being relatively very fast. Therefore there is no reason currently to change this structure.

A. Amino Acid Extraction

1) Setting up for MOROS

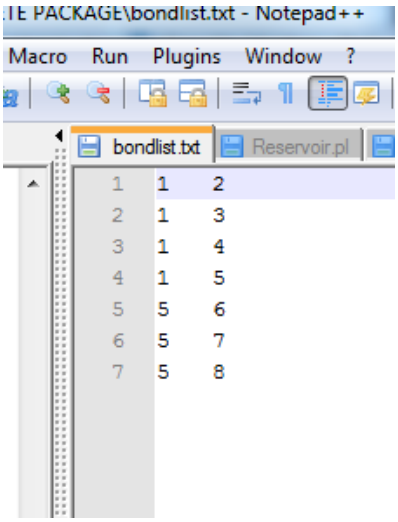
MOROS is designed to be ran on the CSF, using a SSH client such as SSH Secure Shell. The directory that the user wishes to run MOROS from should include the following files:

- MOROS.pl
- haad.exe
- A selection of .pdb files that the user wishes to sample from
- A file called bondlist.txt that has simply a list of the bonded atoms (example below).

Haad.exe is a program developed by Li *et al*[1] that is used to add hydrogen atoms to the sampled geometries as .pdb files typically do not include hydrogen atoms. An executable may be downloaded from the web, or may be found on the shared drive of the Popelier group.

There are no special requirements for the .pdb files included, and there is no limit on the number to be sampled from. Obviously, the more .pdb files, the more samples may be extracted. Some residues are very common, such as alanine and serine, and so relatively large numbers of structures may be sampled from a given set of .pdb files. Residues such as cysteine and tryptophan are less common and so require a larger number of .pdb files to obtain large numbers of sampled structures.

Finally, bondlist.txt is required for one of the final stages of MOROS where the sampled .gjf files are screened for unreasonable bond lengths. The contents is simply a list of all bonds between atoms A-X in the system, without double counts (no need for B-A). An example content of a bondlist.txt is shown below.



The observant reader will have realised that in to write the bondlist.txt file one needs to know the structure. This is unavoidable. Therefore it is recommended that a preliminary

[1] Y. Li, A. Roy, Y. Zhang, Haad: A Quick Algorithm for Accurate Prediction of Hydrogen Atoms in Protein Structures, *PLoS ONE*, 2009, **4**, 1-9

run of MOROS is first performed on a single .pdb file with the option to screen structures by bond length turned off. This is to quickly obtain a structure that can be used to write a bondlist.txt file. MOROS can then be ran on a larger number of .pdb files with the screening turned on to obtain a large number of high quality structures.

2) Initial input and new directories

Upon running MOROS, the user is asked to input three items of information:

1) The first item is the three letter code of the amino acid code that the user wishes to sample, for example SER, TRP, ALA, GLY. This must be in capitals because each line of the .pdb file is searched for a matching string of text and .pdb files are upper case.
2) The user is then asked how many atoms the complete amino acid residue including methyl caps will have. This is used later to check the completeness of the extracted structures. Examples are 23 atoms for serine, and 22 for alanine.
3) The third question asks the user if they wish to screen the output .gjf files for any unrealistically short or long bonds. Either enter "1" for yes or "2" for no. Doing this requires the bondlist.txt file to be present.

At this stage a number of directories are created. The directory "NEWPDBS" is where intermediate .pdb files for each structure are written to and where much of the work of MOROS takes place. The haad.exe executable is moved into this directory also, and the new .gjf files are found in this directory once MOROS is finished.

A directory for specific haad.exe output files is made: NEWPDBS/HHFILES. Details are outlined in the runHAAD subroutine explanation.

A number of new directories are also created:

- NEWPDBS
  - HHFILES
  - XYZFILES
  - GJFFILES
  - PDBFILES
  - PDBHFILES
  - BADBONDS
  - BADATOMS

These directories will be discussed in their relevant sections.

3) Subroutine: ExtractResidue

OVERVIEW:

The first subroutine, ExtractResidue, searches all .pdb files in the directory from which MOROS is being ran for the residue input by the user. When a residue of the correct type is found, the lines of the .pdb corresponding to the desired amino acid are written into a new .pdb file ready to have hydrogen atoms added by haad.exe. The atoms from the preceding and following amino acids that correspond to the methyl caps are included in the new .pdb file.

DISCUSSION:

In this subroutine, MOROS loops though the lines of a number of "parent" .pdb files searching for lines that match the format:

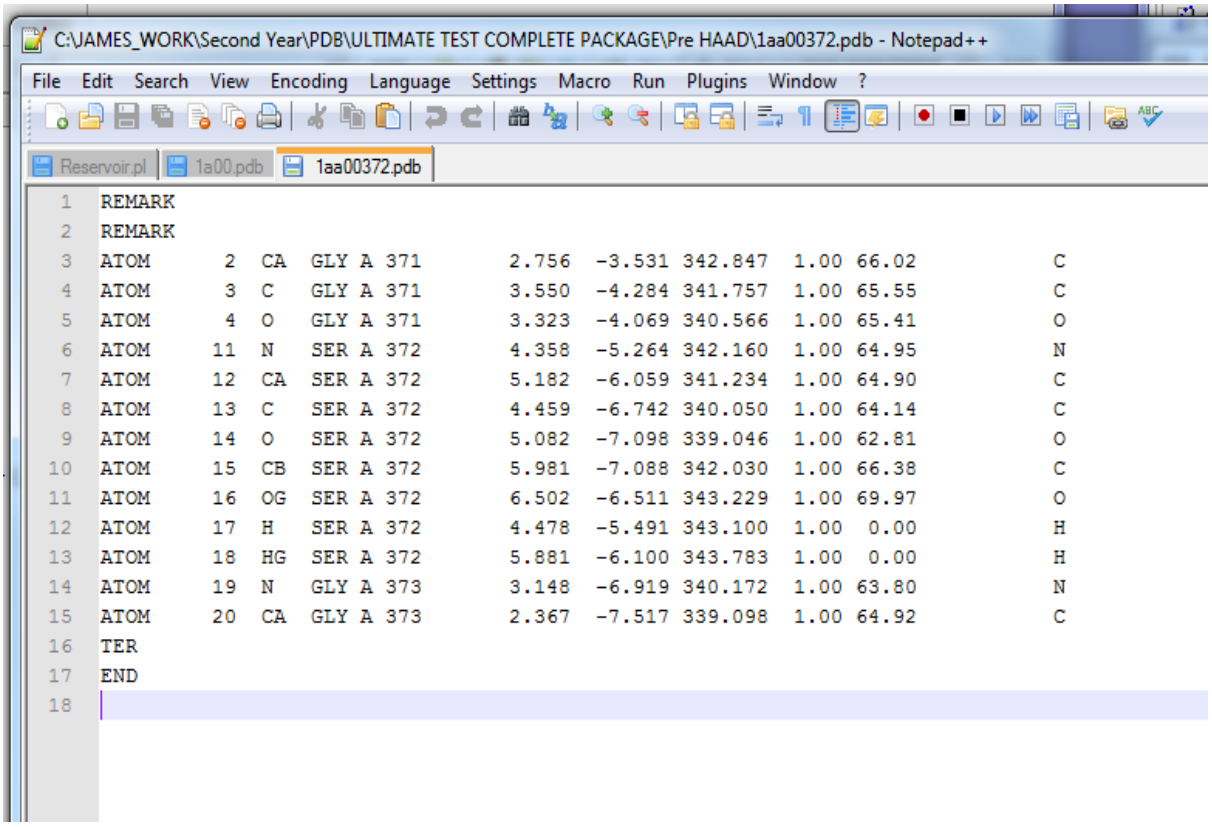
ATOM atomnumber atomtype residuetype strand residuenumber X coordinate Y coordinate Z coordinate 1.00 53.29 N

When a line is found where the residue type matches the three letter code of the amino acid input by the user, a new .pdb file is created which has a name corresponding to the "parent" pdb file that the amino acids was found in, and the residue number of the amino acid in that protein. For example, if MOROS was looping through a .pdb file called 1a09.pdb looking for serine residues, and the 23rd residue was a SER, a .pdb file called 1a090023.pdb would be made. Into this file MOROS copies the following lines of the "parent" .pdb file:

- The lines of the parent .pdb file that have the same residue number as the first line where the desired amino acid was found. Working by residue number (rather than residue type) is required to prevent other residues in the parent .pdb of the same type being included in the new .pdb file.
- The lines of the parent .pdb file that have a residue number one less than the residue number of the desired amino acid, and with atom types CA, C and O. This is to complete the N-terminus peptide bond and methyl cap. The line for the amide nitrogen is also added. Although this will be rewritten as a proton later; including it here gives a greater degree of sampling around the methyl rotation.
- The lines of the parent .pdb file that have a residue number one more than the residue number of the desired amino acid, and with atom types N and CA. This is complete the C-terminus peptide bond and methyl cap. The line for the acidic carbon is also added. Although this will be rewritten as a proton later, including it here gives a greater degree of sampling around the methyl rotation.

A technical note here is that the residue type in the lines of the parent file with residue number one more or one less than the desired residue is substituted with "GLY". By forcing the residues to be a glycine, during the next stage when haad.exe is used to add the hydrogen atoms, the alpha carbon atoms are seen as glycine residues and so two hydrogen atoms and a nitrogen are added, rather than one hydrogen, a side chain, and a nitrogen. This is an issue because as well as adding hydrogen atoms, if haad sees an incomplete residue it tries to "complete it" by adding the missing heavy atoms as well. Because the alpha carbon will become the methyl carbon in the cap, having the maximum number of hydrogen atoms bonded to it is desirable, and not having unwanted side chain heavy atoms and hydrogen atoms added in arbitrary places is also beneficial.

An example of an output .pdb file from this subroutine can be seen below. Note the first and last two lines written by MOROS. These are required to maintain the format of a .pdb file.



4) Subroutine: RunHAAD

OVERVIEW:

Run haad.exe to add hydrogen atoms to new .pdb files. Haad.exe output files have the extension .pdb.h.

DISCUSSION:

Haad, standing for Hydrogen Atom ADdition, is an algorithm written by Li *et al* that MOROS uses to add missing hydrogen atoms to the .pdb files output by the previous subroutine. See the appended reference for details on the specifics of how haad.exe work. Factors such as minimising steric clashes and the possibility of forming a hydrogen bond are considered when deciding where to place hydrogen atoms. If there are two possible positions for a hydrogen, then a second output file is written with the extension .pdb.h.h. Currently these files are moved to the "NEWPDBS/HHFILES" directory and are unused.

As previously stated, haad.exe also adds any missing heavy atoms in the .pdb files. Because the two residues at either end of the structure have been forced to be glycine residues, the only heavy atoms deemed to be missing are the oxygen atoms of the acid group of the following residue. These atoms along with unwanted hydrogen atoms are removed in the next subroutine.

5) Subroutine: ChangeAndRemove

OVERVIEW:

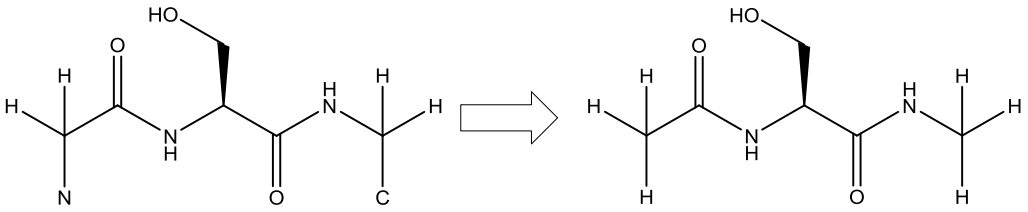
Reads .pdb.h files and writes a .pdb file which has only the atoms desired for the final structure. The methyl caps are completed by changing non-hydrogen atom on each methyl carbon to hydrogen and scaling the bond length to a C-H bond length. Moves .pdb.h.h files

to “HHFILES” directory. Output .pdb files share the name of the .pdb and .pdb.h file, with the addition that “_two” is added to the name, for example “1a090032_two.pdb”.

DISCUSSION:

The ChangeAndRemove code has a very similar structure to the ExtractResidue subroutine because the task which it performs is very similar. ChangeAndRemove loops though each line of the .pdb.h files and extracts the atoms that correspond the methyl capped amino acid. It does this in a similar method to ExtractResidue, using the residue number and atom type specified in each line to identify the desired atoms to write into the output file.

The notable difference to ExtractResidue is that ChangeAndRemove rewrites the nitrogen and carbon atoms bonded to the terminal methyl carbon atoms and shortens the bond length. This is illustrated below.



When the loop matches either the nitrogen atom or the carbon atom that needs replacing, the new distance is calculated by the subroutine ReDistance. This calculates new Cartesian coordinates that will give a bond length of 0.985 Å (as this is the default C-H bond length used within the haad.exe software).

- Initially the distance between the methyl carbon and the atom of interest is calculated
- A scaling factor, n, of $\frac{0.985}{distance}$ is then calculated.
- The new coordinates are then calculated using the following relationships:
- $x_{new} = n(x_{old} - x_{methyl\ carbon}) + x_{methyl\ carbon}$
- $y_{new} = n(y_{old} - y_{methyl\ carbon}) + y_{methyl\ carbon}$
- $z_{new} = n(z_{old} - z_{methyl\ carbon}) + z_{methyl\ carbon}$

The new distance of 0.985 Å has been chosen for consistency with the two methyl hydrogen atoms added by haad.exe, and therefor is open for discussion. If the user wishes to change this, simply changing the value of 0.985 Å in the code will do this. When writing the line for the output file, the old coordinates are substituted for the new ones and the old atom type is substituted for H.

6) Subroutine: makeXYZ

OVERVIEW:

makeXYZ writes a .xyz file for each _two.pdb file and also a .xyz file containing all sampled geometries. Incomplete geometries ignored.

DISCUSSION:

A very straightforward subroutine. The only point worthy of discussion is that upon reading the .pdb file the number of atoms is compared to the expected value input by the user when initially running MOROS. The .xyz file is not written if the number of atoms is incorrect.

7) Subroutine: makeGJF

OVERVIEW:

Writes .gjf files from the .xyz file written above, numbered from SYSNAME0001.gjf to the total number of sampled structures. It also translates the Cartesian coordinates to the origin to prevent problems occurring later in the Pipeline process where .wfn files can be written incorrectly if the Cartesian coordinates have both three digits before the decimal place **and** a minus sign. Large coordinates are a consequence of large sampling geometries from large proteins where residues may exist large distances from the origin.

DISCUSSION:

The key point of discussion here is the translation of the structures to the origin. This has to be done otherwise Gaussian writes wavefunctions with coordinates ********** if they exceed the character limit. This is a problem when a sampled residue is far from the origin of the protein along the negative direction of the Cartesian axes, i.e y = -100.362 Å. Because Gaussian converts the coordinates into Bohr, coordinates in Angstroms that do not appear to be problematic will be upon conversion (Bohr are smaller than Angstrom!).

To centre the molecule, the coordinates of the first atom are read and then all the atoms have the coordinates of the first atom subtracted from their own coordinates. To prevent all structures from having their first atom superimposed, a variable is calculated that is dependent upon the initial x-coordinate,x, and this is added to the x-coordinate of each atom. The variable, v, is calculated as:

$$v = \sqrt{\frac{n \times x}{5}}$$

Where $n = -1$ if x is negative and $n = 1$ if x is positive.

The default .gjf header line is written as:

B3LYP/aug-cc-pVDZ integral=ultrafine 6D 10F nosymm out=wfn scf=tight

This is open to discussion. Options to include B3LYP/apc-1 or HF/6-311G** could be added if desired, however, the Pipeline gives the option to compile the .gjf files with these levels of theory already, so this is not strictly necessary.

8) Screening .gjf files, subroutines: checkAtoms and checkBonds

SUMMARY:

The subroutine checkAtoms ensures that all of the .gjf files have the same elements written in the same order in all .gjf files. Files with inconsistencies are moved to the "NEWPDBS/BADATOMS" directory.

checkBonds calculates the distance between all bonded atoms (as specified in the input file bondlist.txt), and if the distance does not lie within a range of accepted values it the .gjf file is moved to the "NEWPDBS/ BADBONDS" directory.

DISCUSSION:

The two subroutines discussed here are included in MOROS because of the imperfect nature of .pdb files which can cause spurious geometries to be sampled. Example causes of these undesirable structures are incomplete residues, inclusion of hydrogen atoms, and averaged/low resolution structures that give unrealistic bond lengths.

BADATOMS reads the 0001.gjf file into an array and loops through the lines, writing a list of the elements present in the order that they are read. It then loops through all the .gjf files, reads the order of the elements in the same way as above, and then compares the list with that of the 0001.gjf list. If there are any differences then the .gjf file is moved to the directory "NEWPDBS/BADATOMS".

BADBONDS reads the bondlist.txt file and to get the list of the bonded atoms. It then loops through all the .gjf files and reads the Cartesian coordinates and calculates the distance between all bonded atoms. The distance is calculated by a separate subroutine, calcDistance. Once the distance is calculated a counter is added to i the value is greater than 1.7 Å or less than 0.8 Å. If the counter for a .gjf file is greater than 0, the .gjf file is moved to the directory "NEWPDBS/BADBONDS". The upper and lower limits are arbitrary and may be changed.

9) Post subroutines file handling

The final task that MOROS takes is to move all output files into their relative folders and report how many files of each type are present. This is useful not only because the final number of sampled structures is written, but it also breaks down at what stage different files are being screened out.

Nothing beyond simple file handling is involved at this stage and so no detailed discussion is warranted.

The number of sampled structures before any screening is printed on the screen, followed by the number of files removed during the BADATOMS subroutine, the number removed during the BADBONDS subroutine, and finally the remaining number of complete, ready to use .gjf files.

B. Horizon Sphere

The second role of MOROS.pl is to build fragments of a protein around a central atom, including all atoms within an increasing radius. The default fragment radii are 1.5 Å up to 10 Å, with a step size of 0.5 Å. When selecting to run a horizon sphere experiment, the user is asked for the .pdb file name and the central atom for the fragments to build around.

The .pdb file used should already have hydrogen atoms added, however once the fragments have been build, they must then have hydrogen atoms added around the edge of the fragment to satisfy the valence of the outer atoms (which were bonded to atoms outside of the fragment radius).

Final Comments

As it currently exists, MOROS is able to extract amino acid residues from a number of .pdb files. The output structures are written as .gjf files. The structures have methyl caps constructed from the relevant atoms of the residues either side of the sampled residue. This is completed with no apparent bugs.

MOROS.pl is also able to perform hydrogen sphere experiments, building protein fragments around a central atom.

Appendix E

A List of Atomic Charges for all Systems Studied in Chapter 5

The following table contains the range, standard deviation and average value of the charge for all atoms studied in **Chapter 5**.

	Charged			Neutral		
	Range	Standard Deviation	Average	Range	Standard Deviation	Average
Glutamic acid						
N1	0.671	0.110	-1.074	0.710	0.110	-1.080
H2	0.344	0.058	0.394	0.312	0.055	0.400
C3	0.710	0.103	0.355	0.721	0.099	0.360
H4	0.440	0.061	0.070	0.403	0.047	0.075
C5	0.454	0.066	0.033	0.381	0.060	0.039
C6	1.122	0.203	1.118	1.124	0.195	1.121
H7	0.409	0.062	0.010	0.394	0.049	0.033
H8	0.424	0.059	0.026	0.370	0.049	0.046
C9	0.427	0.066	0.019	0.380	0.058	0.027
O10	0.658	0.117	-1.006	0.577	0.113	-0.998
H11	0.341	0.047	0.008	0.388	0.043	0.056
H12	0.498	0.058	0.022	0.350	0.050	0.068
C13	1.222	0.229	1.181	1.179	0.217	1.173
O14	0.545	0.101	-1.017	0.518	0.105	-0.964
O15	0.567	0.115	-1.096	0.656	0.133	-0.960
C16	1.036	0.170	1.172	1.111	0.175	1.167
C17	0.415	0.062	-0.043	0.410	0.062	-0.037
O18	0.592	0.101	-1.033	0.543	0.100	-1.011
H19	0.256	0.040	0.038	0.291	0.042	0.050
H20	0.354	0.042	0.040	0.340	0.040	0.050
H21	0.290	0.040	0.037	0.286	0.041	0.049
N22	0.736	0.112	-1.051	0.800	0.108	-1.051
C23	0.691	0.118	0.276	0.691	0.110	0.277
H24	0.345	0.059	0.410	0.334	0.053	0.415
H25	0.316	0.050	0.031	0.265	0.043	0.043
H26	0.427	0.055	0.030	0.375	0.046	0.042
H27	0.433	0.056	0.052	0.373	0.051	0.062
H28				0.261	0.051	0.550
Aspartic Acid						
N1	0.766	0.120	-1.033	0.758	0.121	-1.041

H2	0.384	0.062	0.401	0.316	0.054	0.411
C3	0.684	0.109	0.342	0.658	0.108	0.357
H4	0.387	0.059	0.070	0.320	0.045	0.082
C5	0.407	0.060	0.014	0.399	0.051	0.014
C6	1.165	0.210	1.115	1.135	0.211	1.105
H7	0.414	0.055	0.034	0.381	0.044	0.072
H8	0.447	0.054	0.030	0.444	0.048	0.079
C9	1.292	0.242	1.206	1.208	0.225	1.213
O10	0.682	0.126	-1.019	0.683	0.122	-0.996
O11	0.563	0.111	-0.992	0.776	0.115	-0.982
O12	0.582	0.123	-1.108	0.677	0.135	-0.965
N13	0.898	0.122	-1.005	0.743	0.119	-0.998
C14	0.756	0.122	0.239	0.757	0.114	0.243
H15	0.358	0.066	0.404	0.327	0.057	0.417
H16	0.473	0.056	0.033	0.331	0.048	0.051
H17	0.487	0.058	0.031	0.377	0.048	0.052
H18	0.439	0.057	0.058	0.432	0.052	0.067
C19	1.061	0.191	1.144	1.177	0.191	1.147
C20	0.425	0.058	-0.054	0.396	0.052	-0.051
O21	0.577	0.113	-1.028	0.628	0.114	-1.016
H22	0.308	0.047	0.046	0.294	0.039	0.066
H23	0.385	0.048	0.040	0.334	0.040	0.055
H24	0.402	0.045	0.033	0.411	0.043	0.054
H25				0.337	0.055	0.563
Lysine						
N1	0.657	0.103	-1.059	0.635	0.101	-1.058
C2	0.651	0.098	0.351	0.657	0.094	0.347
C3	0.851	0.118	1.242	0.843	0.123	1.246
O4	0.416	0.053	-1.098	0.381	0.053	-1.099
C5	0.351	0.049	0.040	0.334	0.047	0.040
C6	0.337	0.048	0.037	0.316	0.048	0.036
H7	0.358	0.051	0.393	0.336	0.051	0.392
H8	0.345	0.041	0.059	0.317	0.040	0.055
H9	0.383	0.047	0.013	0.350	0.041	0.009
H10	0.320	0.042	0.033	0.295	0.036	-0.001
H11	0.295	0.041	0.014	0.265	0.039	-0.007
N12	0.658	0.101	-1.041	0.646	0.101	-1.042
H13	0.331	0.051	0.414	0.322	0.052	0.404
C14	0.646	0.100	0.293	0.669	0.101	0.297
H15	0.354	0.042	0.040	0.343	0.042	0.033
H16	0.366	0.040	0.051	0.382	0.042	0.041
H17	0.373	0.044	0.047	0.359	0.046	0.039
C18	0.896	0.116	1.256	0.905	0.120	1.261
O19	0.406	0.060	-1.085	0.380	0.056	-1.098
C20	0.333	0.046	-0.020	0.320	0.046	-0.019

H21	0.272	0.035	0.044	0.257	0.032	0.037
H22	0.272	0.039	0.051	0.251	0.032	0.042
H23	0.236	0.036	0.051	0.229	0.033	0.037
H24	0.352	0.042	0.020	0.308	0.039	0.008
C25	0.405	0.052	0.036	0.364	0.046	0.051
H26	0.251	0.038	0.009	0.212	0.030	-0.009
H27	0.263	0.038	0.012	0.229	0.030	-0.008
C28	0.531	0.075	0.198	0.657	0.084	0.261
H29	0.251	0.039	0.089	0.224	0.035	-0.004
H30	0.235	0.036	0.080	0.249	0.034	-0.002
N31	0.453	0.061	-1.001	0.562	0.091	-0.927
H32	0.266	0.039	0.477	0.283	0.057	0.319
H33	0.238	0.038	0.479	0.294	0.057	0.322
H34	0.240	0.038	0.478			
Histidine						
N1	0.748	0.115	-1.050	0.662	0.111	-1.046
H2	0.403	0.057	0.391	0.406	0.056	0.388
C3	0.655	0.109	0.345	0.620	0.106	0.340
H4	0.384	0.049	0.082	0.350	0.044	0.072
C5	0.307	0.049	0.030	0.319	0.049	0.039
C6	0.830	0.134	1.223	0.815	0.136	1.228
H7	0.419	0.048	0.079	0.413	0.044	0.043
H8	0.358	0.050	0.076	0.364	0.047	0.036
C9	0.562	0.104	0.395	0.565	0.109	0.329
O10	0.478	0.062	-1.070	0.415	0.058	-1.084
N11	0.720	0.123	-1.070	0.650	0.118	-1.121
C12	0.610	0.115	0.348	0.756	0.134	0.321
H13	0.274	0.038	0.518	0.299	0.055	0.443
C14	0.854	0.148	0.888	0.803	0.157	0.798
H15	0.233	0.033	0.127	0.285	0.034	0.045
H16	0.185	0.027	0.167	0.211	0.029	0.074
C17	0.891	0.128	1.239	0.894	0.133	1.235
C18	0.306	0.048	-0.033	0.304	0.047	-0.032
O19	0.482	0.065	-1.074	0.422	0.063	-1.084
H20	0.281	0.042	0.058	0.258	0.037	0.045
H21	0.291	0.041	0.060	0.257	0.035	0.051
H22	0.297	0.042	0.063	0.268	0.037	0.048
N23	0.765	0.111	-1.016	0.753	0.109	-1.028
C24	0.615	0.102	0.270	0.648	0.106	0.274
H25	0.365	0.064	0.403	0.328	0.060	0.405
H26	0.403	0.041	0.048	0.331	0.041	0.035
H27	0.300	0.039	0.056	0.320	0.043	0.041
H28	0.368	0.045	0.076	0.342	0.048	0.060
N29	0.666	0.117	-1.137	0.792	0.137	-0.955
H30	0.176	0.028	0.507			

Arginine						
N1	0.309	0.056	-1.125	0.621	0.097	-1.072
H2	0.141	0.017	0.426	0.311	0.045	0.397
C3	0.288	0.041	0.367	0.618	0.083	0.355
H4	0.156	0.022	0.038	0.252	0.031	0.051
C5	0.151	0.022	0.083	0.278	0.046	0.047
C6	0.647	0.097	1.379	1.060	0.182	1.193
H7	0.187	0.028	0.000	0.331	0.040	0.020
H8	0.192	0.029	-0.028	0.279	0.039	-0.004
C9	0.130	0.021	0.081	0.354	0.046	0.047
O10	0.302	0.039	-1.166	0.613	0.107	-1.055
H11	0.161	0.024	0.002	0.311	0.030	-0.004
H12	0.193	0.030	0.009	0.328	0.036	0.011
C13	0.286	0.045	0.366	0.569	0.089	0.346
H14	0.131	0.022	0.029	0.226	0.031	0.005
H15	0.129	0.021	0.017	0.219	0.034	0.009
N16	0.440	0.073	-1.087	0.785	0.106	-1.018
H17	0.105	0.013	0.456	0.290	0.046	0.384
C18	0.770	0.124	1.428	1.060	0.180	1.147
N19	0.403	0.063	-1.138	0.737	0.117	-0.957
H20	0.140	0.019	0.487	0.647	0.104	-1.030
H21	0.154	0.016	0.480	0.307	0.055	0.317
N22	0.402	0.063	-1.124	0.320	0.053	0.382
C23	0.298	0.046	0.371	0.303	0.050	0.388
H24	0.146	0.022	0.423	0.724	0.102	-1.052
H25	0.201	0.035	0.025	0.620	0.094	0.297
H26	0.222	0.030	0.029	0.307	0.042	0.409
H27	0.208	0.033	0.024	0.255	0.040	0.043
C28	0.575	0.081	1.392	0.294	0.043	0.040
C29	0.125	0.020	0.049	0.370	0.043	0.041
O30	0.297	0.036	-1.186	0.959	0.148	1.246
H31	0.163	0.028	0.024	0.244	0.043	-0.010
H32	0.126	0.023	0.013	0.558	0.087	-1.080
H33	0.127	0.020	0.029	0.218	0.032	0.038
N34	0.354	0.063	-1.133	0.275	0.032	0.030
H35	0.111	0.015	0.480	0.186	0.029	0.041
H36	0.100	0.015	0.481			

Appendix F

Published Work

PCCP

PERSPECTIVE

View Article Online
View Journal | View Issue

Multipolar electrostatics

Salvatore Cardamone,[ab] Timothy J. Hughes[ab] and Paul L. A. Popelier*[ab]

Cite this: *Phys. Chem. Chem. Phys.*, 2014, **16**, 10367

Received 26th November 2013,
Accepted 5th April 2014

DOI: 10.1039/c3cp54829e

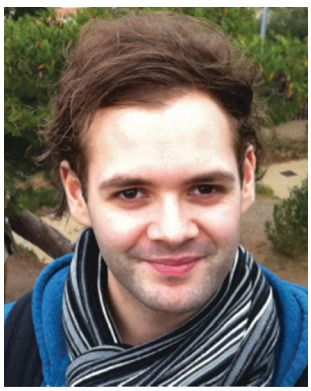
www.rsc.org/pccp

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 π-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

[a] *Manchester Institute of Biotechnology (MIB), 131 Princess Street, Manchester, M1 7DN, UK. E-mail: pla@manchester.ac.uk*
[b] *School of Chemistry, University of Manchester, Oxford Road, Manchester, M13 9PL, UK*



Salvatore Cardamone

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.

Timothy J. Hughes

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.

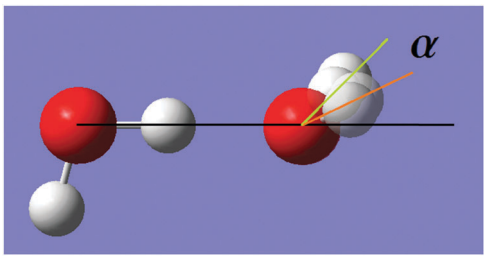


Fig. 1 A schematic geometry of the global minimum of the water dimer obtained by a point-charge model ($\alpha \sim 25^\circ$, orange) and an *ab initio* calculation ($\alpha \sim 45^\circ$, green).

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.[1] 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" α of about 45°, while a point-charge model (*i.e.* single charge for each atom) will generate an α angle about 25°. Disregarding the irrelevant details of the level of theory used or the exact nature of this point charge-model,[2] 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[4] 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.[3,5,6] 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 **R**. The position vector **r** describes the left distribution by sweeping its volume, while the vector **r**′ does the same for the right distribution while being based at **R**.

One can think of an infinitesimal bit of electronic charge density, located at **r**, interacting with an infinitesimal bit of charge density located at **r**′ + **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

Paul L. A. Popelier

Paul Popelier was educated in Flanders up to PhD level and is a full Professor of Chemical Computation. He has published more than 170 contributions, containing 3 books, a commercially released computer program, and 25 single-author items. He is an EPSRC Established Career Fellow and a Fellow of the Royal Society of Chemistry. Currently his group mainly develops a novel force field based on Quantum Chemical Topology (QCT). In his spare time he plays modern jazz piano and composes.

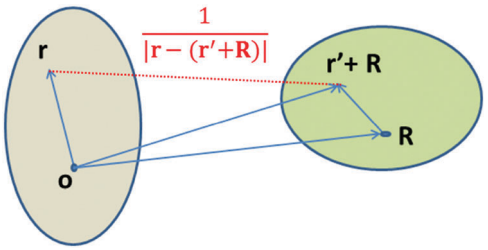


Fig. 2 A schematic representation of the Coulomb interaction between two charge distributions (left and right). Each distribution is described by a vector (**r** or **r**′ + **R**) that sweeps the volume of the respective distribution. The vector **r** marks the position of an infinitesimal charge density that interacts with another infinitesimal charge density at position **r**′ + **R**. The two infinitesimal charge distributions are separated by a distance $|\mathbf{r} - (\mathbf{r}' + \mathbf{R})|$. The inverse of this distance can be separated into factors depending only on **r**, on **r**′ or on **R**.

infinitesimal charged density. This full summation is in fact a six-dimensional (6D) integral running over the three-dimensional coordinate space of the left charge distribution and that of the right distribution. The use of so-called addition theorems[3] enables the expression $1/|\mathbf{r} - (\mathbf{r}' + \mathbf{R})|$ to be factorised into factors that depend on a single variable only, that is **r**, **r**′ or **R**. This factorisation leads to multipole moments. They can be *pre-computed*, that is, calculated separately for the electron distribution described by **r**, and separately for that described by **r**′. It is important to realise that this pre-computation of the multipole moments is done independently of the geometry of their interaction. This explains the enormous advantage of this pre-computation because the 6D Coulomb integral does not have to be calculated anymore. Instead, it can be replaced by two 3D integrals, each yielding the numerical values of the respective multipole moments. However, the price paid for this huge advantage is possible divergence of the multipolar series expansion. In any event, multipole moments describe the full complexity of the charge distribution at hand, whether it is molecular, atomic, ionic, covalent or metallic. With increasing multipolar rank ℓ the multipole moments "pick up" an increasing number of features of this charge distribution. The more complicated the deviation from isotropy of the charge distribution the more multipole moments are necessary, exactly in order to capture this anisotropy, which point-charges miss.

The atomic charge, which is intrinsically isotropic, corresponds to the zeroth moment or the monopole moment ($\ell = 0$). It should be clear from the discussion above that the atomic charge offers only the simplest of descriptions of two interacting electron distributions. Higher rank moments (*i.e.* dipole moment or $\ell = 1$, quadrupole moment or $\ell = 2$, *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 $L = \ell_A + \ell_B + 1$. 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 is 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 description 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.[7] 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[8] by Ritchie and Copenhaver published in 1995 compared the electrostatic potential generated by an atom-centered multipole expansion (up to $\ell = 3$) with that generated by potential-derived charges surrounding some natural and synthetic nucleic acid bases. The multipolar electrostatics always improved the *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[9] or much more recent work[10] 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 implementation 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 multipolar 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[11] today in spite of both suffering from the same known pitfalls, such as accurately reproducing the experimentally observed radial distribution function (RDF) for O$\cdots$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[2,12] and the ST2 potential[13] 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 °C, 100 °C]). TIP4P and SCP/E predicted maximum densities at −15 °C and −38 °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 ~7.5 was also reproduced by QCTwater.

QCTwater also outperformed[29] 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.[31] 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.[32] 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.[36–42] 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 hydrides 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 $\sim 100^\circ$ and $\sim 260^\circ$, 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 functions[54] 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 simulations[56] 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[58] 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[59] 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[58] 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 $N_H\cdots O$

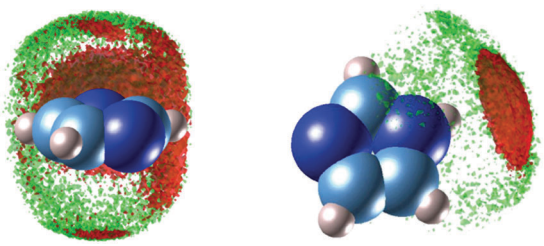


Fig. 3 Comparison of QCT (green) and AMBER (red) in terms of Spatial Distribution Functions (SDFs) of $N_H\cdots O$ (isovalue = 2.0 (left) and 3.0 (right)). The carbon atoms are shaded in light blue. [Source: *J. Phys. Chem. B*, 2011, **115**, 11389.]

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[58] also revealed pronounced differences in the number and ratio of stacked *versus* hydrogen-bonded 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$\cdots$N/O[61] or C–H$\cdots\pi$.[62,63] 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 be both as strong as hydrogen bonding, ranging from $\sim$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.[68–71] Torii and Yoshida showed that the quadrupole moment Θ_{zz} 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.[69] Halogen bonding was proven to be dictated primarily by electrostatic effects through the work of Tsuzuki *et al.*[68] who studied C_6F_6X and C_6H_6X 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$\cdots$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 readably available, QM/MM calculations were suggested as an alternative to force fields.[73] According to very recent work[74] 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^+, Mg^{2+} and Ca^{2+}. 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 CHARM27 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[79] do not account for this type of system. The electrostatic properties of saturated hydrocarbons were modelled[80] 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[81] demonstrated that the use of DMA convincingly exposed deficiencies in this "separated point-charge" model.

A study[82] 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 $C_6H_6\cdots C_6H_5X$, one may postulate relative interaction energies based upon the identity of X. An electronegative group seizes electronic population from the π-system in C_6H_5X. This effect results in a decreased electrostatic repulsion between the two interacting π-systems, and enhances the net electrostatic interaction. An electropositive group, on the other hand, would contribute towards the π-system, thus inducing a net electrostatic repulsion by the opposing mechanism. Such a simplistic model has been criticised by several groups,[83] claiming that an enhanced interaction is observed relative to the benzene dimer, regardless of the identity of X. It was, however, confirmed that electron-withdrawing groups enhanced the interaction more than those which were electron-donating.

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.[84] Such distinct electrostatic features may by captured by the implementation of multipole moments, shown in work[85] where three benzenoids with large negative, neutral and large positive quadrupole moments were complexed with a small molecule (HF, H_2O, NH_3 and CH_4) geometry optimised. The multipole moments clearly governed the energetically favoured geometries of the various complexes.

Past work[86] 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[87] by a point-charge model and a full DMA up to hexadecapole 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 *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 *ab initio* derived conformations of benzene dimers has been reiterated in the work of Koch and Egert.[88] 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,[89] proteins[90] or carbohydrate–protein interactions in immune complexes,[91] to name a few. Biological processes such as molecular recognition and catalysis are frequently stabilised by interactions between the π-density of aromatic systems. Point-charges provide a poor description of the electronic distribution of aromatic systems, and the XED force field aimed[79] 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.*[92] 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[10] 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 *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,[93] 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)[94] values.

Marshall *et al.*[95] ran simulations on β-hairpin structures of model polypeptides involving cation–π interactions between cationic $(Me)_n$-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

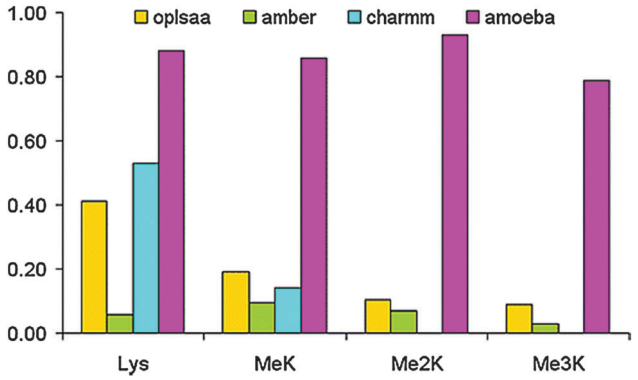


Fig. 4 Summary of the percentage of experimentally observed NOEs for the four model β-hairpin peptides (lysine residues) predicted by 100 ns MD simulation in explicit solvent by the four force fields compared. [Source: *J. Am. Chem. Soc.*, 2012, **134**, 15970.]

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.[96]

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-PP[101,102] 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

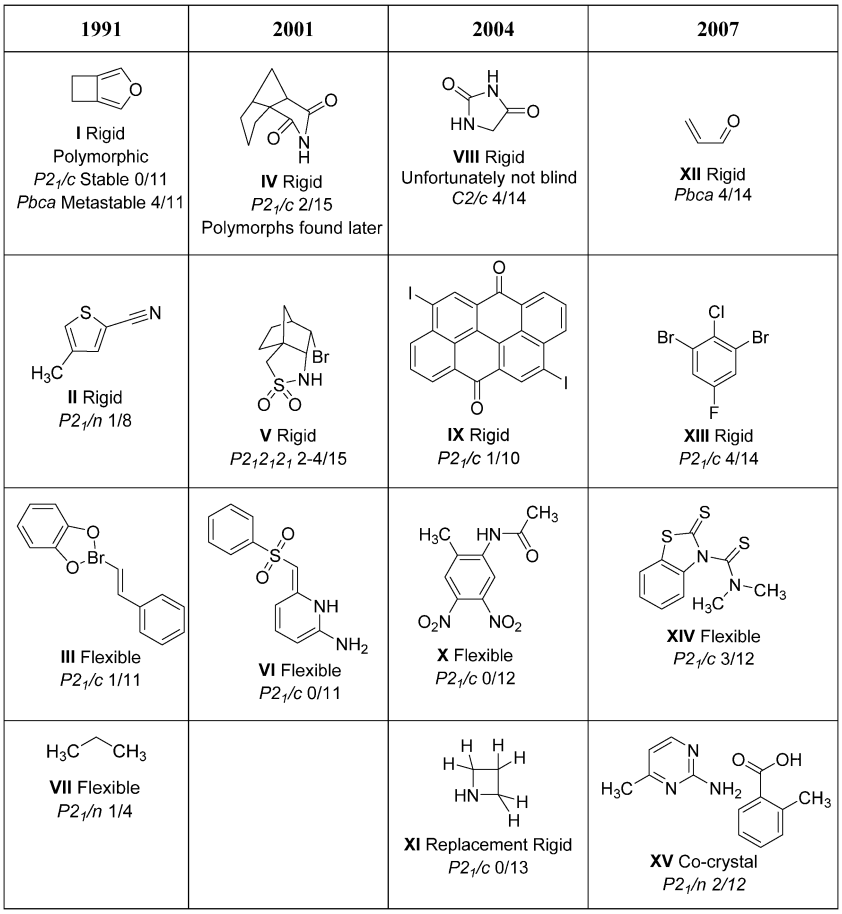


Fig. 5 The structures used in the first four blind tests on crystal structure prediction (CSP). Source: *Int. Rev. Phys. Chem.*, 2008, **27**, 541.

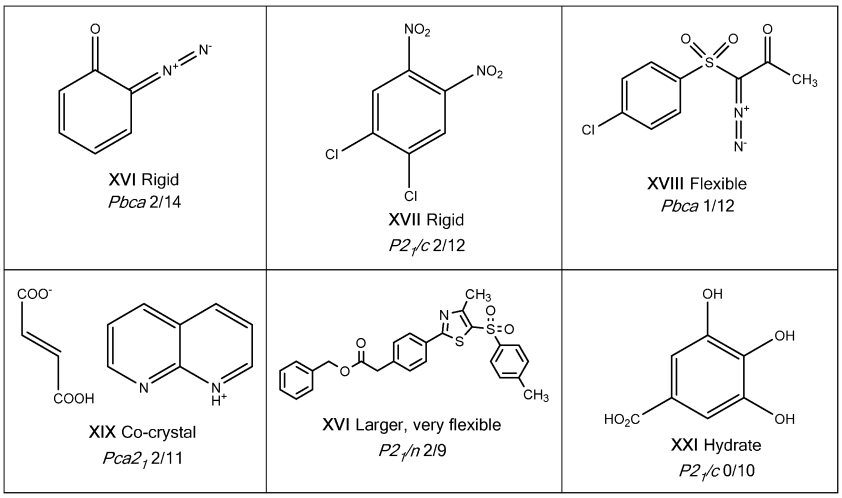


Fig. 6 The six structures used in the CSP2010 blind test. [Source: *Acta Crystallogr., Sect. B: Struct. Sci.*, 2011, **67**, 535.]

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.* 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.[112] 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.* observed[113] that for 50 organic molecules with many polymorphic 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-amino-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 α-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.* 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.[125] Due to

Table 1 The number of successful predictions of the crystal structure of the 21 molecules used in the CSP blind tests using different electrostatic models (point charges, multipole moments or other). The number of methods corresponds to the maximum number of groups using an electrostatic method in the above test. This number may vary between structures within a blind test as not all groups attempted to predict all structures. The numbers outside of parentheses are the number of successful predictions within the top three structures provided by a method, and the numbers within parentheses are the number of correctly predicted structures outside of the top three structures. Only successful top three results are included in CSP1999 study

	Point charge	Multipole	Other
CSP1999			
Number of methods	6	2	3
I	4	0	0
II	1	0	0
III	1	0	0
VII	0	1	0
CSP2001			
Number of methods	11	3	5
IV	1(8)	1(0)	0(0)
V	3(5)	1(1)	0(1)
VI	0(4)	0(0)	0(0)
CSP2004			
Number of methods	11	3	4
VIII	1(4)	2(2)	1(1)
IX	0(5)	1(1)	0(2)
X	0(4)	0(2)	0(1)
XI	0(0)	0(2)	0(1)
CSP2007			
Number of methods	5	4	5
XII	0(2)	1(2)	3(0)
XIII	0(2)	3(1)	1(0)
XIV	0(1)	2(1)	1(0)
XV	0(0)	1(1)	1(0)
CSP2010			
Number of methods	5	3	6
XVI	0(2)	1(2)	1(0)
XVII	0(2)	1(2)	1(0)
XVIII	0(3)	0(0)	1(0)
XIX	1(2)	0(1)	1(0)
XX	0(0)	2(0)	0(1)
XXI (i)	0(0)	0(3)	0(1)
XXI (ii)	0(1)	2(1)	0(1)
XXI (iii)	0(2)	0(3)	0(1)
XXI (iv)	0(1)	0(3)	0(1)

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.[126] 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 $CH_3(CH_2)_xCH_3$, 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 $E = 2E(CH_3) + xE(CH_2)$ is accurate to approximately 0.06 kcal mol^{-1}.[127] 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.[128]

Armed with this, the demonstration that multipole moments possess[129] some amenability to atom typing should follow. In one case study,[130] 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 then ATOM.

Many years later, Mooij *et al.*[131] 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.[132]

Mooij *et al.*[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 ~ 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,[133] CAMM[134] or QCT partitioning[135]). Considering our group's research interests, we focus here on QCT-based techniques. Focusing on energy, Bader and Beddall[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[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,[138] in a method denoted Transferable Atom Equivalents (TAEs).[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-*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 α-conformations. The TAE multipole analysis (TAE-MA) reproduced the features of the electrostatic potential generated at *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[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 (*e.g.* too many for atom types for N) or too coarse-grained (*e.g.* C atom types not diverse enough). Later, an extensive investigation[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,[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 α-helical segment of the protein crambin was studied. The electrostatic energy was probed by consideration of the multipole moment expansion (up to rank $\ell = 4$) centred at a C_α. A new set of multipole moments for C_α was calculated for each structure dictated by the growing horizon sphere. The interaction energy between C_α 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 ~ 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,[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,[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.*,[147] for example, demonstrated the capacity of transferring experimental density parameters for small peptides, based upon the Hansen–Coppens formalism,[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.[149] A more computationally-orientated route has been developed in parallel to the one above,[143] 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.[150] 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).[151] Multipole moment expansions for non-hydrogen atoms were truncated at ranks $\ell = 4$ and $\ell = 2$ 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 Ser, Leu and Gln, 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, which is in keeping with the sign of ΔE 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.[152]

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[153] 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 $\ell = 4$) 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.[154] 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 ($\ell = 0, 1, 2$). Secondly, for the higher-order Hansen–Coppens multipole moments ($\ell = 3, 4$) 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,[154–156] 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.[157] 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.[158] However, this study implemented techniques not optimised for the output of particularly accurate results, using a highly parallelisable CHARM++ interface in conjunction with the AMBER force field and the NAMD molecular dynamics package,[159] *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.[7,160] It should be noted that the interaction involving 'higher order' multipole moments ($\ell \geq 1$) 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.[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.[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,[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.[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[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.[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[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[170] and is called DL_MULTI.

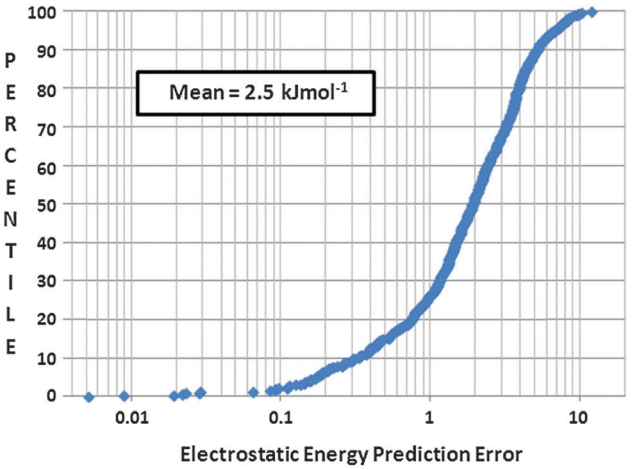


Fig. 7 Error (kJ mol^{-1}) in the total electrostatic energies predicted by machine learning (kriging) for a set of 24 local energy minima of capped histidine. The mean energy error of the sum of 328 atom–atom electrostatic energy values in capped histidine is 2.5 kJ mol^{-1}. The maximum error is 11.9 kJ mol^{-1}. One can read off the curve that ~80% (of the 539 test configurations) have an error of less than 1 kcal mol^{-1} (~4 kJ mol^{-1}). [Source: *J. Comput. Chem.*, 2013, **34**, 1850.]

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.[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.[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 *et al.*[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 Multipole Optimised Energetics for Biomolecular Applications).[52] AMOEBA has been proven effective in a variety of biomolecular simulations, ranging from solvated ion systems[77,174,175] to organic molecules[176] and peptides.[177–179] 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 $\ell = 2$), 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 ΔE values across all conformations at the MP2/TQ, ωB97/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. Kaminský and Jensen,[180] 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 *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.*,[181] 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 2 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

Serine	AMBER94	MM2	MM3	MMFFs	OPLS_2005	AMBER99	CHARMM27	AMOEBA
Number of correct minima	21	19	15	27	23	21	19	34
Number of erroneous minima	1	2	3	6	3	7	2	7
Percentage erroneous	4.8	10.5	20.0	22.2	13.0	33.3	10.5	20.6
MAD [kJ mol^{-1}]	8.9	10.7	14.0	7.4	4.1	10.9	11.1	4.2
Cysteine	**AMBER94**	**MM2**	**MM3**	**MMFFs**	**OPLS_2005**	**AMBER99**	**CHARMM27**	**AMOEBA**
Number of correct minima	23	23	21	28	25	29	21	44
Number of erroneous minima	1	1	3	7	3	5	5	1
Percentage erroneous	4.3	4.3	14.3	25.0	12	17.2	23.8	2.3
MAD [kJ mol^{-1}]	6.3	10.0	13.9	5.4	4.6	5.4	6.6	3.1

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.[182,183] Much like the generic AMOEBA force field, atomic multipole moment expansions up to $\ell = 2$ 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) $\sim$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 $\ell = 2$) 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 heteroatom lone pairs and bond barycentres.

Table 4 Dimerisation energies for formamide and glycyl dipeptide systems computed at the SCF level of theory and SIBFA force field (kJ mol^{-1}). Also given are the energies resulting from individual multipole–multipole interaction ranks

	Formamide	Glycyl dipeptide	Difference, δ
E_{Coulomb}(SCF)	−22.1	−14.3	7.8
E_{Coulomb}(SIBFA)	−20.9	−13.2	7.7
E_{00}(SIBFA)	−16.2	−20.0	−3.8
E_{10}(SIBFA)	−1.6	5.8	7.4
E_{11}(SIBFA)	0.5	−0.6	−1.1
E_{20}(SIBFA)	−3.1	1.8	4.9
E_{21}(SIBFA)	0.1	−1.0	−1.1
E_{22}(SIBFA)	−0.5	0.8	1.3

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 glycyl-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 glycyl-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 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}

	Ethylbenzene	*p*-Cresol	Isopropanol	Imidazole	Methylethyl sulfide	Acetic acid	RMSD
AMOEBA	−0.73	−7.26	−5.58	−10.11	−1.78	−5.69	0.68
TIP3P-like	−0.89	−10.72	−5.29	−10.87	−2.94	−7.46	1.96
Experiment	−0.70	−6.10	−4.70	−9.6	−1.50	−6.70	

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 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 Zn^{2+} and the most basic amino acid, glycine, Rogalewicz *et al.*[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 entireties. 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_{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.[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 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 C–X···P interactions for the various halobenzenes (where X = F, Cl, Br or I and P = Mg^{2+}, H or 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 σ-hole. It is also immediately evident that the multipolar electrostatics implemented in SIBFA directly superimposes on these curves. As such, the effect of

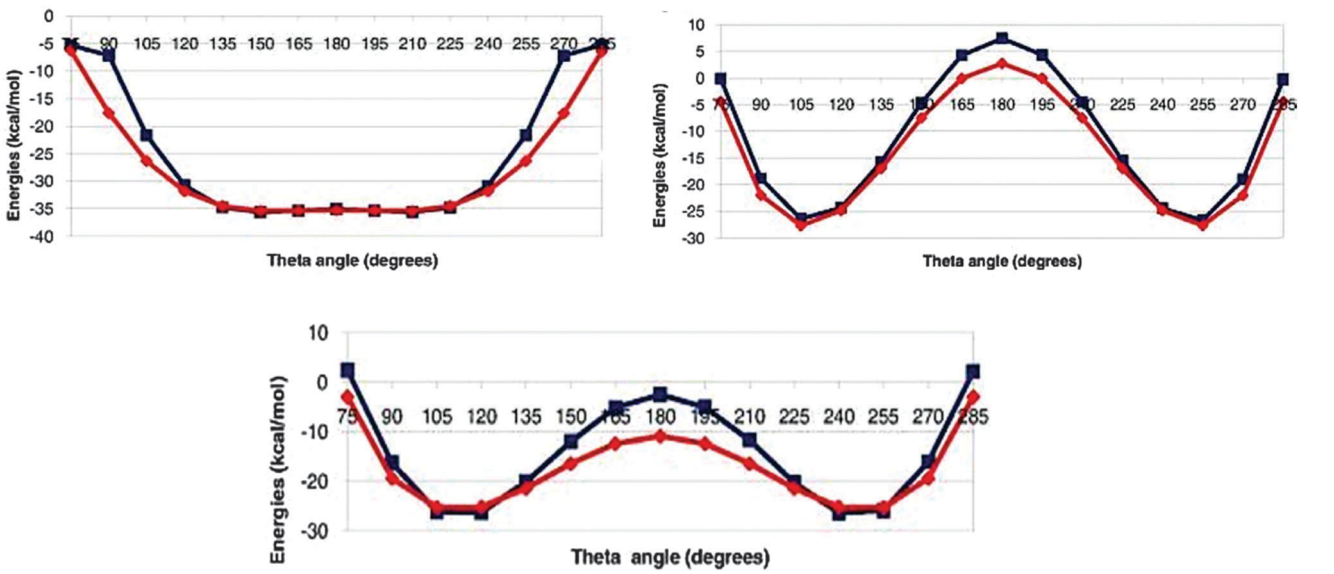


Fig. 8 Energetic profiles as a function of the C–X···P angle for (from top left, clockwise) fluorobenzene-, bromobenzene-, and chlorobenzene-Mg^{2+} systems. Energy from the SIBFA force field is marked in red while Coulombic energy from the Reduced Variational Space (RVS) scheme is in blue. Energetic 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 PES 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 experiment for 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 specifics. 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

1 V. Hanninen, T. Salmi and L. Halonen, *J. Phys. Chem. A*, 2009, **113**, 7133–7137.
2 M. W. Mahoney and W. L. Jorgensen, *J. Chem. Phys.*, 2000, **112**, 8910–8922.
3 A. J. Stone, in *The Theory of Intermolecular Forces*, ed. J. S. Rowlinson, Clarendon Press, Oxford, 1st edn, 1996.
4 A. Raval, S. Piana, M. P. Eastwood, R. O. Dror and D. E. Shaw, *Proteins: Struct., Funct., Bioinf.*, 2012, **80**, 2071–2079.
5 G. B. Arfken and H. J. Weber, *Mathematical Methods for Physicists*, Academic Press, San Diego, California, USA, 1995.
6 P. L. A. Popelier and A. J. Stone, *Mol. Phys.*, 1994, **82**, 411–425.
7 G. A. Cisneros, M. Karttunen, P. Ren and C. Sagui, *Chem. Rev.*, 2014, **114**, 779–814.
8 J. P. Ritchie and A. S. Copenhaver, *J. Comput. Chem.*, 1995, **16**, 777–789.

9 L. Joubert and P. L. A. Popelier, *Phys. Chem. Chem. Phys.*, 2002, **4**, 4353–4359.
10 D. Ghosh, D. Kosenkov, V. Vanovschi, C. F. Williams, J. M. Herbert, M. S. Gordon, M. S. Schmidt and L. V. Slipchenko, *J. Phys. Chem. A*, 2010, **114**, 12739–12754.
11 B. Guillot, *J. Mol. Liq.*, 2002, **101**, 219–260.
12 W. L. Jorgensen and J. Tirado-Rives, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 6665–6670.
13 F. H. Stillinger and A. Rahman, *J. Chem. Phys.*, 1974, **60**, 1545–1557.
14 C. Millot and A. J. Stone, *Mol. Phys.*, 1992, **77**, 439–462.
15 A. J. Stone, *Chem. Phys. Lett.*, 1981, **83**, 233–239.
16 P. Cieplak, P. Kollman and T. Lybrand, *J. Chem. Phys.*, 1990, **92**, 6755–6760.
17 U. Niesar, G. Corongiu, E. Clementi, G. R. Kneller and D. K. Bhattacharya, *J. Phys. Chem.*, 1990, **94**, 7949–7956.
18 P. Barnes, J. Finney, J. Nicholas and J. Quinn, *Nature*, 1979, **282**, 459–464.
19 C. Millot, J. C. Soetens, M. T. C. M. Costa, M. P. Hodges and A. J. Stone, *J. Phys. Chem. A*, 1998, **102**, 754–770.
20 M. P. Hodges, A. J. Stone and S. S. Xantheas, *J. Phys. Chem. A*, 1997, **101**, 9163–9168.
21 R. S. Fellers, C. Leforestier, L. B. Braly, M. G. Brown and R. J. Saykally, *Science*, 1999, **284**, 945–948.
22 F. N. Keutsch and R. J. Saykally, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 10533–10540.
23 S. Y. Liem, P. L. A. Popelier and M. Leslie, *Int. J. Quantum Chem.*, 2004, **99**, 685–694.
24 P. L. A. Popelier and É. A. G. Brémond, *Int. J. Quantum Chem.*, 2009, **109**, 2542–2553.
25 P. L. A. Popelier and F. M. Aicken, *ChemPhysChem*, 2003, **4**, 824–829.
26 P. L. A. Popelier, in *Quantum Chemical Topology: on Bonds, Potentials. Structure, Bonding. Intermolecular Forces and Clusters*, ed. D. J. Wales, Springer, Heidelberg, Germany, 2005, pp. 1–56.
27 R. F. W. Bader, *Atoms in Molecules. A Quantum Theory*, Oxford Univ. Press, Oxford, Great Britain, 1990.
28 P. L. A. Popelier, *Atoms in Molecules. An Introduction*, Pearson Education, London, Great Britain, 2000.
29 S. Y. Liem and P. L. A. Popelier, *J. Chem. Theory Comput.*, 2008, **4**, 353–365.
30 E. R. Batista, S. S. Xantheas and H. Jonsson, *J. Chem. Phys.*, 2000, **112**, 3285–3292.
31 O. Engkvist and A. J. Stone, *J. Chem. Phys.*, 2000, **112**, 6827–6833.
32 M. C. Foster and G. E. Ewing, *J. Chem. Phys.*, 2000, **112**, 6817–6826.
33 R. Taylor, O. Kennard and W. Versichel, *J. Am. Chem. Soc.*, 1983, **105**, 5761–5766.
34 J. P. M. Lommerse, S. L. Price and R. Taylor, *J. Comput. Chem.*, 1997, **18**, 757–774.
35 I. Nobeli, S. L. Price, J. P. M. Lommerse and R. Taylor, *J. Comput. Chem.*, 1997, **18**, 2060–2074.
36 H. Umeyama and K. Morokuma, *J. Am. Chem. Soc.*, 1977, **99**, 1316–1332.
37 P. A. Kollman, *Acc. Chem. Res.*, 1977, **10**, 365–371.
38 G. J. B. Hurst, P. W. Fowler, A. J. Stone and A. D. Buckingham, *Int. J. Quantum Chem.*, 1986, **29**, 1223–1239.
39 A. P. L. Rendel, G. B. Bacskay and N. S. Hush, *Chem. Phys. Lett.*, 1985, **117**, 400–413.
40 T. W. Rowlands and K. Somasundram, *Chem. Phys. Lett.*, 1987, **135**, 549–552.
41 P. Hobza and C. Sandorfy, *J. Am. Chem. Soc.*, 1987, **109**, 1302–1307.
42 G. Alagona and A. Tani, *J. Chem. Phys.*, 1981, **74**, 3980–3988.
43 A. D. Buckingham and P. W. Fowler, *J. Chem. Phys.*, 1983, **79**, 6426–6428.
44 A. D. Buckingham, P. W. Fowler and J. M. Hutson, *Chem. Rev.*, 1988, **88**, 963–988.
45 A. D. Buckingham and P. W. Fowler, *Can. J. Chem.*, 1985, **63**, 2018–2025.
46 J.-H. Lii and N. L. Allinger, *J. Phys. Org. Chem.*, 1994, **7**, 591–609.
47 J.-H. Lii and N. L. Allinger, *J. Comput. Chem.*, 1998, **19**, 1001–1016.
48 P. Cieplak, J. Caldwell and P. Kollman, *J. Comput. Chem.*, 2001, **22**, 1048–1057.
49 J. Kong and J.-M. Yan, *Int. J. Quantum Chem.*, 1993, **46**, 239–255.
50 M. S. Shaik, M. Devereux and P. L. A. Popelier, *Mol. Phys.*, 2008, **106**, 1495–1510.
51 P. Ren, C. Wu and J. W. Ponder, *J. Chem. Theory Comput.*, 2011, **7**, 3143–3161.
52 J. W. Ponder, C. Wu, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. J. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson and T. Head-Gordon, *J. Phys. Chem. B*, 2010, **114**, 2549–2564.
53 R. J. Wheatley and S. L. Price, *Mol. Phys.*, 1990, **71**, 1381–1404.
54 M. A. Freitag, M. S. Gordon, J. H. Jensen and W. J. Stevens, *J. Chem. Phys.*, 2000, **112**, 7300–7306.
55 P. L. A. Popelier and D. S. Kosov, *J. Chem. Phys.*, 2001, **114**, 6539–6547.
56 S. Liem and P. L. A. Popelier, *J. Chem. Phys.*, 2003, **119**, 4560–4566.
57 M. S. Shaik, S. Y. Liem, Y. Yuan and P. L. A. Popelier, *Phys. Chem. Chem. Phys.*, 2010, **12**, 15040–15055.
58 S. Y. Liem, M. S. Shaik and P. L. A. Popelier, *J. Phys. Chem. B*, 2011, **115**, 11389–11398.
59 S. Y. Liem and P. L. A. Popelier, *Phys. Chem. Chem. Phys.*, 2014, **16**, 4122–4134.
60 S. K. Panigrahi and G. R. Desiraju, *Proteins: Struct., Funct., Bioinf.*, 2007, **67**, 128–141.
61 G. G. R. Desiraju and T. Steiner, *The weak hydrogen bond: in structural chemistry and biology*, Oxford University Press on Demand, 2001.
62 T. Steiner and G. Koellner, *J. Mol. Biol.*, 2001, **305**, 535–557.
63 M. Levitt and M. F. Perutz, *J. Mol. Biol.*, 1988, **201**, 751–754.

64 S. Melandri, *Phys. Chem. Chem. Phys.*, 2011, **13**, 13901–13911.
65 P. Auffinger, S. Louise-May and E. Westhof, *J. Am. Chem. Soc.*, 1996, **118**, 1181–1189.
66 A. Maris, S. Melandri, M. Miazzi and F. Zerbetto, *ChemPhysChem*, 2008, **9**, 1303–1308.
67 N. Ramasubbu, R. Parthasarathy and P. Murray-Rust, *J. Am. Chem. Soc.*, 1986, **108**, 4308–4314.
68 S. Tsuzuki, A. Wakisaka, T. Ono and T. Sonoda, *Chem. – Eur. J.*, 2012, **18**, 951–960.
69 H. Torii and M. Yoshida, *J. Comput. Chem.*, 2010, **31**, 107–116.
70 I. Alkorta, F. Blanco, M. Solimannejad and J. Elguero, *J. Phys. Chem. A*, 2008, **112**, 10856–10863.
71 P. Politzer, J. S. Murray and T. Clark, *Phys. Chem. Chem. Phys.*, 2010, **12**, 7748–7757.
72 M. A. A. Ibrahim, *J. Comput. Chem.*, 2011, **32**, 2564–2574.
73 Y. Lu, T. Shi, Y. Wang, H. Yang, X. Yan, X. Luo, H. Jiang and W. Zhu, *J. Med. Chem.*, 2009, **52**, 2854–2862.
74 A. J. Stone, *J. Am. Chem. Soc.*, 2013, **135**, 7005–7009.
75 Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang and P. Kollman, *J. Comput. Chem.*, 2003, **24**, 1999–2012.
76 D. Jiao, C. King, A. Grossfield, T. A. Darden and P. Ren, *J. Phys. Chem. B*, 2006, **110**, 18553–18559.
77 A. Grossfield, P. Ren and J. W. Ponder, *J. Am. Chem. Soc.*, 2003, **125**, 15671–15682.
78 A. Grossfield, *J. Chem. Phys.*, 2005, **122**, 024506–024510.
79 G. Chessari, C. A. Hunter, C. M. R. Low, M. J. Packer, J. G. Vinter and C. Zonta, *Chem. – Eur. J.*, 2002, **8**, 2860–2867.
80 D. E. Williams and T. L. Starr, *Comput. Chem.*, 1977, **1**, 173–177.
81 S. L. Price, *Chem. Phys. Lett.*, 1985, **114**, 359–364.
82 C. A. Hunter and J. K. M. Sanders, *J. Am. Chem. Soc.*, 1990, **112**, 5525–5534.
83 S. Grimme, *Angew. Chem., Int. Ed.*, 2008, **47**, 3430–3434.
84 C. D. Sherrill, *Acc. Chem. Res.*, 2012, **46**, 1020–1028.
85 B. K. Mishra and N. Sathyamurthy, *J. Phys. Chem. A*, 2007, **111**, 2139–2147.
86 P. W. Fowler and A. D. Buckingham, *Chem. Phys. Lett.*, 1991, **176**, 11–18.
87 S. L. Price and A. J. Stone, *J. Chem. Phys.*, 1987, **86**, 2859–2868.
88 U. Koch and E. Egert, *J. Comput. Chem.*, 1995, **16**, 937–944.
89 C. A. Hunter and X.-J. Lu, *J. Mol. Biol.*, 1997, **265**, 603–619.
90 C. A. Hunter, J. Singh and J. M. Thornton, *J. Mol. Biol.*, 1991, **218**, 837–846.
91 C. A. Hunter, *Chem. Soc. Rev.*, 1994, **23**, 101–109.
92 G. Hill, G. Forde, N. Hill, W. A. Lester Jr., W. Andrzej Sokalski and J. Leszczynski, *Chem. Phys. Lett.*, 2003, **381**, 729–732.
93 M. Tafipolsky and B. Engels, *J. Chem. Theory Comput.*, 2011, **7**, 1791–1803.
94 B. Jeziorski, R. Moszynski and K. Szalewicz, *Chem. Rev.*, 1994, **94**, 1887–1930.
95 X. Zheng, C. Wu, J. W. Ponder and G. R. Marshall, *J. Am. Chem. Soc.*, 2012, **134**, 15970–15978.
96 S. L. Price, *Int. Rev. Phys. Chem.*, 2008, **27**, 541–568.
97 J. P. M. Lommerse, W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, W. T. M. Mooij, S. L. Price, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer and D. E. Williams, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2000, **56**, 697–714.
98 W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyanchenko, P. Erk and A. Gavezzotti, *et al*, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 647–661.
99 G. M. Day, W. D. S. Motherwell, H. L. Ammon, S. X. M. Boerrigter, R. G. Della Valle, E. Venuti, A. Dzyabchenko, J. D. Dunitz, B. Schweizer, B. P. van Eijck, P. Erk, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, F. J. J. Leusen, C. Liang, C. C. Pantelides, P. G. Karamertzanis, S. L. Price, T. C. Lewis, H. Nowell, A. Torrisi, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt and P. Verwer, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2005, **61**, 511–527.
100 G. M. Day, T. G. Cooper, A. J. Cruz-Cabeza, K. E. Hejczyk, H. L. Ammon, S. X. M. Boerrigter, J. S. Tan, R. G. Della Valle, E. Venuti, J. Jose, S. R. Gadre, G. R. Desiraju, T. S. Thakur, B. P. van Eijck, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, M. A. Neumann, F. J. J. Leusen, J. Kendrick, S. L. Price, A. J. Misquitta, P. G. Karamertzanis, G. W. A. Welch, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt, J. van de Streek, A. K. Wolf and B. Schweizer, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2009, **65**, 107–125.
101 P. Verwer and F. J. J. Leusen, in *Reviews in Computational Chemistry*, ed. K. B. Lipkowitz and D. B. Boyd, New York, Wiley-VCH, 1998, pp. 327–365.
102 H. Karfunkel, F. J. Leusen and R. Gdanitz, *J. Comput.-Aided Mater. Des.*, 1994, **1**, 177–185.
103 D. J. Willock, S. L. Price, M. Leslie and C. R. Catlow, *J. Comput. Chem.*, 1995, **16**, 628.
104 D. A. Bardwell, C. S. Adjiman, Y. A. Arnautova, E. Bartashevich, S. X. M. Boerrigter, D. E. Braun, A. J. Cruz-Cabeza, G. M. Day, R. G. Della Valle, G. R. Desiraju, B. P. van Eijck, J. C. Facelli, M. B. Ferraro, D. Grillo, M. Habgood, D. W. M. Hofmann, F. Hofmann, K. V. J. Jose, P. G. Karamertzanis, A. V. Kazantsev, J. Kendrick, L. N. Kuleshova, F. J. J. Leusen, A. V. Maleev, A. J. Misquitta, S. Mohamed, R. J. Needs, M. A. Neumann, D. Nikylov, A. M. Orendt, R. Pal, C. C. Pantelides, C. J. Pickard, L. S. Price, S. L. Price, H. A. Scheraga, J. van de Streek, T. S. Thakur, S. Tiwari, E. Venuti and I. K. Zhitkov, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2011, **67**, 535–551.
105 S. L. Price, M. Leslie, G. W. A. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis and G. M. Day, *Phys. Chem. Chem. Phys.*, 2010, **12**, 8478–8490.
106 G. M. Day, W. D. S. Motherwell and W. Jones, *Phys. Chem. Chem. Phys.*, 2007, **9**, 1693–1704.

107 T. G. Cooper, K. E. Hejczyk, W. Jones and G. M. Day, *J. Chem. Theory Comput.*, 2008, **4**, 1795–1805.
108 B. P. v. Eijck, *J. Comput. Chem.*, 2002, **23**, 805–815.
109 A. V. Kazantsev, P. G. Karamertzanis, C. S. Adjiman, C. C. Pantelides, S. L. Price, P. T. A. Galek, G. M. Day and A. J. Cruz-Cabeza, *Int. J. Pharm.*, 2011, **218**, 168–178.
110 M. A. Neumann and M. A. Perrin, *J. Phys. Chem. B*, 2005, **109**, 15531–15541.
111 M. A. Neumann, *J. Phys. Chem. B*, 2008, **112**, 9810–9829.
112 W. T. M. Mooij and F. J. J. Leusen, *Phys. Chem. Chem. Phys.*, 2001, **3**, 5063–5066.
113 G. M. Day, J. Chisholm, N. Sham, W. D. S. Motherwell and W. Jones, *Cryst. Growth Des.*, 2004, **4**, 1327–1340.
114 S. R. Cox, L.-Y. Hsu and D. E. Williams, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 1981, **37**, 293–301.
115 D. E. Williams and S. R. Cox, *Acta Crystallogr., Sect. B: Struct. Sci.*, 1984, **40**, 404–417.
116 D. E. Williams, *J. Mol. Struct.*, 1999, **485–486**, 321–347.
117 D. E. Williams, *J. Comput. Chem.*, 2001, **22**, 1–20.
118 D. E. Williams, *J. Comput. Chem.*, 2001, **22**, 1154–1166.
119 S. L. Mayo, B. D. Olafson and W. A. Goddard, *J. Phys. Chem.*, 1990, **94**, 8897–8909.
120 M. J. Hwang, T. P. Stockfisch and A. T. Hagler, *J. Am. Chem. Soc.*, 1994, **116**, 2515.
121 J. R. Maple, M. J. Hwang, T. P. Stockfisch, U. Dinur, M. Waldman, C. S. Ewig and A. T. Hagler, *J. Comput. Chem.*, 1994, **15**, 162–182.
122 Z. W. Peng, C. S. Ewig, M. J. Hwang, M. Waldman and A. T. Hagler, *J. Phys. Chem. A*, 1997, **101**, 7243–7252.
123 H. Sun, *J. Phys. Chem. B*, 1998, **102**, 7338–7364.
124 V. Marcon and G. Raos, *J. Phys. Chem. B*, 2004, **108**, 18053–18064.
125 S. Brodersen, S. Wilke, F. J. J. Leusen and G. Engel, *Phys. Chem. Chem. Phys.*, 2003, **5**, 4923–4931.
126 J. D. Yesselman, D. J. Price, J. L. Knight and C. L. Brooks, *J. Comput. Chem.*, 2012, **33**, 189–202.
127 R. F. W. Bader, A. Larouche, C. Gatti, M. T. Carroll, P. J. MacDougall and K. B. Wiberg, *J. Chem. Phys.*, 1987, **87**, 1142–1152.
128 R. F. W. Bader, T. A. Keith, K. M. Gough and K. E. Laidig, *Mol. Phys.*, 1992, **75**, 1167–1189.
129 S. L. Price, C. H. Faerman and C. W. Murray, *J. Comput. Chem.*, 1991, **12**, 1187–1197.
130 C. H. Faerman and S. L. Price, *J. Am. Chem. Soc.*, 1990, **112**, 4915–4926.
131 W. T. M. Mooij, F. B. van Duijneveldt, J. G. C. M. van Duijneveldt-vande Rijdt and B. P. van Eijck, *J. Phys. Chem. A*, 1999, **103**, 9872.
132 R. Kumar, F.-F. Wang, G. R. Jenness and K. D. Jordan, *J. Chem. Phys.*, 2010, **132**, 014309–014312.
133 A. J. Stone and M. Alderton, *Mol. Phys.*, 1985, **56**, 1047–1064.
134 W. A. Sokalski and R. A. Poirier, *Chem. Phys. Lett.*, 1983, **98**, 86–92.
135 D. S. Kosov and P. L. A. Popelier, *J. Chem. Phys.*, 2000, **113**, 3969–3974.
136 R. F. W. Bader and P. M. Beddall, *J. Chem. Phys.*, 1972, **56**, 3320–3329.
137 K. E. Laidig, *J. Phys. Chem.*, 1993, **97**, 12760–12767.
138 C. E. Whitehead, C. M. Breneman, N. Sukumar and M. D. Ryan, *J. Comput. Chem.*, 2003, **24**, 512–529.
139 C. M. Breneman, T. R. Thompson, M. Rhem and M. Dung, *Comput. Chem.*, 1995, **19**, 161–179.
140 P. L. A. Popelier and F. M. Aicken, *ChemPhysChem*, 2003, **4**, 824–829.
141 P. L. A. Popelier, M. Devereux and M. Rafat, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 2004, **60**, 427–433.
142 Y. Yuan, M. J. L. Mills and P. L. A. Popelier, *J. Comput. Chem.*, 2014, **35**, 343–359.
143 T. Koritsanszky, A. Volkov and P. Coppens, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 2002, **58**, 464–472.
144 M. J. L. Mills and P. L. A. Popelier, *Comput. Theor. Chem.*, 2011, **975**, 42–51.
145 M. J. L. Mills and P. L. A. Popelier, *Theor. Chem. Acc.*, 2012, **131**, 1137–1153.
146 S. M. Kandathil, T. L. Fletcher, Y. Yuan, J. Knowles and P. L. A. Popelier, *J. Comput. Chem.*, 2013, **34**, 1850–1861.
147 C. Jelsch, V. Pichon-Pesme, C. Lecomte and A. Aubry, *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 1998, **54**, 1306–1318.
148 N. K. Hansen and P. Coppens, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 1978, **34**, 909–921.
149 N. Muzet, B. Guillot, C. Jelsch, E. Howard and C. Lecomte, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 8742–8747.
150 A. Volkov, T. Koritsanszky and P. Coppens, *Chem. Phys. Lett.*, 2004, **391**, 170–175.
151 F. Allen, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 380–388.
152 A. Volkov, X. Li, T. Koritsanszky and P. Coppens, *J. Phys. Chem. A*, 2004, **108**, 4283–4300.
153 M. Woińska and P. M. Dominiak, *J. Phys. Chem. A*, 2011, **117**, 1535–1547.
154 P. M. Dominiak, A. Volkov, X. Li, M. Messerschmidt and P. Coppens, *J. Chem. Theory Comput.*, 2007, **3**, 232–247.
155 V. Pichon-Pesme, C. Lecomte and H. Lachekar, *J. Phys. Chem.*, 1995, **99**, 6242–6250.
156 K. N. Jarzembska and P. M. Dominiak, *Acta Crystallogr., Sect. A: Found. Crystallogr.*, 2012, **68**, 139–147.
157 P. Dominiak, E. Espinosa and J. Ángyán, Intermolecular Interaction Energies from Experimental Charge Density Studies, in *Modern Charge-Density Analysis*, ed. C. Gatti and P. Macchi, Springer, Netherlands, 2012, pp. 387–433.
158 K. Y. Sanbonmatsu and C. S. Tung, *J. Struct. Biol.*, 2007, **157**, 470–480.
159 L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan and K. Schulten, *J. Comput. Phys.*, 1999, **151**, 283–312.
160 A. Y. Toukmaji and J. A. Board Jr., *Comput. Phys. Commun.*, 1996, **95**, 73–92.
161 S. L. Price and A. J. Stone, *J. Chem. Soc., Faraday Trans.*, 1992, **88**, 1755–1763.
162 P. Kdzierski and W. A. Sokalski, *J. Comput. Chem.*, 2001, **22**, 1082–1097.

163 N. Plattner and M. Meuwly, *Biophys. J.*, 2008, **94**, 2505–2515.
164 U. Koch, P. L. A. Popelier and A. J. Stone, *Chem. Phys. Lett.*, 1995, **238**, 253–260.
165 N. Plattner and M. Meuwly, *J. Mol. Model.*, 2009, **15**, 687–694.
166 P. G. Karamertzanis and S. L. Price, *J. Chem. Theory Comput.*, 2006, **2**, 1184–1199.
167 C. Hättig, *Chem. Phys. Lett.*, 1997, **268**, 521–530.
168 P. L. A. Popelier and A. J. Stone, *Mol. Phys.*, 1994, **82**, 411–425.
169 P. L. A. Popelier, A. J. Stone and D. J. Wales, *Faraday Discuss.*, 1994, **97**, 243–264.
170 M. Leslie, *Mol. Phys.*, 2008, **106**, 1567–1578.
171 D. M. Elking, L. Perera, R. Duke, T. Darden and L. G. Pedersen, *J. Comput. Chem.*, 2010, **31**, 2702–2713.
172 M. J. L. Mills and P. L. A. Popelier, submitted.
173 C. Sagui, L. G. Pedersen and T. A. Darden, *J. Chem. Phys.*, 2004, **120**, 73–87.
174 J. C. Wu, J.-P. Piquemal, R. Chaudret, P. Reinhardt and P. Ren, *J. Chem. Theory Comput.*, 2010, **6**, 2059–2070.
175 J.-P. Piquemal, L. Perera, G. A. Cisneros, P. Ren, L. G. Pedersen and T. A. Darden, *J. Chem. Phys.*, 2006, **125**, 054511–054517.
176 J. W. Ponder and D. A. Case, Force Fields for Protein Simulations, in *Advances in Protein Chemistry*, ed. D. Valerie, Academic Press, 2003, pp. 27–85.
177 P. Ren and J. W. Ponder, *J. Comput. Chem.*, 2002, **23**, 1497–1506.
178 T. Liang and T. R. Walsh, *Phys. Chem. Chem. Phys.*, 2006, **8**, 4410–4419.
179 T. Liang and T. R. Walsh, *Mol. Simul.*, 2007, **33**, 337–342.
180 J. Kaminský and F. Jensen, *J. Chem. Theory Comput.*, 2007, **3**, 1774–1788.
181 T. D. Rasmussen, P. Ren, J. W. Ponder and F. Jensen, *Int. J. Quantum Chem.*, 2007, **107**, 1390–1395.
182 P. Ren and J. W. Ponder, *J. Phys. Chem. B*, 2003, **107**, 5933–5947.
183 P. Ren and J. W. Ponder, *J. Phys. Chem. B*, 2004, **108**, 13427–13437.
184 Y. Shi, C. Wu, J. W. Ponder and P. Ren, *J. Comput. Chem.*, 2011, **32**, 967–977.
185 N. Gresh, G. A. Cisneros, T. A. Darden and J.-P. Piquemal, *J. Chem. Theory Comput.*, 2007, **3**, 1960–1986.
186 F. Vigné-Maeder and P. Claverie, *J. Chem. Phys.*, 1988, **88**, 4934–4948.
187 D. R. Garmer and W. J. Stevens, *J. Phys. Chem.*, 1989, **93**, 8263–8270.
188 N. Gresh, H. Guo, D. R. Salahub, B. P. Roques and S. A. Kafafi, *J. Am. Chem. Soc.*, 1999, **121**, 7885–7894.
189 W. J. Stevens, H. Basch and M. Krauss, *J. Chem. Phys.*, 1984, **81**, 6026–6033.
190 N. Gresh and J. Šponer, *J. Phys. Chem. B*, 1999, **103**, 11415–11427.
191 N. Gresh, J. E. Šponer, N. Špačková, J. Leszczynski and J. Šponer, *J. Phys. Chem. B*, 2003, **107**, 8669–8681.
192 G. Tiraboschi, M.-C. Fournié-Zaluski, B.-P. Roques and N. Gresh, *J. Comput. Chem.*, 2001, **22**, 1038–1047.
193 N. Gresh and D. R. Garmer, *J. Comput. Chem.*, 1996, **17**, 1481–1495.
194 N. Gresh and B.-P. Roques, *Biopolymers*, 1997, **41**, 145–164.
195 F. Rogalewicz, G. Ohanessian and N. Gresh, *J. Comput. Chem.*, 2000, **21**, 963–973.
196 K. E. Hage, J.-P. Piquemal, Z. Hobaika, R. G. Maroun and N. Gresh, *J. Comput. Chem.*, 2013, **34**, 1125–1135.
197 O. Engkvist, P.-O. Åstrand and G. Karlström, *J. Phys. Chem.*, 1996, **100**, 6950–6957.
198 On the evaluation of intermolecular potentials, in *Proceedings of the 5th Seminar on Computational Methods in Quantum Chemistry*, ed. G. Karlström, Max-Planck-Institut für Physik und Astrophysik, Groningen, 1981.
199 O. Engkvist, P. Hobza, H. L. Selzle and E. W. Schlag, *J. Chem. Phys.*, 1999, **110**, 5758–5762.
200 A. B. Holt, J. Boström, G. Karlström and R. Lindh, *J. Comput. Chem.*, 2010, **31**, 1583–1591.

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 136 (2015) 32–41

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

journal homepage: www.elsevier.com/locate/saa

Accurate prediction of polarised high order electrostatic interactions for hydrogen bonded complexes using the machine learning method kriging

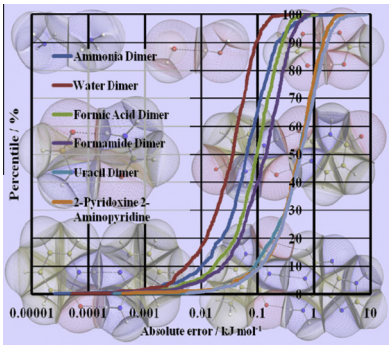
Timothy J. Hughes, Shaun M. Kandathil [1], Paul L.A. Popelier *

Manchester Institute of Biotechnology (MIB), 131 Princess Street, Manchester M1 7DN, United Kingdom
School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom

HIGHLIGHTS

- Charge transfer and dipolar polarisation treated on a common footing.
- Elimination of polarisation catastrophe.
- No need for penetration corrections.
- Kriging successfully predicts multipole moments directly from coordinates.
- High rank multipole moments guarantee accurate electrostatics.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Available online 5 November 2013

Keywords:
Multipole moments
Hydrogen bonding
Force fields
QTAIM
QCT
Kriging

ABSTRACT

As intermolecular interactions such as the hydrogen bond are electrostatic in origin, rigorous treatment of this term within force field methodologies should be mandatory. We present a method able of accurately reproducing such interactions for seven van der Waals complexes. It uses atomic multipole moments up to hexadecupole moment mapped to the positions of the nuclear coordinates by the machine learning method kriging. Models were built at three levels of theory: HF/6-31G**, B3LYP/aug-cc-pVDZ and M06-2X/aug-cc-pVDZ. The quality of the kriging models was measured by their ability to predict the electrostatic interaction energy between atoms in external test examples for which the true energies are known. At all levels of theory, >90% of test cases for small van der Waals complexes were predicted within 1 kJ mol^{-1}, decreasing to 60–70% of test cases for larger base pair complexes. Models built on moments obtained at B3LYP and M06-2X level generally outperformed those at HF level. For all systems the individual interactions were predicted with a mean unsigned error of less than 1 kJ mol^{-1}.

Introduction

Until the development of more powerful computers, the simulation of chemical and biochemical systems will require the use of molecular mechanics force fields. These methods calculate the energy of a system by a sum of both bonded and non-bonded terms. The bonded terms include bond stretches, angle bends and torsional terms, whereas the non-bonded terms include the electrostatic and van der Waals contributions. Chemical systems are dominated by intermolecular interactions such as the hydrogen bond which are typically electrostatic in origin [1,2], and so the electrostatic term should receive special attention. The accuracy of the electrostatic term typically suffers from at least one of two limiting assumptions. The first is an atomic point charge description and the second a lack of polarisation.

The majority of force fields currently in use model the electrostatic contribution to the total energy of a system through simple

* Corresponding author. Tel.: +44 1613064511.
E-mail address: pla@manchester.ac.uk (P.L.A. Popelier).
[1] Current address: Michael Smith Building, the University of Manchester, Manchester M13 9PT, United Kingdom.

1386-1425/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.saa.2013.10.059

atom centred point charges, modelled with a $1/r$ dependence. Due to their isotropic nature, point charges provide a poor description of the electronic distribution around an atom and hence the corresponding force field introduces a prediction error. For example, hydrogen bonding of the general form O–H...O=C is a non-linear interaction resulting from the partially positively charged hydroxyl hydrogen interacting with the lone pairs on the carbonyl oxygen [3–5]. Using only an isotropic atomic point charge, the observed geometry is not reproduced. Force fields typically incorporate such effects in one of two ways: either by the addition of additional non-atomic point charges [6–8], or by the use of atomic multipole moments [9,10]. The former approach is heavily dependent on the parameterisation of the additional charge sites and is not always successful when applied to new systems. Despite their shortcomings, additional charge site models are still widely used, for example the TIP4P and TIP5P water potentials [11,12]. The use of atomic multipole moments, however, is much more justifiable. The point charge, or monopole moment, is simply the zeroth order term in a multipolar expansion of electrostatic interaction. Hence, including higher order terms such as the atomic dipole and quadrupole moments is chemically more rigorous in that it describes more features of the inevitably anisotropic (i.e. non spherical) atomic electron density. Phenomena such as the previously described non-linearity of hydrogen bonding are captured when multipole moments are used rather than point charges, along with other anisotropic effects such as describing the σ-hole in halogen bonding. The polarisable multipolar force field AMOEBA [13] is the most widely used multipolar force field to date, and this uses multipole moments up to quadrupole to describe each atom, obtained via Stone's distributed multipole analysis (DMA) [14]. The way conformational changes influence DMA multipole moment was studied [15] a long time ago.

Incorporating polarisation into a force field is typically done in one of four ways, namely the drude oscillator approach [16,17], fluctuating charges [18,19], induced dipoles with associated isotropic atomic polarisabilities [20], and finally by fitting effective charges to induced dipoles [21]. The method presented here, currently called QCTFF, differs from any of the above. Instead, the machine learning method kriging [22] is used to build models that directly map atomic multipole moments (up to the hexadecupole moment) to the nuclear coordinates of all other atoms in the system. It should be emphasised that no explicit polarisability is ever obtained here; the proposed method focuses immediately on the effect of polarisation. In other words, it is capable of predicting the multipole moments after the polarisation process has been completed. The atoms used are defined by Quantum Chemical Topology (QCT) [23,24], which partitions a molecule or a complex of molecules in a minimal way, without invoking a reference system, by letting the gradient vector reveal a pattern of subspaces, each subspace corresponding to a (topological) atom.

In this work, we show that by using kriging, we are able to build models capable of accurately reproducing the *ab initio* multipole moments of topological atoms for a range of geometries of hydrogen bonded complexes taken from the so-called S22 dataset of Hobza et al. [25] . This dataset consist of a list of 22 molecules that is commonly used as a benchmark for new computational methods to describe intermolecular interactions. This data set is divided into three subsets: hydrogen bonded complexes, dispersion bound complexes, and mixed complexes. Only the hydrogen bonded molecules are considered in this work.

Kriging has been shown to yield highly accurate reproduction of the multipole moments on the small pilot system ethanol [26], and single amino acids such as alanine [27] and histidine [28]. This is validated by accurate intramolecular atomic interaction energies predicted for all atom pairs separated by more than three covalent bonds. This work is the first example of its application to intermolecular systems, other than for an early study on water clusters [29], which demonstrated the superiority of kriging compared to neural networks. Our lab had used [30–32] the latter even earlier for small clusters or single molecules. The long term aim is to incorporate kriging models in the exploration of potential energy surfaces [33], going beyond rigid-body multipolar electrostatics [34,35], and also in molecular dynamics simulations [36–41], which so far have only been achieved with unpolarised topological atoms, again in the rigid body formalism.

In this work, we will apply our kriging method to a set of seven van der Waals complexes, each consisting of two molecules each. Five of these consist of a single molecular species; we will refer to these systems as dimers.

Theory

Quantum Chemical Topology (QCT)

A key concept underpinning QCTFF is that of QCT, which defines the atom in an arbitrarily large system. QCT puts the use of the language of dynamical systems (e.g. basin, attractor, critical point, separatrix, etc.) at the heart of its approach. The well-known quantum theory of atoms in molecules [42] is then a special case of this approach, applied to the electron density and its Laplacian only. The electronic density of a molecule, or a van der Waals complex for that matter, naturally partitions itself into non-overlapping topological atoms by means of gradient paths. A gradient path is a trajectory in 3D space, which can be seen as consisting of infinitesimal vectors, each one orthogonal to an envelope of constant electron density ρ. A gradient path follows the direction of increasing ρ, until it terminates at a critical point. This latter is an attractor, which can only be a nucleus (which is mostly the case), a bond critical point, or a ring critical point. This partitioning method gives rise to well defined, non-overlapping atoms [43] for which multipole moments may then be obtained.

Fig. 1 shows the hydrogen bonded van der Waals complexes of the S22 dataset, where the non-overlapping topological atoms are readily observable. The bond path between the two ammonia molecules is perhaps unexpected in that it suggests a special interaction between the two nitrogen atoms. However, one would typically draw the interaction as a pair of hydrogen bonds between the N and H atoms. The QCT description of the bonding situation is a direct result of the topology of the electron density, and such patterns have been found before elsewhere. For example, in their QTAIM analysis [44] Bone and Bader reported many unusual bond paths in a study of 11 van der Waals complexes. A notable example is that of the CO_2 dimer where a bond paths between the oxygen atoms were observed for the side-on dimer. The meaning and particular appearance of bond critical points and concomitant bond paths cannot be lightly dismissed given their deep connection with the topological energy partitioning [45].

Returning to QCT, the Coulomb interaction [46] energy between two topological atomic basins Ω_A and Ω_B is given by:

$$E_{AB}^{Coul} = \int_{\Omega_A} d\mathbf{r}_1 \int_{\Omega_B} d\mathbf{r}_2 \frac{\rho_{tot}(\mathbf{r}_1)\rho_{tot}(\mathbf{r}_2)}{r_{12}} \quad (1)$$

where ρ_{tot} is equal to the sum of minus the electron density ρ and the nuclear charge density. The expression $1/r_{12}$ in Eq. (1)can be replaced by series expansion involving the spherical harmonics [47,48] to give:

$$\frac{1}{r_{12}} = \sum_{l_A=0}^{\infty}\sum_{l_b=0}^{\infty}\sum_{m_A=-l_A}^{l_A}\sum_{m_B=-l_B}^{l_B} T_{l_A l_B m_A m_B} R_{l_A m_A}(\mathbf{r}_1) R_{l_B m_B}(\mathbf{r}_2) \quad (2)$$

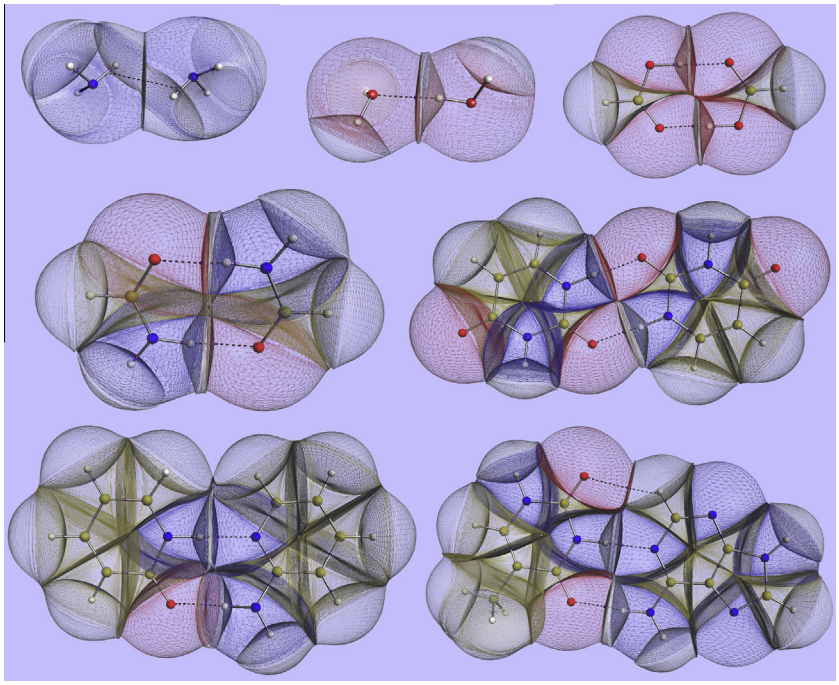


Fig. 1. The seven hydrogen bonded van der Waals complexes studied: the ammonia dimer (top left), the water dimer (top middle), the formic acid dimer (top right), the formamide dimer (middle left), the uracil dimer (middle right), the 2-pyridoxine...2-aminopyridine complex (bottom left) and the adenine...thymine base pair (bottom right). Molecular graphs containing bond paths and critical points are superimposed on the topological atoms, which are capped at their $\rho = 0.0001$ a.u. isosurface.

where $R_{lm}(\mathbf{r})$ is a regular spherical harmonic. The interaction tensor T depends upon the mutual orientation of the two interacting atoms A and B, and their internuclear distance. The simplest interaction term is that two monopole moments (or essentially atomic charges), where T is simply $1/r$. Substituting Eq. (2)into Eq. (1)gives:

$$E_{AB}^{Coul} = \sum_{l_A l_B m_A m_B} Q_{l_A m_A} T_{l_A l_B m_A m_B} Q_{l_B m_B} \tag{3}$$

where Q_{lm} represents a multipole moment:

$$Q_{lm} = \int_{\Omega} \mathrm{d}\mathbf{r} \rho_{\mathrm{tot}}(\mathbf{r}) R_{lm}(\mathbf{r}) \tag{4}$$

that is obtained after a 3D integration over the potentially complicated volume of the topological atom. It is convenient to define an interaction rank L between two multipole moments of order l_A and l_B by:

$$L = l_B + l_B + 1 \tag{5}$$

Previous work [10,36] has shown that an interaction rank of $L = 5$ provides satisfactory description of structural and dynamic characteristics of a system. The value of L is identical to the inverse power in the $1/R^L$ behaviour of an interatomic electrostatic interaction. For example, dipole...dipole interactions behave by the well-known $1/R^3$ law given that $L = 1 + 1 + 1 = 3$. Truncating at $L = 5$ requires monopole, dipole, quadrupole, octopole and hexadecupole moments, meaning that the electron density of a topological atom is described by $1 + 3 + 5 + 7 + 9 = 25$ multipole moments each.

Kriging

Given a set of molecular configuration (geometries) we can calculate the multipole moments for each atom using QCT. Then, assuming that transitions between these configurations occur smoothly, it is a reasonable approximation to interpolate the values of the various multipole moments for each of the intermediate configurations instead of performing the calculations again each time. This is especially true if the configurations are highly similar, since the multipole moments, and hence the total electrostatic energy of the system, can be said to change predictably.

To interpolate the values of the multipole moments for each atom in the system, we use the method of kriging or Gaussian process regression. We have introduced this method elsewhere [26,27,31], but for convenience, we will summarise the key points of this strategy. First, a "training set" is created for each molecular system of interest (details see next Section on the AUTOLINE procedure), which contains the atomic multipole moments as obtained from *ab initio* wave functions. Training data sets are then constructed for each atom, with each training data point consisting of inputs and outputs (or response values). The inputs are the internal coordinates of all the atoms for each molecular geometry in the training set, while the outputs consist of the multipole moments of the given atom corresponding to each geometry. These training data sets are then used to construct kriging models for each multipole moment on each atom. A kriging model is a numerical way of expressing the variation in the values of a multipole moment as a function of the atomic coordinates of the surrounding atoms. This can be imagined as constructing a d-dimensional hypersurface of best fit that passes through the training points. To do this, we model the correlation or covariance between each pair of the n training points in an $n \times n$ correlation matrix $\boldsymbol{R}$, whose elements are given by:

$$R_{ij} = \mathrm{cor}[\epsilon(\boldsymbol{x}^i), \epsilon(\boldsymbol{x}^j)] = \exp\left[-\sum_{h=1}^{d} \theta_h |x_h^i - x_h^j|^{p_h}\right] \tag{6}$$

where the vectors x^i and x^j are any two training points composed of d components or so-called *features* in the language of machine learning. In our case these are essentially the atomic coordinates of the molecular system. The details of the exact way in which we define these features is given elsewhere [28] as well as the details of the local atomic frames installed on each nucleus in order to fix the orientation of the multipole moments. In Eq. (6), θ_h and p_h

are two parameters to be determined, which convey the relevance of each dimension '*h*'. Since there are *d* values of θ and *p*, we can write them as *d*-dimensional vectors $\boldsymbol{\theta}$ and $\boldsymbol{p}$. Eq. (6)expresses the simple idea that, for a well-conditioned function, if two training points are close together, their response values relative to some datum are likely to be similar. This is expressed as "a correlation between errors", which is represented by $\mathrm{cor}[\epsilon(\boldsymbol{x}^i), \epsilon(\boldsymbol{x}^j)]$.

The task now is to derive a maximum-likelihood model or function over the training data set that produces the response values observed. This is done by maximising the logarithm of a likelihood function.

$$\log L = -\frac{n}{2}\log(\hat{\sigma}^2) - \frac{1}{2}\log(|\boldsymbol{R}|) \tag{7}$$

where

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \mathbf{1}\hat{\mu})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \mathbf{1}\hat{\mu})}{n} \tag{8}$$

$$\hat{\mu} = \frac{\mathbf{1}'\boldsymbol{R}^{-1}\boldsymbol{y}}{\mathbf{1}'\boldsymbol{R}^{-1}\mathbf{1}} \tag{9}$$

where $\boldsymbol{y}$ is a vector made up of the response values corresponding to the *n* training data points, and $\mathbf{1}$ is a vector of ones (rank *n*).

Because *n* and $\boldsymbol{y}$ are constants, and $\boldsymbol{R}$ depends only on θ and $\boldsymbol{p}$, log*L* depends only on these parameters. The task is therefore, to find the optimal values of the θ and $\boldsymbol{p}$ vectors that maximise log*L*. We employ particle swarm optimisation (PSO) to achieve this. Once the optimal θ and $\boldsymbol{p}$ vectors have been derived for each moment, we can make predictions of the moment values for a new geometry x^* through the following equation:

$$\hat{y}(\boldsymbol{x}^*) = \hat{\mu} + \sum_{i=1}^{n} a_i \cdot r_i \tag{10}$$

where $\hat{y}$ is the response value, a_i is the *i*th element of the vector $\boldsymbol{a} = \boldsymbol{R}^{-1}(\boldsymbol{y} - \mathbf{1}\hat{\mu})$, and r_i is the *i*th element of $\boldsymbol{r}$, which is calculated as:

$$\mathbf{r} = \{\mathrm{cor}[\epsilon(x^*), \epsilon(\boldsymbol{x}^1)], \mathrm{cor}[\epsilon(\boldsymbol{x}^*), \epsilon(\boldsymbol{x}^2)], \ldots, \mathrm{cor}[\epsilon(\boldsymbol{x}^*), \epsilon(\boldsymbol{x}^n)]\}' \tag{11}$$

and each term of $\boldsymbol{r}$ is calculated using Eq. (6).

Note that each component of each multipole moment constitutes a separate set of response values. Hence, each component of each multipole moment must be trained for and predicted separately. The process of deriving the optimal values of the $\boldsymbol{\theta}$ and $\boldsymbol{p}$ vectors is achieved by PSO. This is implemented in our in-house application FEREBUS (see Section 2.3), which has been improved with OpenMP parallelisation for the calculation of the correlation matrix $\boldsymbol{R}$, thus reducing execution time. We have described this approach in earlier publications [26–28,49]. Briefly, PSO achieves the optimisation of a single objective function (here the log likelihood) through the evaluation and comparison of several concurrent candidate solutions. This "swarm" of candidate solutions then evolves and finds its way to an optimum by learning both from their own experiences, as well as the best solution found by the swarm as a whole, until no improvement in the objective function value is realised for a number of consecutive iterations.

The AUTOLINE procedure

The AUTOLINE procedure was followed for the building of the kriging models. This has been discussed in more detail elsewhere [28] and is outlined in Fig. 2. The Cartesian coordinates for each of the complexes were taken directly from the Benchmark Energy and Geometry Database (BEGDB) [50]. The training set geometries for each of the complexes were obtained by normal mode sampling. This was a two-stage process. Initially, the second derivatives of the potential energy surface for each complex were calculated at the correct level of theory (using GAUSSIAN03 [51] for HF/6-31G** and B3LYP/aug-cc-pVDZ, and GAUSSIAN09 [52] for M06-2X/aug-cc-pVDZ), and then 2000 geometries were sampled for each system by randomly pumping energy into the normal modes and taking "snapshots".

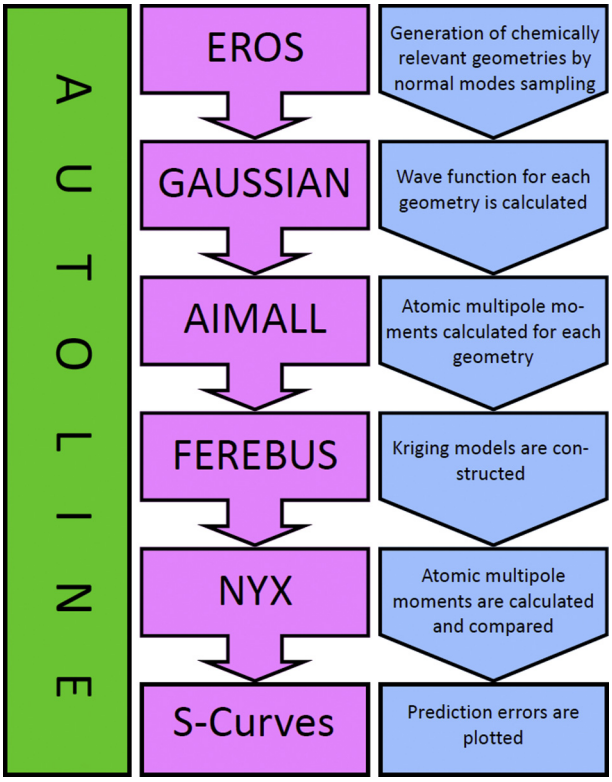


Fig. 2. The highly automated AUTOLINE procedure followed to build and test kriging models for the prediction of multipole moments.

Training set geometries are generated by the in-house program EROS. The basic principle involves taking either the geometry of the global energy minimum or a local energy minimum, and then inputting quasi-random amounts of energy into the normal vibrational modes. The energy is spread evenly over all vibrational modes of the system. The motion of the vibrational modes is modelled harmonically, and 'snapshots' of the nuclear coordinates at random points during the vibrational motion are used as the training set geometries. If the input energy is too high, then bonds may dissociate and atoms fly apart. Therefore an iterative process takes place where initially a maximum input energy is defined, typically about 200 kJ mol^{-1}, and geometries are generated. If broken bonds are present then the maximum energy is lowered and the process is repeated until no broken bonds are present. A bond is defined as broken if:

$$R_{AB} > k(A_{vdW} + B_{vdW}) \tag{12}$$

where R_{AB} is the internuclear distance between bonded atoms *A* and *B*, *k* is a constant, and A_{vdw} and B_{vdw} are the van der Waals radii of atoms *A* and *B*, respectively. In this work $k = 1.2$ was used as this is the default value used by GAUSSIAN.

For systems such as amino acids multiple energy minima may appear in the same training set. Assuming that the geometries of the energy minima that are distorted are chemically relevant, ensures that the kriging models are constructed and later used in a chemically relevant conformational space. In this work, only one

minimum geometry was used for each van der Waals complex, given by the Cartesian coordinates published elsewhere [50].

Wave functions were obtained for each of the geometries using GAUSSIAN, and then AIMALL [53] was used to calculate the atomic multipole moments for each of the 2000 geometries. Candidate geometries were discarded if any atom in the geometry had an integration error [54] L(Ω) larger than 0.001 a.u.. Subsequently, kriging models relating each multipole moment to the nuclear coordinates of the system were built using the in-house software FEREBUS, with a training set of 600 randomly selected geometries. Twelve particles were used for the FEREBUS PSO in all systems, and the exponent parameters p were optimised, rather than all set fixed to a value of 2. The performance of the kriging models for each complex was then tested on 600 *external* test geometries. It is important that none of the test geometries appear in the training set, in order to simulate and test proper predictivity.

One could assess the performance of a kriging model by comparing a multipole moment that it predicts with its *true* calculated value. However, the ultimate arbiter of performance is the error in interaction energy rather than errors in multipole moments. Therefore we assess the performance of all 25 kriging models for a given atom, one for each multipole moment, by making this atom interact with other atoms, and monitor the total interaction energy. Note that energies were calculated up to interaction rank $L = 5$. One possible way is to probe a given kriged atom with "true" atoms. However, it is more realistic to probe a kriged atom with other kriged atoms because when QCTFF is applied to molecular simulation "true" atoms will not be present. The in-house program NYX performs the task of energy error calculation and assessment. Expressed mathematically, the total absolute error of the predicted interaction energies can be written as:

$$\left|\Delta E_{\text{system}}\right| = \left|E_{\text{system}}^{\text{true}} - E_{\text{system}}^{\text{predicted}}\right| = \left|\sum_{AB} E_{AB}^{\text{true}} - \sum_{AB} E_{AB}^{\text{predicted}}\right| = \left|\sum_{AB} \Delta E_{AB}\right| \tag{13}$$

where

$$\Delta E_{AB} = E_{AB}^{\text{true}} - E_{AB}^{\text{predicted}} \tag{14}$$

Eq. (14)is illustrated by means of Fig. 3, which shows how an atom B probes a given atom A. Of course, one could equally state that atom A probes a given atom B, which alerts one not to double count the energy error expressed in Eq. (14).

The double sum in Eq. (14)needs to be specified. All atom–atom interaction energies are purely inter-molecular. In other words, only atom–atom electrostatic energies were calculated where one atom is part of one monomer and the other atom part of the other monomer; intra-molecular atom–atom interactions were not assessed. Put more precisely, the subscript A in eq 13 runs over all N atoms in molecule 1, and the subscript B includes all M atoms in molecule 2.

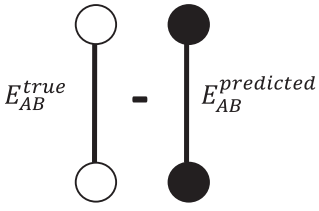


Fig. 3. A schematic representation of ΔE_{AB} as expressed in eq 14 and calculated by the program NYX. Shaded atoms represent the predicted values of the multipole moments from the kriging models and white atoms represent the true values of the multipole moments.

Finally, the energy error $|\Delta E_{\text{system}}|$ for all test configurations (i.e. geometries) are plotted in a single curve, which we call a S-curve. These curves appear for the first time in Fig. 4 where they will be discussed in more detail. Essentially, an S-curve explicitly displays the overall performance of a kriging model as tested on all the external test geometries it predicted for.

The AUTOLINE procedure is a fully automated process. The time taken for completion is dependent on a number of factors, the two most significant being the level of theory of the *ab initio* GAUSSIAN calculation and the number of training examples input to FEREBUS, of which the latter takes the longest.

Computational methods

Kriging models were built at three different levels of theory: HF/6-31G**, B3LYP/aug-cc-pVDZ, and M06-2X/aug-cc-pVDZ. This allows comparison of how well kriging performs at different levels of theory. Unpublished work has shown that B3LYP/apc1 consistently outperforms HF/6-31G**, in that kriging models lead to interaction energies closer to the true energy. This is due to the higher levels of theory including electron correlation, which produces atomic monopole moments of smaller absolute value. One can show mathematically [27] why for the same kriging settings (e.g. number of data points in the training set) the Hartree–Fock level will produce worse errors.

The reason for including the Hartree–Fock level of theory in this work requires some justification in the light of its well-known limited accuracy. Firstly, some popular force fields such as AMBER or CHARMM include parameterisation from Hartree–Fock level data. Showing that our method is able to produce accurate predictions relative to the "true" Hartree–Fock value proves we can compete with, and eventually possibly supersede, the methodologies currently in place. Secondly, assuming that the multipole moments obtained from HF wave functions behave similarly to the multipole moments obtained at correlated levels of theory, using HF obtained multipole moments is still a valid proof of concept. Indeed, when future kriging models are to be implemented into a force field and applied to molecular dynamics simulation it will be desirable to use kriging models that have been built on the most chemically accurate data. However, such models come at the cost of lengthy and expensive *ab initio* calculations, which is why the latter have to be carried out only when all the parameters for the process of constructing the kriging models have been fine-tuned. To prove

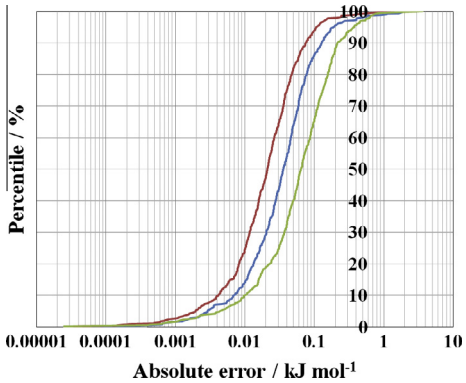


Fig. 4. Comparison between the effect of the level of theory of the PES and the level of theory of the wave functions obtained to build kriging models for the ammonia dimer. Blue: training set geometries obtained from HF PES and training set wave functions obtained at HF; Red: training set geometries obtained from HF PES and training set wave functions obtained at B3LYP; Green: training set geometries from B3LYP PES and training set wave functions obtained at B3LYP. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for the first time that kriging is capable of modelling intermolecular interactions (other than in water clusters), i.e. the essence of the current work, using a lower level of theory is justified.

A common criticism of the B3LYP density functional is that it does not describe long-range electron correlation effects, which play a key role in the binding of many van der Waals complexes [55–58]. The M06-2X functional [59] has been specifically designed to provide accurate interaction energies for a range of intermolecular interaction types, in particular van der Waals complexes. In this work we use both the B3LYP and M06-2X functionals to see if improved modelling of the long-range electron correlation lowers the magnitude of the prediction errors of intermolecular interactions provided by our kriging models. In recent work [60] by Friesner et al. a database of highly accurate CCSD(T) noncovalent interaction energies was assembled. The database was then used to fit a correction term to be added to the B3LYP density functional to allow for accurate intermolecular interactions. This was tested using the aug-cc-pVDZ and LACVP* basis sets, and was compared with both the B3LYP-D3 method [61], and the M06-2X hybrid functional. In an effort to maintain some level of consistency with the work of Friesner et al. the aug-cc-pVDZ Dunning basis set was chosen in this work for building of the B3LYP and M06-2X kriging models.

As a final comment we stress that all errors presented in this paper are relative to the correct value given at a specific level of theory, rather than relative to an experimental or high level of theory *ab initio* value such as CCSD(T)/CBS.

Results and discussion

Effect of level of theory on the training set

The training set geometries are sampled by putting energy into the normal modes of vibration of the system. These normal modes are calculated directly from the derivatives of the potential energy surface, and so are affected by the level of theory used to construct the Potential Energy Surface (PES). Therefore, one must keep in mind that true comparisons cannot be made between the performances of kriging models at different levels of theory. To generate the training set geometries at each level of theory, the maximum amount of energy is pumped into the sample without breaking any bonds. This maximum amount of energy changes when the PES is built at a different level of theory. For example, Hartree–Fock theory is known to predict bonds to be too polar. Subsequently, the force constants for these bonds are higher than those at B3LYP level, for example. This means that less vibrational motion may take place when pumping a large amount of energy into a HF PES compared to pumping a smaller amount of energy into a B3LYP PES. Table 1 shows the amount of energy put into the systems at different levels of theory. Hartree–Fock does indeed show the greatest tendency to have the most energy pumped in, although is noted that this is not seen throughout. This is in part due to the random way in which energy is put into the vibrational modes.

Table 1
Energy (kJ mol^{-1}) pumped into the hydrogen bonded complexes. The highest values are in bold and the lowest values in italics. Numbers in brackets indicate the first lowest average prediction error across 600 external test examples, the second lowest and the highest.

System	M062X	B3LYP	HF
Ammonia dimer	**150** (1)	120 (3)	*90* (2)
Water dimer	**90** (1)	*40* (2)	69 (3)
Formic acid dimer	50 (2)	*46* (1)	**90** (3)
Formamide dimer	60 (2)	*50* (1)	**110** (3)
Uracil dimer	180 (1)	180 (2)	**225** (3)
2-Pyridoxine…2-aminopyridine	200 (3)	*190* (2)	**240** (1)
Adenine…thymine	*150* (1)	**210** (2)	180 (3)

As stated above, previous unpublished work of our group has shown that for *the same* training set geometries, B3LYP/apc-1 consistently outperforms HF/6-31G**, yielding kriging models that generate more accurate predictions of the electrostatic interaction between two topological atoms. To confirm that B3LYP/aug-cc-pVDZ also outperforms HF/6-31G**, kriging models were built for the ammonia dimer at the B3LYP/aug-cc-pVDZ level using the geometries sampled from the HF/6-31G** PES. The use of the latter "mixed level" in the construction of a training (and test) set of geometries needs a comment. We are at liberty to construct any training set, as long as it consists of a representative range of chemically relevant geometries. The training set geometries obtained from the HF/6-31G** PES fulfil both of these criteria. Therefore we may proceed with using these geometries for a training (and test) set that will be built at a higher level of theory, in this case B3LYP/aug-cc-pVDZ. This allows the direct comparison of the effect on the accuracy of the kriging models, of different levels of theory at which the atomic multipole moments are obtained.

Fig. 4 shows the S-curves of the effect of changing level of theory. The following example guides the interpretation of a typical S-curve. At a given point on the S-curve the *y*-value corresponds to the percentage of the test set geometries that have a total prediction error within the value on the *x*-axis. For example, it can be seen for the red line of Fig. 4 that 50% of the test set of geometries had absolute prediction errors of less than 0.02 kJ mol^{-1}. Thus it follows that an S-curve that lies to the left is superior to one right of it. The results seen in Fig. 4 confirm that B3LYP outperforms HF methods when the same training geometries and test geometries are used. The results also show that the training set geometries obtained from a PES calculated at the B3LYP/aug-cc-pVDZ level (Fig. 4, green line) lead to higher prediction errors for the two curves corresponding to the HF/6-31G** PES sampled training sets (Fig. 4, red and blue lines). Table 1 shows that more energy was put into the B3LYP PES than into the HF PES. Hence, one would expect the training set geometries to span a larger conformational space for kriging to capture in its models, and hence prediction errors will be slightly higher.

Prediction of the total electrostatic energy of the hydrogen bonded complexes

Fig. 5 shows the S-curves obtained, for all seven hydrogen bonded complexes of the S22 data set at all three levels of theory, and using 600 training examples. Looking at Eq. (13), we emphasise that the individual interaction errors (for each test geometry) are summed *before* the absolute value of this sum is taken. Hence, "cancellation of errors" is possible and indeed likely for each point on the S-curve. This cancellation is justified as the Coulomb law is itself additive. In other words, there is no summation of absolute values of atom–atom interactions when calculating a total electrostatic energy, but a summation of the actual values themselves (whether positive or negative). Analysis of the individual interactions is dealt with in Section 4.3.

Fig. 5 shows that, for all three levels of theory used, the smaller systems lie furthest to the left, with a lower error, and the larger systems lie to the right. This is partially due to increased number of interactions present in the larger systems, and this is an almost linear relationship. Despite this increase in error with number of interactions, even the larger aromatic complexes are predicted within 1 kJ mol^{-1} for 70% of the test geometries, both at B3LYP and M06-2X level. For the ammonia dimer and the water dimer, almost 100% of test structures were predicted within 1 kJ mol^{-1}. None of the complexes have a single test geometry with an error

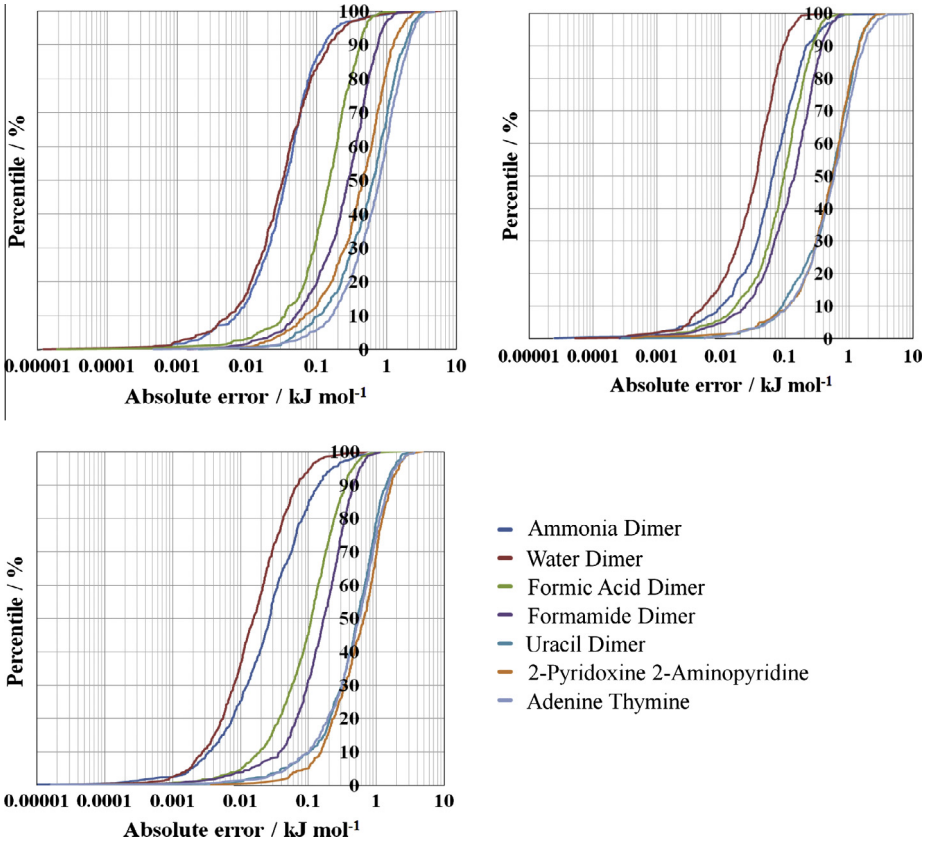


Fig. 5. S-curves of the prediction error for the seven hydrogen bonded dimers in this work at the HF/6-31G** (top left), B3LYP/aug-cc-pVDZ (top right) and M06-2X/aug-cc-pVDZ (bottom left) levels of theory.

greater than 9 kJ mol^{-1}. Almost all interactions are predicted within 1 kcal mol^{-1}, which is often referred to as "chemical accuracy".

The errors for the Hartree–Fock complexes are on average higher than the error of the same complex at either B3LYP or M06-2X levels of theory, as expected. This is a consequence of the improved description of electron correlation as previously mentioned in the Section on Computational Methods, and extensively explained in Ref. [27]. Fig. 6 shows the mean absolute prediction errors of the seven hydrogen bonded systems plotted against the number of intermolecular atomic interactions, for three levels of theory (wave functions and PES obtained at the same level). Fig. 6 demonstrates that neither of the two density functionals consistently outperforms the other. Plotting a trend line through the values of the average prediction error of each system against total number of interactions for B3LYP and M06-2X levels of theory yields overlapping lines (red[2] and blue lines). Plotting a similar line for the HF level of theory (green line) shows that one can expect the average error to increase with a higher number of interactions at a faster rate. The R^2 value of 0.93 for the B3LYP data is higher than that of both HF and M06-2X (R^2 = 0.88 for both), suggesting that there is a stronger correlation between average error and number of interactions. However, due to the random sampling of the geometries this cannot be stated with certainty.

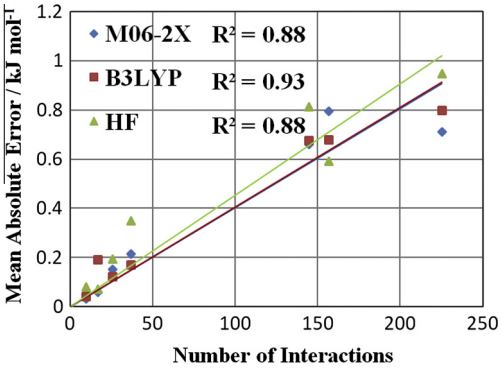


Fig. 6. Mean absolute prediction errors of the seven hydrogen bonded systems plotted against the number of intermolecular atomic interactions.

Prediction of intermolecular interactions

Inspection of the prediction errors for individual interactions (Fig. 7 and related unpublished figures) shows that for all systems the majority of interactions are predicted within ±2 kJ mol^{-1} of the true value with the exception of the adenine…thymine base pair where the errors are mostly within ±4 kJ mol^{-1}. This is within the bounds of the so-called chemical accuracy.

The global trend observed for all interactions is a general increase in accuracy with range. This is explained by the magnitude of the long-range interactions being smaller and therefore errors in magnitude being smaller. Also, the $1/r^L$ dependence of higher order multipole moments, such as quadrupole and octopole moments, results in the electrostatic interaction tending to zero at longer range for these terms. This is most clearly seen in the larger systems of the uracil dimer, the 2-pyridoxine…2-aminopyridine complex, and the adenine…thymine base pair. At large distances, the main contribution is the monopole…monopole interaction

[2] For interpretation of color in Fig. 6, the reader is referred to the web version of this article.

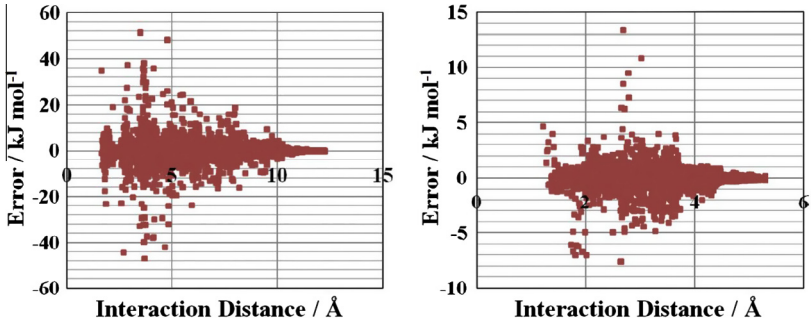


Fig. 7. Prediction errors for individual interactions for the adenine...thymine base pair (left) and the ammonia dimer (right).

(L = 1). This interaction is predicted with a low error as the isotropic nature of the monopole moments means that there is no angular dependence. Therefore the interaction energy between two distant atoms is mainly dependent only on the distance, which presents an easier problem for kriging to model. Additionally, the longest-range interactions are between two hydrogen atoms, for which their moments vary only by small amounts due to being on the outer edges of the system away from the highly polarising heteroatoms on the interior.

There are some poorly predicted interactions for all systems. However, in the geometries for which poor predictions are found, errors can cancel so that when all interactions are summed for that geometry, the net error is very low. This explains the absence of high-error geometries on the total energy S-curves in Fig. 5. Due to the cancellation of errors in all systems, the mean unsigned error (MUE) for the prediction of all the individual interactions in a given system is always higher than the average error for the total energy S-curve.

It is not always the case that a system that has cancellation of large errors will correspond to a point on the total energy S-curve with a high error. For example, for the ammonia dimer, there is an interaction that is predicted with an error of −7 kJ mol^{-1} (Fig. 7). However, the geometry that this interaction belongs to has a total error of -0.02 kJ mol^{-1}, which is the 69th best overall prediction out of 600 test geometries. Likewise, for the adenine...thymine complex, there is a geometry in which an individual interaction is predicted with an error of −38 kJ mol^{-1} (Fig. 7) but the total error for the test geometry is only 1 kJ mol^{-1}, which is only the 173rd worst predicted test geometry out of 600. However, for the adenine...thymine complex, the second worst predicted test geometry does contain the worst predicted individual atom–atom error out of any in the test set.

Table 2
Mean unsigned errors of total net system error and of individual interactions (kJ mol^{-1}).

System	Mean unsigned error of total system interaction energies	Mean unsigned error of individual interaction energies
Ammonia dimer	0.20	0.36
Water dimer	0.05	0.18
Formic acid dimer	0.13	0.23
Formamide dimer	0.18	0.32
Uracil dimer	0.68	0.68
2-Pyridoxine...2-aminopyridine	0.68	0.43
Adenine...thymine	0.80	0.88

The cancellation of errors seen in some of the large complexes is not as large as it first appears. Inspection of Table 2 shows that for all systems the mean unsigned error for the interaction predictions is less than 1 kJ mol^{-1}. For adenine...thymine there are 225 atomic interactions between bases, so over 600 test geometries there are a total of 135,000 interactions, of which 65,500 points are plotted in Figs. 7 and 8. The great majority of predictions lie within ±4 kJ mol^{-1} (Fig. 8).

The symmetry of the predictions is also an interesting feature of the results. Kriging is free from chemical intuition and does not discriminate between different bond types such as hydrogen bonds or C...C interactions. Therefore, if the models built by kriging provide a good answer to the problem asked of them, in this case the values of the atomic multipole moments with respect to the geometry of the system, then the models are successful. As a result there is no preference for over- or underpredicting a specific interaction between two atoms as long as the sum of all predictions in the test geometry sums to the correct intermolecular interaction energy.

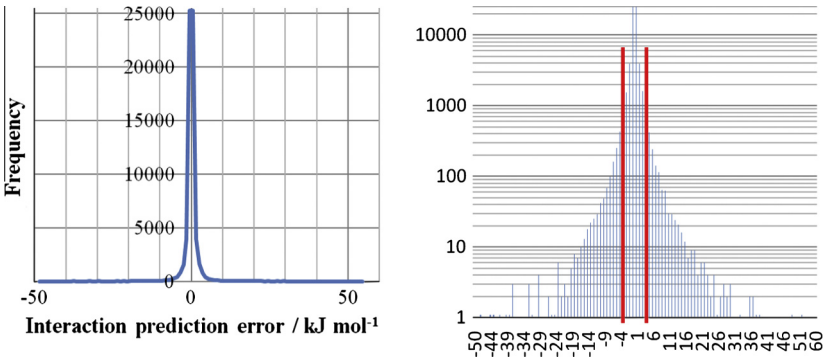


Fig. 8. Spread of prediction errors for 65,500 individual interactions of the adenine...thymine base pair. The graph on the right shows a logarithmic spread of errors (in kJ mol^{-1}) with the ±1 kcal mol^{-1} error bound being marked by the red lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

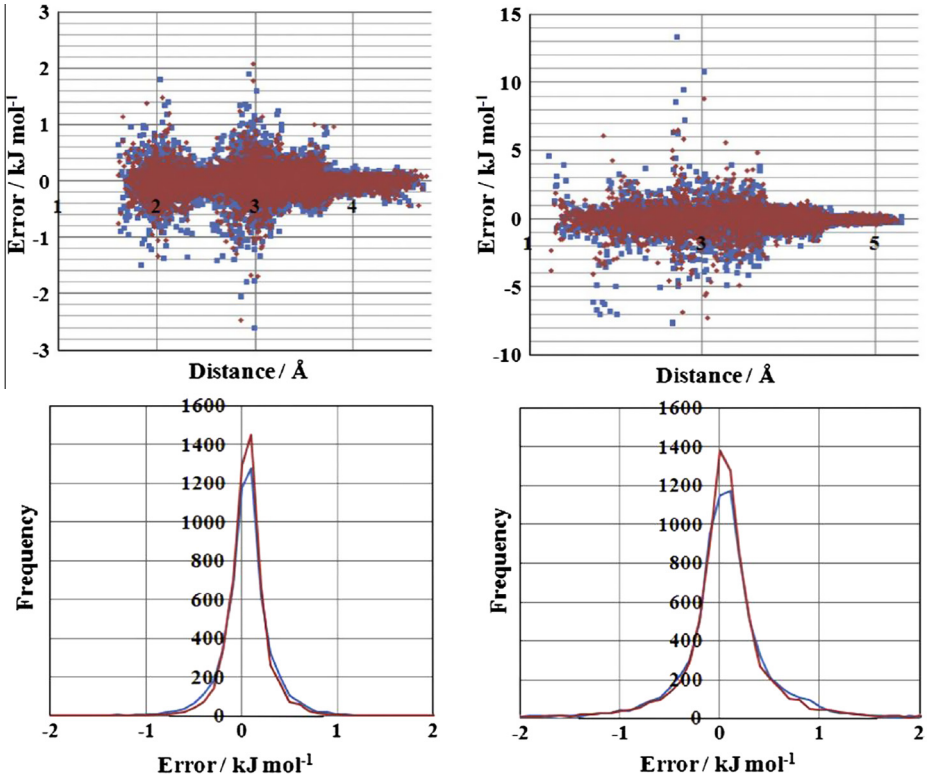


Fig. 9. Scatter plots of and histograms for the individual interaction prediction errors for the B3LYP water dimer (left) and the B3LYP ammonia dimer (right) given by kriging models built with 600 examples (blue) and 1000 examples (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Standard deviations and mean unsigned errors (MUE) (in kJ mol^{-1}) for the prediction of individual interactions for the ammonia and water dimers with training set sizes of 600 geometries and 1000 geometries.

System	Number of training examples	Standard deviation	MUE
Ammonia	600	0.70	0.36
Ammonia	1000	0.60	0.32
Water	600	0.27	0.18
Water	1000	0.23	0.15

The error of the interactions can be lowered by increasing the training set size. This is computationally expensive due to the N^3 scaling where N is the number of training geometries. However, the computational tractability is currently being tackled in our group with early signs of success. Models were rebuilt for both the B3LYP ammonia dimer and the B3LYP water dimer to see the effect of increased training set size. For both systems the unsigned mean error and standard deviation decreased upon moving from 600 to 1000 training set geometries (Fig. 9 and Table 3). For both the ammonia and water dimer the highest errors (“worst offenders”) disappear in going from 600 (blue) to 1000 (red). The CPU time for training the kriging models went from between 2 and 3 days per atom to 6–7 days per atom for 600 examples and 1000 examples, respectively.

Conclusions

The current work shows that high-rank multipole moments up to hexadecupole can be modelled by kriging as a function of nuclear coordinates to high accuracy for intermolecular hydrogen bonded systems. As such systems are ubiquitous within chemistry, and the accurate modelling of intermolecular interactions is of great importance in the design of a next-generation force field. As the models are built on *ab initio* values for the moments, kriging allows for near-*ab initio* electrostatic interaction energies to be obtained in a fraction of the time. The models are able to model intermolecular interactions, including hydrogen bonding, mostly within ±2 kJ mol^{-1}, and the standard deviation and mean unsigned error of intermolecular interactions are shown to decrease with an increase in training set size. In general, models built from moments obtained at the Hartree–Fock level of theory lead to larger errors in the prediction of electrostatic interactions than models built at B3LYP and M06-2X. There is no observable difference between the accuracy of our results for the two density functionals.

Future studies aim to lower the errors further by making more technical changes to the kriging process, such as the number of particles in the PSO, and also by improving the efficiency of the program to allow greater training set sizes. The current work, however, delivers proof-of-concept that machine learning can be used to accurately describe intermolecular interactions. Kriging models for the dispersion bound and mixed dispersion/hydrogen bonded complexes of the S22 set will be built in future work.

References

[1] H. Umeyama, K. Morokuma, J. Am. Chem. Soc. 99 (1977) 1316–1332.
[2] G.J.B. Hurst, P.W. Fowler, A.J. Stone, A.D. Buckingham, Int. J. Quantum Chem. 29 (1986) 1223–1239.
[3] I. Nobeli, S.L. Price, J.P.M. Lommerse, R. Taylor, J. Comput. Chem. 18 (1997) 2060–2074.
[4] J.P.M. Lommerse, S.L. Price, R. Taylor, J. Comput. Chem. 18 (1997) 757–774.
[5] P.Y. Ren, C.J. Wu, J.W. Ponder, J. Chem. Theory. Comput. 7 (2011) 3143–3161.
[6] J.-H. Lii, N.L. Allinger, J. Phys. Org. Chem. 7 (1994) 591–609.
[7] J.-H. Lii, N.L. Allinger, J. Comput. Chem. 19 (1998) 1001–1016.
[8] P. Cieplak, J. Caldwell, P. Kollman, J. Comput. Chem. 22 (2001) 1048–1057.
[9] J. Kong, J.-M. Yan, Int. J. Quantum. Chem. 46 (1993) 239–255.
[10] M.S. Shaik, M. Devereux, P.L.A. Popelier, Mol. Phys. 106 (2008) 1495–1510.

[11] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, J. Chem. Phys. 79 (1983) 926.
[12] M.W. Mahoney, W.L. Jorgensen, J. Chem. Phys. 112 (2000) 8910–8922.
[13] J.W. Ponder, C. Wu, V.S. Pande, J.D. Chodera, M.J. Schnieders, I. Haque, D.L. Mobley, D.S. Lambrecht, R.A.J. DiStasio, M. Head-Gordon, G.N.I. Clark, M.E. Johnson, T. Head-Gordon, J. Phys. Chem. B 114 (2010) 2549–2564.
[14] A.J. Stone, Chem. Phys. Lett. 83 (1981) 233–239.
[15] U. Koch, P. Popelier, A. Stone, Chem. Phys. Lett. 238 (1995) 253–260.
[16] J. Cao, B.J. Berne, J. Chem. Phys. 99 (1993) 6998.
[17] G. Lamoureux, B. Roux, J. Chem. Phys. 119 (2003) 3025.
[18] S.W. Rick, S.J. Stuart, B.J. Berne, J Chem. Phys. 101 (1994) 6141–6156.
[19] R.T. Sanderson, Science 114 (1951) 670–672.
[20] J. Applequist, J.R. Carl, K.-K. Fung, J. Am. Chem. Soc. 94 (1972) 2952.
[21] P.J. Winn, G.G. Ferenczy, C.A. Reynolds, J. Comput. Chem. 20 (1999) 704–712.
[22] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, The MIT Press, Cambridge, USA, 2006.
[23] P.L.A. Popelier, É.A.G. Brémond, Int. J. Quantum. Chem. 109 (2009) 2542–2553.
[24] P.L.A. Popelier, F.M. Aicken, ChemPhysChem 4 (2003) 824–829.
[25] P. Jurecka, J. Sponer, J. Cerny, P. Hobza, Phys. Chem. Chem. Phys. 8 (2006) 1985–1993.
[26] M.J.L. Mills, P.L.A. Popelier, Comput. Theor. Chem. 975 (2011) 42–51.
[27] M.J.L. Mills, P.L.A. Popelier, Theor. Chem. Acc. 131 (2012) 1137–1153.
[28] S.M. Kandathil, T.L. Fletcher, Y. Yuan, J. Knowles, P.L.A. Popelier, J. Comput. Chem. 34 (2013) 1850–1861.
[29] C.M. Handley, G.I. Hawe, D.B. Kell, P.L.A. Popelier, Phys. Chem. Chem. Phys. 11 (2009) 6365–6376.
[30] S. Houlding, S.Y. Liem, P.L.A. Popelier, Int. J. Quantum. Chem. 107 (2007) 2817–2827.
[31] C.M. Handley, P.L.A. Popelier, J. Chem. Theory Comput. 5 (2009) 1474–1489.
[32] M.G. Darley, C.M. Handley, P.L.A. Popelier, J. Chem. Theory Comput. 4 (2008) 1435–1448.
[33] L. Joubert, P.L.A. Popelier, Phys. Chem. Chem. Phys. 4 (2002) 4353–4359.
[34] P.L.A. Popelier, A.J. Stone, Mol. Phys. 82 (1994) 411–425.
[35] P.L.A. Popelier, A.J. Stone, D.J. Wales, Faraday Discuss. 97 (1994) 243–264.
[36] S.Y. Liem, P.L.A. Popelier, M. Leslie, Int. J. Quantum Chem. 99 (2004) 685–694.
[37] S.Y. Liem, M.S. Shaik, P.L.A. Popelier, J. Phys. Chem. B 115 (2011) 11389–11398.
[38] S. Liem, P.L.A. Popelier, J. Chem. Phys. 119 (2003) 4560–4566.
[39] S.Y. Liem, P.L.A. Popelier, J. Chem. Theory Comp. 4 (2008) 353–365.
[40] M.S. Shaik, S.Y. Liem, P.L.A. Popelier, J. Chem. Phys. 132 (2010) 174504.
[41] M.S. Shaik, S.Y. Liem, Y. Yuan, P.L.A. Popelier, Phys. Chem. Chem. Phys. 12 (2010) 15040–15055.
[42] R.F.W. Bader, Atoms in Molecules. A Quantum Theory, Oxford Univ. Press, Oxford, Great Britain, 1990.
[43] P.L.A. Popelier, Atoms in Molecules An Introduction, Pearson Education, London, Great Britain, 2000.
[44] R.G.A. Bone, R.F.W. Bader, J. Phys. Chem. 100 (1996) 10892–10911.
[45] V. Tognetti, L. Joubert, J. Chem. Phys. 138 (2013) 024102.
[46] P.L.A. Popelier, D.S. Kosov, J. Chem. Phys. 114 (2001) 6539–6547.
[47] P.L.A. Popelier, L. Joubert, D.S. Kosov, J. Phys. Chem. A 105 (2001) 8254–8261.
[48] A.J. Stone, The Theory of Intermolecular Forces, first ed., vol. 32, Clarendon Press, Oxford, 1996.
[49] G.I. Hawe, P.L.A. Popelier, Can. J. Chem. 88 (2010) 1104–1111.
[50] J. Řezáč, P. Jurečka, K.E. Riley, J. Černý, H. Valdes, K. Pluháčková, K. Berka, T. Řezáč, M. Pitoňák, J. Vondrášek, Coll. Czech. Chem. Commun. 73 (2008) 1261–1270.
[51] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, J.A.J. Montgomery, J.T. Vreven, K.N. Kudin, J.C. Burant, J.M. Millam, S.S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G.A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J.E. Knox, H.P. Hratchian, J.B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, P.Y. Ayala, K. Morokuma, G.A. Voth, P. Salvador, J.J. Dannenberg, V.G. Zakrzewski, S. Dapprich, A.D. Daniels, M.C. Strain, O. Farkas, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J.V. Ortiz, Q. Cui, A.G. Baboul, S. Clifford, J. Cioslowski, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, C. Gonzalez, J.A. Pople, GAUSSIAN03, Gaussian Inc., Pittsburgh PA, 2003.
[52] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.J.A. Montgomery, J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, N.J. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Ö. Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox, Gaussian 09, Gaussian, Inc., Wallingford CT, 2009.
[53] T.A. Keith, AIMAll. 11.04.03 ed., 2011 aim.tkgristmill.com.
[54] F.M. Aicken, P.L.A. Popelier, Can. J. Chem. 78 (2000) 415–426.
[55] E.R. Johnson, R.A. Wolkow, G.A. DiLabio, Chem. Phys. Lett. 394 (2004) 334–338.
[56] F.O. Kannemann, A.D. Becke, J. Chem. Theory Comput. 5 (2009) 719–727.
[57] A.K. Rappé, E.R. Bernstein, J. Phys. Chem. A 104 (2000) 6117–6128.
[58] S. Tsuzuki, H.P. Luthi, J. Chem. Phys. 114 (2001) 3949–3957.
[59] Y. Zhao, D. Truhlar, Theor. Chem. Acc. 120 (2008) 215–241.
[60] S.T. Schneebeli, A.D. Bochevarov, R.A. Friesner, J. Chem. Theory Comput. 7 (2011) 658–668.
[61] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, J. Chem. Phys. 132 (2010) 154104–154119.

Computational and Theoretical Chemistry 1053 (2015) 298–304

Contents lists available at ScienceDirect

Computational and Theoretical Chemistry

journal homepage: www.elsevier.com/locate/comptc

Where does charge reside in amino acids? The effect of side-chain protonation state on the atomic charges of Asp, Glu, Lys, His and Arg

Timothy J. Hughes, Paul L.A. Popelier *

Manchester Institute of Biotechnology (MIB), 131 Princess Street, Manchester M1 7DN, United Kingdom
School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, United Kingdom

ARTICLE INFO

Article history:
Received 3 July 2014
Received in revised form 31 July 2014
Accepted 31 July 2014
Available online 8 August 2014

Keywords:
Quantum Theory of Atoms in Molecules (QTAIM)
Amino acids
Quantum Chemical Topology (QCT)
Atomic charge
Protonation

ABSTRACT

Quantum topological atomic charges have been calculated at B3LYP/apc-1 level to identify where the charge is located on amino acid residues when the side-chain has been either protonated (Arg, Lys, His) or deprotonated (Glu, Asp). All energy local energy minima in the Ramachandran map of each (neutral) amino acid were populated with a number of distorted molecular geometries, summing up to a thousand geometries for each amino acid. The majority of the molecular charge is found on the side-chain (81–100%), with a large percentage of the charge located on the functional group undergoing protonation/deprotonation. Each side-chain (or residue) methylene group was found to act as an insulator between an amino acid's backbone and its side-chain because it accepts the majority of charge not located on the side-chain. As a result there is no significant charge on backbone atoms relative to the neutral molecule. In the case of His^+ and Arg^+, where the charge is spread over a large number of atoms due to resonance, the influence of the positive charge on the backbone atoms is reduced.

1. Introduction

The complex mechanisms of enzymatic catalysis have been studied intensively for decades. A common feature in these mechanisms is the protonation and deprotonation of the active site amino acid side-chains involved in the catalysis. For example, the rate limiting step in the conversion of $CO_2 + H_2O \rightarrow HCO_3^- + H^+$ by the enzyme carbonic anhydrase is a proton transfer involving the residue His64 [1,2]. Similarly, a proton transfer mechanism involving a Glu residue in the active site is found to be the rate-determining step in the mechanism of the enzyme glutaminylcyclase [3]. The subtle changes in the electronic charge of the active site atoms of glutaminylcyclase play a role in determining the path that the reaction follows. This effect arises through strengthening of hydrogen bonds within the active site upon proton transfer. The mechanism employed by enzyme horseradish peroxidase includes a nucleophilic attack by the hydroxyl oxygen of Ser195. However, this step requires activation through the deprotonation of the hydroxyl group [4]. Deprotonation results in the charge of the oxygen atom becoming more negative and hence more nucleophilic.

The above examples show that when developing a computational model to describe enzymatic reactions, any changes in electronic structure must be captured. Early potentials that enabled the modelling of reactions include the empirical valence bond approach [5] and the "ReaxFF" force field [6]. The popularity of QM/MM approaches is increasing in the study of such systems due to increases in computer power [7]. Currently under development in our lab is the quantum chemical topological force field (QCTFF). This is a novel approach to building a molecular mechanics force field, in which machine learning is used to map quantum mechanical properties (such as atomic multipole moments [8–10], kinetic energy [11] and exchange-repulsion) directly to the coordinates of the system. Preliminary work has shown that this methodology enables the modelling of changes in atomic charge as a reaction path is followed.

There is a perhaps surprising lack of literature detailing the changes in the atomic charges of amino acids upon a change of the side-chain protonation state, with studies [12–15] typically focusing on the zwitterionic states of amino acids. To address this gap in the literature, a thousand geometries for each of a total of five amino acids that most commonly undergo changes in protonation state (Asp, Glu, His, Lys, Arg) have been sampled for both the protonated and deprotonated state, and the changes in average atomic charges have been compared. In this work, charges have been obtained from the Quantum Theory of Atoms in Molecules

* Corresponding author. Tel.: +44 161 3064511; fax: +44 161306 4559.
E-mail address: pla@manchester.ac.uk (P.L.A. Popelier).

http://dx.doi.org/10.1016/j.comptc.2014.07.020
2210-271X/

(QTAIM) [16–18]. The extensive QTAIM work [19–21] of Matta and Bader on all natural amino acids, provides a rich background to the current work but does not specifically address the question of where an excess or depletion of a formal unit charge resides compared to the neutral amino acid.

There are many methods of obtaining atomic charges but the question of which protocol produces the "best" atomic charges is contentious. Arguments for and against the different charge methods typically fall into one of two competing schools of thought. The first supports a belief that the atomic charge should be capable of reproducing the electrostatic potential around an atom. The second asserts that the charge should correctly describe the charge transfer within a molecule. We subscribe to the latter assertion. Indeed, an atomic charge should do no more than what the name says: describe, and indeed correctly represent, the charge on an atom. Demanding that the single number that is the atomic charge also reproduces the electrostatic potential generated by the atom, is problematic because it ignores the non-spherical features of an atom in a molecule, which are prominent at short (and even medium) range [22]. Atomic charges that are fitted to best reproduce an electrostatic potential are just numbers [23], one of many possible solutions, and by no means guaranteed to capture true charge transfer effects. QTAIM charges fall into the second category of charges: they do reproduce well the charge transfer in a molecule, even if they have been criticised for being "unrealistically" high [24]. At the other end of the spectrum, the Hirshfeld population analysis produces very small charges, which ironically become more QTAIM-like, when corrected for the arbitrariness of the promolecule, as done in Hirsheld-I [25]. The criticism that QTAIM charges do not reproduce the electrostatic potential is remedied by performing a multipolar expansion (of which the QTAIM charge is the first term of the expansion, the monopole moment) where it was shown [26] a long time ago that reproduction of the *ab initio* electrostatic potential was achieved at a so-called interaction rank of $L = 5$. To quote from this work, "*This work makes clear that the atomic population (or rank zero multipole moment) is just one term of the expansion of a physically observable quantity, namely the electrostatic potential. Hence, QTAIM populations (and thus charges) cannot be judged on their reproduction of the electrostatic potential. Instead, they must be seen in the context of a multipolar expansion of the exact electrostatic potential of a topological atom*".

2. Background and computational details

2.1. Topological atoms

The electron density of a system partitions itself, without the need for setting any parameter values, or calculation through an iterative procedure. The only concept required is that of the gradient vector of the electron density, denoted $\nabla\rho$, which traces paths of steepest ascent. The vast majority of these so-called gradient paths terminate at a nuclear position, thereby carving out one subspace for each nucleus. These subspaces are identified with topological atoms. More details can be found in a very recent, didactic, historic and refreshing account [27] of QTAIM. The central idea of partitioning a quantum system by means of gradient paths was first carried out [16] on ρ by the group of Bader, and later [28,29] on the Laplacian of ρ. Meanwhile several other three-dimensional quantum property density functions were investigated (for a list of examples see Box 8.1 in [27]) justifying [30] the overarching name Quantum Chemical Topology (QCT) [31].

Fig. 1 shows an example, relevant to the current work, of a protonated lysine falling apart into topological atoms, as generated [32,33] by in-house software. The latter can be seen as bubbles, touching each other without overlapping or leaving gaps; they are malleable boxes that change their shape in response to a change in the nuclear skeleton.

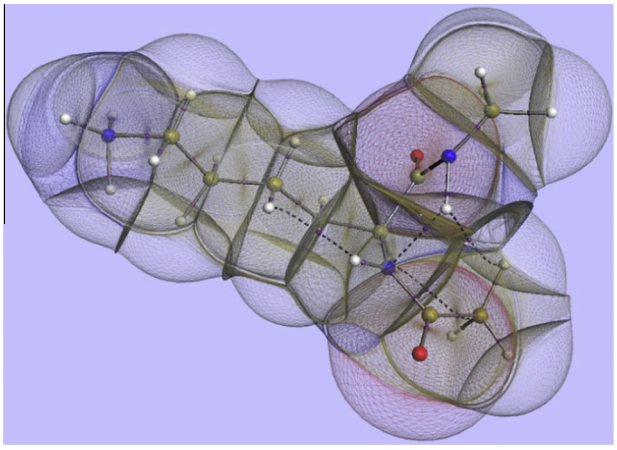


Fig. 1. Finite-element representation of a molecular geometry of protonated lysine.

2.2. Geometry generation

Each amino acid was capped by a [$CH_3C{=}O$] group at the N-terminus, and by a [$NHCH_3$] group at the C-terminus to create the so-called "dipeptide". The minimum energy geometries for each neutral amino acid were obtained through a comprehensive search of the potential energy surface [34]. The number of energetic minima for each amino acid is given in Table 1.

A thousand geometries for each amino acid were obtained by distributing energy into the normal vibrational modes of each local energy minimum for each neutral amino acid. These geometries were generated by the in-house computer program EROS. Quasi-random amounts of energy are put into the normal vibrational modes, which then spreads evenly over all vibrational modes. The motion of the vibrational modes is modelled harmonically, and 'snapshots' of the nuclear coordinates at random points during the vibrational motion formed the set of geometries. Bonds may dissociate and atoms fly apart if the input energy is too high, which is corrected by an iterative process where initially a maximum input energy is defined, typically about 200 kJ mol^{-1} and geometries are generated. If broken bonds are present then the maximum energy is lowered and the process is repeated until no broken bonds are present. More details can be found in Ref. [35].

All charged amino acid residues except Arg were obtained by direct addition or removal of a proton on the side-chain of the distorted geometries. For each of the thousand sampled neutral Asp and Glu residues, the acidic proton was removed in order to obtain the geometries of the Asp^- and Glu^-, respectively. A similar approach was also taken in the case of Lys^+, where a proton was added to the primary amine to give the positively charged tetrahedral ammonium group. His^+ was similarly obtained by protonating the lone pair position of N29 to give the positively charged imidazolium group. Due to the more complex structural changes that take place in Arg upon protonation a different approach was taken

Table 1
Number of local energy minima for each amino acid studied in this work.

Amino acid	No. minima
Asp	36
Glu	36
His	24
Lys	39
Arg	61

in obtaining the Arg^+ geometries. In particular, the neutral Arg has a guanidine system with two pyramidal nitrogens (N19 and N16) and one planar nitrogen (N34). However, in Arg^+ this group formally becomes a guanidinium group, which has three planar nitrogens. The addition of a proton to N34 causes the geometrical change between guanidine and guanidinium. Therefore, an alternative approach was taken; a proton was added to each of the minimum energy geometries and then the guanidinium group alone ($[-NH-C(NH_2)_2]^+$) was allowed to relax by partial geometry optimisation. These new "minima" were then input to EROS to sample the thousand distorted Arg^+ geometries.

2.3. Computational details

Normal modes sampling was performed by the in-house FORTRAN code EROS. All *ab initio* calculations were performed by GAUSSIAN09 [36] at the B3LYP/apc-1 [37] level of theory, taking advantage of a basis set with polarisation and diffuse functions optimised for use with density functionals. QTAIM charges for all atoms were calculated with the program AIMAll [38], and are listed in the Supporting Information, as averages over all configurations (corresponding to all local energy minima), along with ranges and standard deviations. An atomic charge is an atomic property the least sensitive to integration error [39].

3. Results and discussion

Numbered geometries for all five protonated capped amino acids (Asp, Glu, Lys^+, His^+ and Arg^+) are provided in Fig. 2. For convenience, both the protonated and deprotonated amino acids share a common numbering system. The following discussion refers to the amino acids as consisting of both side-chain atoms and backbone atoms. The set of side-chain atoms consist of all atoms starting with C_β (including its methylene hydrogens), whereas the backbone corresponds to the C_α, the two peptide groups and the methyl caps, as well as all associated hydrogen atoms.

Tables containing the average value, range and standard deviation of all atomic charges for both the protonated and deprotonated systems studied in this work is provided as Supporting Information. With the exception of Arg/Arg^+ (due to the different method of obtaining the structures), patterns in both the standard deviation and the range of atomic charge for similar atom types are observed.

A number of general observations can be made, across the various systems. For an atom in a neutral amino acid, both the range of the atomic charge and its standard deviation maintain a similar value in the charged amino acid. Within an amino acid, there is also no clear distinction in the behaviour of the range of the charge or the standard deviation between the side chain atoms and backbone atoms. Peptide nitrogen atoms have a range between 0.65 and 0.80 a.u. around the average value. Peptide oxygen atoms have a smaller range, between 0.45 and 0.65 a.u. around the average. Peptide hydrogen atoms have the smallest range in atomic charge of the peptide group atoms, with a range of 0.30–0.40 a.u. around the average value. Peptide carbon atoms show the largest range of atomic charges with a range often over 1 a.u. around the average value. Alpha carbon atoms have a smaller range of roughly 0.65 a.u. The side chain methylene carbon atoms exhibit smaller ranges in atomic charge than both the peptide and C_α atoms. The standard deviation of the atomic charge does not show any correlation with the magnitude of the charge. However, a larger range of charges does correlate to a larger standard deviation. The standard deviation of the hydrogen atoms is always less than 0.05 a.u., which is considerably smaller than most carbon, nitrogen and oxygen atoms. Hydrogen atoms within the functional groups of positively charged amino acid side chains (for example H32, H33 or H34 of Lys^+) exhibit a decrease in the range of charges and the standard deviation relative to the same hydrogen atoms when present in the neutral side chain. This is due to less charge being available to these atoms (overall charge of +1). The standard deviation of the carbon, oxygen and nitrogen typically lie in the range of 0.5–1.2 a.u.

3.1. Acidic amino acids (Asp and Glu)

The atomic charge (averaged over all thousand geometries) for all atoms of both Asp and Asp^- can be seen in Fig. 3. The difference between the atomic charges in the neutral and in the charged amino acid is also plotted. Atom H25 is the acidic proton that is removed upon deprotonation. In the neutral molecule, the acidic proton has a charge of +0.56 a.u. (see Fig. 3), which means that upon deprotonation a charge of $(-1) + 0.56 = -0.44$ a.u. is left over

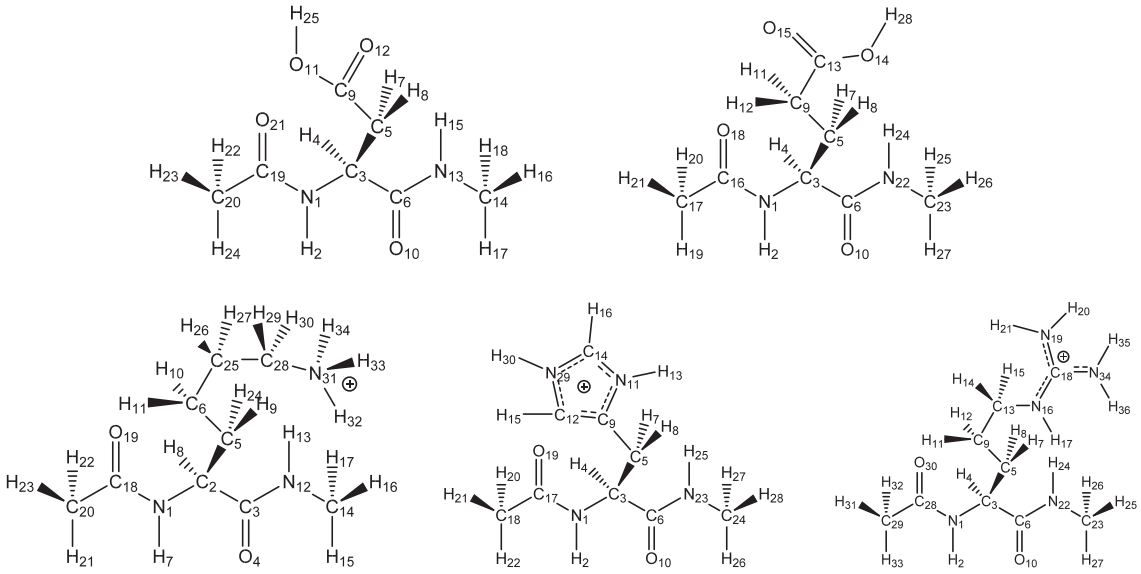


Fig. 2. Numbered geometries for capped amino acids Asp (top left), Glu (top right), Lys^+ (bottom left), His^+ (bottom middle) and Arg^+ (bottom right). The numerical labels of the atoms ("atom number") of the deprotonated geometries are the same. In all five cases the proton removed upon deprotonation is the highest numbered proton.

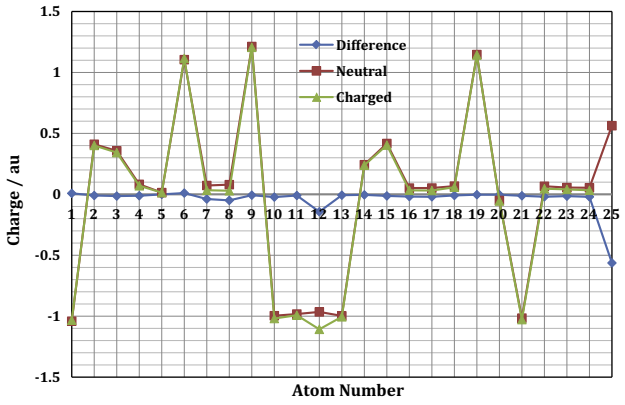


Fig. 3. The averaged atomic charges of both Asp (green) and Asp^- (red) and the difference (blue) between the neutral and charged atomic charges. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to be distributed over the remaining geometry. Of this 0.44 of an electron, 57% (−0.25 a.u.) moves onto the side-chain atoms. The remaining 43% (−0.19 a.u.) is found on the backbone atoms.

Despite the even spread of H25's charge over the whole molecule upon deprotonation, the total molecular charge is highly concentrated on the side-chain of the molecule of Asp^-. Upon deprotonation the sum of all side-chain atomic charges (including H25 for the neutral side-chain) decreases from −0.01 a.u. to −0.81 a.u. meaning that 81% of the total molecular charge is found on the side-chain atoms. The carboxylate group of Asp^- has a summed charge of −0.89 a.u. (89% of the molecular charge). The methylene group of the side-chain increases in charge from Asp to Asp^-, with a summed (group) atomic charge of 0.08 a.u. (= |−0.89 − (−0.81)|). There are no chemically significant changes in atomic charges of the backbone atoms. Curiously, one of the most significant changes in backbone charge is that the hydrogen atoms on the methyl capping groups undergo a difference in summed charge of −0.10 a.u. when going from Asp to Asp^-.

Similar results are found for the deprotonation of Glu to Glu^-. The differences in average atomic charge over a thousand conformations are shown in Fig. 4. Atom H28 corresponds to the acidic proton that is removed when going from Glu to Glu^-. The charge of H28 in Glu is 0.55 a.u. meaning that in Glu^- only −0.45 a.u. of additional negative charge is available to the molecule for redistribution. A value of −0.33 a.u. of the additional charge (73%) remains on the side-chain atoms, and the remaining −0.12 a.u. is shared by the backbone atoms.

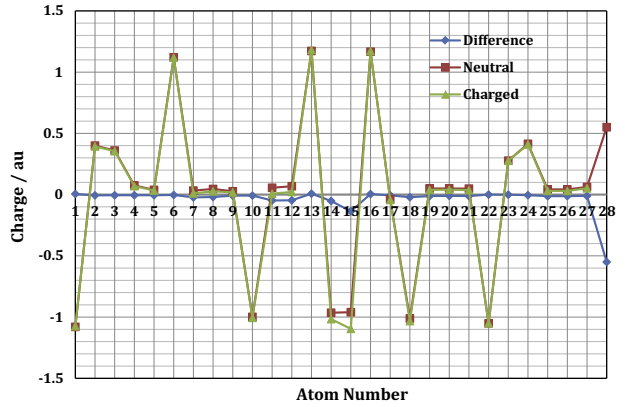


Fig. 4. The averaged atomic charges of both Glu (green) and Glu^- (red) and the difference (blue) between the neutral and charged atomic charges. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Similar to Asp^-, it is apparent that the majority of the negative molecular charge of Glu^- is found on the side-chain atoms (−0.88 a.u., 88% of total molecular charge). A similar situation to that of Asp^- arises where the majority of the side-chain charge of Glu^- is concentrated on the carboxylate group. In Glu^- the carboxylate atoms have a summed charge of −0.93 a.u., which is an increase in charge of −0.73 a.u. relative to the summed charge of the neutral carboxylic acid group. There is no significant change in backbone atom charges. The methyl hydrogens increase in summed charge by −0.7 a.u., which is less than in the case of Asp^-.

There are differences between the changes seen in atomic charges for the two systems Asp^- and Glu^-. Eight percent more charge is located on the side-chain of Glu^- than on the side-chain of Asp^-. Also, less of the additional charge available upon deprotonation is found on the backbone atoms for Glu^- (26%) compared to Asp^- (43%). This observation has led to the idea of a "buffering" methylene group. Methylene groups are neutral fragments in the side-chain that act to separate the polar carboxylic acid/carboxylate group from the rest of the amino acid. The additional methylene group in the side-chain of Glu^- creates a more insulating buffer between the charged carboxylate group and the amino acid backbone. This buffering is responsible for the increased localisation of the charge on the side-chain in Glu^- than in Asp^-.

In summary, the deprotonation of the acidic hydrogen in Asp and Glu, causes the newly available negative charge to predominantly reside on the side-chain atoms (81% and 88% for Asp^- and Glu^-, respectively). In particular, the charge is localised on the three carboxylate atoms (COO^-). Changes in the charge of backbone atoms, when going from the protonated to the deprotonated state, are insignificant due largely to "buffering" methylene groups. The buffering effect is greater in the case of Glu^- where there are two methylene groups.

3.2. Basic amino acids (Lys, His and Arg)

Fig. 5 shows the atomic charges of Lys and Lys^+. The acidic proton in Lys^+ (H33) has a charge of 0.48 a.u. This means that the atoms present in Lys undergo a sum increase in positive charge of 0.52 a.u. when going from neutral Lys to protonated Lys^+ (because 0.52 of an electron has moved onto H33). A positive charge of 0.44 a.u. (85%) is generated on side-chain atoms. As one would expect, the backbone atoms of Lys^+ remain relatively unaffected by the protonation of the amine group due to the four methylene groups "buffering" the ammonium group from the backbone. This explains the summed charge of the backbone atoms increasing by only (0.52–0.44 =) 0.08 a.u. upon protonation.

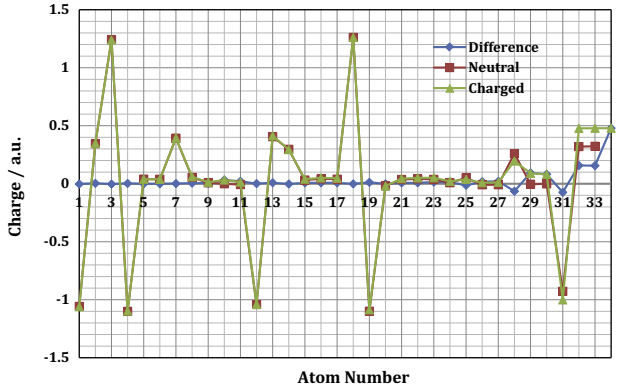


Fig. 5. The averaged atomic charges of both Lys (red) and Lys^+ (green) and the difference (blue) between the neutral and charged atomic charges. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

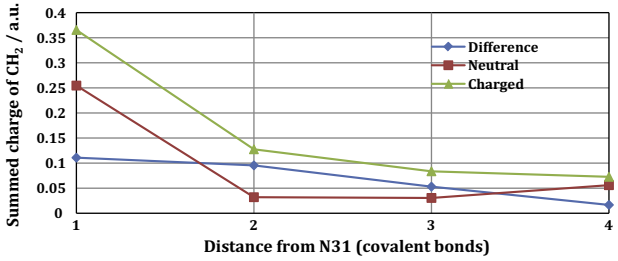


Fig. 6. Summed charges of the methylene groups of Lys (red), Lys^+ (green) and their difference (blue) against the number of covalent bonds from the side-chain nitrogen atom (N31) (1 = C_ε, 2 = C_δ, 3 = C_γ and 4 = C_β). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

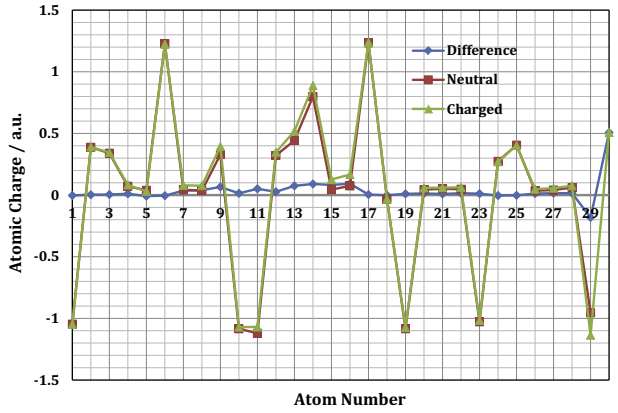


Fig. 7. The averaged atomic charges of both His (red) and His^+ (green) and the difference (blue) between the neutral and charged atomic charges. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

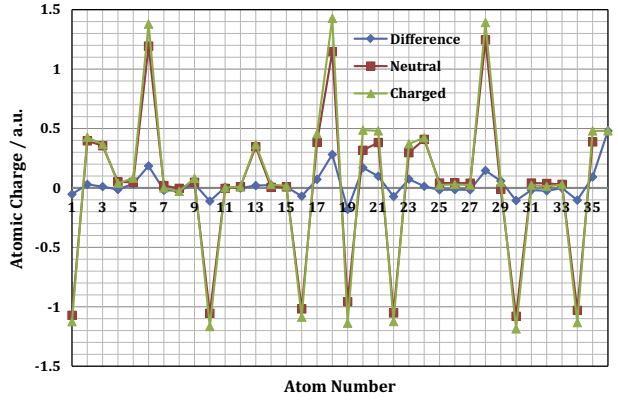


Fig. 8. The averaged atomic charges of both Arg (red) and Arg^+ (green) and the difference (blue) between the neutral and charged atomic charges. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fragmenting the molecule into side-chain atoms and backbone atoms and summing the atomic charges gives a clear illustration of the buffering effect. The summed charge of all side-chain atoms in Lys is 0.09 a.u., whereas in Lys^+ the summed charge is 1.01 a.u. (an increase of 0.92 a.u.), whereas the backbone atoms have a summed charge of −0.01 a.u. This shows that all of the positive molecular charge is found on the side-chain. The ammonium atoms ($[-NH_3]^+$) of Lys^+ have a summed charge of 0.43 a.u., which is the largest contribution to the molecular charge. The remaining charge resides on the methylene groups. The summed charge of

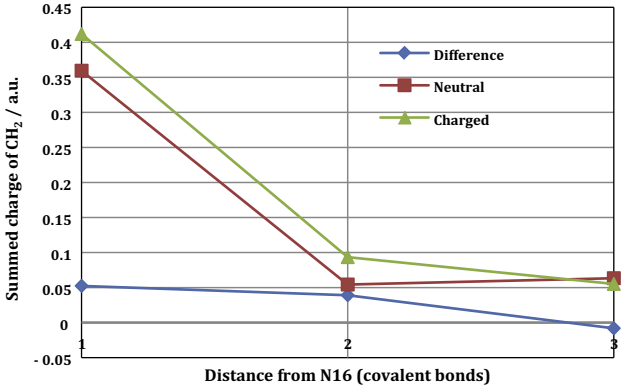


Fig. 9. Summed charges of the methylene groups of Arg (red), Arg^+ (green) and their difference (blue) against the number of covalent bonds counting from the side-chain nitrogen atom (N16) (1 = C_δ, 2 = C_γ and 3 = C_β). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

each methylene group is plotted in Fig. 6 against the number of covalent bonds between the carbon atom and the ammonium nitrogen. The summed charge of the methylene atoms decreases as the number of covalent bonds away from the ammonium nitrogen increases. The summed charge of the methylene groups in the neutral Lys molecule are also plotted in Fig. 6. From left to right, the gap between the neutral and charged values narrows, and by the fourth methyl carbon the difference between the charged and neutral methylene groups is only a summed charge of 0.02 a.u. This illustrates clearly the "buffering" effect of the methylene groups; the backbone atoms are almost unaware of the protonation of the amine group.

The atomic charges of His and His^+ can be seen in Fig. 7. The acidic proton of His^+ (H30) has a charge of 0.51 a.u. meaning that 0.49 a.u. of positive charge much be built up on the atoms present only in His (0.49 of an electron has moved onto H30). Of this charge, 76% (0.37 a.u.) lies on the side-chain atoms. The summed charge of the side-chain atoms is 0.93 a.u, which is 0.88 a.u more positive than the neutral side-chain. This again shows that the molecular charge is predominantly located on the side-chain, with the backbone atoms of His^+ undergoing a change in summed charge of 0.12 a.u. The only other amino acid that only has a single methylene group to protect the side-chain from the effects of side-chain protonation is Asp/Asp^-. The backbone atoms of Asp^- experience a greater change in summed charge (−1.9 a.u.). An incorrect assumption would be at the methylene (C5H7H8, Fig. 2) in Asp^- is a worse "buffer" than the methylene (C5H7H8) in His^+. This is not true. Instead, in His^+ the positive charge is delocalised over the imidazolium and therefore its methylene group is no longer directly bonded to a charged atom but rather a group of atoms charged to a lesser extent. Thus, the methylene group in His^+ is only 0.07 a.u. more positive than the methylene in the neutral His, compared to a difference of −0.16 a.u. for the methylene of Asp and Asp^-.

The atomic charges of Arg and Arg^+ can be seen in Fig. 8. The acidic proton of Arg^+ (H36) has a charge of 0.48 a.u. meaning that 0.52 a.u. of positive charge is built up on the atoms present in the neutral Arg molecule. The side-chain atoms of Arg increase by a total summed charge of 0.44 a.u. when the proton is added, which accounts for 86% of the charge build up. The small contribution to this charge by the backbone atoms is due to a combination of the factors previously discussed. Firstly, there are three buffering methylene groups to separate the protonated guanidinium group from the backbone. The summed charge of the methylene groups can be seen in Fig. 9. By the second methyl group (C_γ) the

difference between the charged and neutral methylene groups is less than 0.05 a.u. The second reason for the low increase in backbone charge is that the positive charge is stabilised by the delocalised – system of the guanidinium group. The eight guanidine atoms present in Arg account for 81% of the total side-chain increase in summed charge of Arg^+.

The side-chain atoms of Arg^+ have a summed charge of 1.01 a.u., accounting for all of the positive charge of the molecule. The backbone atoms have a summed charge of −0.01 a.u. due to the three "buffering" methylene groups and the spread of the charge over the guanidinium group. The guanidinium group has a summed charge of 0.45 a.u, which is the largest contribution to the molecular charge. The next largest contributor to the molecular charge is the methylene group adjacent to the guanidinium group, with summed charge of 0.41 a.u.

4. Conclusions

The atomic charges of five amino acids that undergo protonation (Lys, His and Arg) and deprotonation (Asp and Glu) have been studied. The QCT atomic charges of all atoms, averaged over a thousand conformations, for both charged and neutral amino acids have been compared. For Asp and Glu, which are deprotonated to form Asp^- and Glu^-, the majority of the negative charge is located on the side-chain atoms (81% and 88% respectively). Less charge is found on the backbone of Glu^- than Asp^- due to the additional side-chain methylene group "buffering" the charge. The buffering effect of methylene groups is more apparent in the positively charged amino acids Lys^+, His^+ and Arg^+ due to the large number of methylene groups in Lys^+ and Arg^+. By the third methylene group counting from the site of protonation, the summed charge of the CH_2 group is comparable to that of the neutral molecule. Spread of the charge over multiple side-chain atoms (such as in the imidazolium ring of His^+ and the guanidinium group of Arg^+) also reduces the effect of the charge on backbone atoms.

Acknowledgements

TJH is grateful to AstraZeneca for the provision of top-up funding his BBSRC CASE Ph.D. studentship.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.comptc.2014.07.020.

References

[1] C.M. Maupin, G.A. Voth, Proton transport in carbonic anhydrase: insights from molecular simulation, Biochim. Biophys. Acta (BBA) – Proteins Proteomics 1804 (2010) 332–341.
[2] R.L. Mikulski, D.N. Silverman, Proton transfer in catalysis and the role of proton shuttles in carbonic anhydrase, Biochim. Biophys. Acta (BBA) – Proteins Proteomics 1804 (2010) 422–426.
[3] M. Calvaresi, M. Garavelli, A. Bottoni, Computational evidence for the catalytic mechanism of glutaminyl cyclase, DFT Invest. Proteins: Struct., Funct., Bioinformatics 73 (2008) 527–538.
[4] H.B. Dunford, Mechanisms of horseradish peroxidase and alpha-chymotrypsin, Prog. React. Kinet. Mech. 38 (2013) 119–129.
[5] A. Warshel, R.M. Weiss, An empirical valence bond approach for comparing reactions in solutions and in enzymes, J. Am. Chem. Soc. 102 (1980) 6218–6226.
[6] A.C.T. van Duin, S. Dasgupta, F. Lorant, W.A. Goddard, ReaxFF: a reactive force field for hydrocarbons, J. Phys. Chem. A 105 (2001) 9396–9409.
[7] M.W. van der Kamp, A.J. Mulholland, Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology, Biochemistry 52 (2013) 2708–2728.
[8] M.J.L. Mills, P.L.A. Popelier, Polarisable multipolar electrostatics from the machine learning method Kriging: an application to alanine, Theor. Chem. Acc. 131 (2012) 1137–1153.
[9] M.J.L. Mills, G.I. Hawe, C.M. Handley, P.L.A. Popelier, Unified approach to multipolar polarisation and charge transfer for ions: microhydrated Na+, PhysChemChemPhys 15 (2013) 18249–18261.
[10] S.M. Kandathil, T.L. Fletcher, Y. Yuan, J. Knowles, P.L.A. Popelier, Accuracy and tractability of a Kriging model of intramolecular polarizable multipolar electrostatics and its application to histidine, J. Comput. Chem. 34 (2013) 1850–1861.
[11] T.L. Fletcher, S.M. Kandathil, P.L.A. Popelier, The prediction of atomic kinetic energies from coordinates of surrounding atoms using Kriging machine learning, Theor. Chem. Acc. 133 (1499) (2014). 1491-1410.
[12] P.I. Nagy, B. Noszal, Theoretical study of the tautomeric/conformational equilibrium of aspartic acid zwitterions in aqueous solution, J. Phys. Chem. A 104 (2000) 6834–6843.
[13] M. Remko, D. Fitz, B.M. Rode, Effect of metal ions (Li+, Na+, K+, Mg2+, Ca2+, Ni2+, Cu2+ and Zn2+) and water coordination on the structure and properties of L-histidine and zwitterionic L-histidine, Amino Acids 39 (2010) 1309–1319.
[14] E. Deplazes, W. van Bronswijk, F. Zhu, L.D. Barron, S. Ma, L.A. Nafie, K.J. Jalkanen, A combined theoretical and experimental study of the structure and vibrational absorption, vibrational circular dichroism, Raman and Raman optical activity spectra of the L-histidine zwitterion, Theor. Chem. Acc. 119 (2008) 155–176.
[15] F. Weixin, K.R. Amareshwar, K. Rai, Z. Lu, Z. Lin, Structural stabilities of metallated histidines in gas phase and existence of gaseous zwitterionic histidine conformers, J. Mol. Struct. 895 (2009) 65–71.
[16] R.F.W. Bader, Atoms in Molecules. A Quantum Theory, Oxford Univ. Press, Oxford, Great Britain, 1990.
[17] P.L.A. Popelier, Atoms in Molecules. An Introduction, Pearson Education, London, Great Britain, 2000.
[18] P.L.A. Popelier, in: D.J. Wales (Ed.), Structure and Bonding. Intermolecular Forces and Clusters, vol. 115, Springer, Heidelberg, Germany, 2005, pp. 1–56.
[19] C.F. Matta, R.F.W. Bader, Atoms-in-molecules study of the genetically encoded amino acids. III. Bond and atomic properties and their correlations with experiment including mutation-induced changes in protein stability and genetic coding, Proteins: Struct., Funct. Genet. 52 (2003) 360–399.
[20] C.F. Matta, R.F.W. Bader, Atoms-in-molecules study of the genetically encoded amino acids. II. Computational study of molecular geometries proteins: structure, Funct. Genet. 48 (2002) 519–538.
[21] C.F. Matta, R.F.W. Bader, An atoms-in-molecules study of the genetically-encoded amino acids: I. Effects of conformation and of tautomerization on geometric, atomic, and bond properties, Proteins: Struct. Funct. Genet. 40 (2000) 310–329.
[22] S. Cardamone, T.J. Hughes, P.L.A. Popelier, Multipolar electrostatics, Phys. Chem. Chem. Phys. 16 (2014) 10367–10387.
[23] P.L.A. Popelier, New insights in atom-atom interactions for future drug design, Curr. Top. Med. Chem. 12 (2012) 1924–1934.
[24] C. Fonseca Guerra, J.W. Handgraaf, E.J. Baerends, F.M. Bickelhaupt, Voronoi deformation density (VDD) charges: assessment of the Mulliken, Bader, Hirshfeld, Weinhold, and VDD methods for charge analysis, J. Comp. Chem. 25 (2004) 189–210.
[25] P. Bultinck, C. Van Alsenoy, P.W. Ayers, R. Carbó-Dorca, Critical analysis and extension of the Hirshfeld atoms in molecules, J. Chem. Phys. 126 (2007) 144111.
[26] D.S. Kosov, P.L.A. Popelier, Atomic partitioning of molecular electrostatic potentials, J. Phys. Chem. A 104 (2000) 7339–7345.
[27] P.L.A. Popelier, in: G. Frenking, S. Shaik (Eds.), The Nature of the Chemical Bond Revisited, Wiley-VCH, 2014, pp. 271–308 (Chapter 8).
[28] N.O.J. Malcolm, P.L.A. Popelier, An algorithm to delineate and integrate topological basins in a three-dimensional quantum mechanical density function, J. Comp. Chem. 24 (2003) 1276–1282.
[29] N.O.J. Malcolm, P.L.A. Popelier, The full topology of the Laplacian of the electron density: scrutinising a physical basis for the VSEPR model, Faraday Discuss. 124 (2003) 353–363.
[30] P.L.A. Popelier, F.M. Aicken, Atomic properties of amino acids: computed atom types as a guide for future force field design, ChemPhysChem 4 (2003) 824–829.
[31] P.L.A. Popelier, É.A.G. Brémond, Geometrically faithful homeomorphisms between the electron density and the bare nuclear potential Int, J. Quant. Chem. 109 (2009) 2542–2553.
[32] M. Rafat, M. Devereux, P.L.A. Popelier, Rendering of quantum topological atoms and bonds, J. Mol. Graph. Modell. 24 (2005) 111–120.
[33] M. Rafat, P.L.A. Popelier, Visualisation and integration of quantum topological atoms by spatial discretisation into finite elements, J. Comput. Chem. 28 (2007) 2602–2617.
[34] Y. Yuan, M.J.L. Mills, P.L.A. Popelier, F. Jensen, Comprehensive analysis of energy minima of the twenty natural amino acids, J. Phys. Chem. A, in press, http://dx.doi.org/10.1021/jp503460m.
[35] T.J. Hughes, S.M. Kandathil, P.L.A. Popelier, Accurate prediction of polarised high order electrostatic interactions for hydrogen bonded complexes using the machine learning method Kriging, Spectrochim. Acta Part A (2014), http://dx.doi.org/10.1016/j.saa.2013.10.059.
[36] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G.A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H.P. Hratchian, A.F. Izmaylov, J. Bloino, G. Zheng, J.L. Sonnenberg, M.

Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. Montgomery, J. A., J.E. Peralta, F. Ogliaro, M. Bearpark, J.J. Heyd, E. Brothers, K.N. Kudin, V.N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, N. Rega, N.J. Millam, M. Klene, J.E. Knox, J.B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R.E. Stratmann, O. Yazyev, A.J. Austin, R. Cammi, C. Pomelli, J.W. Ochterski, R.L. Martin, K. Morokuma, V.G. Zakrzewski, G.A. Voth, P. Salvador, J.J. Dannenberg, S. Dapprich, A.D. Daniels, Ö. Farkas, J.B. Foresman, J.V. Ortiz, J. Cioslowski, D.J. Fox, Gaussian Inc., Wallingford, CT, 2009.

[37] F. Jensen, Polarization consistent basis sets. III. The importance of diffuse functions, J. Chem. Phys. 117 (2002) 9234–9240.

[38] T.A. Keith, AIMAll (Version 10.07.25), 2010, aim.tkgristmill.com.

[39] F.M. Aicken, P.L.A. Popelier, Atomic properties of selected biomolecules. Part 1. The interpretation of atomic integration errors, Can. J. Chem. 78 (2000) 415–426.