Quantum Mechanically Derived Biomolecular Force Fields

Alice Allen

Department of Physics
University of Cambridge

March 2019

This dissertation is submitted for the degree of
Doctor of Philosophy

Robinson College

Supervisor: Dr. Daniel Cole
Prof. Mike Payne

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Alice Allen
March 2019

Acknowledgements

I would like to thank my supervisor, Dr. Daniel Cole, for his advice and guidance throughout the project. Additionally, I would like to thank Dr. Michael Robertson for allowing the use of his files and scripts to assist with the torsional parameter part of this work and Joshua Horton for performing the MM simulations in Chapter 5 and testing the modified Seminario software. I would also like to thank Prof. Mike Payne for his help and provision of facilities and resources, and the EPSRC for funding this project.

Abstract

Molecular mechanics force fields are used to understand and predict a wide range of biological phenomena. However, current biomolecular force fields assume that parameters must be fit to the properties of small molecules and subsequently transferred to model large proteins. Here, we look to challenge this assumption and create a new class of QUantum mechanical BEspoke (QUBE) biomolecular force fields. QUBE is based around the use of atoms-in-molecule electron density partitioning to derive the non-bonded component of the force field. This thesis focusses on the derivation and validation of compatible bonded parameters that enable QUBE to be used in protein modelling.

Whilst parametrizing the bond and angle components of the new force fields, the inadequacy of current parametrization schemes became apparent. This led to the development of a new bond and angle parametrization method that relies on only the quantum mechanical Hessian of a molecule. The new method resulted in the accurate recreation of the normal modes for a set of small molecules, heterocyclic molecules, dipeptides and a large osmium containing complex. The new method had an overall error in the normal mode frequency recreation of 6.3%, which is below that of the popular force field OPLS (7.4%).

Torsional parameters were also calculated for our protein force field and the conformational preferences of peptides and proteins were subsequently tested. Comparable accuracy to standard transferable force fields was achieved for simulations of short peptides, and this was demonstrated by the simulations' J coupling errors, rotamer populations and backbone distributions. The J coupling errors remained at an acceptable level for protein simulations of ubiquitin and GB3, and two of the five proteins tested retained their experimental structure well during the MD simulations. In certain regions, particularly those with no clear secondary structure or a turn, three of the proteins exhibited some deviations from the experimental structure as the simulations progressed. However, given that this is the first generation of our QUBE force field, with future version envisaged, we view the results as promising.

Additionally, improvements to the electrostatic potential of system-specific small molecule force fields were investigated. A new method was developed to add off centre point charges. The extra charges led to a reduction in the error of an atom's electrostatic potential of 65.8%, as well as improvements to the free energy of hydration, for a benchmark set of molecules.

The methods and software developed in this thesis have the potential to improve the accuracy and accessibility of force field derivation, particularly for applications in biomolecular modelling.

Chapter 1

Introduction

Quantum mechanical (QM) calculations have been used to study a wide variety of systems and processes.[19,20,98,118] For example, calculations have been used to predict the structure of hydrogen at high pressures, to better understand the electronic and optical properties of photovoltaic materials, and to improve the design of heterogeneous catalysts.[64,87,90] However, systems consisting of more than a few hundred atoms are considered outside the reach of traditional QM methods.[107] This is because traditional QM methods have a computational time that scales cubically, or even higher, with the number of atoms.[107] The development of linear scaling density functional theory (DFT) has allowed systems that were once considered to be beyond the capability of QM calculations to be routinely studied.[19,71] Protein complexes are an example of this, with linear scaling DFT used to advance our understanding of the function and structure of such systems.[19] Although linear scaling DFT has increased the size of the systems that can be studied, limitations on the timescales that can be investigated still remain.

Classical molecular mechanics (MM) simulations are the natural alternative to DFT if longer simulations, on the timescale of nanoseconds, as opposed to picoseconds, are required.[67,112] However, with the increase in the timescale that can be studied by moving to MM methods there is a reduction in the accuracy of the simulation. This is primarily due to the approximations made in the functional form and the parameters used to describe interactions in and between molecules.[84,121] The potentials that characterize these molecular interactions are known as force fields. The AMBER force field is one of a number of force fields (e.g. CHARMM, GROMOS and OPLS) which began development in the 1980's.[54,95,125] The majority of force fields use a similar functional form:[95]

$$V = V_{stretching} + V_{bending} + V_{torsional} + V_{LJ} + V_{elec} \tag{1.1}$$

with components of the potential energy function describing the bonded interactions (the stretching, bending and torsional terms) and the non-bonded terms (the Lennard-Jones interactions and electrostatic effects).

The Lennard-Jones potential includes an attractive component, which is due to the dispersion effects that occur as a result of fluctuation in the electron density, and a repulsive component, which is present to prevent atoms coming too close together.[95] In most force fields, the electrostatic interactions are described with atom centered points charges interacting via Coulomb's law.[95] The bonded components of a force field are generally described by harmonic potentials.[54,125] The torsional component is the exception to this, and instead is usually an anharmonic function composed of trigonometric terms.[95]

Both quantum mechanical and experimental data is used in current force field parameterization, with more accurate data and better fitting techniques, such as applying regularization and weighting functions, employed to calculate parameters for new generations of force fields.[54,99,121,125] Improvements in parameters are increasingly being carried out with systematic and automatic techniques.[11,15,70,84,114,121] For example, the ForceBalance program can automatically optimize a set of a force field parameters to recreate a set of user specified input data, whilst the Force Field ToolKit (ffTK) can parametises the bond, angle, dihedral and charge terms of a small molecule using data supplied from QM calculation.[84,121] Changes to force field parameters can also be coupled with simultaneous changes to the functional form.[11,15,36,70,84,114,121] However, current work often focuses on small molecule parametrization, and not the development of new biomolecular force fields, and frequently neglects reparametrization of the bond, angle and van der Waals components of the force field.[84,114] Additionally, biomolecular force fields aim to be transferable to multiple systems, and this need for transferability can restrict accuracy as local polarization effects are not included in non-polarizable force fields.[61,99] Although transferable force fields have been successful for multiple applications, they can fail to be sufficiently accurate for certain practical considerations.[22,85] Using free energy calculations for computer aided drug design has been investigated in numerous studies, however it has been shown that free energy calculations can fail to recreate experimentally determined trends.[22,85,101] For example, in Ref 101, the experimental binding affinity of a set of ligands to plasmin, an enzyme involved in blood clotting, was compared to the results from free energy calculations. No correlation was observed between the experimental and computational results, demonstrating the need to further improve the methods and force fields used in the free energy calculations.

In traditional force fields, every atom in a molecule is assigned a type based on its atomic number, bonding and chemical environment.[113] The atom's type then dictates the parameters that are used to model the intermolecular and intramolecular interactions.[113] The force field terms for each atom type are stored as a library and are parametrized with experimental and computational data from a set of small molecules. There is therefore an assumption made that the parameters found for a set of small molecules can be used for a wide range of systems.[83] However, it has been shown that the non-bonded parameters in a force field, the electrostatic and Lennard-Jones terms, are dependent on additional effects that are not included in the atom's type, such as local polarization effects.[60,111] This means that for certain systems the charge distribution is not adequately reproduced by standard force fields that use atom types.[68] If the charges used in the MM simulation were specifically fit for the system under study, and not taken from atom types, improvements in the charge distribution may be expected. The use of system specific charges for small molecule force fields has already been investigated by multiple groups, however system specific biomolecular force fields have not been previously developed.[28,83]

In Ref. 22 the use of linear scaling DFT for system specific force field parameters began to be investigated. The charges for each atom were found using density derived electrostatic and chemical (DDEC) electron density partitioning.[68] DDEC divides the electron density of a molecule between individual atoms using a weighting scheme. The Lennard-Jones parameters used in Ref. 22 were found with the equations of Tkatchenko and Scheffler, which are commonly employed to add dispersion effects to DFT calculations and relate the electron density to the van der Waals parameters.[111] The success of this method was evidenced by the high level of agreement between the experimental and MM liquid properties for a set of small molecules.[22] Experimental liquid densities, heats of vaporization, and free energies of hydration of around 40 molecules were used to benchmark the methods, which resulted in mean unsigned errors of 0.014 g/cm^3, 0.65 kcal/mol and 1.03 kcal/mol respectively.[22] Additional simulations demonstrated the feasibility and advantages of deriving bespoke parameters for a protein-ligand complex. The computed relative binding free energy of indole and benzofuran to the lysozyme protein using the environment-specific force fields (-0.4 kcal/mol) was in excellent agreement with experiment (-0.6 kcal/mol), and was substantially more accurate than the values calculated with standard force fields (-2.4 kcal/mol).

However, clear areas of improvements remained to be made to the force field used in Ref. 22. The most pressing changes required were to the bonded parameters, which were taken from the OPLS force field. As bonded parameters are dependent on the non-bonded parameters, reparametrization should be performed when non-bonded terms are updated.[116] Reparametrization of the bonded terms, and testing the performance of our new protein force field, QUBE (QUantum BEspoke), forms the focus of this thesis. The QUBE force field uses the non-bonded parametrization methodologies developed in Ref. 22 which results in system specific terms with transferable bonded terms that are parametrized using QM calculations of dipeptides. As the non-bonded terms are specific to the system of interest, QUBE will therefore include polarization effects that are not present in traditional transferable force fields, and the expectation is that this may lead to an improvement in the accuracy of the force field.

The bond and angle parameters control the fluctuations of the bonds and angles around their equilibrium positions. During the development of our force field, it became apparent that a bond and angle parameterization approach did not exist that relied only on QM data and accurately reproduced the normal modes of a molecule.[11,24,72,84,103] We have altered an existing approach, the Seminario method, to develop a scheme that results in accurate normal mode prediction and relies only on the QM Hessian matrix of a molecule.[1,103] Our new approach was then tested on 70 molecules, including a set of dipeptides, which provided the bond and angle parameters for use in our biomolecular force field.

With new bond and angle terms derived, the torsional parameters remained the only component of our force field that had not been reparameterized. New torsional parameters were fitted using QM dihedral energy scans, and the accuracy of these parameters was verified through simulations of peptides and proteins.

The final section of this thesis focuses on improving the electrostatic potential of a molecule. The addition of off center point charges is investigated as a method to correct for anisotropy in the electrostatic potential, with a new, automated approach implemented in the linear scaling DFT code ONETEP. This work not only shows the improvements that can be made to the electrostatic potential, but also demonstrates how systematic improvements can easily be made to our force field, due to the workflow developed and the use of only QM data in the parametrization process.

Chapter 2

Theory

In this chapter, the necessary theory required for the rest of the thesis is presented. The chapter begins with a brief overview of proteins and peptides, before moving on to a description of the drug discovery process. The Hartree-Fock approximation and density functional theory are subsequently covered, and the final section details the components and non-bonded parametrization of the QUBE force field.

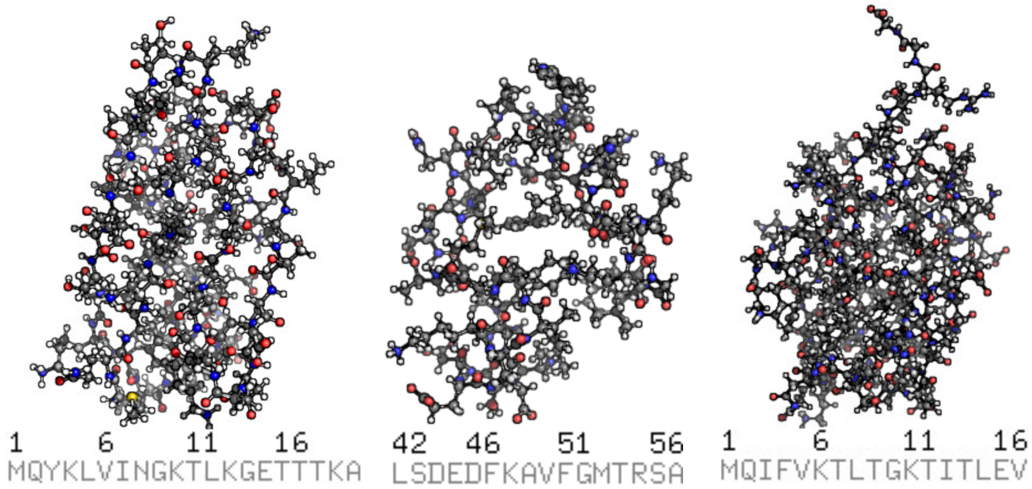


Fig. 2.1: An example of three protein's (1P7E, 2F4K, 1UBQ) structure with a section of their amino acid sequence. One letter codes are used to identify each residue present. Carbon atoms are shown in grey, hydrogen in white, oxygen in red and nitrogen in blue.

2.1 Proteins

Proteins are an essential component of human life and are involved in a vast number of processes.[37] Many medicines work by changing the behaviour of proteins in our body, and therefore further developing our understanding of proteins is essential for future developments in medicine.

Proteins consist of a linear sequence of amino acids and are large biological molecules containing thousands of atoms. Three proteins are shown in Fig.2.1, with a section of their amino acid sequence also shown.[13] An amino acid is a compound consisting of a central carbon atom attached to a sidechain group, an amine group and a carboxyl group.[37] Twenty different types of amino acids occur naturally and are distinguished from one another by their side chain groups.[37] The amino acids with the simplest forms are glycine, with a sidechain consisting of just one hydrogen atom, and alanine, with a methyl sidechain group, see Fig. 2.2.[37] Different amino acids have varying properties and they can be classified according to the characteristics they possess, i.e. they can be charged or uncharged, polar or nonpolar, and hydrophobic or hydrophilic.

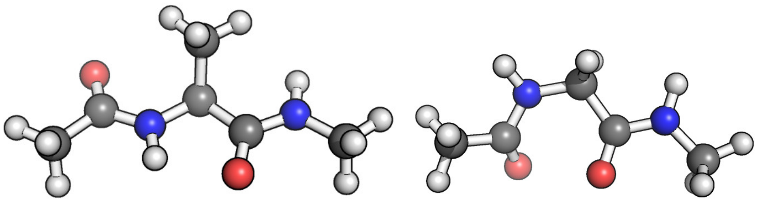


Fig. 2.2: Alanine and glycine dipeptides. Carbon atoms are shown in grey, hydrogen in white, oxygen in red and nitrogen in blue.

2.1.1 Peptides

Short sequences of amino acids are not classified as proteins but instead are referred to as peptides, and consist of amino acids joined together by a peptide bond.[13] Dipeptides are used in this work for parameterising and testing our force fields. Throughout this thesis, a dipeptide refers to an amino acid (X) blocked with an acetyl (CH_3CO) and a N-methyl group (Ace-X-NMe).[99]

2.1.2 Secondary Structure

Peptides can fold into multiple structures. These structures can be identified by the backbone dihedral angles ϕ and ψ.[37] The atom names and main dihedral angles of an amino acid are

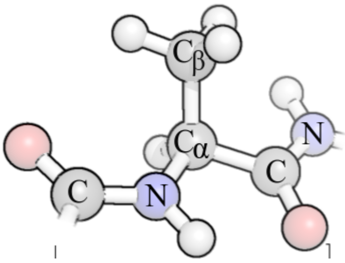


Fig. 2.3: An alanine amino acid showing the naming convention for atoms used. The main dihedral angles reparametrized are ϕ (C-N-C$_\alpha$-C), ϕ' (C-N-C$_\alpha$-C$_\beta$) ψ (N-C$_\alpha$-C-N) , ψ' (C$_\beta$-C$_\alpha$-C-N), χ_1 (N-C$_\alpha$-C$_\beta$-X$_\gamma$),χ_1' (C-C$_\alpha$-C$_\beta$-X$_\gamma$) and χ_2 (C$_\alpha$-C$_\beta$-X$_\gamma$-Y$_\delta$) X$_\gamma$ and Y$_\gamma$ correspond to heavy atoms that are not shown). The dihedral angle is the angle between the plane containing the first three atoms and the plane containing the last three atoms.

described in Fig.2.3. The main conformations seen in proteins are usually illustrated by the ψ/ϕ distribution, or Ramachandran plot, in Fig 2.4.

Two of the most predominant structures found in proteins are α helices and β sheets.[13] These structures can be easily identified in a protein using a ribbon diagram. An example is given in Fig 2.5 with the α and β regions labeled. A turn is also identified, this is a region where the protein backbone changes its overall direction.[37]

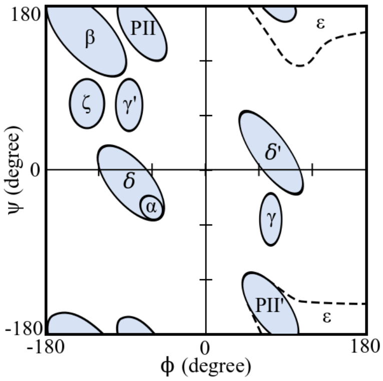


Fig. 2.4: The major conformations observed in protein structures as proposed in Ref. 43.

2.1.3 Computer Aided Drug Discovery

Free energy calculations are performed for a range of applications, including computer aided drug discovery, as they can provide information about properties of a system such as solubility or protein-ligand binding.[38] The absolute free energy of a system is considered expensive to compute, and instead the relative free energy of different systems is generally used.[38] A number of methods are available for calculating this relative free energy difference, such as free energy perturbation and thermodynamic integration.[38] The accuracy of free

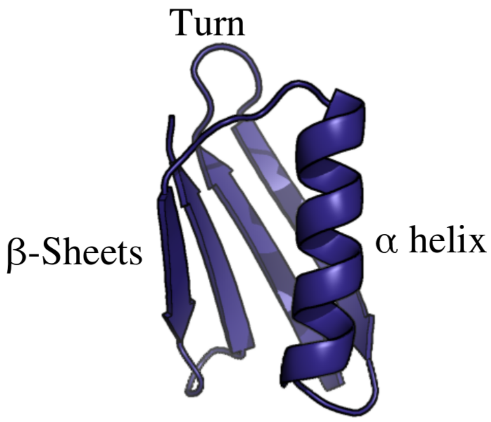


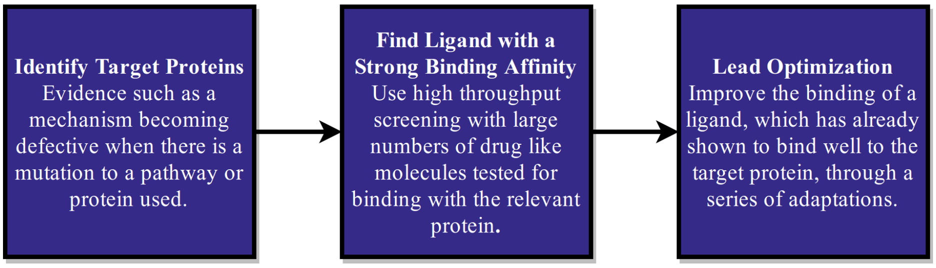
Fig. 2.5: A ribbon diagram of GB3(1P7E) with the secondary structures shown.

energy calculations is dependent on the sampling of the configurational space of the systems and the force field used.[16,17] While not the main focus of this thesis, one of the motivations for creating our new force field is to improve free energy calculations, particularly those used in computer-aided drug discovery. A brief overview of the drug discovery process is therefore given in this section.

The drug discovery process consists of a number of stages, with the workflow illustrated in Fig. 2.6. The first stage in the drug development process consists of identifying a target protein whose function influences the progression of a disease.[48] Various types of evidence can be used in identifying this target protein and the mechanisms it is responsible for in the cell.[20]

Once the relevant protein target has been established, the next phase is finding a ligand that has a strong binding affinity to the protein and modulates its activity. This is commonly carried out using high throughput screening (HTS) techniques and involves adding large numbers of drug like molecules, available in libraries of compounds, to grids with small wells containing the relevant protein.[47,48] The ability of the smaller molecules to bind to the protein is then measured using an observable such as the fluorescence. Further measurements can then be performed on molecules that appear desirable, with confirmation of results and synthesis of analogs carried out.[48]

From these results, lead compounds can be identified, these are molecules that are promising candidates for use as a potential drug molecule.[48] In the lead optimization component of the drug discovery process, a molecule that has been shown to bind well to a target protein is adapted in order to improve its binding affinity.[17,48] This is carried out using experimental methods and can be a long and expensive task.[48] Therefore, being able to perform computational free energy calculations, with the relative free energy of binding for a ligand indicating which molecules bind well to the protein, could lead to better efficiency and lower costs. The



Identify Target Proteins
Evidence such as a mechanism becoming defective when there is a mutation to a pathway or protein used.

Find Ligand with a Strong Binding Affinity
Use high throughput screening with large numbers of drug like molecules tested for binding with the relevant protein.

Lead Optimization
Improve the binding of a ligand, which has already shown to bind well to the target protein, through a series of adaptations.

Fig. 2.6: The initial stages of the drug discovery process.[17]

accuracy of these calculations depends on both improving ligand and biomolecular force fields.

2.2 Quantum Mechanical Methods

In order to improve free energy calculations, and molecular mechanical simulations generally, we use QM methods to parametrize our force field. The theory behind the QM approaches is given in this section.

Two approximations that can be used to solve the Schrödinger equation are discussed. Both the Hartree-Fock approximation and DFT look to solve the many-electron Hamiltonian by reducing the problem to a set of one-body equations. The Hamiltonian of a many-electron system is given by.[78]

$$H = -\sum_k \frac{\hbar}{2m}\nabla_k^2 + \sum_k V(r_k) + \frac{1}{8\pi\varepsilon_0}\sum_{k\neq k'} \frac{e^2}{|\mathbf{r_k} - \mathbf{r'_k}|} \tag{2.1}$$

Where m is the mass of the electron, $\mathbf{r_k}$ is the position of electron k, e is the charge on an electron and $V(r_k)$ is the electron-ion interaction. The first term corresponds to the kinetic energy of the electrons and the final term is the electron-electron interaction.[82] Equation 2.1, relies on the Born-Oppenheimer method to decouple the electron and ion motion.[100] Throughout this section, atomic units are used with $\hbar = m = e = 1$ and spin labels are omitted.

2.2.1 Hartree-Fock Approximation

The difficulty in solving a many-body quantum system arises from the the electron-electron interactions.[82] If it was not for this interaction, the system could be simply treated as a set of independent particles in an external potential.[78] The Hartree and Hartree-Fock approximations assume a form for the wavefunction which allows the problem to be reduced to a series of one-electron equations.[78]

The original Hartree approach approximated the wavefunction as $\Phi(\mathbf{r_1},...,\mathbf{r_N}) = \phi_1(r_1)\phi_2(r_2)...\phi_N(r_N)$.[78] However, this wavefunction does not obey Pauli's exclusion principle which requires that an exchange of electrons causes a change in sign of the wavefunction.[78] The Hartree-Fock approximation uses the Slater determinant to define the wavefunction and this results in exchange being described exactly:[33]

$$\Phi = (N!)^{-1/2} \begin{vmatrix} \phi_1(\mathbf{r_1}) & \cdots & \phi_N(\mathbf{r_1}) \\ \vdots & & \vdots \\ \phi_1(\mathbf{r_N}) & \cdots & \phi_N(\mathbf{r_N}) \end{vmatrix}$$

Exchanging electrons corresponds to swapping columns in the Slater determinant, which causes a change in sign of the determinant and, hence, the wavefunction as required.[78] This wavefunction can be used to calculate the expectation value of the energy, $E = \langle\Phi|H|\Phi\rangle$.[82] The Hartree-Fock equations can then be derived using the variational principle, with the minimum of $E = \langle\Phi|H|\Phi\rangle$ with respect to the wavefunction found, this minimum E will correspond to the ground state energy in the Hartree-Fock approximation.[82] The difference between the ground state energy calulated via the Hartree-Fock approximation and the exact ground state energy is usually referred to as the correlation energy.[82]

The Hartree approximation, which does not obey Pauli's exclusion principle, can be solved directly.[82] The Hartree-Fock approximation generally cannot, and a basis must be introduced to describe the orbitals. The resultant scaling of the computational cost is approximately N_{basis}^4, where N_{basis} is the number of basis functions (which are generally atom centered Gaussian functions in quantum chemistry methods).[82,97]

2.2.2 Density Functional Theory

Density functional theory maps the many-body problem onto a non-interacting particle problem, and writes the energy of the system as a functional of the single particle electron density.[82] The theorems required to achieve this are briefly described below.

The Hohenberg–Kohn theorems establish a relationship between a system's ground state energy and its groundstate electron density.[82] The first thereom states that for a many-body system the total energy is a unique functional of the electron density.[82]

$$E[n(\mathbf{r})] = \int V_{ext}(\mathbf{r})n(\mathbf{r})d^3r + E_{II} + F[n(\mathbf{r})] \geq E_{GS} \tag{2.2}$$

where $n(\mathbf{r})$ is the electron density, $V_{ext}(\mathbf{r})$ is the external potential, E_{II} is the interaction between the nuclei, $F[n(\mathbf{r})]$ is a universal functional which does not refer to a specific system or external potential, and E_{GS} is the ground state energy.[82] The second theorem states that the electron density that minimizes the energy functional is the ground state electron density.[82]

$$E[n_{GS}(\mathbf{r})] = E_{GS} \tag{2.3}$$

The Hohenberg–Kohn theorems allow the energy to be expressed in terms of the electron density and establishes the existence of the functional F. However, the form of this functional is unknown and it cannot be easily approximated.[82] The Kohn-Sham approach rewrites the unknown functional F as $F[n(\mathbf{r})] = T_S[n(\mathbf{r})] + E_{Hartree}[n(\mathbf{r})] + E_{xc}[n(\mathbf{r})]$, where T_S is the independent particle kinetic energy which can be expressed in terms of the Kohn-Sham orbitals. This results in the following expression for the energy of the system:[65]

$$E[n(\mathbf{r})] = T_S[n(\mathbf{r})] + \int n(\mathbf{r})[V_{ext}(\mathbf{r}) + \frac{1}{2}\Phi(\mathbf{r})]d^3r + E_{II} + E_{xc}[n(\mathbf{r})] \tag{2.4}$$

where $E_{xc}[n(\mathbf{r})]$ is the exchange correlation functional and $\Phi(\mathbf{r})$ is the Hartree potential given by:

$$\Phi(\mathbf{r}) = \int \frac{n(\mathbf{r}')}{\mid \mathbf{r}-\mathbf{r}' \mid} d^3r' \tag{2.5}$$

This is equivalent to an independent particle moving in the potential:[91]

$$V(\mathbf{r}) = V_{ext}(\mathbf{r}) + \Phi(\mathbf{r}) + \frac{\delta E_{xc}[n(\mathbf{r})]}{\delta n(\mathbf{r})} \tag{2.6}$$

The solution to this independent particle problem can be found by solving the Schrödinger equation for non-interacting particles:[52]

$$[-\frac{1}{2}\nabla^2 + V(\mathbf{r})]\psi_i(\mathbf{r}) = \varepsilon_i \psi_i(\mathbf{r}) \tag{2.7}$$

where ψ_i is the Kohn-Sham orbital i and ε_i is the Kohn-Sham eigenvalue. With the electron density being given by:

$$n(\mathbf{r}) = \sum_{i=1}^{N} \mid \psi_i(\mathbf{r}) \mid^2 \tag{2.8}$$

The potential used in the Kohn-Sham equations is a function of the electron density. Therefore, equation 2.7 must be solved self consistently so that the Kohn-Sham orbitals generate an electron density that in turn reproduces the Hartree and exchange-correlation potentials used.[82] As well as this, the Kohn-Sham eigenstates must be orthogonalised to satisfy:[82]

$$\langle \psi_i \mid \psi_j \rangle = \int \psi_i(\mathbf{r})\psi_j(\mathbf{r})d\mathbf{r} = \delta_{ij} \tag{2.9}$$

Whilst both the Hartree potential and external potential given in equation 2.6 are well defined, the exchange-correlation term is not.[82] The Kohn-Sham equations would give the true ground state energy if the exact form of the exchange-correlation functional was known.[52] Whereas we previously has an unknown functional $F[n(\mathbf{r})]$, the reformulation of the Hohenberg-Kohn theorems using the Kohn-Sham approach has resulted in a new unknown functional $E_{xc}[n(\mathbf{r})]$. However, whilst $F[n(\mathbf{r})]$ is very difficult to approximate, $E_{xc}[n(\mathbf{r})]$ can be calculated with remarkably simple approximations which are surprisingly accurate for many physical and chemical properties. One such approximation is the local density approximation (LDA). This states that the energy density at a given point can be approximated by the energy density of a homogeneous electron gas with the same electron density:[82]

$$E_{xc}[n(\mathbf{r})] = \int \varepsilon_{xc}^{hom}[n(\mathbf{r})]n(\mathbf{r})d^3r \tag{2.10}$$

where $\varepsilon_{xc}^{hom}[n(\mathbf{r})]$ is the electron-correlation per electron for a homogeneous electron gas with electron density $n(\mathbf{r})$.[52] The LDA does not account for nearby inhomogeneities in electron density, yet despite this works well for many properties in a large number of systems (although it does have a number of well documented flaws).[91] An improvement can be made to this approximation by additionally considering the gradient of the electron density at each point i.e. $E_{xc}[n(\mathbf{r})] = \int \varepsilon_{xc}(n(\mathbf{r}), |\nabla n(\mathbf{r})|, ...)n(\mathbf{r})d^3r.$ This is known as the generalized gradient approximation (GGA) and is employed in the PBE functional, which is used in this work.[65]

An alternative type of functional that can be used is a hybrid functional. One respect in which the Hartree-Fock approximation differs from DFT is that it gives an explicit form for the exchange energy.[82] Hybrid functionals use a fraction of the exchange energy from Hartree-Fock theory to improve the exchange-correlation functional.[82]

As the number of atoms in a system increases, the number of Kohn-Sham eigenstates increases linearly. The orthonormality of these eigenstates must be enforced and this results in conventional Kohn-Sham DFT methods having approximately $\mathrm{O(N^3)}$ scaling.[8] Although this is better than the Hartree-Fock approach, for large systems the computational cost of DFT codes with $\mathrm{O(N^3)}$ scaling becomes overly expensive.[97] Linear scaling DFT reformulates the underlying theory and this results in an O(N) scaling method.[97] This allows QM calculations to be carried out on systems consisting of thousands of atoms.[19]

2.2.3 Linear Scaling Density Functional Theory

There are a number of different approaches to achieve linear scaling in DFT.[97] In this section, we will just discuss the density matrix formulation used in ONETEP. Further details of alternative methods can be found in Ref. 97.

ONETEP achieves linear scaling by using a density matrix approach and then exploiting the decay properties of the density matrix.[40] The density matrix is defined as:[107]

$$\rho(\mathbf{r},\mathbf{r}') = \sum_i f_i \psi_i(\mathbf{r}) \psi_i^*(\mathbf{r}') \tag{2.11}$$

where f_i gives the occupancy of the Kohn-Sham eigenstate $\psi_i(\mathbf{r})$.

By using density matrices, the orthonormality of the eigenstates can be enforced by imposing idempotency on the density matrix:[107]

$$\rho^2(\mathbf{r},\mathbf{r}') = \int \rho(\mathbf{r},\mathbf{r}'')\rho(\mathbf{r}'',\mathbf{r}')d^3r'' = \int \sum_i f_i \psi_i(\mathbf{r})\psi_i^*(\mathbf{r}'') \sum_j f_j \psi_j(\mathbf{r}')\psi_j^*(\mathbf{r}'')d^3r'' = \rho(\mathbf{r},\mathbf{r}') \tag{2.12}$$

An idempotent matrix is a matrix, M, which obeys the relationship $\mathbf{M} = \mathbf{MM}$. This characteristic is desirable as ensuring orthonormality of the eigenstates by alternative methods leads to less favourable scaling properties.[8]

The charge density and energy expressed in terms of a density matrix are given by the following equations:[107]

$$n(\mathbf{r}) = 2\rho(\mathbf{r},\mathbf{r}) = 2\sum_i f_i \mid \psi_i(\mathbf{r}) \mid^2 \tag{2.13}$$

where the factor of 2 is due to spin degeneracy.[107]

$$E[n(\mathbf{r})] = -\int [\nabla^2_{\mathbf{r}'}\rho(\mathbf{r},\mathbf{r}')]_{\mathbf{r}'=\mathbf{r}} d^3r + \int V_{ext}(\mathbf{r})n(\mathbf{r})d^3r + E_{Hartree}[n(\mathbf{r})] + E_{xc}[n(\mathbf{r})] \tag{2.14}$$

The minimum of equation 2.14, with respect to the density matrix, gives the ground state total energy of the system. Additionally, the density matrix must be idempotent, as previously discussed, and describe the correct number of electrons:[107]

$$2\int \rho(\mathbf{r},\mathbf{r})d^3r = N_e \tag{2.15}$$

Although the density matrix approach improves scaling, as the orthonormality of eigenstates no longer has to be enforced, the scaling still remains as $O(N^2)$.[40] However, by using the decay properties of the density matrix linear scaling can be achieved. The decay properties can be expressed as:[107]

$$\rho(\mathbf{r},\mathbf{r}') \backsim \exp(-\gamma \mid \mathbf{r}-\mathbf{r}' \mid) \tag{2.16}$$

where γ is a constant which depends on the band gap of the material.

Therefore, as $\mid \mathbf{r}-\mathbf{r}' \mid$ tends towards infinity the density matrix will tend towards zero.[107] This decay can be explained using Walter Kohn's theory of nearsightedness, which states that changes at large distances from a region have negligible effects on that region's properties.[19] In the density matrix form, regions that have a large spatial separation correspond to elements far off the diagonal of the matrix. Due to the decay properties mentioned, these elements can be set to zero and this results in a sparse band diagonal matrix.[19]

The nearsightedness properties are also exploited with the basis set that is used in ONETEP. Non-orthogonal generalized Wannier functions (NGWFs) are used which are atom-centered, localized and non-orthogonal. The NGWFs are optimized in situ when calculating the minimum energy in eq. 2.14.[41] This is to prevent errors introduced due to the small basis set used and thus gives near-complete basis set accuracy using a minimal basis.[107] The density matrix represented in terms of NGWFs is:[42]

$$\rho(\mathbf{r},\mathbf{r}') = \sum_{\alpha\beta} \phi_\alpha(\mathbf{r}) K^{\alpha\beta} \phi^*_\beta(\mathbf{r}') \tag{2.17}$$

where $\phi_\alpha(\mathbf{r})$ is a NGWF basis function and $K^{\alpha\beta}$ is the density kernel.[40]

In ONETEP, the NGWFs are localised to a spherical region and are expanded in terms of a set of functions known as periodic cardinal sine (PSINC) functions.[40,41] By ensuring the NGWFs are set to zero outside an atom centered spherical region, the density matrix is truncated further. The reformulation of the Kohn-Sham equations into a density matrix format, along with the truncation of the density matrix and the localization of the NGWFS, allow linear scaling to be achieved.

2.3 Force Fields

Density functional theory using a density matrix approach allows electronic structure calculations to be carried out for large biological systems.[19] However, DFT cannot provide us with information about biological process which take place on timescales longer than picoseconds due to the computational expense.[19] MM force fields can be used to simulate timescales up to milliseconds, but are less accurate than DFT.[67] Therefore, to retain some of the accuracy of DFT calculations, whilst being able to access longer timescales, we aim to use the electron density from DFT calculations to parametrize the non-bonded components of a classical biomolecular force fields for use in molecular dynamics simulations. The components of a force field and parametrization approach used for our force field is described in this section.

2.3.1 Components of a Force Field

A force field describes the potential energy function used in a MM simulation. There is some variation in the functional form used, such as the AMOEBA force field which alters the electrostatic component to account explicitly for polarization effects, however generally most biochemical force fields share a similar functional form.[96] The components of the OPLS force field, and the QUBE force field developed in this thesis, are as follows:[115]

$$V = V_{stretching} + V_{bending} + V_{torsional} + V_{vdW} + V_{elec} \tag{2.18}$$

The first three components describe covalent interactions between atoms.

- Bond stretching

$$V_{stretching}(r) = \sum_{bonds} k_b(b-b_0)^2 \tag{2.19}$$

  where k_b is the force constant, b is the bond length and b_0 is the equilibrium length.[95] The bond stretching component is therefore approximated by a simple harmonic function.

- Bond bending

$$V_{bending}(r) = \sum_{angles} k_\theta(\theta-\theta_0)^2 \tag{2.20}$$

  where k_θ is the force constant, θ is the angle between a triplet of bonded atoms and θ_0 is the equilibrium angle.[95] The form is again approximated by the harmonic function.

- Bond torsion

$$V_{torsional}(r) = \sum_i \frac{V_1^i}{2}(1+cos(\phi)) + \frac{V_2^i}{2}(1+cos(2\phi)) + \frac{V_3^i}{2}(1+cos(3\phi)) + \frac{V_4^i}{2}(1+cos(4\phi)) \tag{2.21}$$

  where V_1, V_2, V_3, V_4 are the constant parameters and ϕ is the dihedral angle.[99] The dihedral angle is defined by a set of four atoms.

The non-bonded components are:

- Electrostatic

$$V_{elec}(r) = \sum_{\substack{nonbonded \\ pairs}} \left[\frac{q_i q_j}{r_{ij}} \right] \tag{2.22}$$

  where q is usually the atom centered point charge and r_{ij} is the distance between atoms i and j.[108] The form comes from Coulomb's law and describes the classical electrostatic interaction between charges.

- Lennard-Jones (LJ)

$$V_{LJ}(r) = \sum_{\substack{nonbonded \\ pairs}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right] \tag{2.23}$$

  where A_{ij} and B_{ij} are parameters that are fit to QM or experimental data. The first part corresponds to the repulsive component of the force on the atoms due to Pauli's exclusion principle and the second part is the van der Waals/dispersion term. This van der Waals term is due to spontaneous fluctuations in the electron density.[108] This can be described by a series of terms with decreasing powers of r. However, it tends to be truncated to just the first r_{ij}^{-6} term.[109]

Force fields consist of a library of parameters, with parameters for unknown molecules fit via analogy with the database. They are used for a wide variety of systems that may differ significantly from the original set of molecules used to parametrize the force field. The parametrization approaches used in traditional force fields use both QM and experimental data.[54,96,99] For example, atom charges tend to be parametrized using QM electrostatic information whilst LJ terms can be fit to reproduce experimental liquid properties.[54,96] Further details of bonded parametrization methods are given in the relevant chapters of this thesis.

2.3.2 Parameterizing Non-bonded Components

We are looking to move away from using a library of predetermined parameters to describe our force field and instead derive non-bonded terms that are specific to the system of interest. To find new parameters for the QUBE environment specific force field a number of stages are carried out. These stages are shown in Figure 2.7 and consist of three main steps:

- A DFT calculation is performed using ONETEP and this results in the electron density of the whole molecule.

- A partitioning scheme is then used to assign the electron density of the molecule to individual atoms.

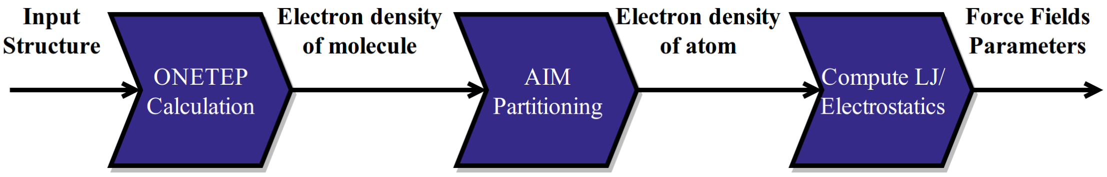


Fig. 2.7: The workflow for deriving new parameters for the environment specific force field.

- From the partitioned electron density, new force field parameters can be found using relationships between the non-bonded terms and the electron density.

Further details of the parametrization process for the QUBE force field are now outlined.

Modelling Induction within the Force Field

Induction effects are due to the distortion of the electron density by nearby molecules.[108] In our force field, the effects of solvents can be included by changing the dielectric constant of the implicit solvent used for the DFT calculation. The implicit solvent model used in ONETEP is a minimal parameter model which defines a cavity around the molecule and treats the solvent as a dielectric continuum surrounding this cavity.[29,30] The use of an implicit solvent is important as MM calculations will often be carried out in a water environment and the solvent will polarize the molecule. However, deriving the charges for use in a non-polarizable force field by using an implicit solvent model with a dielectric constant $\varepsilon = 80$ will result in overly polarized charges. This is because the energetic contributions required to distort the wavefunction are not included.[22] This can be corrected by using a dielectric constant less than $\varepsilon = 80$ or by setting the charges on the atoms to be halfway between their values in the vacuum and condensed phase.[22,24]

The approach used in Ref. 22 was to set the dielectric constant to $\varepsilon = 4$. This was justified by accurate free energy of hydration calculations when $\varepsilon = 4$ was used to parametrize the non-bonded terms, and the following arguments. The free energy of hydration is the difference in energy between the molecule in vacuum and the molecule in water and can be expressed quantum mechanically as:

$$\Delta G_{exp} = \langle \Psi' | \hat{H} | \Psi' \rangle - \langle \Psi_0 | \hat{H}_0 | \Psi_0 \rangle \tag{2.24}$$

where $\hat{H}$ is the Hamiltonian in the DFT implicit solvent model, $\hat{H}_0$ is the DFT Hamiltonian in the gas phase, Ψ' is the wavefunction in its polarized state in water and Ψ_0 is the wavefunction in the vacuum ground state.

Inconsistencies in the calculation of ΔG_{exp} arise due to two different wavefunctions appearing in the equation, Ψ_0 and Ψ'.[22] These two wavefunctions correspond to different electron densities and consequently would correspond to different force field parameters. Therefore, in a fixed charge force field it is generally not possible to model this change in wavefunction. However, there must exist an intermediate wavefunction (Ψ) such that:[22]

$$\Delta G_{exp} = \underbrace{(\langle\Psi'|\hat{H}|\Psi'\rangle - \langle\Psi|\hat{H}|\Psi\rangle)}_{\gamma_0} + (\langle\Psi|\hat{H}|\Psi\rangle - \langle\Psi|\hat{H}_0|\Psi\rangle) + \underbrace{(\langle\Psi|\hat{H}_0|\Psi\rangle - \langle\Psi_0|\hat{H}_0|\Psi_0\rangle)}_{\gamma_1} \tag{2.25}$$

$$\gamma_1 + \gamma_0 = \langle\Psi|\hat{H}_0|\Psi\rangle - \langle\Psi_0|\hat{H}_0|\Psi_0\rangle + \langle\Psi'|\hat{H}|\Psi'\rangle - \langle\Psi|\hat{H}|\Psi\rangle = 0 \tag{2.26}$$

This would correspond to a 'half-polarized' wavefunction and the use of this wavefunction would allow the free energy to be calculated with only one wavefunction:[22]

$$\Delta G_{exp} = \langle\Psi|\hat{H}|\Psi\rangle - \langle\Psi|\hat{H}_0|\Psi\rangle \tag{2.27}$$

It has been shown that for small molecules a dielectric value of $\varepsilon = 4$ in the inhomogenous Poisson equation produces the wavefunction Ψ with $\gamma_0 + \gamma_1 \approx 0$.[22]

The use of a 'half-polarized' wavefunction is based on the assumption that the wavefunction that minimizes the difference between the experimental and MM free energy of hydration is the one to use for parametrizing force fields. It has been shown in Ref. 22 that this assumption results in accurate liquid properties for small molecules. However, as the dielectric constant in that work was chosen to minimize the error in the free energy of hydration, it may also be correcting for additional errors in the force field. In Ref. 22, charges calculated in an implicit solvent of $\varepsilon = 4$ had a dipole moment approximately halfway between the vacuum and condensed phase dipole moments. This is consistent with an alternative approach that has been used to model induction effect, the implicitly polarized charge (iPol) method[24,63]

The iPol method was used to parametrize charges for the AMBER ff14ipq and ff15ipq force fields.[15,24,63] To recreate the energy of a polarizable dipole in an external field using a fixed dipole model, the polarization of the fixed dipole model should be approximately halfway between the dipole polarizated in the external field and the dipole in vacuum.[63] This results in the iPol charges being exactly halfway between the charges found in a vacuum and the charges found in a solvent.[24]

The average of the condensed and gas phase parameters allows the polarization energy to be included in a non-polarizable model.[15] This approximation was derived in Ref. 63 by considering the electrostatic energy of a set of dipoles in an electric field. In this work, and Ref. 15, it was shown that the polarization energy could be modeled by a fixed charge representation if dipoles were set to halfway between their value in vacuum and their value in the presence of a field. In practice this means that for each molecule two sets of charges can be calculated, one in vacuum and one in the condensed phase, and the average value of these two sets is then used for the force field parameters.

Electrostatic Components

There are a number of methods available for finding an atom's charge from the output of a DFT calculation. The atom in molecule (AIM) method has been shown to reproduce experimentally measured quantities well, and is therefore used in our force field.[68] The AIM approach partitions the electron density of the molecule to individual atoms.[68] This differs from electrostatic potential (ESP) methods, which assign charges that best recreate the electrostatic potential at a surface around the molecule.[95] The ESP approaches can fail to be sufficiently accurately for large molecules, and is not feasible when there are atoms buried far beneath the surface.[80,81]

The AIM scheme used in this work is the density derived electrostatic and chemical charges (DDEC) method, specifically DDEC/c3.[80,81] DDEC combines desirable qualities from two methods, the iterative-Hirshfeld (IH) scheme and the iterated stockholder atoms (ISA) scheme.[10,68,73] Both methods use a weighting scheme which defines the proportion of the electron density of the molecule each atom receives. The charge assigned to each atom is then defined by:

$$q_A = Z_A - \int W_A(\mathbf{r})n(\mathbf{r})d^3r \tag{2.28}$$

where $n(\mathbf{r})$ is the electron density, Z_A is the charge on the nucleus and $W_A(\mathbf{r})$ is the weighting function.

The IH method is an iterative from of the Hirshfeld method. In the Hirshfeld method, the weighting is based on the atomic densities:[23]

$$W_A(\mathbf{r}) = \frac{n_A^0(\mathbf{r})}{\sum_B n_B^0(\mathbf{r})} \tag{2.29}$$

where $n_A^0(\mathbf{r})$ is the atomic reference density. The atomic reference density used is the neutral gas-phase density although alternative choices could be made.[68] The Hirshfeld method tends to result in atomic populations that are too close to zero and the IH method was developed in an attempt to address this and additional problems.[68]

The first stage in the IH method is calculating the electronic population, $Q_A(i) = \int W_A(\mathbf{r})n(\mathbf{r})d^3r$, using the original Hirshfeld method.[68] The reference densities used are then updated using linear interpolation between the density of the ions with the closest integer number of electrons to $Q_A(i)$:

$$n_A^0(i,\mathbf{r}) = n_A^{0,\tau}(\mathbf{r})(\tau+1-Q_A(i)) + n_A^{0,\tau+1}(\mathbf{r})(Q_A(i)-\tau) \tag{2.30}$$

where τ is the integer below Q_A (floor(Q_A)).

This allows the electron density of an atom with the hypothetical non-integer charge Q_A to be estimated. The process is iterated until the set of atomic charges does not change by more than a set threshold. At each iteration the new weighting used is:[68]

$$W_A(i+1,\mathbf{r}) = \frac{n_A^0(i,\mathbf{r})}{\sum_B n_B^0(i,\mathbf{r})} \tag{2.31}$$

An alternative approach, which does not require reference densities, is the ISA method. The weighting used for the first iteration of partitioning is $n_A^0 = 1$ and the densities are subsequently updated using equation 2.31 and the following equation:[73]

$$n_A^0(i,\mathbf{r}) = \langle n_A(i,\mathbf{r})\rangle_A \tag{2.32}$$

where $\langle\cdots\rangle_A$ denotes the spherical average about the center of the atom.

In ISA, the partitioned electron density used in the weighting function at iteration $i+1$ is given by the spherical average of the electron density at iteration i. However, whilst the ISA method has been shown to produce a better approximation to the electrostatic potential than the IH scheme, the ISA method can be slow to converge.[22]

The DDEC scheme[80,81] is a compromise between convergence time and accuracy and uses a combination of both methods. The weighting function used is:[68]

$$W_A(i+1,\mathbf{r}) = \frac{[n_A^{IH}(i,\mathbf{r})]^{\chi}[n_A^{ISA}(i,\mathbf{r})]^{1-\chi}}{\sum_B [n_B^{IH}(i,\mathbf{r})]^{\chi}[n_B^{ISA}(i,\mathbf{r})]^{1-\chi}} \tag{2.33}$$

where χ is an adjustable parameter which is set at 0.02 in this thesis.[22]

Lennard-Jones Parameters

With the electron density of each atom calculated, the Lennard-Jones terms can then be derived. The coefficient of the r^{-6} component is found by following the scheme of Tkatchenko and Scheffler and rescaling the reference free atom coefficients by the ratio of the AIM volume to the free volume of the atom:[111]

$$B_i = \left(\frac{V_i^{AIM}}{V_i^{free}}\right)^2 B_i^{free} \tag{2.34}$$

where the B_i^{free} parameters were found from a time-dependent DFT calculation and V_i is the volume of the atom given by:[111]

$$V_i = \int r^3 n_i(\mathbf{r}) d^3 r \tag{2.35}$$

where r is the distance to the nucleus of the atom and $n_i(\mathbf{r})$ is the electron density for the free atom or the electron density from the AIM calculation.

The Tkatchenko-Scheffler equations give a relationship between the electron density and the C_6 parameters of an atom.[111] These equations are derived by using the exact expression for the C_6 term between two atoms A and B, $C_{6AB} = \frac{3}{\pi} \int_0^\infty \alpha_A(i\omega)\alpha_B(i\omega)d\omega$, where $\alpha_A(i\omega)$ is the frequency-dependent polarizability of A. Approximations are then made to $\alpha_{A/B}(i\omega)$, which results in the C_{6AB} being expressed in terms of C_{6AA} and C_{6BB} (and a further α_A^0 term that is calculated using TDDFT). These homonuclear C_6 parameters can then be readily approximated using equation 2.34, which is derived using the C_{6AB} equations along with the use of effective volumes (equation 2.35) and further approximations. The validity of these approximations for calculating C_6 terms was assessed in Ref. 111 and shown to be surprisingly good for a large range of molecules. The Tkatchenko-Scheffler relations have been widely used to add the dispersion interaction to DFT energies and were shown to be appropriate for force field parametrization of small molecules in Ref. 22.[111]

By constraining the minimum of the LJ potential to the vdW radii, R_i^{AIM}, of the atom the r^{-12} coefficient can then be calculated.[22]

$$0 = \left. \frac{\partial V_{vdW}}{\partial r_{ij}} \right|_{r_{ij}=2R_i^{AIM}} \tag{2.36}$$

$$A_i = \frac{1}{2} B_i (2R_i^{AIM})^6 \tag{2.37}$$

R_i^{AIM} is estimated using the scaling relationship of Tkatchenko and Scheffler.[111]

$$R_i^{AIM} = \left(\frac{V_i^{AIM}}{V_i^{free}} \right)^{\frac{1}{3}} R_i^{free} \tag{2.38}$$

The R_i^{free} parameters were fit to experimental liquid densities and the parameters for H, C, N, O, S, F and Cl were found in Ref. 22.

The interaction between two different elements are calculated with the geometric combining rule used in OPLS:

$$A_{ij} = \sqrt{A_i A_j} \tag{2.39}$$

This applies to both coefficients A_{ij} and B_{ij}.

Additionally, hydrogen atoms attached to polar atoms are given Lennard-Jones terms equal to zero, with the polar atoms having a new B coefficient given by:

$$\sqrt{B'_x} = \sqrt{B_x} + n_H\sqrt{B_H} \tag{2.40}$$

where B'_x, B_x are old and new terms respectively, n_H is the number of hydrogen atoms and B_H is the hydrogen's dispersion coefficient.

The non-bonded parameters between atoms separated by 3 bonds (1-4 atoms) are scaled by 0.5. This scaling is also used in OPLS.

2.3.3 Molecular Dynamics Simulations

A force field can be used describe the interactions between atoms in MM simulations with molecular dynamics (MD) used to propagate the dynamics of a molecule.

In classical MD simulations, Newton's equations of motions are solved for a system of atoms and this produces the trajectory of the atoms over a period of time.[31] Many different schemes can be used to solve the equations of motion. The most commonly used is the Verlet method, which is derived by a Taylor expansion of $\mathbf{r}(t+\Delta t)$, and uses discrete timesteps of Δt to calculate the position of atoms at time $t+\Delta t$ from their positions at previous timesteps:

$$\mathbf{r}(t+\Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \mathbf{a}(t)\Delta t^2 + O(\Delta t^4) \tag{2.41}$$

$$\mathbf{v}(t) = \frac{\mathbf{r}(t+\Delta t) - \mathbf{r}(t-\Delta t)}{2\Delta t} \tag{2.42}$$

$$\mathbf{a}(t) = \mathbf{f}(t)/m \tag{2.43}$$

where $\mathbf{r}$ is the position, $\mathbf{v}$ is the velocity, t is the time, m is the mass of an atom, and $\mathbf{a}$ is the acceleration.[2]

Additional techniques need to be used in MD simulations in order to keep pressure and temperature at constant values. One method to control the temperature uses an extended Lagrangian, which adds an additional term to the equations of motion so energy can flow between the system and a reservoir.[31] A frictional coefficient is added so as to minimize the difference between the instantaneous and statistical temperatures. A similar approach is used in the Nosé-Hoover method, with a frictional coefficient added to ensure there is the minimum difference between the instantaneous pressure and the pressure of an external reservoir.[31] The molecular dynamics simulations performed in this thesis follow the protocol described in Ref. 99.

Chapter 3

Bond and Angle Parameters

In Ref. 22, the bond and angle terms were taken from the OPLS force field. These values needed to be updated as terms should not be transferred from one force field to another. The development of a bond and angle parametrization method is described in this chapter. Using this new approach, bond and angle terms are calculated for our biomolecular force field.

3.1 Introduction

The bond and angle components of a force field are used to describe vibrations of the bonds and angles around their equilibrium positions, and generally are described by a harmonic potential. Historically, in biomolecular force fields such as OPLS and AMBER, many of the bond and angle force constants were found by fitting MM normal modes and frequencies to experimental or quantum mechanical (QM) studies of small molecules.[54,125] Such an approach creates interdependencies between force field parameters.[114] That is, the computed bond and angle force constants are dependent on the choice of torsional and non-bonded parameters used in the original fitting procedure, and therefore changes to one component of the force field requires a re-fit of all the other parameters. The bond force constants that could not be fit to experiment were estimated by assuming a linear relationship between the force constants and experimental bond lengths,[125] which may limit the achievable accuracy.[72] Given the importance of intramolecular interactions in determining conformational preferences of molecules[24] and reproducing accurate vibrational spectra,[11] biomolecular force field developers are beginning to reparametrize the bond and angle terms as a means to improve the accuracy of MM simulations.[24,122] However, there is no standard approach for bond and angle parametrization that combines both accuracy and ease-of-use, whilst removing the problem of parameter interdependence. Given that the non-bonded parameters used in our force field are system specific, having bond and angle terms that do not depend on the rest of the force field is even more important than for conventional force fields.

A number of methods have recently been developed that are aimed at finding bond and angle parameters with greater ease and accuracy.[11,24,72,84,103] These methods can be divided into fitting approaches, which rely on MM calculations as part of the parametrization process, and non-fitting approaches, which rely only on QM data. The use of multiple iterations to parametrize a MM force field through fitting to the QM Hessian matrix has been shown to give reasonably accurate MM normal modes.[11] However, the dependence of the fitting process on repeated calculations of the MM Hessian matrix, results in interdependencies between force field parameters.[11] This effectively means that bond and angle parameters should be updated when changes to other components of the force field are made. Similarly, in an extension to the CHARMM force field, which was fit to QM frequency spectra, the issue of parameter interdependencies meant that repeated parametrization of bond and angle parameters was required as dihedral and non-bonded components of the force field were updated.[114] This adds time and effort to the parametrization process.[114] Another example of a large scale parametrization approach is the method used for the AMBER ff15ipq force field.[24] Eight generations of improvements were carried out, with repeated MD simulations and QM optimization at each cycle creating tens of thousands of conformations that were used to fit the bond and angle parameters.[24] Automating this process, so that it is suitable for use by inexperienced users to parametrize molecules outside the fitting set, would not be straightforward. Speed is often a factor in fitting methods, not just because of interdependencies, but also due to difficulties in the fitting process. The Force Field Toolkit (ffTk), is a VMD plugin that works with the CHARMM force field to parameterize small molecules. This method fits the MM potential energy surface to the QM potential energy surface and convergence of force constants can be slow.[84] Like all fitting approaches, this method also requires an initial estimate of the force field parameters for the first MM calculation, which relies on the preliminary values being available, and reasonably close to the optimal values. A method to fit parameters using the partial QM Hessian matrix of a molecule has recently been developed and tested on 23 molecules. This gave a mean unsigned error of 73.3 cm^{-1} for the recreation of QM vibrational frequencies[123] which gives an indication of the levels of error that are typically obtained by fitting MM parameters to QM data.

The methods discussed so far all have the disadvantages of non-transferability, interdependency of force field parameters and reliance on an initial parameter estimate. These are inherent characteristics of methods that rely on fitting MM force field parameters to QM or experimental data. Therefore, it is seemingly advantageous to move away from using MM calculations as part of the bond and angle parametrization procedure. Non-fitting methods offer speed, transferability and independence from the other force field components, but currently lack accuracy. The most widely used of the non-fitting methods is the Seminario method[103] which uses projections of the QM Hessian matrix to determine force constants for MM force fields, and is available through the AMBER suite of programs, in the VFFDT plugin or the MCPB.py program. This method has been popularly applied to biomolecular systems containing metals, for which general force field parameters are typically lacking.[5,72,74,103,127] However, this method has been shown to be less accurate than fitting to the QM Hessian matrix for two small test sets.[11,123] In particular, the Seminario method struggled to recreate the normal modes of molecules with more than five atoms, with particular problems recreating the angle bending frequencies.[123] This points to possible inaccuracies in the angle force constants.

In this section, we propose a modification to the Seminario method which substantially improves the computed angle force constants by taking into account the geometry of the systems under study. We extensively test the ability of the modified Seminario method to reproduce QM vibrational frequencies, and compare the accuracy to standard MM force fields and the original Seminario method. The benchmark data set comprises a total of 70 molecules, including small molecules and dipeptides, against which standard force fields have been parametrized, and also more complex organic heterocycles and a metal containing complex. For the majority of the 70 molecules tested, the modified Seminario method is more accurate than the original approach. A program that implements the method proposed is supplied for use by other groups, which allows users to quickly and easily derive bond and angle parameters from the output of a Gaussian09[32] frequency calculation. To prevent repetition of calculations, we have also freely supplied a complete set of bond and angle force field parameters for each of the 20 naturally-occurring amino acids (supporting information of Ref. 1). These terms are used in the following chapter with our biomolecular force fields, and could similarly be used by other groups developing biomolecular force fields.

3.2 Theory

3.2.1 Normal Mode Analysis

The Hessian matrix is a 3N × 3N matrix, where N is the number of atoms, containing the second derivative of the energy with respect to an atom's position, see equation 3.3 for the partial Hessian matrix. The Hessian matrix can be calculated computationally using finite difference methods or analytically.[12,105]

The eigenvectors and eigenvalues of the matrix $F = M^{-1/2}HM^{-1/2}$, where M is a matrix of the atomic masses and H is the Hessian matrix, correspond to the normal modes and frequencies of the molecule.[86] If the atoms in the molecule are displaced in the direction of an eigenvector then sinusoidal motion, with a frequency given by the eigenvalue, will continue in the direction of the eigenvector.

The comparison of QM and MM normal mode frequencies is a common test of the accuracy of the bond and angle force constants.

3.2.2 Seminario Method

The Seminario method was developed by Jorge Seminario in 1996[103] to parametrize harmonic bond and angle force field parameters from the QM Hessian matrix of the molecule. This provided a valuable tool for obtaining intramolecular force field parameters directly from QM data, without the need for empirical input. In this section, we outline the original Seminario methodology.

The reaction force, $\delta\mathbf{F}$, due to a small displacement $\delta\mathbf{r}$ in a system comprising N atoms can be written to second order as:

$$\delta\mathbf{F} = -[\mathbf{k}]\delta\mathbf{r} \tag{3.1}$$

where $[\mathbf{k}]$ is the Hessian matrix of the molecule. For practical applications in MM simulations, the relationship between the total energy of a molecule and its nuclear coordinates are typically expressed in terms of a force field equation in the internal coordinates:

$$V = \sum_{\text{bonds}} \frac{1}{2}k_r(r-r_0)^2 + \sum_{\text{angles}} \frac{1}{2}k_\theta(\theta-\theta_0)^2 + \dots \tag{3.2}$$

where the first term accounts for two-body bond stretching about an equilibrium bond length (r_0), and the second for three-body angle bending about an equilibrium bond angle (θ_0). MM force fields generally also include an anharmonic four-body torsional term, and reparametrization of this term is covered in the following chapter. The objective of the Seminario method is therefore to obtain the MM harmonic force constants, k_r and k_θ, from the full QM Hessian matrix $[k]$.

By analogy with eq 3.1, the force felt by atom A due to displacement of atom B is given by $\delta\mathbf{F_A} = -[\mathbf{k_{AB}}]\delta\mathbf{r_B}$. The 3 × 3 interatomic force constant matrix $[\mathbf{k_{AB}}]$ contains only the elements of the full Hessian matrix relating to atoms A and B:

$$[\mathbf{k_{AB}}] = -\begin{bmatrix} \frac{\partial^2 E}{\partial x_A \partial x_B} & \frac{\partial^2 E}{\partial x_A \partial y_B} & \frac{\partial^2 E}{\partial x_A \partial z_B} \\ \frac{\partial^2 E}{\partial y_A \partial x_B} & \frac{\partial^2 E}{\partial y_A \partial y_B} & \frac{\partial^2 E}{\partial y_A \partial z_B} \\ \frac{\partial^2 E}{\partial z_A \partial x_B} & \frac{\partial^2 E}{\partial z_A \partial y_B} & \frac{\partial^2 E}{\partial z_A \partial z_B} \end{bmatrix} \tag{3.3}$$

The three eigenvalues of $[\mathbf{k_{AB}}]$, λ_i^{AB}, are the force constants in the direction of the three eigenvectors, v_i^{AB}. However, we instead require the force constants for changes in intramolecular bond lengths and angles. To calculate the bond force constant for the bond AB, each eigenvector is projected onto the direction of the bond vector, $\hat{u}^{AB}$:

$$k_r = \sum_{i=1}^{3} \lambda_i^{AB} |\hat{u}^{AB}.\hat{v}_i^{AB}| \tag{3.4}$$

In the original Seminario paper the definition of a bonded atom was determined by the eigenvalues of $[\mathbf{k_{AB}}]$. We have not used this definition and use the conventional definition of bonded atoms specified by the force field.

The angle force constant, k_θ, is more complex as it involves projections onto directions perpendicular to two different bonds AB and CB. Let us define two vectors, $\hat{u}^{PA1}$ and $\hat{u}^{PC1}$ Fig. 3.1(a)), that are perpendicular to the bonds AB and CB respectively and lie in the plane ABC. Then k_{PA} and k_{PC} are defined as the corresponding force constants obtained by projecting the eigenvectors of the partial Hessian matrix onto these two vectors:

$$k_{PA} = \sum_{i=1}^{3} \lambda_i^{AB} |\hat{u}^{PA1}.\hat{v}_i^{AB}| \tag{3.5}$$

$$k_{PC} = \sum_{i=1}^{3} \lambda_i^{CB} |\hat{u}^{PC1}.\hat{v}_i^{CB}| \tag{3.6}$$

Via analogy to two springs connected in series, the angle force constant is then approximated by:

$$\frac{1}{k_\theta} = \frac{1}{R_{AB}^2 k_{PA}} + \frac{1}{R_{CB}^2 k_{PC}} \tag{3.7}$$

where R_{AB} and R_{CB} are the two bond lengths (Fig 3.1(a)).

3.2.3 Modified Seminario Method

The Seminario method for the derivation of harmonic angular force constants assumes that the change in energy associated with the displacement of atom A along the direction $\hat{u}^{PA1}$

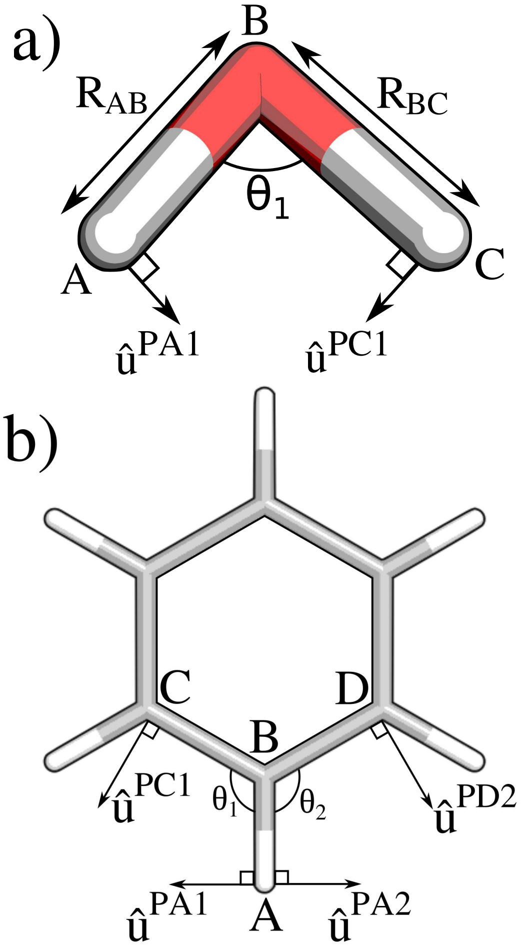


Fig. 3.1: (a) The original Seminario method applied to a water molecule. (b) Extension of the original Seminario method to a larger molecule (benzene). If atom A moves in the direction perpendicular to bond AB, $\hat{u}^{PA1}$, both angles θ_1 and θ_2 are altered. Therefore, the angle force constant obtained via projection of the eigenvector of the matrix $[\mathbf{k_{AB}}]$ onto $\hat{u}^{PA1}$ is over-estimated by a factor of two.

will only change the angle involving atoms A, B and C as in Fig. 3.1(a). However, in larger molecules, neighboring angles may also be altered by a displacement of atom A in the direction of $\hat{u}^{PA1}$. Let us consider how the Seminario method would calculate the angle force constants involving four atoms (A, B, C, D) in the same plane, such as in a benzene molecule (Fig. 3.1(b)). The Seminario method finds the force constant for angle ABC, θ_1, from the projections of the eigenvectors of the partial Hessian matrices on to $\hat{u}^{PA1}$ and $\hat{u}^{PC1}$. Hence the Seminario estimate of k_{θ_1} includes all QM forces on atom A acting in the direction $\hat{u}^{PA1}$. Importantly, however, the equivalent calculation of k_{θ_2} also includes all QM forces on atom A acting in the direction $\hat{u}^{PA1}$. This inevitably results an intramolecular MM force field that is too stiff.

We can think of this problem in terms of the change in energy caused by a small change, Δx along $\hat{u}^{PA1}$, that would be computed using the original Seminario method:

$$k_{PA1}^{Seminario}(\Delta x)^2 = k_{PA1}(\Delta x)^2 + k_{PA2}(\Delta x)^2 \tag{3.8}$$

where k_{PA1} (k_{PA2}) is the hypothetical value of $k_{PA1}^{Seminario}$ ($k_{PA2}^{Seminario}$) that would be computed if ABC (ABD) existed in isolation from all other angles. For the water example $k_{PA1}^{Seminario} = k_{PA1}$. For the benzene molecule , $k_{PA1} = k_{PA2}$ and so the Seminario method over-estimates the change in energy by a factor of two.

If ABC and ABD are not in the same plane, movement in the direction $\hat{u}^{PA1}$ can still cause displacement in the direction of neighboring angles. The change in energy predicted by the original Seminario method may then be approximated by:

$$k_{PA1}^{Seminario}(\Delta x)^2 = k_{PA1}(\Delta x)^2 + k_{PA2}(\Delta x|\hat{u}_{PA1}.\hat{u}_{PA2}|)^2 \tag{3.9}$$

Assuming further that $k_{PA1} \approx k_{PA2}$, which is true when both angles are in the same plane:

$$k_{PA1}^{Seminario} = k_{PA1}(1 + |\hat{u}_{PA1}.\hat{u}_{PA2}|^2) \tag{3.10}$$

$$k_{PA1} = \frac{k_{PA1}^{Seminario}}{(1 + |\hat{u}_{PA1}.\hat{u}_{PA2}|^2)} \tag{3.11}$$

Equation 3.11 rescales the original value of $k_{PA1}^{Seminario}$ by a factor that accounts for the geometry of the molecule. To extend the above analysis to sites B with multiple angles, we have found empirically that the mean of the additional contribution ($|\hat{u}_{PA1}.\hat{u}_{PA2}|^2$) from all neighboring angles gives the most reasonable agreement with the QM vibrational frequency spectra. This results in our modified formula for the angle force constant for ABC when $N > 1$ and $M > 1$, where N (M) is the number of angles in the force field that have a central atom B and involve movement of the bond AB (BC):

$$\frac{1}{k_\theta} = \frac{1+\frac{\sum_{i=1}^{N}|\hat{u}^{PA_1}.\hat{u}^{PA_i}|^2-1}{N-1}}{R_{AB}^2 k_{PA}^{Seminario}} + \frac{1+\frac{\sum_{i=1}^{M}|\hat{u}^{PC_1}.\hat{u}^{PC_i}|^2-1}{M-1}}{R_{CB}^2 k_{PC}^{Seminario}} \tag{3.12}$$

If $N = 1$ ($M = 1$) the left (right) hand component is replaced by the original Seminario method.

3.2.4 Computational Implementation

Our proposed modification to the Seminario method comprises eq 3.4 for the intramolecular harmonic bond force constants and eq 3.12 for the modified angle force constants. The equilibrium bond lengths and angles are obtained from the optimized QM structure. The modified Seminario method is available for use through a MATLAB (or Python) program which may be freely downloaded from *https://github.com/aa840/ModSeminario*

(*https://github.com/aa840/ModSeminario_Py*) along with a short tutorial explaining how to use the program to find the bonded parameters of a benzene molecule. For the example given, the Hessian matrix of the molecule can be converted into bond and angle parameters in a matter of seconds on a standard desktop computer. Larger molecules may also be parametrized in negligible compute times since the method scales approximately linearly with the number of bonds and angles.

The optimized structure and connectivity of the molecule, as well as the QM Hessian matrix, is read in from Gaussian 09[32] output files (specifically .fchk and .log files). Optionally, a BOSS z-matrix,[55] which can be produced using the LigParGen web server,[27,28,56] may be supplied to provide the OPLS atom types. If a z-matrix is supplied as input, the OPLS atom types are used to return the average value for each bond and angle class. However if OPLS atom types are not required, or are unavailable (for example, for molecules containing a metal), the Gaussian 09 output files can be used in isolation, with no bond and angle parameter averaging performed. Thus the program can be used for a wide range of molecules and force fields.

Following standard practice, the QM-derived vibrational frequencies of a molecule can be multiplied by a constant to better fit experimental vibrational spectra.[102] This is incorporated into the modified Seminario method by multiplying the bond and angle force constants by the square of the frequency scaling constant.[89] This scaling constant can be altered by the user according to the level of QM theory employed, or set equal to one.

3.3 Simulation Methods

To test the accuracy of the modified Seminario method, 38 small organic molecules were chosen with a diverse range of chemical structures. Test sets of this nature are commonly used to parametrize MM biomolecular force fields. The molecules contained more than six atoms to ensure that the effect of our angle correction is apparent. Following the small molecule validation set, we also repeated our analysis on a set of ten heterocyclic molecules. The full list of small and heterocyclic molecules is provided in Section A.1 of the appendix.

For each molecule, a structural optimization and frequency calculation was performed using Gaussian 09 with the ωB97XD functional and a 6-311++G(d,p) basis set.[32] The QM vibrational frequencies were re-scaled by a factor of 0.957, which is the value recommended for the ωB97XD/6-311G(d,p) level of theory by the Computational Chemistry Comparison and Benchmark DataBase.[51] The same scaling factor is also used to effectively scale the MM frequencies, as outlined in Section 3.2.4. The level of QM theory chosen for the frequency calculation is the same as that used in the recent re-parametrization of the protein backbone torsional parameters for OPLS-AA/M.[99] To ensure that this choice did not significantly influence the results, our analysis was repeated for a subset of 10 small molecules using the B3LYP/cc-pVTZ level of theory. As reported in Table 3.1, the computed accuracy of our method is not strongly dependent on the choice of underlying QM data.

Bond and angle parameters for each molecule were found, as described above, from the computed QM Hessian matrices, using the original and modified Seminario method. Dihedral and non-bonded parameters were assigned from the OPLS/CM1A force field[28,59] using the LigParGen web server to obtain z-matrices.[27] The MM normal modes and frequencies were calculated using the BOSS general purpose molecular modeling software with Broyden-Fletcher-Goldfarb-Shanno (BFGS) structural optimization .[55] The mean percentage error in each molecule is computed as $\frac{100}{3N-6}\sum_{i=7}^{3N}|\frac{\alpha\nu_i^{QM}-\nu_i^{MM}}{\alpha\nu_i^{QM}}|$ where $\nu_i^{MM/QM}$ is the frequency of the ith MM/QM normal mode, α is the vibrational scaling factor and N is the number of atoms in the molecule. The mean unsigned error (MUE) is also given for comparison and is computed as $\frac{1}{3N-6}\sum_{i=7}^{3N}|\alpha\nu_i^{QM}-\nu_i^{MM}|$. Although the above measures of error are commonly used in assessing the accuracy of bond and angle parameters, we emphasize that the QM frequencies, that are treated as 'ideal' values, are derived from the same QM Hessian matrix used to parameterize the bond and angle terms.

In order to provide direct comparison with the MCPB.py force field parametrization program,[72] we also analyzed the complex tris(1,10-phenanthroline)-osmium(II) (Os$[(\text{phen})_3]^{2+}$)

[1] The QM frequencies used were found from B3LYP calculations and scaled with a vibrational scaling factor of 0.965.

	Modified Seminario (B3LYP) (%) [1]	**Modified Seminario (** $\boldsymbol{\omega}$**B97XD)** (%)
Acetic acid	4.39 (53.75)	4.28 (51.12)
Benzene	4.05 (47.94)	4.51 (50.38)
Cyclohexane	5.77 (60.06)	5.77 (57.76)
Ethane	4.91 (74.50)	4.61 (69.91)
Methyl acetate	5.62 (55.48)	6.73 (53.07)
Methyl benzoate	4.79 (36.41)	5.06 (36.32)
Propene	5.37 (67.15)	6.13 (66.51)
Pyridine	3.67 (41.29)	3.83 (41.01)
Triethylamine	4.04 (50.96)	4.61 (52.59)
Trifluorobenzene	5.08 (40.56)	5.55 (42.93)
Mean	4.77 (52.81)	5.11 (52.16)

Table 3.1: The mean percentage error between the QM and MM normal mode frequencies for a subset of the small molecules. The mean unsigned error is shown in brackets (cm^{-1}). The QM frequencies used in the calculation of the error have been scaled to better reproduce experimental frequencies.[102] The errors for the modified Seminario method using a B3LYP functional and using a ωB97XD functional are given.

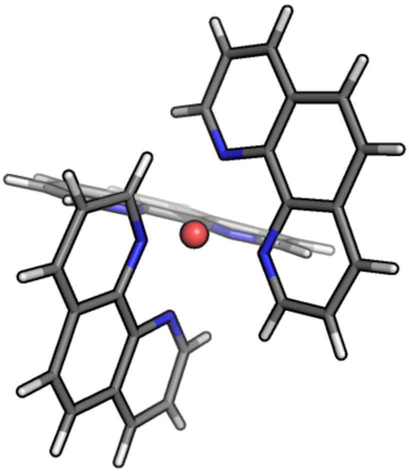


Fig. 3.2: The Os[(phen)$_3$]$^{2+}$ complex. Hydrogen atoms are colored in white, carbon in grey, nitrogen in blue and osmium in red.

shown in Fig. 3.2.[25] The QM Hessian matrix, computed using the B3LYP functional and a 6-31G(d) basis set, was obtained directly from the work by Li et al.[72] Bond and angle parameters were computed using the modified Seminario method. For consistency with the MCPB.py analysis, we computed MM normal modes using AMBER16,[14] with the AMBER ff14SB and GAFF force fields for torsional and non-bonded parameters. A vibrational scaling factor was not applied in this case.

Finally, we computed bond and angle force field parameters for each of the 20 naturally-occurring amino acids with the same methods used for the small molecule and heterocycle data sets. The OPLS-AA/M force field was used for all torsional and non-bonded parameters.[99] The amino acids (X) were blocked with acetyl and N-methyl groups (Ace-X-NMe). Our definition of a dipeptide (a single amino acid with two peptide bonds) is the same as that described in Chapter 2. A total of 80 structures were analyzed to account for variation in backbone and side chain conformations. A minimum of one β-sheet and one α-helical conformation was tested for each dipeptide, which have ψ and ϕ dihedral angles of ($-60°$, $-45°$) and ($-135°$, $135°$) respectively.[99]Additional starting configurations for larger amino acids were generated by fixing the backbone dihedral angles and scanning the side chain dihedral angle (N-Cα-Cβ-Xγ, where the atom type Xγ depends on the amino acid) for local minima. The starting structures were fully optimized as part of the QM frequency calculation, ensuring a representative sampling of low energy structures. The reported bond and angle force field parameters were averaged over the different conformations produced.

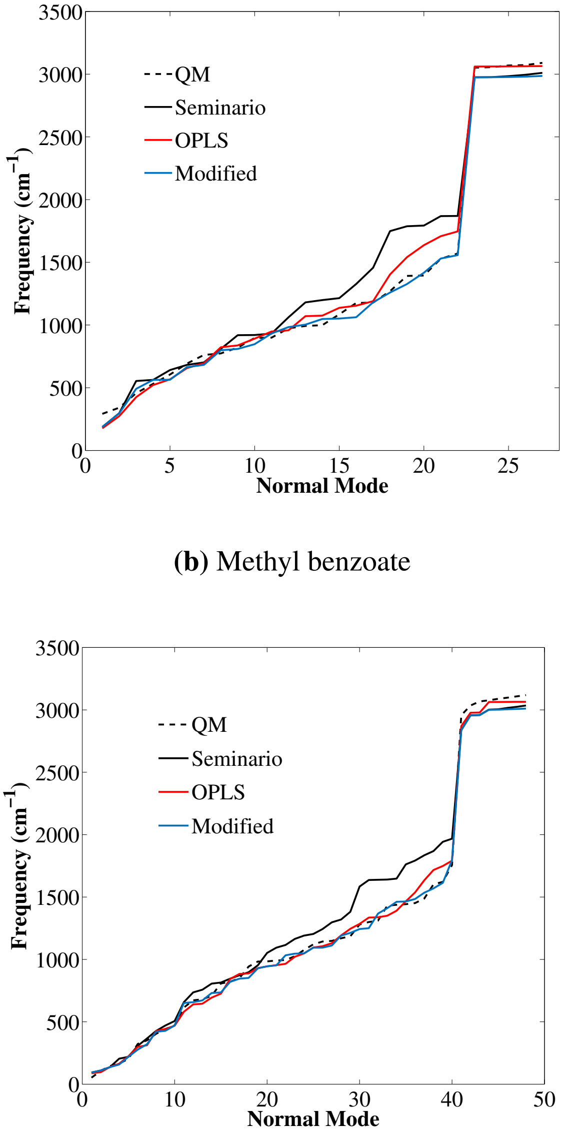
3.4 Results

	OPLS	Original Seminario	Modified Seminario
Small Molecules	7.3% (60.4)	12.3% (119.5)	6.4% (52.3)
Heterocycles	8.6% (82.6)	11.7% (132.3)	6.8% (52.8)
Dipeptides	7.0% (46.6)	12.4% (104.3)	6.1% (39.5)
Average	7.4%(59.4) ± 2.9%(19.3)	12.3%(116.7)± 3.4%(26.6)	6.3%(48.5) ± 2.75%(13.7)

Table 3.2: Mean percentage error (%) in the MM vibrational frequencies for OPLS and the original/-modified Seminario parametrization schemes. The value shown in brackets is the mean unsigned error (cm^{-1}). The QM frequencies used in the calculation of the error have been scaled to better reproduce experimental frequencies.[102] The error shown represents the standard deviation of the results.

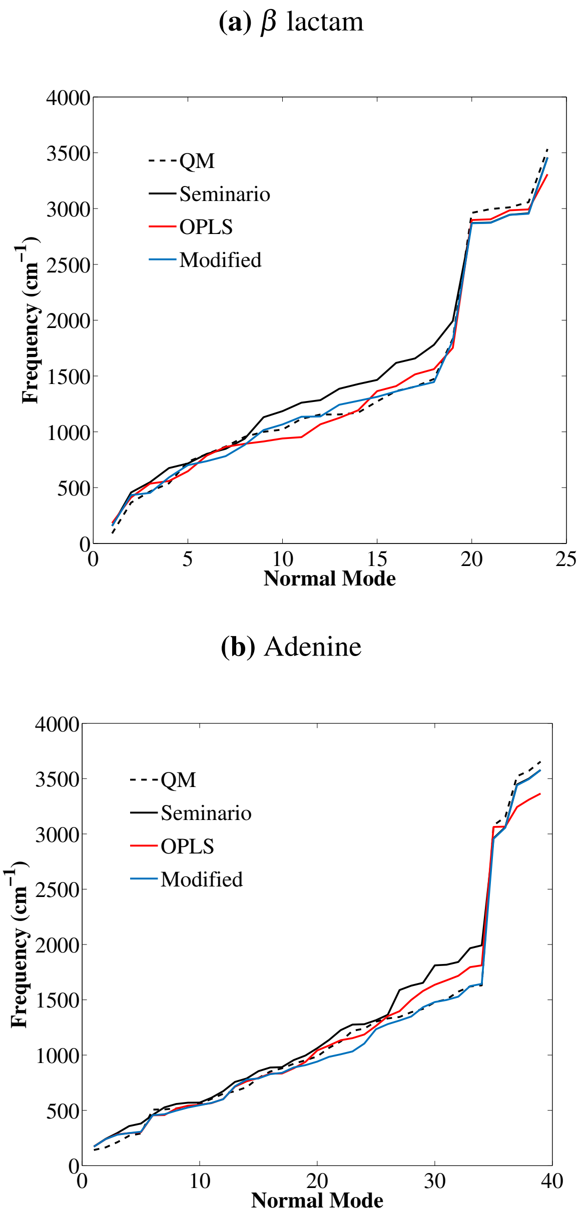
Table 3.2 summarizes the ability of various force field parametrization techniques to reproduce the QM vibrational frequencies of a range of tested molecules. The full list of molecules and their associated errors are given in Appendix A. Focusing first on the small molecule data set, the percentage error for the vibrational frequencies computed using the OPLS force field is 7.3%. As expected, the computed frequencies are very similar to high level QM data because force fields are parameterized to fit experimental vibrational spectra of small molecules such as these.[54,125] Similar accuracy is expected for other standard MM biomolecular force fields, which often employ the same bond and angle parameters. Next, we reparameterize the bond and angle equilibrium values and force constants using the method proposed by Seminario,[103] and combine the parameters with OPLS torsional and non-bonded parameters. The resulting error is almost twice as high as that of the standard force field. Mean unsigned errors of > 100 cm^{-1} have been reported in other studies,[11,123] which casts doubt on the suitability of the Seminario method as an automated parametrization tool.[72] In contrast, the error in our modified parametrization scheme (6.4%) is much lower than the original Seminario method and similar to the OPLS force field. The corrections that we have made to the Seminario method, described in Section 3.2.3, result in a more accurate reproduction of the QM vibrational spectra. Figure 3.3 demonstrates this improvement in the case of the vibrational spectra of phosphorine and methyl benzoate.

As discussed, standard MM force fields are expected to perform well for this small molecule data set. As a more stringent test, we have computed the QM vibrational spectra of ten more complex heterocyclic compounds, which are expected to be less structurally similar to the original parametrization set. Table 3.2 summarizes the average error in the three parametrization methods across all ten molecules, and Fig. 3.5 further breaks down the results by molecule. As expected, the error in the vibrational frequencies computed using OPLS (8.6%) is higher than that computed for the small molecule test set. However, the original Seminario yields even higher errors, again indicating its unsuitability for force field parametrization. Encouragingly, the modified Seminario method maintains a low error (6.8%), which is largely constant across the heterocycle test set and is consistently lower than the original Seminario method (Fig. 3.5). This is can also be seen in Fig. 3.4 , which shows the vibrational spectra of adenine and β-lactam. Closer examination reveals that the majority of the improvement in the accuracy of the normal modes is brought about by the changes to the harmonic force constants, rather than the bond lengths or equilibrium angles. Some parameters have very large deviations from the corresponding OPLS parameters. For example, one of the C–C bond force constants in pyrrole, found using the modified Seminario method, is 25% lower than the corresponding OPLS parameter. Improvements in the optimized structures are also observed with the new parameters for heterocyclic molecules, particularly for the four membered rings. In the QM optimized structures, all the heavy atoms in β-lactam and oxetane are coplanar, which is correctly reproduced by the modified Seminario parameters. However, optimization with OPLS yields slightly twisted


(a) Phosphorine

(b) Methyl benzoate

Fig. 3.3: The vibrational spectrum of phosphorine and methyl benzoate computed using QM, and compared with the original and modified Seminario methods and OPLS. The QM frequencies used in the calculation of the error have been scaled to better reproduce experimental frequencies.[102]


(a) β lactam

(b) Adenine

Fig. 3.4: The vibrational spectrum of β lactam and adenine computed using QM, and compared with the original and modified Seminario methods and OPLS. The QM frequencies used in the calculation of the error have been scaled to better reproduce experimental frequencies.[102]

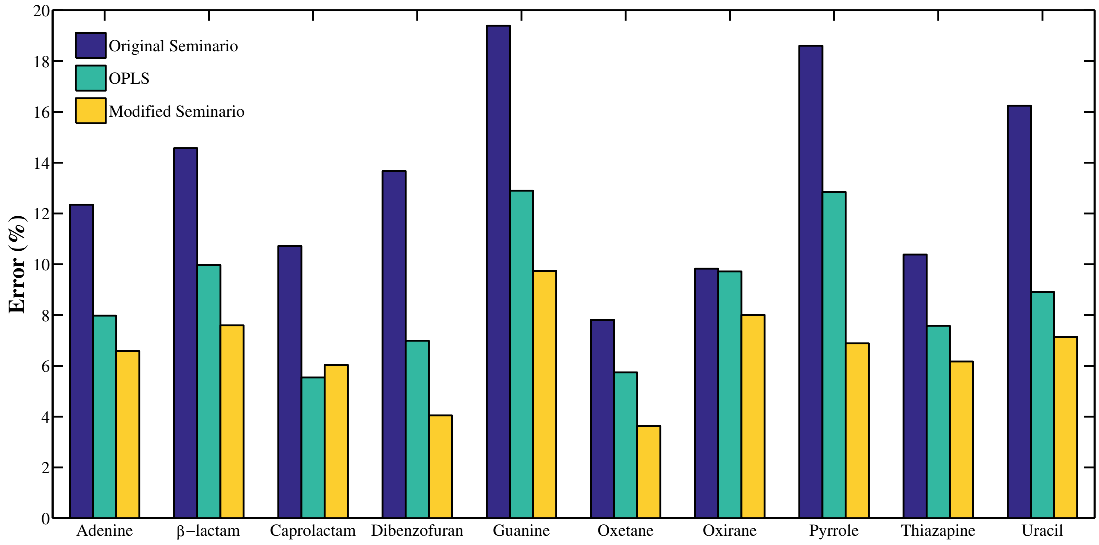


Fig. 3.5: The error in the MM vibrational frequencies for a set of heterocyclic molecules, using bond and angle parameters from the original Seminario method, OPLS and the modified Seminario method. The QM frequencies used in the calculation of the error have been scaled to better reproduce experimental frequencies.[102]

structures, with a computed C–O–C–C dihedral angle of 20.2° for oxetane and a C–N–C–C dihedral angle of 11.5° in β-lactam.

Figure 3.6 shows a comparison between the QM and MM vibrational frequency spectra of $Os[(phen)_3]^{2+}$. The MCPB.py method uses the original Seminario method to parameterize bonds and angles involving the metal ion, and applies the standard AMBER force field elsewhere.[72] For direct comparison, we have used identical dihedral and non-bonded parameters, but replaced all bond and angle parameters with those computed using the original and modified Seminario methods. All methods agree well in the low frequency range 0–1250 cm^{-1}, while reproduction of the very high frequency ($>$ 2500 cm^{-1}) vibrations of bonds involving hydrogen is problematic for all methods. The original Seminario method and MCPB methods clearly over-estimate the vibrational frequencies of normal modes in the intermediate regime (1250–2500 cm^{-1}). These modes largely involve angle bending motions, which are precisely the motions that we set out to correct in the modified Seminario method. The overall error (6.0%) of our modified Seminario method is half that computed using the original Seminario method, and very similar to the errors computed for the small molecule and heterocycle data sets. Furthermore, this additional test case demonstrates that the modified Seminario method works well for relatively large system sizes (67 atoms), and

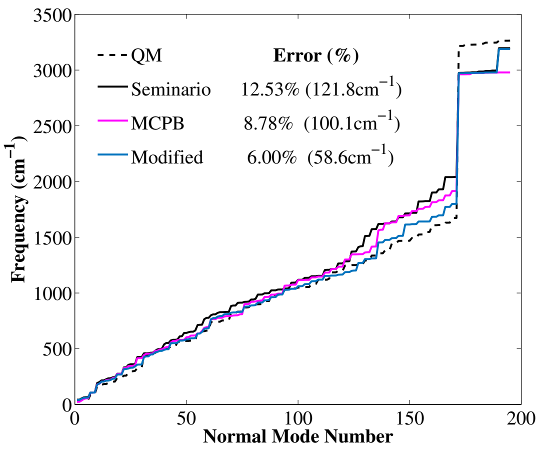


Fig. 3.6: The vibrational spectrum of Os[(phen)$_3$]$^{2+}$ computed using QM, and compared with the original and modified Seminario methods, as well as the bonded parameters reported in Ref. 72 (MCPB). The mean unsigned error for each method is given in the key. No vibrational scaling factor has been applied to the QM frequencies.

is not too dependent on the underlying force field that is used to compute torsional and non-bonded energetics.

3.4.1 Amino Acid Parameter Set

The validation tests described in this chapter reveal that harmonic force constants derived using the modified Seminario method give vibrational spectra that are in very good agreement with QM data across a wide range of molecules. We therefore envisage this method being used as a toolkit for automated parametrization of molecules that are missing force field parameters (for example, metal complexes), or for which the transferability of the standard force field parameters are questionable (for example, heterocyclic molecules). As a further resource, and also to test whether the modified Seminario method is suitable for *transferable* force field parametrization, we have computed a new bond and angle parameter set for the 20 naturally occurring amino acids using our new method. The amino acids are blocked with acetyl and N-methyl groups to form dipeptides. We performed QM calculations on a total of 80 dipeptide structures (including different backbone and side chain conformations) and averaged parameters for each atom type over all structures and amino acids. Using these averaged bond and angle parameters alongside the OPLS-AA/M force field, we computed the vibrational spectra of each of the 20 amino acids and compared our results with QM data. The results are summarized in Table 3.2, and show the expected trend. The OPLS

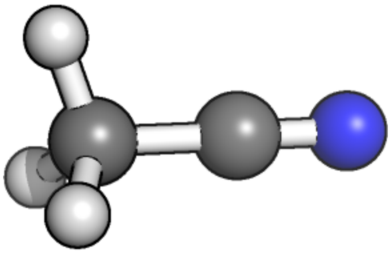


Fig. 3.7: The molecule acetonitrile. The nitrogen atom is shown in blue, carbon in grey and hydrogen in white. The parallel C–N and C–C bonds can be seen in the figure.

force field, which has been parameterized to reproduce experimental vibrational spectra of small molecules, has a low percentage error of 7.0% relative to our QM data. The original Seminario method fails to reproduce the QM vibrational spectra, while the modified method results in a slightly lower error (6.1%) than the OPLS force field. As a comparison, we also computed the vibrational spectra for each amino acid using bond and angle parameters that are *specific* to that molecule (Table A.3). However, the error is virtually identical to the averaged parameter set for the modified Seminario method, indicating that the harmonic bond and angle parameters are indeed transferable.

To better understand how the bond and angle parameters vary, the modified Seminario parameters were compared to the OPLS force field (Section A.2). It was found that the bond lengths and equilibrium angles do not deviate far from the OPLS parameters. However, the modified Seminario bond force constants are generally lower than the OPLS parameters, with the mean bond force constant being 411 kcal/mol/Å^2 for OPLS and 341 kcal/mol/Å^2 for our new parameter set. In contrast, the modified Seminario angle force constants are slightly higher than for the OPLS force field. However, even more apparent is the larger range of force constants that are computed using the modified Seminario method. OPLS angle force constants range between 33–85 kcal/mol/rad^2 for the amino acid set, whilst the corresponding modified Seminario parameters lie between 22–178 kcal/mol/rad^2 (Fig. A.12).

3.4.2 Nitriles

Nitriles are a special case and the modified Seminario method cannot be used in the usual way. Let us consider the acetonitrile molecule shown in Fig.3.7. The vectors defining the bonds between C–N and C–C are parallel to one another. As these vectors are linearly dependent, a plane for the angle C–C–N cannot be defined. This is important as the modified Seminario method relies on u_N which is the normal vector to the angle's plane. Therefore, the angle

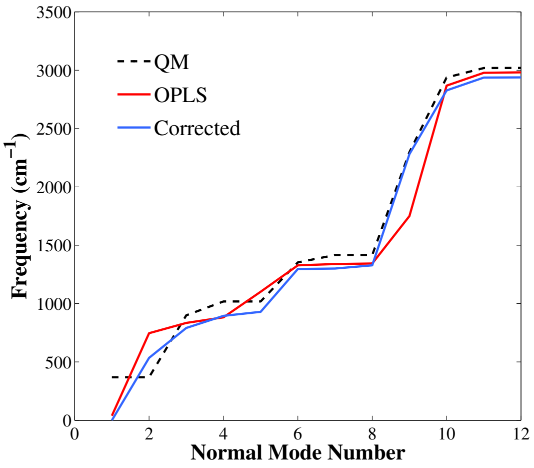


Fig. 3.8: The vibrational spectrum of acetonitrile computed using QM, and compared with the modified Seminario method and OPLS.

force constant for nitriles will cause problems in both the original and modified Seminario method.

A solution to this problem is as follows. As u_N cannot be defined, we instead look to uniformly sample the force constant for motion in all directions. This can be carried out by calculating the force constant for a range of u_N vectors. We chose the u_N vectors to be uniformly distributed across the surface of a unit sphere. For the acetonitrile case, this resulted in angle force constant for C–C–N ranging from approximately 44 to 60 kcal/mol/rad^2, with a mean value of 52.7 kcal/mol/rad^2 after vibrational scaling.

The normal modes for the molecule are shown in Fig. 3.8. OPLS has an error of 21.79% (155.24 cm^{-1}) whilst the modified Seminario method has an error of 17.19% (117.23 cm^{-1}). This error is higher than other molecules tested but lower than that of OPLS. By examining how the error changed with the C–C–N angle force constant, it could be seen that issues with acetonitrile frequency recreation were not due to the C–C–N angle force constant. Therefore, we employ the method outlined in this section for molecules with linear groups.

3.5 Discussion and Conclusion

We have developed a method for the parametrization of harmonic bond and angle force constants for molecular mechanics force fields. The method recreates QM normal mode frequencies with a consistently high level of accuracy, and uses no empirical or MM data in the parametrization process. Use of our bonded parameters results in similar levels of error to standard force fields for a general set of small molecules and dipeptides, and a noticable improvement for heterocyclic molecules. In certain cases, the optimized structures of hetereocyclic molecules are greatly improved using the new approach. The parameters have been computed using a modified version of the widely used Seminario method,[103] in which critical improvements have been made to the angle force constants. For the majority of the 70 molecules tested, the modified Seminario method is more accurate than the original approach.

Although the accuracy of the modified Seminario method is extremely good, Fig 3.6 reveals possible areas for further improvement. All methods tested show quite large errors in the very high frequency bond stretching modes involving hydrogen (although these modes are unlikely to critically affect many computed properties of interest). The modified Seminario method substantially improves the recreation of intermediate modes involving angle-bending motions. It should be emphasized that we do not claim that eq 3.12 is the only method for partitioning k_θ parameters from the full QM Hessian matrix, and other schemes are possible. In fact, we investigated one such scheme during development of the modified Seminario method. Motivated by the observation that the original Seminario method strongly over-estimates the stiffness of larger molecules, we investigated a simple re-scaling of the angle force constants by a constant multiplicative factor. This method gave percentage errors of around 8.5%, which is an improvement over the original Seminario method, but significantly worse than our modified approach, which accounts more rigorously for the molecular environment. A brief summary of the scaling method is given in the final section of this chapter.

With regards to the low frequency portion of the vibrational spectra, we have combined the derived harmonic bond and angle parameters with torsional and non-bonded parameters from standard MM force fields. Figure 3.6 is typical of the vibrational spectra computed in this study and shows that the errors in the low frequency part of the spectrum are low. Nevertheless, further improvements in accuracy could be possible by re-parameterizing the torsional terms using the modified Seminario bond and angle parameters, this is carried out in the following chapters. Finally, we have assumed throughout this study that the QM normal modes and frequencies are an accurate representation of experimental values. This is reliant on the vibrational scaling factors used being suitable for the molecules tested, and not being frequency dependent.[50]

The Seminario method is one of a number of methods that can be used to parameterize harmonic bond and angle force field terms.[11,24,123] The level of accuracy that can be obtained by fitting MM parameters to the QM Hessian matrix has been previously reported as 63.9 cm^{-1} or 73.3 cm^{-1} depending on the details of the fitting procedure.[11,123] These methods appear to be less accurate than the modified Seminario method, though they have only been tested on a small number of molecules, and further testing should ideally be carried out on equivalent data sets with identical error analysis. As well as potentially improved accuracy, the modified Seminario method also has other clear advantages compared to all fitting methods currently in use. Since the force field parameters are derived directly from the QM Hessian matrix, initial estimates of the remaining force field parameters are not required, and interdependencies between the different components of the force field are avoided. Reduction of parameter interdependencies is desirable to prevent the need for several iterations of fitting, and therefore to produce the most efficient parametrization schemes. Recent efforts to improve biomolecular force fields have seen a number of groups reparametrize the bond and angle components of proteins using fitting approaches.[24,122] Therefore, each new iteration of these force fields will require a full reparametrization of the bonded terms. We offer a simple, alternative solution by supplying a library of bond and angle parameters for the set of twenty naturally-occurring amino acids. These parameters can then be used as the basis for any future protein force fields that employ the standard harmonic functional form for bond and angle terms, as well as our own biomolecular force field.

The modified Seminario method has been implemented in a freely available program and offers a means to parameterize bond and angle terms in a fast and automated way. Future work will aim to combine the modified Seminario method with automated fitting of torsional parameters, using the methods described in the following chapter. Developments such as these will be crucial in creating an automated workflow for the accurate parametrization of MM force fields directly from QM data.[22,68,69] This is useful for applications such as parameterizing ligands for free energy calculations of protein-ligand systems.

3.6 Scaled Seminario Results

In this section we briefly discuss an alternative modification that can be made to the Seminario method to improve the normal mode frequency recreation. This simply involves multiplying all angle force constants by a constant value:

$$k_\theta^{Scaled} = \gamma k_\theta^{Seminario} \tag{3.13}$$

where γ is a constant to be determined.

To find the value of γ, the mean error in the normal mode frequencies was calculated for various values of γ for a set of 10 small molecules and the results are shown in Fig 3.9. The minimum error is 72.94 cm^{-1} for $\gamma = 0.68$, compared to 118.44 cm^{-1} when the scaling constant is equal to one.

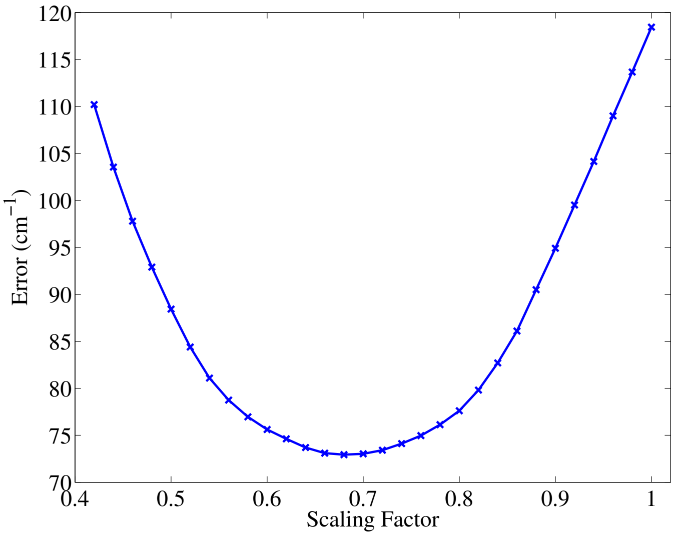


Fig. 3.9: The change in the error in the normal mode frequency recreation with γ.

The optimal value of γ value was tested on a further 20 molecules that were not in the original test set with the results presented in Table 3.3. There is a clear improvement in the results when a scaling constant is used. However, with a mean error of 6.50% the modified Seminario method, described above, outperforms the scaled Seminario values with error 8.48%.

It is interesting to see how the γ value relates to the correction from the modified Seminario method. Figure 3.10 shows the original Seminario parameters against the modified Seminario parameters for the dipeptide set. From this figure, and the error results, it can be seen that applying just one scaling factor is not a sufficient correction as three sets of points can be clearly seen, each set corresponding to a different value of γ. This is not surprising as a range of geometries are present in dipeptides, with planar groups such as phenyl requiring different optimal scaling factors from non-planar groups such as methyl.

However, using a scaling constant may be sufficient for certain applications. It offers existing codes that use the Seminario method a very simple way to improve normal mode frequency recreation with negligible additional coding requirement. If this were carried out, further testing of the most appropriate scaling constant, and the normal mode frequency error would be advised.

	OPLS (%)	Original Seminario (%)	Scaling 0.68 Seminario (%)	Modified Seminario (%)
A-methylstyrene	8.63 (46.95)	15.94 (133.74)	9.16 (56.89)	8.00 (41.18)
Acetaldehyde	5.44 (69.46)	8.08 (102.01)	6.35 (89.16)	4.91 (69.19)
Acetamide	15.46 (82.34)	21.76 (127.61)	16.55 (80.71)	13.70 (63.34)
Acetone	6.17 (48.49)	10.85 (100.74)	8.81 (89.65)	7.02 (67.12)
Aniline	6.40 (69.80)	15.70 (159.46)	8.60 (76.64)	4.73 (39.83)
Anisole	4.65 (40.04)	12.37 (133.92)	5.67 (50.18)	3.90 (36.26)
Benzamide	6.32 (58.83)	15.49 (148.49)	8.76 (72.38)	5.19 (36.68)
Dimethylamine	6.01 (76.29)	9.83 (124.73)	5.67 (73.94)	5.16 (66.50)
Dimethylaniline	6.64 (42.27)	10.53 (122.23)	4.66 (46.32)	4.96 (43.25)
Dimethylether	8.01 (94.84)	10.04 (131.24)	5.96 (80.02)	5.67 (71.71)
Dimethylsulfide	4.32 (45.40)	8.22 (67.11)	11.48 (115.65)	9.06 (76.35)
Ethane	4.47 (55.92)	5.59 (84.69)	7.67 (111.77)	4.61 (69.91)
Methyl_benzoate	5.15 (37.47)	12.21 (121.42)	5.87 (47.20)	5.06 (36.32)
Methylacetamide	8.54 (78.79)	11.94 (130.73)	9.50 (112.17)	6.54 (86.30)
Methylamine	6.62 (81.13)	7.19 (94.94)	7.14 (94.90)	3.79 (51.64)
Phosphorine	7.26 (66.06)	13.70 (145.41)	7.31 (70.95)	5.02 (41.65)
Propane	4.05 (45.52)	7.05 (92.03)	6.92 (94.70)	4.31 (58.54)
Propene	5.41 (52.47)	12.64 (129.80)	7.42 (77.74)	6.13 (66.51)
Trifluorobenzene	7.49 (76.61)	13.02 (131.72)	7.12 (69.10)	5.55 (42.93)
Trifluorotoluene	18.52 (58.58)	24.74 (130.17)	18.95 (65.50)	16.67 (38.47)
Mean	7.28 (61.36)	12.34 (120.61)	8.48 (78.78)	6.50 (55.18)

Table 3.3: The mean percentage error between the QM and MM normal mode frequencies for a set of small molecules using OPLS, the original Seminario, the modified Seminario and the Seminario method with angle force constants scaled by 0.68. The MUE is shown in brackets (cm^{-1}).

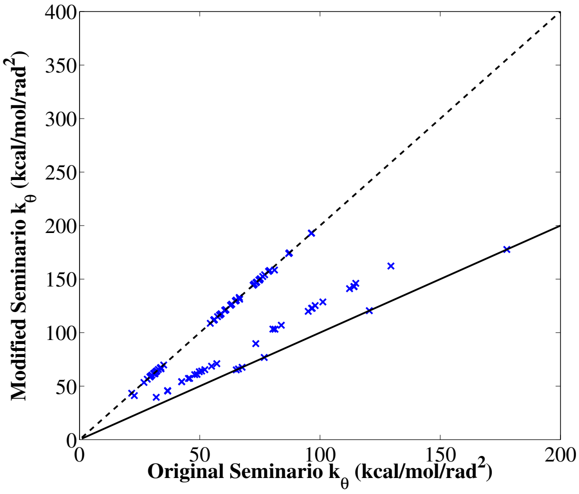


Fig. 3.10: The modified Seminario parameters plotted against the original Seminario parameters. Lines at $y = x$ and $y = 2x$ are also plotted with $y = x$ corresponding to angles with no neighbors and $y = 2x$ corresponding to planar molecules.

Chapter 4

Torsional Parameters

With a set of bond and angle terms derived for the QUBE force field, the torsional parameters remain the only component left to calculate. The torsional component helps to dictate the conformational preferences of a molecule and is a function of a molecule's dihedral angles. The dihedral angle is the angle between two planes, with each plane containing three atoms present in the molecule. The reparameterization of torsional parameters has been shown to be a particularly crucial step in improving the accuracy of force fields, as demonstrated for AMBER ff15ipq, CHARMM36 and OPLS-AA/M.[7,24,99] In Ref. 22, system-specific non-bonded terms were used in combination with standard OPLS-AA bonded parameters. The use of the OPLS parameters limits the accuracy of the force field due to the dependence of the optimal torsional parameters on the non-bonded components of the force field. This section therefore focuses on re-fitting the key torsional parameters that describe an amino acid's backbone and sidechain dynamics.

With the torsional terms reparametrized, the system-specific protein force field is complete. The new torsional parameters are combined with QUBE non-bonded terms derived directly from the electron density of the proteins and peptides studied using linear scaling DFT[22] and bond and angle terms from our modified Seminario dipeptide parameter library. The complete force field is then tested via MD simulations of a variety of systems to investigate the validity of using our parametrization approach. This is the first time a system-specific biomolecular force field has been fully developed and tested.

The methods and validation tests used for the QUBE force field broadly follow the approaches employed in the development of OPLS-AA/M, the latest OPLS force field.[99] Multiple QM energy scans of backbone and sidechain dihedral angles of dipeptides are used to fit torsional parameters by minimizing the difference between QM and MM energy scans. The system-specific nature of our force field adds complications to the torsional parameter fitting process as we do not use atom typing, the energy difference corrected for by the torsional parameters will vary. Regularization is used in an attempt to address this issue.

As well as parametrizing the torsional components, software was required to translate a set of QUBE parameters into a format which could be readily used with existing MD software. The necessary scripts to convert the DFT output to a CHARMM file format are therefore supplied at *https://github.com/aa840/QUBEMAKER*, along with a tutorial. This is to facilitate future investigation of system-specific force fields.

In order to test the performance of the QUBE force field, simulations of dipeptides, Gly_3, Ala_5 and five different proteins were performed. This is similar to the testing carried out when developing protein force fields such as AMBER ff15ipq, AMOEBA, CHARMM36 or OPLS-AA/M.[7,24,99,104]

4.1 Theory

The functional form of standard biomolecular force fields has five components. Covalent interactions between atoms are modelled using harmonic bond stretching and angle bending parameters, while rotations about a bond are described by anharmonic 4-body torsional terms. Non-bonded interactions are described by a sum of Coulombic interactions between (usually) atom-centered point charges and a physically-motivated Lennard-Jones interaction. In what follows, we give an overview of how these various components are parameterized in the QUBE protein force field, and we contrast these approaches to those used in standard transferable force fields. This is to supplement the information given in Section 2.3.

4.1.1 Non-bonded Parameters

The non-bonded components of a molecular mechanics force field aim to describe the quantum mechanical electrostatic, dispersion and exchange-repulsion interactions.[108] The latter two interactions are approximated by a Lennard-Jones potential, which combines a short-range repulsive r^{-12} potential with a longer-range attractive r^{-6} interaction. The charge parameters are generally fit to the quantum mechanical electrostatic potential of small molecules. The Lennard-Jones parameters are then fit to reproduce experimental data, such as liquid densities and heats of vaporization.[15,24,54] However, neither of these parameterization strategies scale to large molecules, such as proteins. Therefore, in standard force fields, parameters are fit to the properties of small molecules and then transferred to larger molecules using assigned atom types.

The aim of QUBE is to move away from the requirement for transferable force field parameters, and instead to derive parameters for individual atoms based on quantum mechanical calculations. To achieve this goal, a different approach to non-bonded parameter fitting is required. The basis of the QUBE approach is atoms-in-molecule (AIM) partitioning of the total QM electron density as described in Section 2.3.2.[22] First, a QM simulation of the molecule under study is performed. From the output of the QM calculation, the total electron density of the molecule is partitioned onto individual atoms using an AIM weighting scheme. There is no unique method to perform this partitioning, but we favor the density derived electrostatic and chemical (DDEC) scheme,[80,81] which is a combination of the iterative stockholder atoms and iterative Hirshfeld approaches.[10,73] With the electron density partitioned to individual atoms, the atom-centered charges can be simply found by integrating the electron density over all space (and adding the nuclear charge). The DDEC approach has been implemented in the linear-scaling density functional theory (DFT) software, ONETEP, which allows us to perform QM calculations of, and assign parameters to, systems comprising thousands of atoms, including entire proteins. The derived charges have been shown to be suitable for use in flexible force field design in multiple works.[22,81] The charges are specific to the system under study and, by performing the QM calculation in the presence of an implicit solvent model, polarization of the charges in the condensed phase can be included in the model.[22]

The dispersion and exchange-repulsion interactions are described using a Lennard-Jones potential with a form $\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}$. The dispersion coefficient, B_{ij}, is calculated from the partitioned electron density by using the Tkatchenko-Scheffler relationship to rescale the free atom dispersion coefficient by the computed volume of the atom in the molecule.[22,111] The standard combination rule, $B_{ij} = \sqrt{B_i B_j}$, can then be used to determine heteroatomic dispersion coefficients. The A_{ij} parameter, which describes the short-range repulsion between overlapping electron clouds, cannot be calculated directly from the electron density. Instead it is computed by requiring that the minimum in the interatomic Lennard-Jones potential coincides with the estimated van der Waals radius of the atom in the molecule.[22] This non-bonded parameter derivation scheme requires just one fitting parameters per atom (corresponding to the van der Waals radii of a free atom in vacuum). Further details of this procedure are given in Section 2.3.

4.1.2 Bond and Angles

The absence of interdependencies between the computed bond and angle parameters and the other terms in the force field means parameters derived using the modified Seminario method, as discussed in Section 3, are suitable for use in the QUBE protein force field. The modified Seminario method could be used to derive bespoke bond and angle parameters for entire proteins, but the computation of the full QM Hessian matrix for large molecules is currently infeasible. Since large-scale polarization effects are expected to be significantly less important for bond and angle parameters than charges, we use the library of bonded parameters derived in the previous chapter[1] (and the same atom types as those used for OPLS-AA/M[99]).

Whilst preparing the protein simulations, it was found that using just dipeptide parameters led to missing terms for the disulfide bridge between pairs of cysteine residues. These bond and angle parameters were therefore derived using the QM Hessian matrix of dimethyl disulfide.

4.1.3 Torsional Parameters

With the force field non-bonded parameters from atoms-in-molecule electron density partitioning,[22] and bond and angle parameters from the QM Hessian matrix,[1] all that remains to complete the QUBE protein force field is to obtain the torsional parameters. Unfortunately, it is infeasible to derive torsional parameters from QM simulations that are specific to each protein. Therefore, a library of bonded parameters for the QUBE protein force field is instead created. We make the assumption that the derived parameters for dipeptides are transferable to proteins, and in Section 4.3 we test the limitations of this assumption by validating the force field against experimental peptide and protein dynamical observables.

The torsional terms in a force field are a function of the dihedral angles (ϕ) in a molecule. Here, we use the same functional form as the OPLS force field:

$$V_{tors}(r) = \sum_i \frac{V_1^i}{2}(1+cos(\phi)) + \frac{V_2^i}{2}(1-cos(2\phi)) + \frac{V_3^i}{2}(1+cos(3\phi)) + \frac{V_4^i}{2}(1-cos(4\phi)) \tag{4.1}$$

where V_{1-4}^i are parameters to be fit. This work is based on the torsional fitting methods and QM benchmark data used in the fitting of the OPLS-AA/M force field with data supplied by M. J. Robertson (Yale University).[99] We focus here on reparameterizing the backbone (ϕ, ϕ', ψ and ψ') and sidechain (χ_1, χ_1', χ_2 and χ_2') torsional parameters, see Fig. 2.3. Re-fitting the remaining torsional parameters is beyond the scope of the current work, and these parameters are instead taken from the OPLS-AA/M force field.[99] However, parallel efforts are being made to develop a toolkit for automated parameterization of small molecules using the QUBE force field, which will facilitate derivation of the remaining parameters in future.[44]

When fitting torsional parameters, the main objective is to minimize the difference between MM and QM gas phase dihedral energy scans. However, weighting schemes and regularization can also be used to change the form of the error function that is minimized. The general form of the error function used in this study is given by:

$$\text{Error} = \sqrt{\frac{\sum_n (E_{MM} - E_{QM})^2 e^{-W/k_b T}}{n}} + \lambda \sum_i (V_i - V_i^0)^2 \tag{4.2}$$

where k_B is the Boltzmann constant, T is a weighting temperature, n is the number of points at which the energy is evaluated, W is the contribution from the weighting scheme, λ is the regularization coefficient, V_i is the torsional parameter being optimized and V_i^0 is an initial estimate of the torsional parameter. Where we have used regularization we have used $V_i^0 = 0$ as the initial guess as previously suggested.[116] The V_4 term was set to zero throughout the fitting procedure to avoid overfitting.[99]

Weighting schemes are used to prioritize accuracy in particular regions of the dihedral scan, for example in the β-sheet region of the Ramachandran plot. A range of weighting schemes has been previously used, including schemes that prioritize the lowest QM energies[99] or that prioritize regions that have been shown experimentally to be most populated by proteins.[7,104] Regularization is a technique generally used to prevent overfitting to data. As we will show, the use of regularization appears to be particularly important for the QUBE force field, where non-bonded parameters used in simulations are not necessarily the same as those used during fitting. The high λ values we use for certain torsional scans ensures that the dihedral parameters only correct for the most significant energy differences between MM and QM, and hence are less specific to the particular non-bonded parameter set used during fitting.

4.2 Methods

4.2.1 Torsional Parameter Fitting

Torsional parameter fitting followed the general strategy employed in the development of the OPLS-AA/M force field,[99] amongst others, in which parameters are fit to reproduce QM gas phase potential energy surfaces. Fitting and validation was performed using dipeptides of the form (Ace-X-NMe), where X is the amino acid, Ace is an acetyl group and NMe is the N-methyl group.

ONETEP Calculation Setup

Four nonorthogonal generalised Wannier functions (NGWFs) were used for all atoms with the exception of hydrogen which used one. The NGWFs had radii of 10 Bohr. The periodic sine (psinc) basis was used to describe the NGWFs, with a grid size employed that corresponds to a plane wave cutoff energy of 1020 eV. The PBE exchange-correlation functional was used with PBE OPIUM norm-conserving pseudopotentials.[92] Partitioning of the polarized ground-state electron density was performed using the DDEC scheme in ONETEP with the mixing parameter (γ) set to 0.02.

QUBE Parameter Derivation

The ground state electron densities of the dipeptides were computed using the ONETEP linear-scaling DFT code[106] and parameter settings given above.[22] Since the reference QM potential energy scans are performed in the gas phase, we have decided to derive QUBE force field charges and Lennard-Jones parameters from the vacuum electron density (rather than in an implicit solvent). The assumption here is that the required correction to the MM potential energy surface is approximately the same in the gas and condensed phases.

Charge and Lennard-Jones parameters were derived from the QM ground state electron density using the DDEC scheme[80,81] as implemented in the ONETEP code[68,69] (Section 4.1.1). To account for conformational dependence,[68] the non-bonded parameters were derived for multiple conformations of each dipeptide and averaged. Non-bonded parameters on identical atoms (for example, hydrogen atoms in a methyl group) were symmetrized. It should be noted that only atom-centered charges were used in this chapter, though off-site charges to model anisotropic electron density distributions have been developed for use with small molecules and this will be described further in Chapter 5. Bonded parameters were assigned to the dipeptides from the library developed with the modified Seminario method using OPLS-AA/M atom typing rules.[1]

Potential Energy Scans

The torsional potential energy scans of alanine, glycine and all sidechains are the same as those used in the development of the OPLS-AA/M force field, as described previously.[99] In brief, structures were relaxed in the gas phase using Gaussian 09 with a ωB97X-D functional and a 6-311++G(d,p) basis set. Dihedral angles were scanned in 15° increments from -180° to 180°. A single point energy calculation was then performed on the optimized structure using the double hybrid functional B2PLYP-D3BJ and the Dunning basis set aug-cc-pVTZ. A 2D scan of ϕ (C-N-C$_\alpha$-C) and ψ (N-C$_\alpha$-C-N) was carried out for alanine and glycine.

The sidechain energy scans follow the same methods as the backbone scans, except that the single point B2PLYP calculation was not performed for all χ_2 scans, as ωB97X-D was shown to give sufficiently accurate results.[99] These 1D scans give the energy as a function of the χ_1 dihedral angle (N-C$_\alpha$-C$_\beta$-X$_\gamma$) or the χ_2 dihedral angle (C$_\alpha$-C$_\beta$-X$_\gamma$-Y$_\delta$). The ψ and ϕ values were constrained to their values in an α helical ($\phi = -60^\circ, \psi = -45^\circ$) or a β-sheet ($\phi = -135^\circ, \psi = 135^\circ$) conformation. All scans used in this work can be found in the Supporting Information for Ref. 99.

In this work, an additional 2D scan of the ϕ and ψ dihedral angles of serine was found to be necessary for accurate torsional parameters for serine and threonine, which both have a polar oxygen atom at the X$_\gamma$ position. This followed similar protocols to those previously described, however the sidechain χ_1 angle now had to be taken into account. Scans were performed at 30° increments of ϕ/ψ , for χ_1 initially set to -60°, 60° and 180°. This gave three 2D energy scans for the main rotamers of serine. The minimum energy structure for each ϕ/ψ angle was then used to construct the overall minimum ϕ/ψ potential energy surface, see Figure 4.1.

Fitting Dipeptide Torsional Parameters

Torsional parameters were optimized by minimizing the error function shown in equation 4.2 using a steepest descent algorithm. MM potential energy surfaces were computed by scanning dihedral angles in 15° increments using the BOSS software.[55] The backbone torsional parameters for all dipeptides tested, excluding serine and threonine, were fit to the alanine and glycine scans previously described. The total error for the two scans was given by:

$$\text{Error}_{\text{Total}} = 0.928 \times \text{Error}_{\text{Ala}} + 0.072 \times \text{Error}_{\text{Gly}} \tag{4.3}$$

with the prefactors corresponding to the relative frequency of each amino acid in the human proteome. Preliminary testing, which is detailed in Appendix B.1, showed that a weighting function and regularization did not improve the configurations sampled during the dipeptide MD simulations and so were not used ($\lambda = 0$, $W = 0$).

Both QM energies and experimental Ramachadran plots were tested as weighting functions. The plots shown in Fig. 4.2 are produced using data given from Ref. 110 and give the probability density as a function of the ϕ and ψ dihedral angles of the residue. This probability density is neighbor dependent and the plots shown are when all neighboring residue types are used. These densities are derived from X-ray crystallography data of proteins.

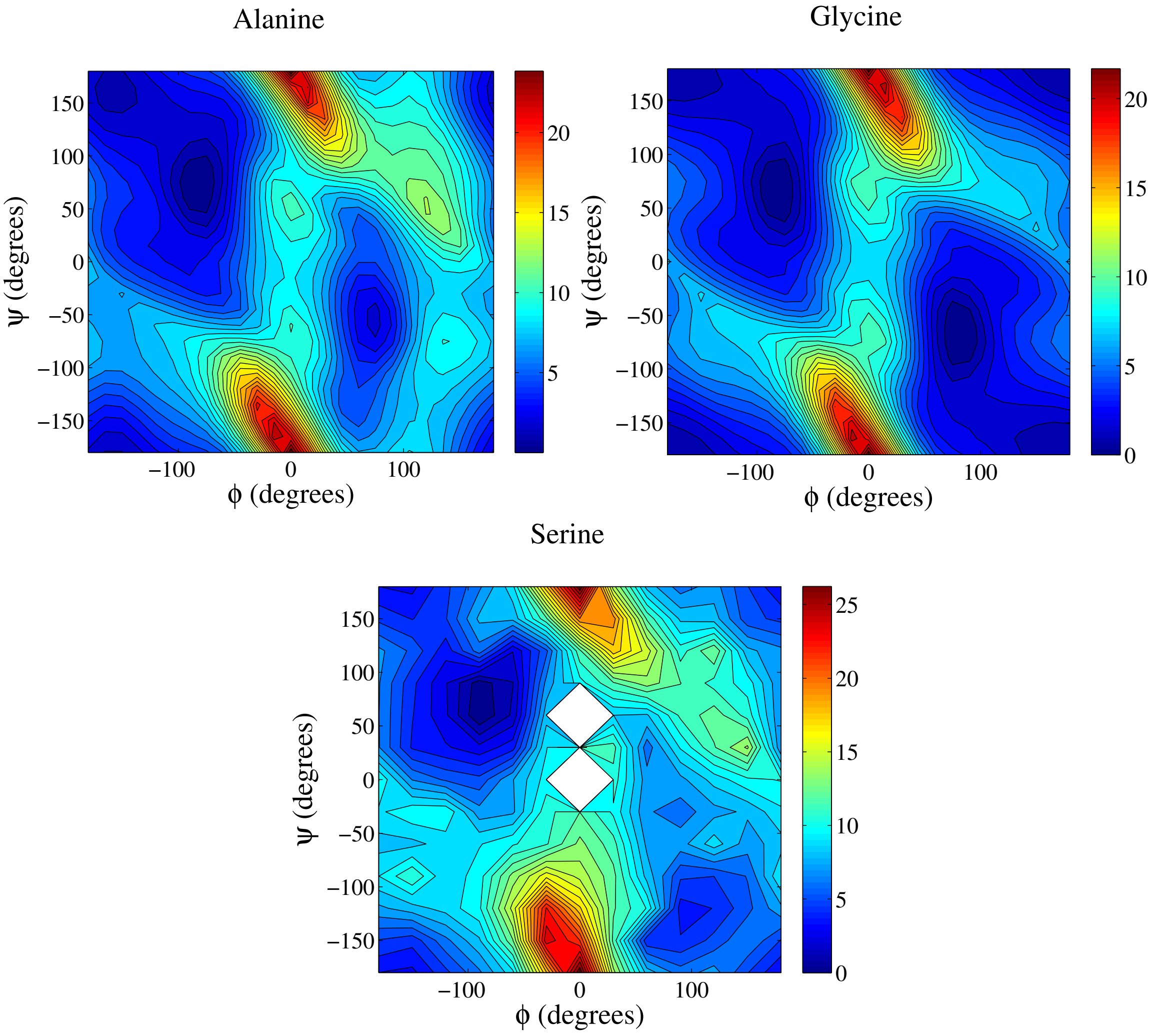
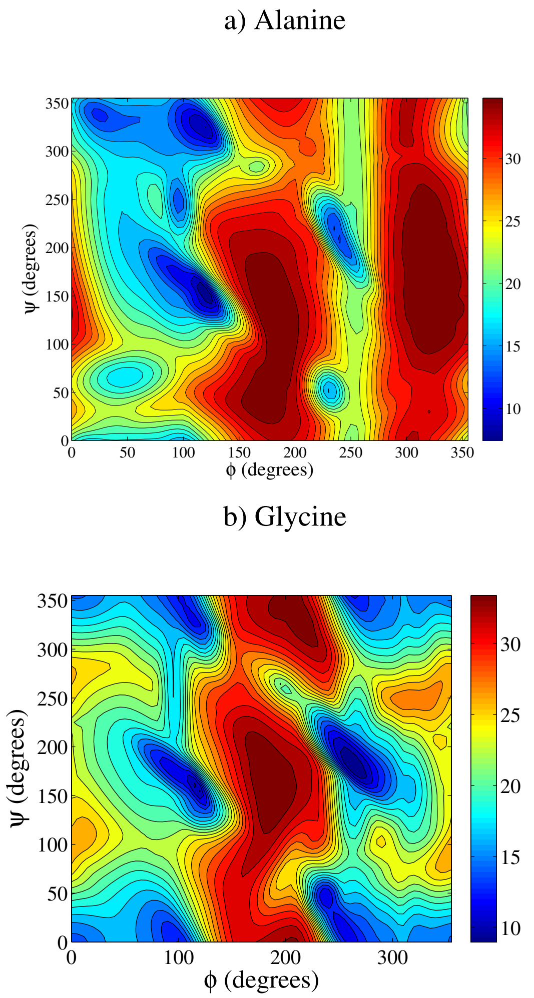


Fig. 4.1: The 2D energy scan of ψ and ϕ dihedral angles for a set of dipeptides. For serine the minimum value of the three sidechain scans is shown. The white areas correspond to regions with unavailable data due to convergence issues with the QM calculations.


a) Alanine

b) Glycine

Fig. 4.2: The probability density, $-log(p_{\psi,\phi})$, of ψ and ϕ dihedral angles of a) alanine and b) glycine residues in protein structures. The data used is from Ref. 110.

The remaining dipeptides, threonine and serine (both of which contain aliphatic hydroxyl groups in their sidechain), were assigned identical backbone parameters that were parameterized to fit the QM scans of serine. For these scans regularization and weighting were shown to be necessary to produce dipeptide dynamics in agreement with experiment. An investigation of how the simulation error changes with regularization is given in Table B.4 in the Appendix. The regularization was set to $\lambda = 0.05$.

The sidechain scans for all dipeptides followed the same fitting process as the alanine/glycine backbone with no weighting or regularization used. As atom-typing is not used for the non-bonded parameter assignment in the QUBE force field, each set of sidechain torsional parameters is also residue-specific. This differs from the approach used in OPLS-AA/M in which sidechain torsional parameters with the same set of atom types are generally assigned the same parameters.

In a number of cases it was necessary to make manual changes to the fitting process. This was restricted to setting a number of torsion parameters to zero (that is, $\lambda = \infty$) and reducing the number of scans used in the fitting process. In particular, the asparagine ϕ/ψ distribution was improved by setting the χ_2 torsional parameters to zero. Additionally, using just the QM energy scan with the lowest minimum energy in the fitting process was shown to result in an improvement in the MD simulations of the dipeptides for the χ_1 torsional parameters of cysteine, methionine, serine and threonine. The need for manual input in the fitting process was also required for developing OPLS-AA/M[99] and is likely due to the restrictive functional form of the torsional potential and the conformational dependence of the energy scans. The full set of manual changes involved are listed with the torsional parameters in Section B.2 in the Appendix.

Torsional parameters, and improper terms, that have not been reparametrized in this work are the same as those used in OPLS-AA/M .

4.2.2 MD Simulations

Alanine Pentapeptide and Glycine Tripeptide

As the non-bonded parameters used are specific to the system under study, they are not the same for an alanine dipeptide molecule as for the alanine pentapeptide (Ala$_5$). The alanine residue in the dipeptide is blocked by acetyl and N-methyl groups whereas the central three alanine residues in Ala$_5$ have neighboring alanine residues on both sides. Therefore, varying environments exist for alanine residues in the different molecules. Consequently, the parameters found for the alanine dipeptide were found to be unsuitable for MD simulations of Ala$_5$ (Table B.6). Fortunately, the use of regularization in the fitting process resulted in torsional parameters that were sufficiently accurate. Alanine and glycine backbone torsional parameters were refit to the QM energy scans with $\lambda = 0.50$, no weighting was used. The λ value used for the regularization were found by minimizing the differences between simulated and J Coupling for Ala$_5$ (Table B.6). Separate parameters are used for alanine pentapeptide and glycine tripeptide (Gly$_3$), as residue-specific parameters should result in a more accurate force field.

Protein Torsional Parameters

It is expected that optimal backbone torsion parameters for protein simulations are more similar to those developed for Ala$_5$ and Gly$_3$ than for the set of dipeptides. We therefore use

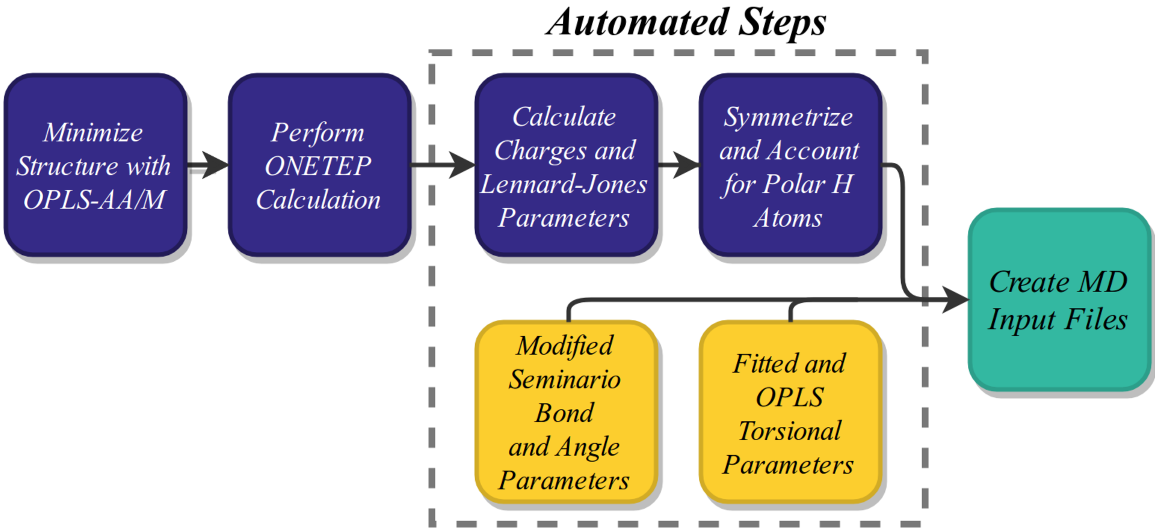

Automated Steps
Minimize Structure with OPLS-AA/M
Perform ONETEP Calculation
Calculate Charges and Lennard-Jones Parameters
Symmetrize and Account for Polar H Atoms
Modified Seminario Bond and Angle Parameters
Fitted and OPLS Torsional Parameters
Create MD Input Files

Fig. 4.3: A flowchart illustrating the steps required to assign the QUBE protein force field. Blue is used for non-bonded terms, yellow is used for bonded terms.

a regularization $\lambda = 0.50$ for all protein backbone torsional parameter derivation. Alanine, glycine, serine, and proline torsional parameters are fit to available QM potential energy surfaces and are therefore residue-specific. Threonine uses torsional parameters fit to the serine torsional scan, and all other amino acid use torsional parameters fit to joint alanine/glycine energy scans. These backbone parameters are combined with the dipeptide sidechain torsional parameters to give the full QUBE protein force field parameter set.

Bond and Angle Parameters for Cysteine

The S–S bond parameters and S–S–CT angle parameters needed to describe the disulphide bridge between two cysteine residues were derived from dimethyl disulfide. This method follows the same approach as used in Chapter 3. A structural optimization and frequency calculation was performed using Gaussian 09 with the ωB97XD functional and 6-311++G(d,p) basis set. The QM Hessian was extracted from the Gaussian output files and the modified Seminario method used to find the bond and angle terms. A scaling factor of $(0.957)^2$ was used to scale the bond and angle terms. The terms calculated were S–S, a force constant of 146.29 kcal/mol/Å^2 and bond length of 2.070Å, and S–S–CT, a force constant 99.50 kcal/mol/rad^2 and equilibrium angle of 102.05°.

Protein MD Simulations

Following a number of previous force field validation studies,[7,24,99,104] the QUBE force field was validated by molecular dynamics (MD) simulations of a benchmark test set of five proteins. The structures chosen (with the PDB codes shown in bracket) were ubiquitin (1UBQ), GB3 (1P7E), BPTI (5PTI), binase (1BUJ) and a villin headpiece subdomain (2F4K).

Figure 4.3 shows the steps required to set up a QUBE protein force field for a MD simulation. As in Section 4.1.1, the ONETEP linear-scaling DFT software is used to compute the ground state electron density of the five proteins, and assign the charge and Lennard-Jones parameters from the partitioned atomic electron densities. Consistent with the QUBE small molecule approach, every atom in the protein is assigned bespoke non-bonded parameters derived from the quantum mechanical electron density. To model polarization effects in the condensed phase, the electron density is computed using an implicit solvent model[29,30] with a dielectric constant of 80. For the set of proteins, the non-bonded parameters are then set halfway between their vacuum and condensed phase values, applying the iPol method described in Chapter 2. This method was chosen as it has previously been used in the AMBER ff14ipq and ff15ipq biomolecular force fields.[24] Preliminary work investigated using a dielectric of $\varepsilon = 80$, $\varepsilon = 10$ and $\varepsilon = 4$, but there was little change to the protein dynamics. Typical computational requirements for the protein ONETEP calculations are approximately 2000 cpuhrs. The sensitivity of the non-bonded parameters to the choice of input structure is investigated in Section 4.3.5 for the GB3 structure. In order to provide a consistent and computationally efficient approach to assigning the non-bonded parameters, we recommend minimizing the experimental structure using a standard transferable force field in explicit water prior to the DFT calculation. In this study, we used the OPLS-AA/M force field for the initial minimization.

Following non-bonded parameter assignment, bond, angle and torsion parameters were assigned based on the OPLS-AA/M atom types. For torsion types not re-parameterized in this study, OPLS-AA/M parameters are retained.[58,99] All parameters, including atom-specific non-bonded parameters, are written to a CHARMM-style parameter file. Preparation of the parameter files is fully automated, and scripts and step-by-step tutorials are available. MD simulations were performed using the NAMD software, using input parameters detailed in Section 4.2.2. The necessary scripts to convert the DFT output to a CHARMM file format are supplied at *https://github.com/aa840/QUBEMAKER*, along with a tutorial.

Statistics were collected over a period of 200 ns for dipeptides and Ala_5 and Gly_3, and 500 ns for the proteins, and three simulations of this length were performed for each peptide and protein to enhance the sampling of configuration space.

MD Simulation Properties

Simulations of the dipeptides, the alanine and glycine peptides, and the protein structures were all carried out with the following protocol. The molecules were placed in a water box, with the total length ranging from 25 Å, for small dipeptides, to 62 Å, for the largest protein. The TIP3P force field was used to model water molecules.[57,79] The charge of the system was then neutralized with chlorine and sodium ions at a salt concentration of around 150 mM. The temperature was set to 300 K and the pressure to 1 atm. These parameters were maintained throughout the simulation by using a Nose-Hoover Langevin piston barostat. The piston period was set to 100 fs and the damping timescale was 500 fs. The Langevin thermostat had a damping coefficient of 1 ps^{-1}. Long ranged electrostatics were treated with particle mesh Ewald. The cutoff for non-bonded terms was set at 11 Å with the smoothing function starting at 9 Å.

The system was first equilibrated for 5 ns using a timestep of 2 fs. The simulation was then run for a further 200 ns, again using a timestep of 2 fs with use of the SHAKE and SETTLE algorithms. These properties are the same as those used in Ref. 99. The protein simulations employed a 400 ps equilibration period in which the system was gradually heated to 300 K. This heating period and the 5ns of equilibration were not used in the subsequent analysis of the simulation.

The proteins were prepared using the automated psf generator, which is incorporated into VMD, and the OPLS-AA/M topology files. The psf generator added missing hydrogen atoms to the structures, as well as creating pdb files and psf files with OPLS-AA/M parameters. VMD plugins were also used to solvate the proteins and add the appropriate number of sodium and chlorine atoms. The psf file created using VMD was then used as an input to our program QUBEMAKER. This program produces a new psf with the QUBE charge parameters along with an additional NAMD input file which contains all other necessary force field parameters. The files produced at each stage were manually checked to ensure no mistakes were presence.

Simulation Analysis

The backbone and sidechain dihedral angles sampled during the dipeptide simulations were analyzed and compared with experimental data. The sidechains sampled were separated into p(+60°), t(180°) and m(-60°) rotamers and the populations of each were then compared to protein coil library data.[99] The backbone J coupling term $^3J(H_N,H_\alpha)$ was calculated from the dipeptide simulations using the Karplus parameters proposed in Ref. 45 and the ϕ dihedral angles sampled, with the experimental J coupling values taken from Ref. 4. J coupling occurs due to an atom's nuclear spin perturbing the spin of electrons nearby, which in turn perturbs the nuclear spins of other atoms. J couplings can be found experimentally using the NMR spectrum of a molecule and from simulations by the Karplus equations, which link the dihedral angles sampled by the molecule to the J coupling. The form of the Karplus equations is:

$$J(\theta) = Acos(2\theta) + Bcos(\theta) + C \tag{4.4}$$

where A, B, C are the Karplus parameters and θ is the dihedral angle.

The J coupling values from the alanine and glycine simulations could be compared to multiple experimental values given in Ref. 34. Three separate Karplus parameters sets were used and are those used in Ref. 6 and Ref. 99, with the measurement of error also being the same:

$$\chi^2 = N^{-1} \sum_{j=1}^{N} (\langle J_j \rangle_{sim} - J_{j,exp})^2 / \sigma_j^2 \tag{4.5}$$

where N is the total number of J coupling parameters, $\langle J_j \rangle_{sim}$ is the average J coupling in the simulation, $J_{j,exp}$ is the experimental J coupling, and σ_j is the systematic error in the Karplus equations.

The Karplus parameters used for the protein simulations were taken from multiple sources with both Ref. 45 and Ref. 119 used for backbone Karplus parameters. Methyl sidechain Karplus parameters are also supplied in Ref. 119, and Ref. 93 was used for all other sidechain Karplus parameters. Experimental J coupling terms for ubiquitin and GB3 were taken from a number of sources.[18,45,75,93,119]

J Coupling Analysis

J coupling values are commonly used to validate the accuracy of force fields.[24,99] However, if they are used in isolation as a validation test (without ϕ/ψ distribution analysis) results can be misleading. For example, the J Coupling value for the left-handed α-helical conformation calculated using the Karplus parameters is close to the experimental J Coupling for the dipeptides. Therefore, a low J coupling error for the simulation can occur if there is a high proportion of left-handed α-helical conformations. However, experimental results show that left handed helical populations should remain low,[46] and this demonstrates the need to analyze the ϕ/ψ distribution. Additionally, the experimental J coupling value is similar for all the dipeptides in the set. Therefore, similar ϕ/ψ distributions for the set of dipeptides could result in low J coupling errors. However, experimental measurements indicate that this is not the case and different dipeptides occupy varying ϕ/ψ regions.[35] This points to additional issues with using the J coupling values as a means to analyze a force field's accuracy.

Additionally, there may be problems with using the Karplus parameters with glycine, as previously suggested in Ref. 99. To demonstrate the unsuitability of these Karplus parameters, ${}^2J(N,C_\alpha)$ as a function of the dihedral angle is shown in Fig. 4.4. The dashed red line is the experimental J coupling, and it can be seen that no dihedral angle can recreate the experimental measurement. This issue has also been previously discussed in Ref. 88.

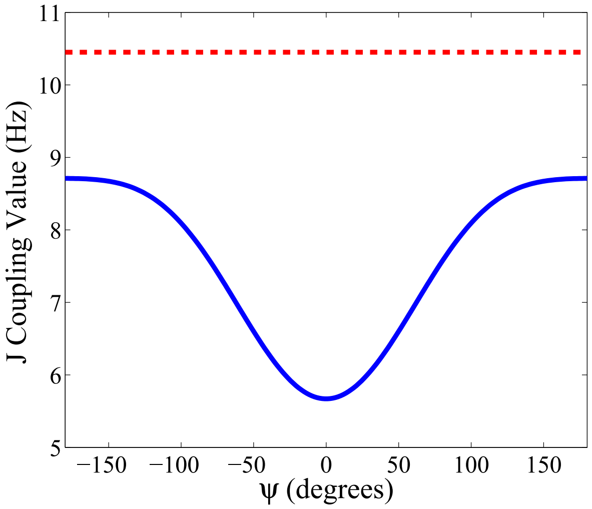


Fig. 4.4: The J coupling value for ${}^2J(N,C_\alpha)$ as a function of ψ is shown in blue. The red dashed line shows the experimental value.

In addition to the issues with the glycine J coupling values discussed above, the second set of Karplus parameters used in Ref. 6 have also been problematic for a variety of force fields.[24,99] Simulations of Ala$_5$ struggle to recreate the experimental ${}^2J(N,C_\alpha)$ coupling with the second set of Karplus parameters. Figure 4.5 shows the J coupling for all three sets of Karplus parameters along with the experimental value. From this figure, the conformation populations that would result in a low J coupling error can be estimated. For the first set of Karplus parameters a high PPII ($\phi = -75.0°$, $\psi = 145°$) conformation would result in a low error. For the second and third set a sizeable β population would be required along with the PPII conformation, however this would negatively impact other J coupling terms. For instance, with the second set of Karplus parameters ${}^3J(C',C')$ has an error of 160 Hz at $\phi = -135°$ (the β conformation). This points to problems with the Karplus parameters used for ${}^2J(N,C_\alpha)$ as also discussed in Ref. 34. It is useful to consider the issues discussed in this section when analyzing the results of the simulations.

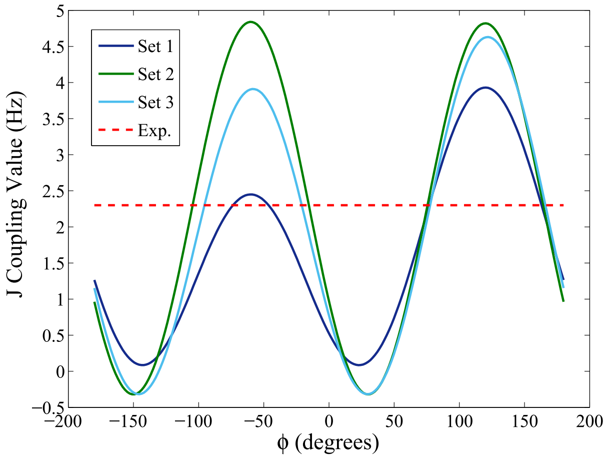


Fig. 4.5: The J coupling value for $^{2}J(N,C_{\alpha})$ as a function of ψ is shown in blue. The red dashed line shows the experimental value.

4.3 Results

4.3.1 Torsional Parameter Fitting

The backbone torsional parameters and associated errors in the recreation of the QM energy scans are given in the Appendix B.2. For the alanine and glycine scans, the error for the QUBE force field is 1.25 kcal/mol compared to 0.93 kcal/mol for OPLS-AA/M, which is a reasonable level of agreement. For proline and serine, the errors remain comparable to OPLS-AA/M.

Although a low error is clearly the desired result, it does not necessarily correspond to accurate non-bonded parameters. The degree to which torsional parameters can improve the fit between MM and QM scans depends not only on the accuracy of the non-bonded and bond and angle parameters, but also on the shape of the energy difference between the QM and MM scans. The functional form used in classical MM force fields is very restrictive. However, the energy difference between the QM and MM energy scans must be recreated by the functional form for low errors to be achieved. Therefore, although we use errors as a guide to the performance they cannot be relied upon as a measure of accuracy of a force field.

For the sidechain torsion parameters, the mean value of the error in the recreation of the QM energy scans for our force field is 1.29 kcal/mol compared to 1.12 kcal/mol for OPLS-AA/M. Particularly high errors occur for both the χ_1 and χ_2 glutamate scans and glutamine's χ_1 term. For glutamate, the error was also high for OPLS-AA/M but the rotamer populations remained close to the experimental values and this may be due to a problem with the functional form used in classical force fields.[99] The OPLS-AA/M error for glutamine is roughly half that of our force field. Therefore, if the dipeptide simulations had also shown inaccuracies this would cause some concern about the non-bonded parameters. However, as this was not the case, it is not investigated further in this work.

Four of the χ_1 dihedral angle scans for asparagine are shown in Fig 4.6. From the figure the change in the energy scan with conformation can clearly be seen. Additionally, the figure shows the difference between the QM and MM energy scan varies for the conformations.

4.3.2 Dipeptide Simulations

J coupling values allow a quantative measure of the accuracy of the ϕ distribution occupied during a dipeptide simulation. The results are summarized in Table B.7. The QUBE force field's mean RMSE of 0.42 Hz can be compared to 0.35 Hz for OPLS-AA/M. The error for our force field is far lower than OPLS-AA and OPLS-AA/L with errors of 0.97 Hz and 0.79 Hz respectively. The arginine dipeptide is primarily responsible for the higher error for

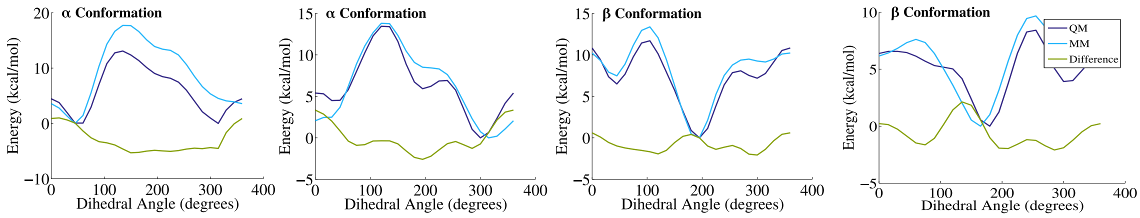


Fig. 4.6: Four of the χ_1 dihedral angle scans for asparagine. The QM and MM energies are shown, as well as the difference between the two scans. The MM energies scans use the optimized torsional parameters.

QUBE, with the arginine dipeptide excluded the error drops to 0.33 Hz. Residue specific arginine backbone torsional parameters could be calculated, however given that the ϕ/ψ distribution of arginine occupies the main conformations expected, this is not investigated in this work.

Figure 4.7 shows the collective ϕ/ψ distribution of the dipeptide simulations, this can be compared to the main protein conformation shown in Fig. 2.4. As discussed in Section 4.2.2, it is important to consider the dihedral distributions present as well as the J coupling values. Encouragingly, the ϕ/ψ distribution for the dipeptides show that the major conformations present in protein structures are present in the dipeptide simulations. The ζ conformation does have a slightly lower ψ angle than suggested, and there is an additional region with very low occupancy to the right of the γ conformation. However, these are very small discrepancies.

The individual ϕ/ψ distributions are shown in Section. B.3.2 of the Appendix. Generally, similar areas of the ϕ/ψ distribution are occupied by all the dipeptides. The serine and threonine dipeptides do not occupy identical regions to the other dipeptides, which is not unexpected given that they have a separate set of backbone torsional parameters. There are several dipeptides which show populations of left handed α helical conformation. High left handed helical populations have previously caused problems for other force fields.[46] However, since the PPII and β conformations always remain a highly occupied region, the populations of the left handed α helical region were not seen as a concern. The right handed α helical populations are small for all the dipeptides. This is in agreement with experimental results.[4]

As well as the backbone conformations sampled, the sidechain rotamer populations were also analyzed. In Fig.4.8, the simulated rotamer populations are compared to experiment values taken from protein coil library data.[45] Given that the experimental data is not for dipeptides, perfect agreement is not expected. However, populations at extreme values would cause concern and a correlation between the experimental and simulated values is favorable.

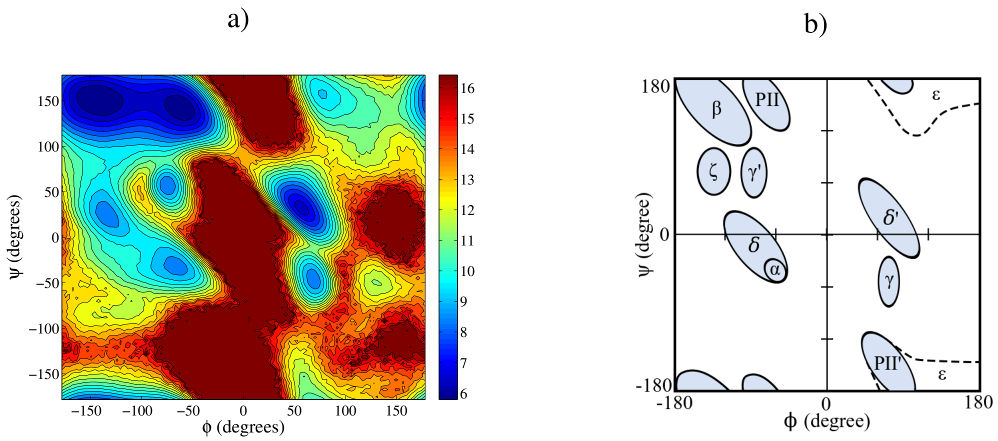


Fig. 4.7: a) The ϕ/ψ distribution plotted in the form $-log(p_{\psi,\phi})$ (where $p_{\psi,\phi}$ is the probability of a region being occupied) for our dipeptide simulations. The dark red regions correspond to low probability areas including conformations that are not sampled during the simulation. b) The major conformations observed in protein structures as proposed in Ref. 43.

Figure 4.8 shows that no dipeptides have populations consisting of just one type of rotamer and there are no extremely high values (as was seen for OPLS-AA and OPLS-AA/L). The rotamer M populations are slightly low than expected. However, given the issues previously mentioned with the experimental data used, further changes were not made to adjust the outliers shown in the figure.

The rotamer data, which was used to construct Fig. 4.8, is shown as a table of errors in Table B.8. With a MUE of 14%, QUBE performs better than both OPLS-AA and OPLS-AA/L with error of 23% and 21% respectively. The error is not as low as OPLS-AA/M, which has an error of 10%, however with further empirical changes to the torsional parameters this error could probably be reduced. Looking at individual dipeptide errors, protonated histidine and asparate have the highest errors. The protonated histidine experimental data includes all ionization states of histidine and therefore may not be accurate, which would explain the high error. Asparate's high error is more problematic and in future versions of our force fields further changes to its sidechain parameters may be considered.

4.3.3 Peptide Simulations

The J coupling errors for the alanine pentapeptide are shown in Table 4.1, with the ϕ/ψ distribution shown in Fig.4.9 and further results given in Section B.4 of the appendix. Three sets of Karplus parameters are used to evaluate the error and the value in brackets excludes the $^2J(N,C_\alpha)$ coupling term. Issues with the $^2J(N,C_\alpha)$ coupling Karplus parameters are discussed in Ref. 99, Ref. 24 and in Section 4.2.2.

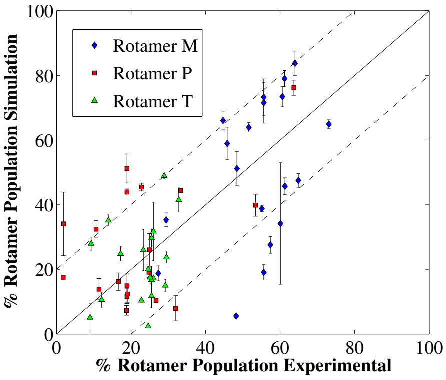


Fig. 4.8: A comparison of the rotamer populations for the dipeptide MD simulations and experimental values from the protein coil library. The dashed lines show the populations that fall within ±20% of the experimental values.

J Coupling Error		
Set 1	**Set 2**	**Set 3**
0.90 ± 0.03 (0.86 ± 0.03)	4.16 ± 0.01 (0.81 ± 0.03)	1.51 ± 0.02 (0.87 ± 0.03)

Table 4.1: The J Coupling error for the alanine pentapeptide simulation. The Karplus parameters used are the same as those used in Ref. 99. The values shown in brackets correspond to the J Coupling errors excluding $^2J(N,C_\alpha)$.

The J coupling error for set 1 is very encouraging and is lower than both the OPLS-AA/M value (1.16 ± 0.02 Hz) and the AMOEBA values (0.99 Hz).[99,104] The values for set 2 and 3 with the excluded $^{2}J(N,C_{\alpha})$ term are similar in value. In the simulations carried out in this work, as well as the work of Amber ff15ipq and OPLS-AA/M, the low β populations present result in a high $^{2}J(N,C_{\alpha})$ error for the second and third set of Karplus parameters.[24,99]

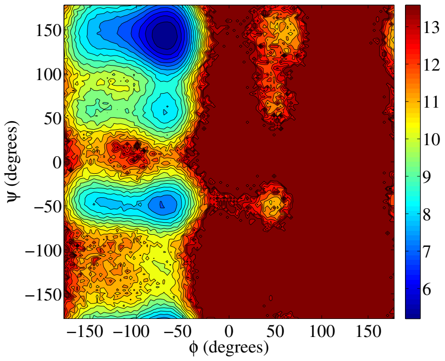


Fig. 4.9: The ψ and ϕ distribution of the central residues of the alanine pentapeptide, plotted in the form $-log(p_{\psi,\phi})$ (where $p_{\psi,\phi}$ is the probability of a region being occupied). The lighter regions correspond to low probability areas including conformations that are not sampled during the simulation.

The conformation with the largest population present is PPII with 62.10 ± 2.23 % of the simulation spent in this conformation, see Table B.12 in the Appendix. This is similar to the results seen for OPLS-AA/M (53.5 ± 0.2 %). Both force fields also have a low α helical population as expected from experimental data.[7]

For the alanine pentapeptide, the role of the torsional parameters is to correct for problems in the reproduction of the energy surface due to the parameters and functional form used. In preliminary work, it was shown that even with the backbone torsional terms set to zero the J coupling error for the peptide remained below that of OPLS-AA/M with a value of 0.97 Hz. This demonstrates the accuracy of the non-bonded parameters used in the force field.

The problems associated with using the Karplus parameters for Gly_3 are discussed in Section 4.2.2, Ref. 99, Ref. 88 and Ref. 34. Therefore, we evaluate the backbone conformations of Gly_3 through its ϕ/ψ distribution alone. In Fig, 4.10 the OPLS-AA/M distribution is compared to the distribution using the QUBE force field. A lower α population is occupied by the QUBE force field, but otherwise both distributions are very similar.

The dipeptide and peptide simulations have demonstrated that the QUBE force field, and the parametrization methods used to create it, are sufficiently accurate to recreate conformational properties of short peptides. The error for these systems is comparable to OPLS-AA/M and the ϕ/ψ distribution demonstrates that the main conformations seen

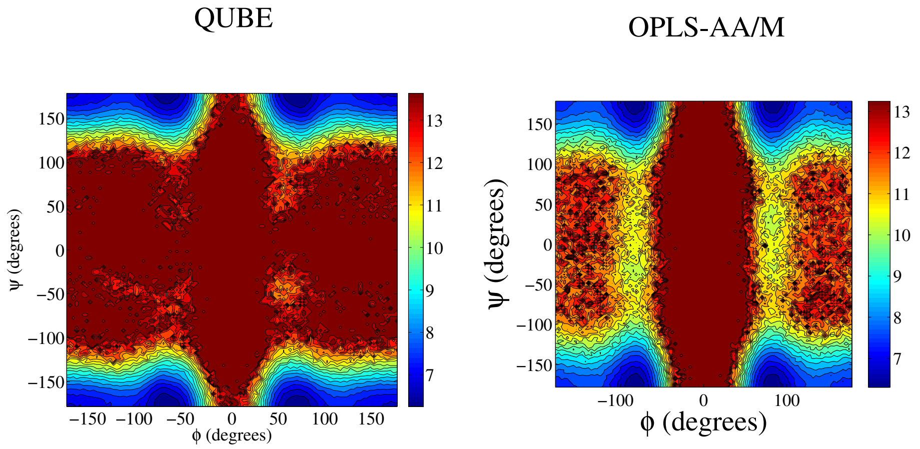
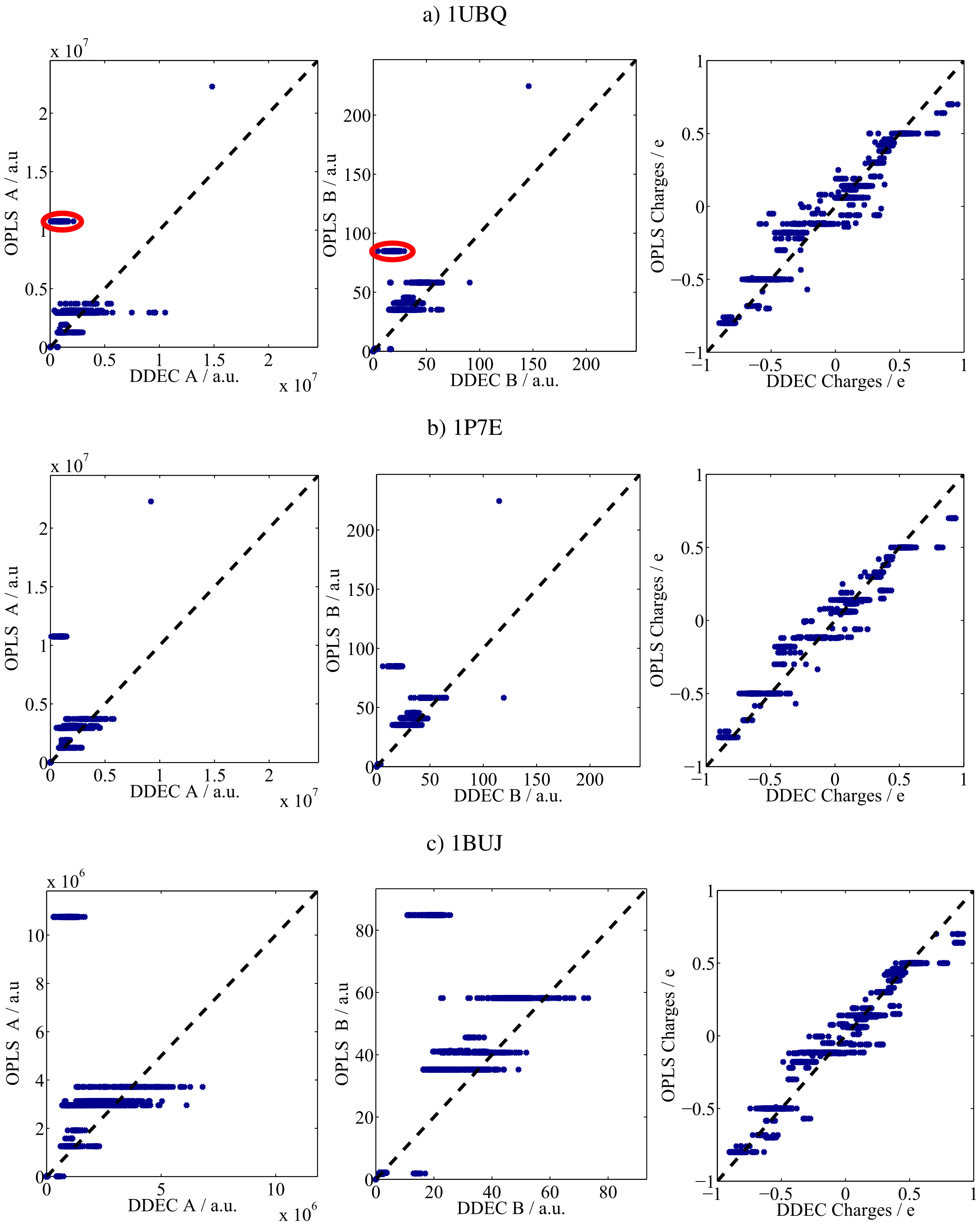


Fig. 4.10: The ψ and ϕ distribution for the glycine tetrapeptide (all residues are included).

in protein structures are populated. Issues with the transferability of torsional parameters have already been identified from the longer peptide simulations and solved by applying regularization. The performance of our force field for entire proteins was investigated to further understand the intricacies of fitting torsional parameters to a system-specific force field.

4.3.4 Protein Non-Bonded Parameters

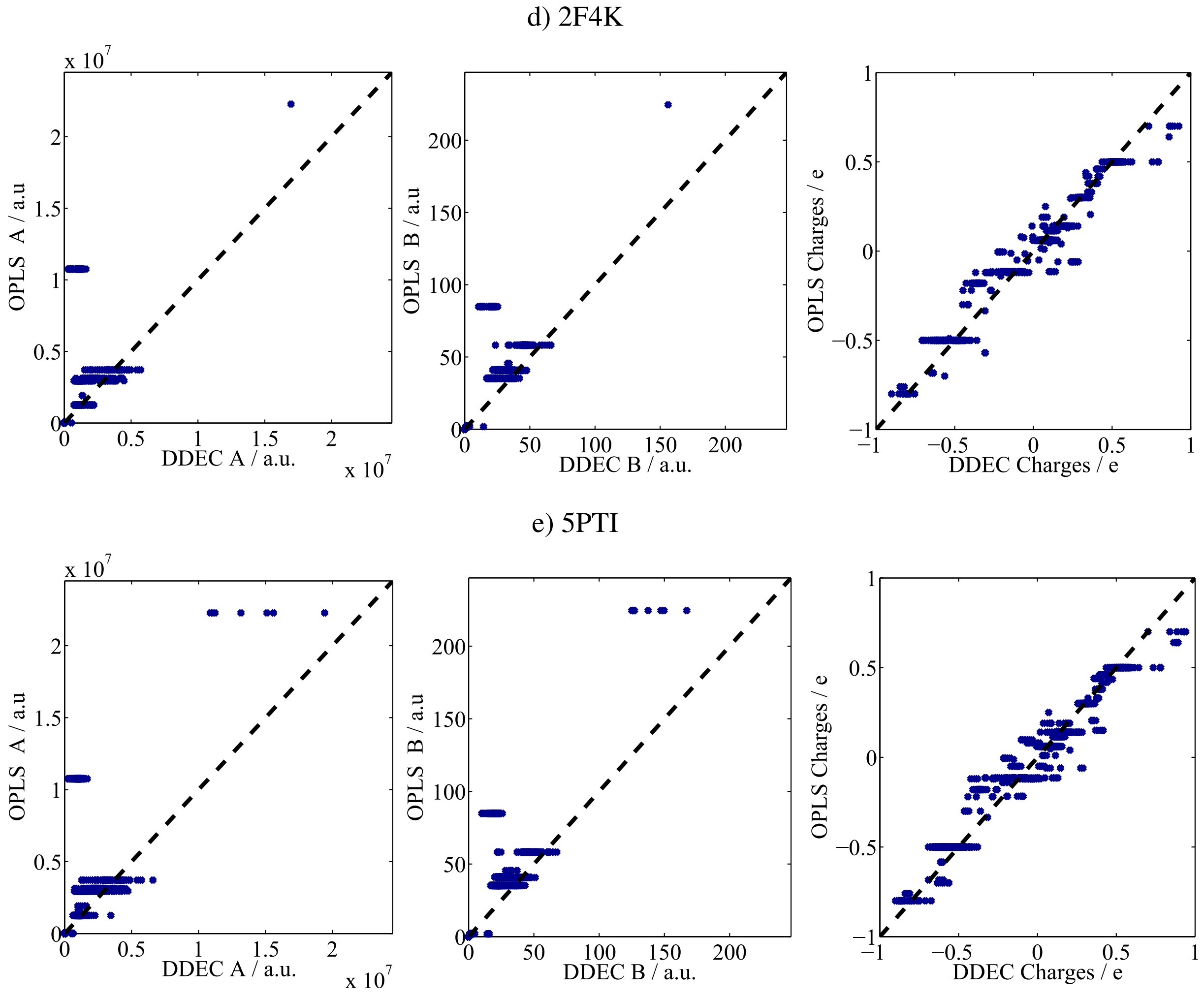
Before the results of the molecular dynamic simulations are presented, the QUBE and OPLS non-bonded parameters for each atom of the five proteins tested are compared. The QUBE non-bonded parameters are calculated in an implicit solvent with a dielectric of $\varepsilon = 80$. Comparable trends are seen for iPol or $\varepsilon = 4$ non-bonded parameters, see Section 2.3.2.



a) 1UBQ

b) 1P7E

c) 1BUJ


d) 2F4K

e) 5PTI

Fig. 4.11: A comparison between the QUBE and OPLS non-bonded parameters. The dashed line corresponds to $y = x$.

From Fig. 4.11 it can be seen that all the proteins studied show a similar relationship between the OPLS and QUBE terms. The QUBE and OPLS charges correlate well with no clear outliers. However, both the A and B coefficients have a group of QUBE terms that deviate from the OPLS parameters, as highlighted in Fig. 4.11a). This was also observed in Ref. 22, and the cluster of terms associated with this deviation corresponds to carbonyl carbon atoms which are electron deficient. Additionally, the OPLS non-bonded terms have a small number of possible values, as atom types are used, whilst the QUBE terms have a far greater level of variation. This is expected as the non-bonded terms will change depending on the local polarization environment.

4.3.5 Change in Non-bonded Parameters with Protein Conformation

The non-bonded parameters used in the QUBE force field will be calculated for one specific input structure. However, during the course of the MD simulation many difference conformations will occur. In order to see the change in non-bonded terms with the structure, a short MD simulation of GB3 in a water box was carried out using OPLS-AA/M. Ten structures were then taken from this simulation and the non-bonded parameters for these ten structure were calculated.

Figure 4.12a) shows the distribution of the difference in atom charges from their mean values. It can be seen that some charges show discrepancy from the mean values, with a maximum difference of 0.161 e. However, as the distribution has a standard deviation of 0.0243 e the majority of atom charges do not change considerably. This shows that whilst the non-bonded parameters do respond to their environment the changes are not excessive, which would be problematic. Figure 4.12c) demonstrates that this difference is independent of the charge. Both figures clearly show the importance of considering the input structure used. The charge of the residue is considered in Fig. 4.12b) and changes to the input structure are seen to result in the charge varying. Therefore, the differences are not entirely due to a redistribution of the electron density within the residue.

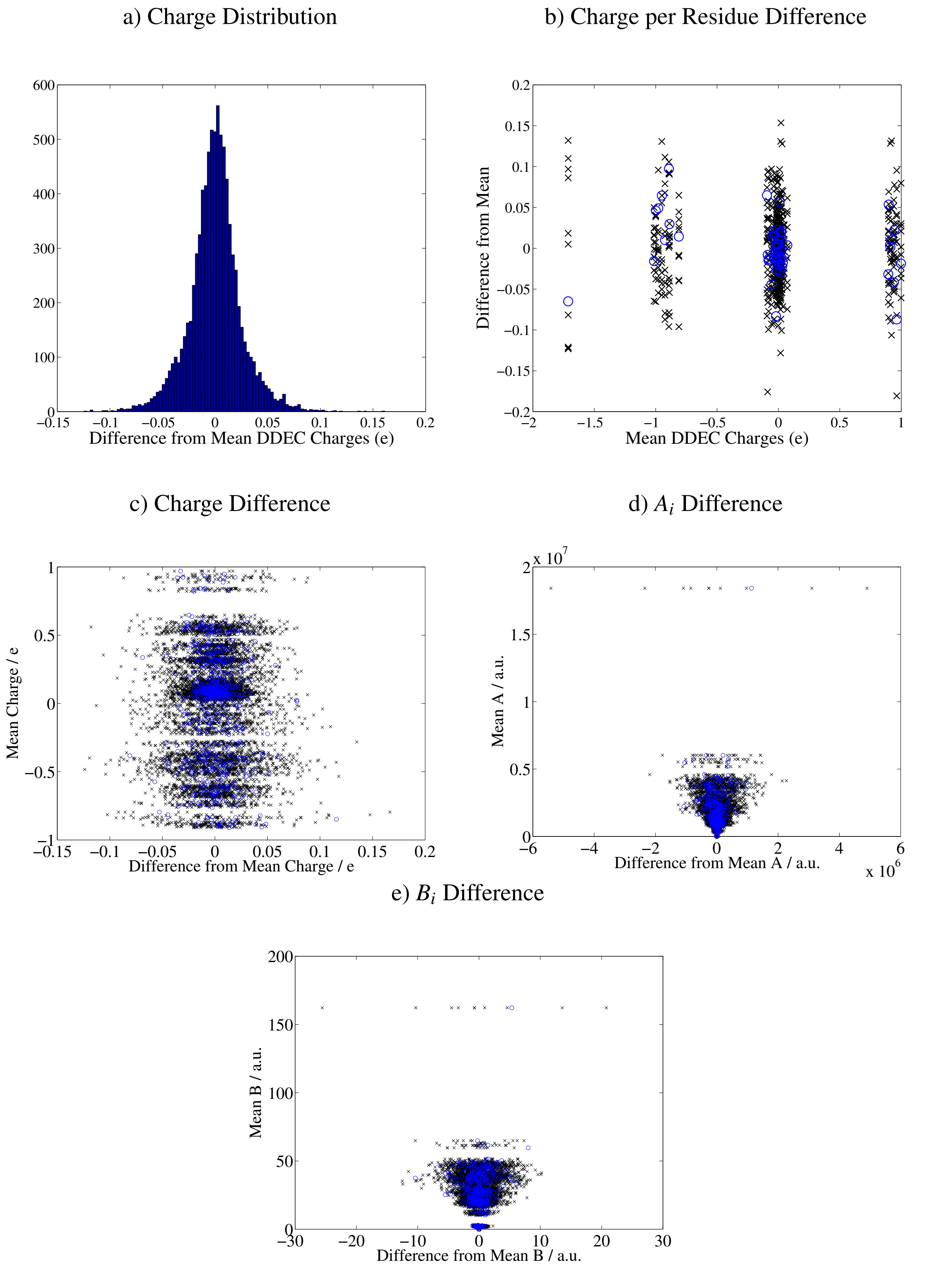
The difference in the Lennard-Jones parameters from the mean value is also considered in Fig. 4.12. The non-bonded parameters clearly change with the input structure. Again, this shows the importance of the input structure used for the ONETEP calculation.

	Run 1	**Run 2**	**Run 3**	**Run 4**	**Run 5**	**Run 6**	**Run 7**	**Run 8**	**Run 9**	**Run 10**	**Minimum**
Charge	0.026	0.026	0.023	0.026	0.025	0.025	0.025	0.025	0.025	0.025	0.017
$\mathbf{A_i}$	2.93×10^5	2.37×10^5	2.87×10^5	2.58×10^5	2.35×10^5	2.33×10^5	2.44×10^5	2.14×10^5	2.34×10^5	2.44×10^5	1.71×10^5
$\mathbf{B_i}$	1.87	1.74	1.81	1.83	1.68	1.65	1.69	1.54	1.59	1.70	1.29

Table 4.2: The standard deviation of the difference between the non-bonded parameters from one structure and the mean values of the non-bonded parameters is shown in this table.

The standard deviation of the difference in the nonbonded parameters from the mean, shown in Table 4.2, was also used to evaluate the minimized structure's non-bonded parameters. A low standard deviation is considered preferable as we require nonbonded terms to be representative of as large a range of structures as possible, and therefore close to the mean values. For all three coefficients, the minimum structure non-bonded parameters have the lowest standard deviation. This is encouraging as it shows that the minimum structure does not have many extreme values and the standard deviations are low.

To choose the most appropriate input structure, it is important to consider not only the accuracy of the non-bonded terms but also the time required to find these non-bonded terms. We ideally want to create a force field that can be readily used by other research groups. If for every protein force field we wanted to create, it was necessary to take multiple snapshots of a MD simulation, calculate the non-bonded values and then average these terms, the cost of the QM calculations would start to become prohibitive. Although, as the computational cost of these calculations decreases further this may be the preferred option. Additionally, taking structures from an MD simulation will weight the structures sampled according to the force field used to carry out the simulation. It would also mean that two separate groups could create a protein force field for the same protein and end up with different non-bonded parameters as the structures sampled happened to vary. Therefore, we chose the input structure to the ONETEP calculation to be the experimental structure minimised with OPLS-AA/M in a water box. This provides a quick and consistent way to obtain an input structure. In preliminary work, the mean non-bonded parameters of the ten structures of GB3 were used to perform a 200 ns MD simulation. The conformational preference of the molecule did not greatly vary from the minimized structure results.



a) Charge Distribution

b) Charge per Residue Difference

c) Charge Difference

d) A_i Difference

e) B_i Difference

Fig. 4.12: The change in non-bonded QUBE parameters with structure. Figure a) shows the distribution of all the atoms charges from the mean values whilst c), d) and e) are the charge, A and B terms as a function of the mean values. Figure b) shows the difference between the mean residue charge and the residue charge of each structure. The black crosses correspond to the ten structures from the MD simulation whilst the blue circles are the non-bonded terms for the structure minimized in water.

4.3.6 Protein Simulations

The use of system-specific non-bonded parameters for biomolecular force fields allows local polarization effects to be included and this may result in improvements in the accuracy of the force field. Although the conformational preferences for the peptides tested are promising, it is not known whether the torsional parameters will continue to be appropriate for use with proteins. As the non-bonded parameters change with the system studied, the transferability of torsional parameters cannot be as readily assumed. To assess this we first look at simulations of the proteins ubiquitin and GB3.

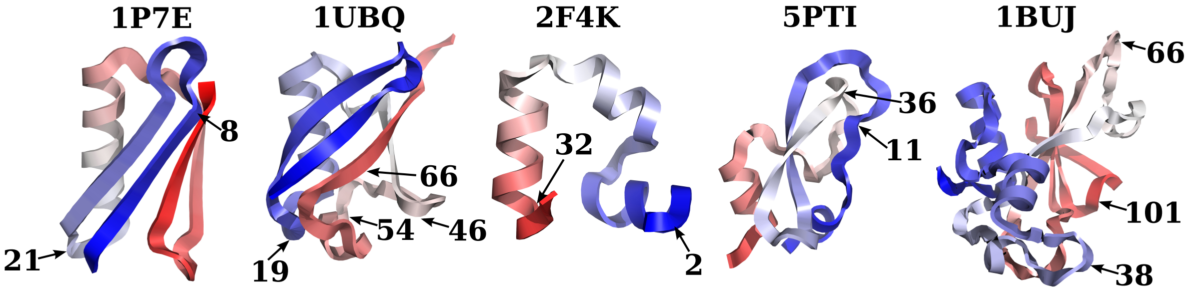


Fig. 4.13: The experimental structures of the proteins tested. Regions that showed the most significant deviation from the experimental structure in the simulations are labelled. The red-white-blue color gradient represents the residue number.

The J coupling errors for ubiquitin and GB3 are summarised in Table 4.3. With an overall RSME of 1.54 Hz, the error using the QUBE force field for ubiquitin is worse than OPLS-AA/M, which has a value of 1.12 Hz, but better than OPLS-AA and OPLS-AA/L with values of 1.70 Hz and 1.84 Hz respectively.[99] GB3 follows the same trend with an RMSE of 1.10 Hz for the QUBE force field, compared to the error for OPLS-AA/M of 0.90 Hz, whilst OPLS-AA and OPLS-AA/L both have an error of 1.46 Hz.[99]

The J coupling results suggest that whilst the transfer of torsional parameters from dipeptides to proteins may cause some issues, the QUBE force field remains more accurate than OPLS-AA and OPLS-AA/L. This is promising when we consider that OPLS has been in development for many years with multiple iterations and changes carried out. In contrast, this is the first version of the QUBE force field. The $^3J(H_\alpha,H_\beta)$ coupling term is the main reason for the difference between the force field's RMSE. For GB3,the $^3J(H_\alpha,H_\beta)$ error for the QUBE force field is 1.80 Hz, this is well below the errors for OPLS-AA and OPLS-AA/L of 3.71 Hz and 3.38 Hz respectively.

However, as discussed in Section 4.2.2, the J coupling error should not be used as the only measure of a force field's accuracy. To further test the performance of our force field,

	Backbone Couplings (Hz)		Sidechain Coupling (Hz)			Overall
	1997 Values	2007 Values	^{3}J(H$_\alpha$,H$_\beta$)	^{3}J(C$'$,C$_\gamma$)	Methyl C$_\gamma$	RMSE
Ubiquitin	0.94 ± 0.07	1.15 ± 0.05	2.40 ± 0.11	1.16 ± 0.20	1.20 ± 0.05	1.54 ± 0.07
GB3	0.93 ± 0.05	1.03 ± 0.05	1.80 ± 0.05	-	0.84 ± 0.02	1.10±0.04

Table 4.3: J coupling errors for the proteins ubiquitin and GB3.

we analyzed the ϕ/ψ distribution and RMSD of the C_α atoms of each residue from the crystal structure for five proteins. The dihedral angles of the experimental structure are shown on each ϕ/ψ plot and these experimental points, along with the previous data for AMBER ff15ipq (the ϕ/ψ plots are given in the SI of Ref. 24) are used to evaluate the performance of our force field.[24]

Figure 4.13 shows the residue number and structure of the five proteins in the testing set.

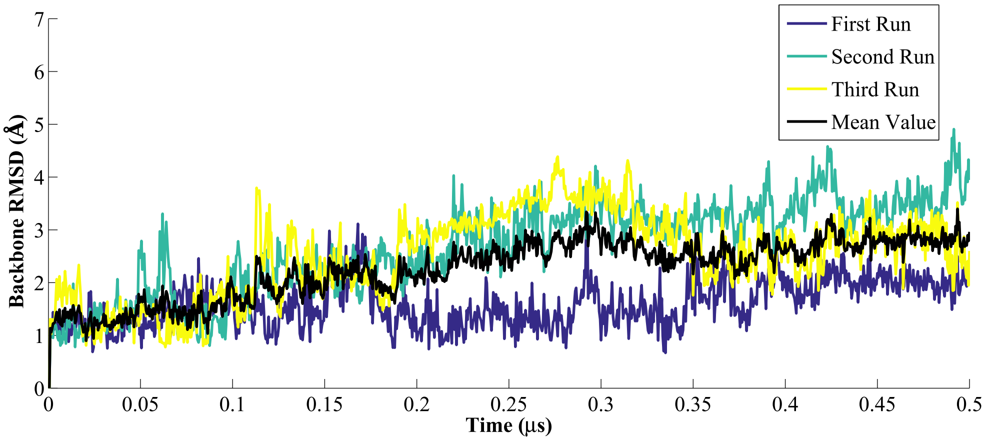


Fig. 4.14: The RMSD, relative to the experimental structure 1P7E, for the three simulations of the protein GB3.

Figure 4.14 shows the RMSD of GB3 over the course of the simulation. The RMSD for the first simulation run generally remains below 2 Å with reasonable retention of the experimental structure. The second and third simulation runs also generally remain within 2 Å of the experimental structure for the first 100 ns of the simulation. However, by 250 ns one of the β sheets present in the protein begins to separate and this causes a rise in the RMSD. The ϕ/ψ distributions of GB3, Figure B.5 in the Appendix, and the RMSD per residue, Figure B.4, help us to further analyze the results. Residues which deviate from the experimental structure, and the ff15ipq results, tend to have high J coupling errors. For example, residues 8 - 21 have a large discrepancy from the experimental structure and this is reflected by high J coupling errors in this region. The backbone J coupling error, using the 2007 Karplus parameters, for residues 8 - 21 is 2.00 Hz which is almost double the total backbone error. This region corresponds to the β sheet which separates. Aside from this region, the only other residues which show noticeable deviation from the crystal structure are Val 39, Asp 40, Gly 41 and Thr 55. However, the small deviations that are present in these four residues are also observed for AMBER ff15ipq.[24]

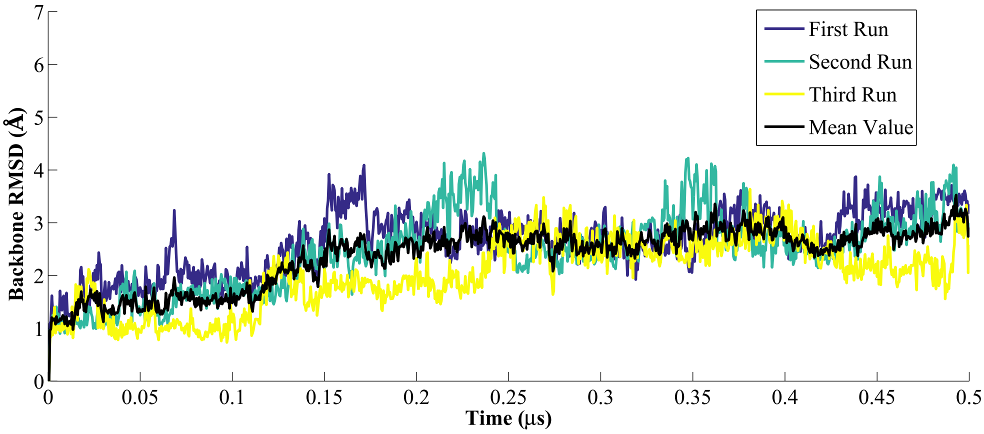


Fig. 4.15: The RMSD, relative to the experimental structure 1UBQ, for the three simulations of the protein ubiquitin.

The RMSD of ubiquitin is shown in Figure 4.15. As was seen for GB3, the first 100 ns of the simulation show reasonable retention of the experimental structure for all three simulation runs. However, as the simulation progresses the RMSD of the protein rises to give a mean values of approximately 3 Å. Ubiquitin shows a deviation from the experimental values in a number of regions. These regions tend to be those without a clear secondary structure or with a turn. This is the case for residues 7-11, a turn connecting two β sheets, and for residues 72-74 which are at the end of the protein. The flexibility in these regions is not of concern since the results of ff15ipq and experimental NMR measurements suggest deviations from the crystal structure in these portions of the protein. A deviation is also observed for Asp 52 and Gly 53 which are α helical in the crystal structure. The PPII and γ regions that are occupied instead are also seen in the ff15ipq. However, between residues 54-66, which show no secondary structure or a turn, the deviation from the experimental structure continues. The J coupling error between residues 62 and 65 also reflects issues in this region. Residues 19-20 and residues 46-47 are also turns or without a clear structure and again show deviations from the experimental structure which all contributes to the rising RMSD of the protein.

The villin headpiece has the lowest RMSD of all the proteins tested, Figure 4.16. Excellent agreement with the ϕ/ψ distribution of ff15ipq is also seen with the ϕ/ψ distribution for QUBE presented in Figure B.7. We observe residues 10-12, a loop linking two helices,

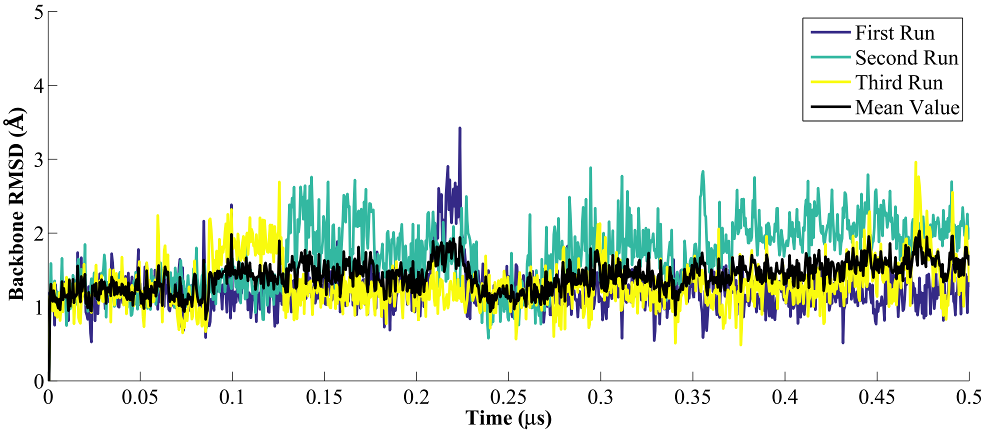


Fig. 4.16: The RMSD, relative to the experimental structure 2F4K, for the three simulations of the villin headpiece.

taking an alternative conformation from the crystal structure, however this is also present in ff15ipq simulations. The only small difference between the ϕ/ψ distributions for QUBE and ff15ipq is for residue 2 and residues 32-34, which are located at the beginning and end of the headpiece and have no clear secondary structure. The three α helices present in the protein are retained throughout the simulations.

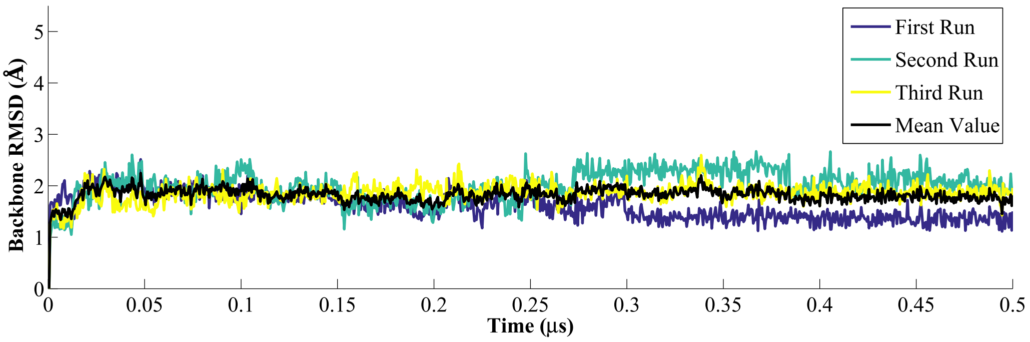


Fig. 4.17: The RMSD, relative to the experimental structure 5PTI, for the three simulations of the protein BPTI.

With a mean RMSD of 1.82 Å, BPTI retains its experimental structure better than all of the other proteins tested in the set, with the exception of villin which has a mean RMSD of 1.41 Å. Figure 4.17 shows the RMSD over the course of the simulation and we can see that RMSD remains relatively stable, and consistently below 3 Å. The slight deviations that are present follow the same trends previously seen. From the RMSD per residue, Fig. B.10, and the ϕ/ψ distributions, Fig. B.11 it can be seen that residues 11-17 and residues 36-40 show noticeable differences from the crystal structure. Again, these correspond to regions with no fixed secondary structure or a bend. In regions with a helical or β sheet structures the agreement is sufficiently accurate.

Binase has an overall RSMD that is much higher than the other proteins. This was also seen for ff15ipq which had a mean RMSD of 3.4 Å. This is in part due to the multiple loops that are present in the structure. As the 20 structures of the NMR ensemble are shown in the ϕ/ψ distributions, Fig. B.9, multiple conformations sampled by a residue can be seen. Whilst some residues are clearly more representative of the experimental structure using ff15ipq, such as Hie 101 where the QUBE force field occupies alternative structures, for other residues it is more ambiguous which ϕ/ψ distribution is more appropriate. For example, the NMR ensemble for Lys 38 shows β and α conformations being present, and this is shown for both ff15ipq and our force field, but the proportion of each conformation is different. The ensemble of Ser 66 is not well represented with ff15ipq but with our force field all structures in the ensemble are captured to some degree, although an additional α helical conformation is also seen. The two α helices present in 1BUJ, around residue 10 and residue 30, are generally well represented with our force field but small conformational changes do occur. Towards the end of the third simulation run the protein does not retain the experimental structure sufficiently well. In the regions with no structure, a bend or a turn we again see a difference from ff15ipq and the experimental structure. This causes the RMSD to reach extremely high values as the protein does not retain its structure.

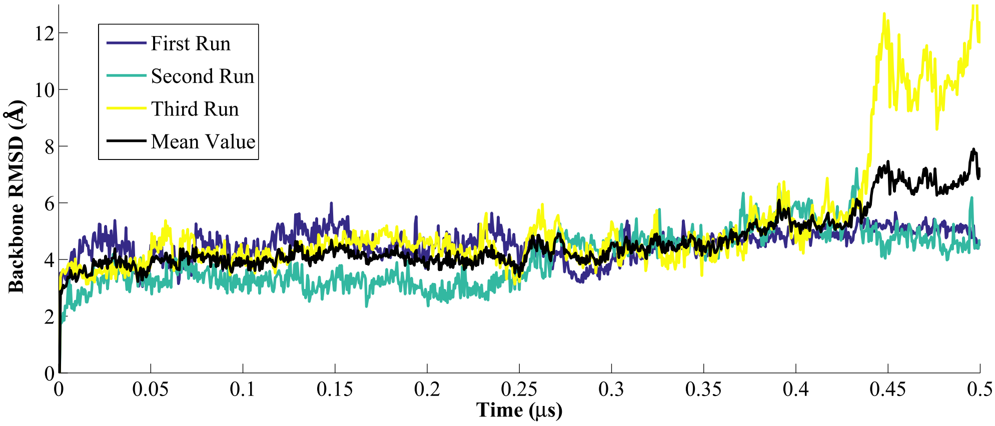


Fig. 4.18: The RMSD, relative to the experimental structure 1BUJ, for the three simulations of the protein binase.

4.4 Discussion and Conclusion

The assumption that biomolecular force fields must be parametrized using small molecules has persisted since MM simulations began and remains in all force fields currently used.[54,95,125] Within this work, we look to challenge this assumption by deriving system-specific non-bonded parameters, from linear scaling DFT, for use in the QUBE force field. These non-bonded terms were used with the derived bond and angle parameters from Chapter 3 and reparametrized torsional terms.

We have shown that using a system-specific QM force field can result in accurate conformational preferences for short peptides. Rotamer populations and J coupling errors for the dipeptide molecules were in good agreement with experimental values and compared favorably with the latest OPLS force field. For longer peptide molecules, the problems associated with fitting torsional parameters to a system-specific force field became more apparent. Fortunately, using regularization in the fitting process was shown to overcome these issues and resulted in Ala$_5$ having a J coupling error of 0.90 ± 0.03. Further work investigating disordered peptides will ascertain how general this fix is. The accuracy of the peptide simulation not only further validated the non-bonded parametrization strategy, but also the modified Seminario method developed in Chapter 3. For the protein simulations, the RMSD of the protein remained low, below 2 Å, for two of the five proteins tested. The α helices present in all of the proteins generally stayed close to their experimental structure, but the β sheets did not always remain together and regions with no clear structure or exhibiting a turn regularly deviated from the starting structure. Despite this, the majority of the regions in the proteins retained their experimental structure and the J coupling values for GB3 and ubiquitin were below that of OPLS-AA and OPLS-AA/M.

Whilst fitting torsional parameters, we found that using dihedral energy scans alone is not sufficient. This was also seen in the development of OPLS-AA/M with empirical changes required to the sidechain torsional parameters, and we can infer from this that automatically fitting backbone and sidechain torsional parameters using dihedral energy scans may not be possible. Whilst developing QUBE, manual adjustments to the torsional parameter fitting process were required, and the most obvious failure of dihedral energy scan fitting was that for a number of sidechains it was more accurate to set the torsional parameters to zero than use the originally proposed terms. This is in part due to the functional form used, with improvements to this discussed below, but also due to the poor sampling of relevant energy structures by dihedral energy scans. This problem is reduced by the iterative fitting methods used in AMBER ff14ipq and ff15ipq.[15,24] These methods sample the energy structures used for torsional parameter fitting by performing MD simulations with the current iteration of the force field. This was not a viable approach for this work as it is necessary to derive an initial set of torsional parameters. However, in future versions of the force field it will be considered and may lead to improvements in the backbone torsional parameters.

There are also additional considerations involved in using torsional parameters for a system-specific force field. The torsional parameters are fit with one set of non-bonded parameters but then have to be used for a wide range of non-bonded terms. It may be possible to address this issue by changing the functional form of the torsional component. The functional form currently used is inaccurate due to both the shape of the function used (with its constraining requirement of symmetry being just one problem) and torsional parameter dependency on only one dihedral angle. This problem has already be addressed for the backbone torsional parameters in CHARMM22 by using CMAP, a grid based correction.[9] Extending the CMAP correction so it is dependent on the χ_1 dihedral angle has also recently been carried out.[62] CMAP corrections could be added to our force field, however due to the system-specific nature of QUBE they may not be as effective as in previous studies. An alternative improvement that could be made to our force field would be to assign torsional parameters based on a greater number of properties. For instance, during the side chain torsional parameter fitting process, it was seen that the optimized torsional parameters for α helical and β sheet conformations could greatly vary. This problem could be easily addressed with our force field by assigning torsional parameters based on secondary structure information from the crystal structure. The additional flexibility that can be offered when atom types are not used in the force field parameter files has been an unexpected benefit of a system-specific force field. More complex torsional parameter assignment could also help to deal with the intrinsic difficulties of using torsional parameters for a system-specific force field. For example, the torsional parameters could be dependent on the non-bonded parameters used, with adjustments made for highly polarized regions.

However, the torsional parameters may not be the only factor responsible for the inaccuracies in the protein simulations. The non-bonded parameters themselves may be accurate but the interaction between the QUBE force field and the water model (TIP3P) could be incorrect. This would require a new compatible water model being found or developed. A protein's accurate interaction with water is crucial and problems in these interactions may be responsible for some of the instabilities in structures that we have seen. Alternatively, the non-bonded parameters could be unsuitable. The parametrization strategy employed may be a source of error. Changes could be made to the QM calculation performed, i.e. the exchange-correlation functional or implicit solvent method used, or to the strategy used to obtain force field parameters from the QM data, i.e. the partitioning method used or the method to model induction could be altered. The functional form could also be improved with the addition of a r^{-8} term to model dispersion or the addition of off center charges to model anisotropy effects, as discussed in the following chapter.

Chapter 5

Anisotropy

5.1 Introduction

Modeling the electrostatic interactions in force fields is often carried out by using a point charge at the center of each atom.[117] However, point charge models can be too simplistic to accurately recreate the electrostatic potential (ESP) of nitrogen, sulfur or halogens.[66] This is generally due to the presence of lone pairs or σ holes.[66,117] Therefore, methods need to be introduced to model anisotropy in the electrostatic distribution. An example of anisotropy is shown by the ESP in Fig.5.1.

In Ref. 22, the addition of extra sites for electrostatic charge was shown to improve calculations of the free energy of an indole and benzofuran molecule binding to the L99A mutant of T4 lysozyme, as well as improve liquid properties for a set of small molecules. The addition of off center charges has also been shown to improve the MM structures predicted, and trends in the interaction energy.[26] The method used in Ref. 22 relied on the partitioned dipole and quadrupole moments, calculated from the atomic electron density, to fit optimal positions for the charges. However, the method employed did not consistently converge and resulted in a large number of off center point charges. Improvements were required to correct these issues and improve usability of the method.

A number of different approaches have been previously developed to add off center point charges. Existing methods tend to be applicable to only a certain type of bonding, such as Ref. 53 and Ref. 49 which are limited to halogen bonding, or a specific element, such as Ref. 126 which investigates anisotropy in sulphur. Additionally, Ref. 126 uses experimental data to parameterize the off center charges which also makes it unsuitable for our purposes. We require a method that relies only on the electron density, can be used for a variety of different atoms, can be used for large biological molecule and is automated. The biological molecule

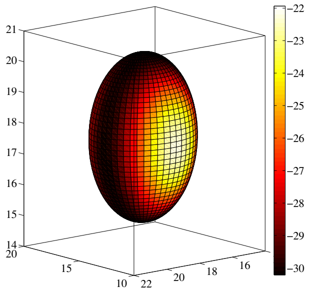


Fig. 5.1: An example of the anisotropy in the electrostatic potential energy (of a 1 eV test charge) of the Cl atom of CH_3Cl. The potential energy is measured at 1.40 times the van der Waals radius and is in kcal/mol. The anisotropy is due to a σ hole, which is a positively charged region that can occur in covalently bonded Group IV–VII atoms.[94]

requirement is because we plan to use the approach not only for small molecules, but also in future versions of our protein force field.

In this section, a new method is implemented for calculating off center point charges. It is based on optimizing the recreation of the QM ESP with additional point charges. The method is validated on small molecules by the accuracy of the recreation of the atom's QM ESP, as well as the molecules' liquid properties.

In our force field parametrization approach, the electron density for each atom is found by partitioning the molecule's electron density according to the DDEC scheme described in Section 2.3.2. Adding off center point charges can be readily added into our parametrization workflow by using the electron density from the QM calculation as the parametrization data and we have implemented this method into ONETEP.

5.2 Off Center Point Charge Optimization

5.2.1 One Additional Off Center Charge

In addition to the requirements previously mentioned, the method also needed to result in off center point charges that maintain the symmetry around the atom and its bonding environment. This prevented us from using an existing approach, such as that of Ref. 3 or Ref. 39, as constraints on the positions of the additional charges are necessary. Therefore, we used the symmetry requirements to dictate the directions along which extra charges could be placed. The vectors for one additional off center point charge are shown in Fig. 5.2. These are the only directions that preserve symmetry. The vector direction is governed by the number of bonds the atom exhibiting anisotropy has:

- One bond - The atom A (which exhibits anisotropy) has one neighbor, atom B. The vector for the extra charge is $\mathbf{r_1} = \lambda_1 \mathbf{r_{AB}}$, where $\mathbf{r_{AB}}$ is a vector between atom A and atom B and λ_1 is to be determined, which is along the bond axis.
- Two bonds - The atom A has two neighbors, atom B and atom C. The vector for the extra charge is $\mathbf{r_1} = \lambda_1(\mathbf{r_{AB}} + \mathbf{r_{AC}})$, which is along the bisector of the two bond vectors.
- Three bonds - The atom A has three neighbors, atom B, atom C and atom D. The vector for the extra charge is $\mathbf{r_1} = \lambda_1(\mathbf{r_{AB}} - \mathbf{r_{AC}}) \times (\mathbf{r_{AD}} - \mathbf{r_{AC}})$, which makes an equal angle with all three bond vectors.

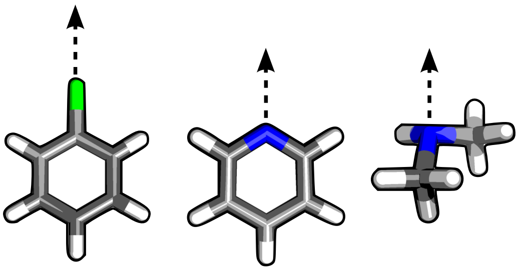


Fig. 5.2: The directions along which off center point charges are placed for an atom with one, two or three bonds.

After the vector is assigned, the optimal position along the vector and the charge of the off center point are determined. This is carried out using a grid search of parameters to find the values which best recreate the QM ESP. Assigning a direction the charges are allowed to be placed along reduces the number of variables that need to be optimized from four (the x, y, z coordinates and the charge) to two (the distance along the vector and the charge). This simplification is particularly important when multiple off center point charges are added, this is described in the following section. An example of the change in the ESP error as a function of the two variables is shown in Fig. 5.3. The QM ESP energy is calculated from the partitioned atomic electron density with a reference charge of 1 e. This results in a method which can be divided into tasks for individual atoms and is parallelisable. The ESP is taken at a series of points on sets of spheres with radii between 1.4-2.0 times the van der Waals radius of the atom. The error is calculated using equation 5.1, where Φ_i^{ref} is the ESP energy from the QM calculation, Φ_i is the ESP for the point charges and M is the number of sampling points. The search is constrained by a maximum distance of 1.00 Å (1.50 Å for chlorine), as charges outside the van der Waals radius are problematic in molecular dynamic simulations, and a maximum charge of ±2.0 e, to prevent extreme values occurring. The atom centered point charge is given a value so that the overall charge for the atom is unchanged. The method is summarized with the flowchart in Fig. 5.5.

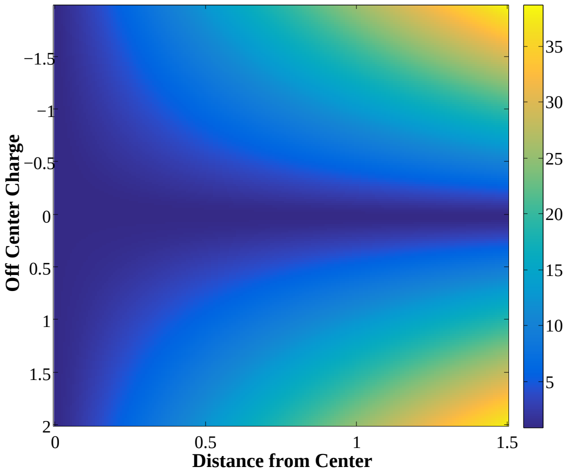


Fig. 5.3: The error in the ESP energy (kcal/mol) for CH3Cl as a function of the off center charge and distance from the center. The minimum error value corresponds to the position of the off center sites and is at 0.02 eV and 1.5Å.

$$F(\Phi,\Phi^{ref}) = \sum_{i=1}^{M} \frac{|\Phi_i - \Phi_i^{ref}|}{M} \tag{5.1}$$

An additional threshold parameter was required to distinguish between atoms that showed anisotropy and those that did not. This parameter was assigned a value of 0.9025 kcal/mol. Above this value the anisotropy method is used, below the value no off center charges are added. An exception to this is hydrogen, which never has off center charges added. Furthermore, charges are only included in the final output when there is an improvement in error of 0.0625 kcal/mol. (*These parameters were suggested from preliminary liquid simulation tests by Joshua Horton at Newcastle University.*[44])

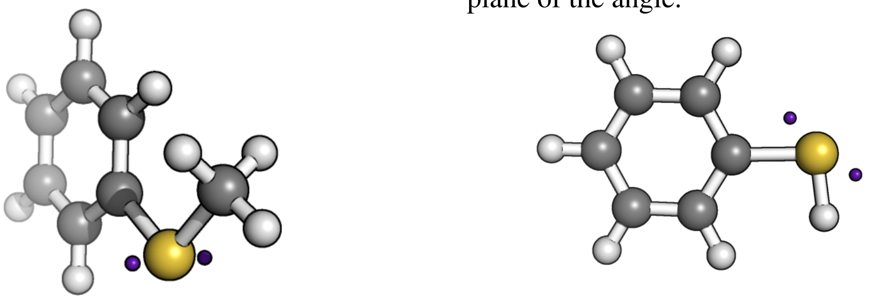
5.2.2 Multiple Off Center Charges

In Ref. 22, it was often necessary to add more than one off center point charge to recreate the anisotropy seen in the QM ESP. Therefore, our approach was extended to add multiple charges. Again, the method depends on the number of bonds the atom exhibiting anisotropy has:

- One bond - A second off center charge is placed along the $\mathbf{r_1}$ vector, $\mathbf{r_2} = \lambda_2 \mathbf{r_{AB}}$.

- Two bonds - If two extra point charges are used, the original vector is a line of symmetry. The two charges are then placed in the same plane as the vectors that point from the atom to the neighboring atoms, $\mathbf{r_{1,2}} = \lambda_{\|}(\mathbf{r_{AB}} + \mathbf{r_{AC}}) \pm \lambda_{\perp}(\mathbf{r_{AB}} + \mathbf{r_{AC}}) \times (\mathbf{r_{AB}} \times \mathbf{r_{AC}})$, or perpendicular to this plane, $\mathbf{r_{1,2}} = \lambda_{\|}(\mathbf{r_{AB}} + \mathbf{r_{AC}}) \pm \lambda_{\perp}(\mathbf{r_{AB}} \times \mathbf{r_{AC}})$. An example is shown in Fig. 5.4. A third extra charge can also be added and is placed along the bisector vector $\mathbf{r_3} = \lambda_3(\mathbf{r_{AB}} + \mathbf{r_{AC}})$.

- Three bonds - A second off center charge is placed along the $\mathbf{r_1}$ vector, $\mathbf{r_2} = \lambda_2(\mathbf{r_{AB}} - \mathbf{r_{AC}}) \times (\mathbf{r_{AD}} - \mathbf{r_{AC}})$. An exception is made for primary amine groups with the second off center charge placed along the bisector of the NH2 angle $\mathbf{r_2} = \lambda_2(\mathbf{r_{NH_1}} + \mathbf{r_{NH_2}})$. This is necessary as the regions between the nitrogen and hydrogen atoms exhibit anisotropy in ESP. This is not due to a physical characteristic but a consequence of using the individual atom's electron density to fit additional charges which includes areas between bonds.



(a) Two extra off center charges placed perpendicular to the plane of the angle.

(b) Two extra off center charges in the plane of the angle.

Fig. 5.4: An example of the directions off center charges are placed for the two bond case.

When additional charges are added, the first site's position and charge continue to be updated. Consequently, the position and charge of the first point is dependent on the position and charge of the subsequent points. This results in the problem growing exponentially more expensive with the number of extra charges. To accommodate this, the increments in the parameter search are made larger when multiple charges are needed.

An additional consideration when multiple charges are used is at what point to increase the number of charges added. If the ESP error falls below the error threshold of 0.9025 kcal/mol no further charges are added. Furthermore, there must also be an improvement in the ESP error of at least 0.0625 kcal/mol before an additional charge is added.

5.3 Methods

A set of 39 small molecules were selected to test the anisotropy method. The linear scaling DFT code ONETEP was used to calculate the ground-state electron density.[106] Four nonorthogonal generalised Wannier functions (NGWFs) were used for all atoms with the exception of hydrogen which used one. The NGWFs had a radii of 10 Bohr. The periodic sine (psinc) basis was used to describe the NGWFs, with a grid size ($0.45a_0$) employed that corresponds to a plane wave cutoff energy of 1020 eV. The PBE exchange-correlation functional was used with PBE OPIUM norm-conserving pseudopotentials.[92] The calculation was carried out in an implicit solvent of $\varepsilon = 4$ to include the induction effects described in Chapter 2. Off center point charges were added using the method described in Section 5.2.1. An atom was determined to be a neighboring atom if it was positioned within the free van der Waals radius, taken from Ref. 124. The grid spacing used for the electron density was $0.225a_0$, half the psinc spacing.

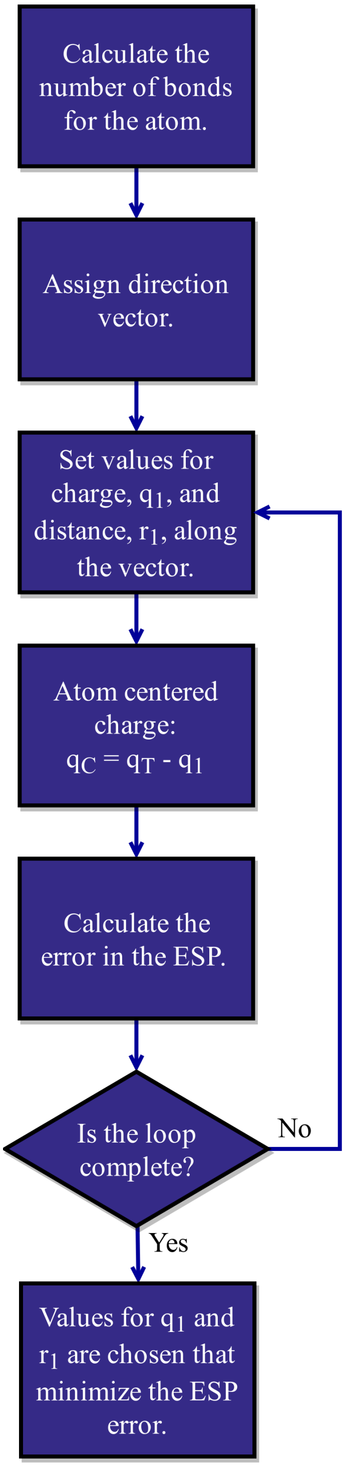
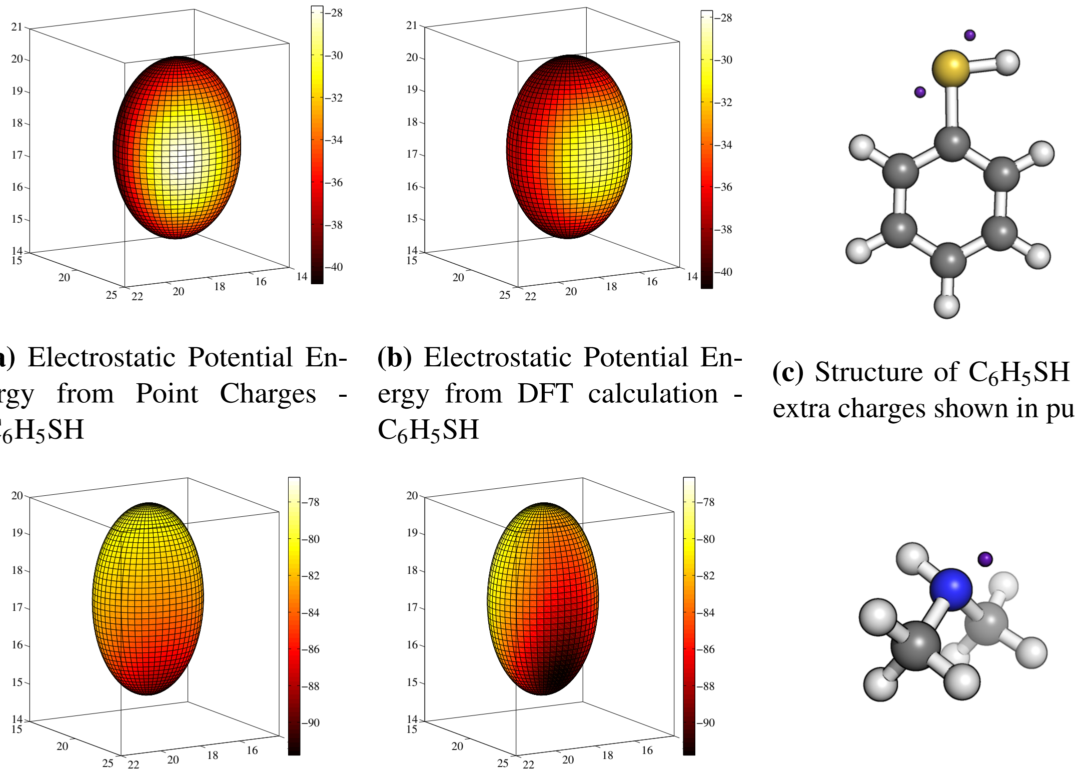

Calculate the number of bonds for the atom.
Assign direction vector.
Set values for charge, q_1, and distance, r_1, along the vector.
Atom centered charge: $q_C = q_T - q_1$
Calculate the error in the ESP.
Is the loop complete?
No
Yes
Values for q_1 and r_1 are chosen that minimize the ESP error.

Fig. 5.5: Workflow of method used to find optimal positions for off center point charges. Where q_T is the total charge on the atom and q_C is the atom centered charge.

5.4 Results

Table 5.1 shows a summary of the results. When extra charges are added, a mean decrease in the ESP energy error of 65.8% occurs. The error in the ESP energy recreation of these anisotropic atoms becomes comparable to ‘isotropic’ atoms with the addition of off center point charges. It is worth noting that the remaining error may be of little consequence. Differences between the ESP will not matter in certain areas as the regions will not be accessible to solvents. Additionally, the error is the difference in potential felt by a unit charge, in real solvents the charge carried will be less.



(a) Electrostatic Potential Energy from Point Charges - C_6H_5SH

(b) Electrostatic Potential Energy from DFT calculation - C_6H_5SH

(c) Structure of C_6H_5SH with extra charges shown in purple.

(d) Electrostatic Potential Energy from Point Charges - $(CH3)_2NH$

(e) Electrostatic Potential Energy from DFT calculation $(CH3)_2NH$

(f) Structure of $(CH3)_2NH$ with extra charges shown in purple.

Fig. 5.6: Two examples of the ability of the method to recreate the anisotropy in the ESP. The potential is shown at 1.4 times the van der Waals radius and is in kcal/mol.

To illustrate how the method qualitatively reproduces the ESP, two examples of the electrostatic potential with off center point charges are shown in Fig. 5.6. The ESP from the DFT calculation is given for comparison. For both the cases shown, the anisotropy is caused by the presence of lone pairs. The method generally results in chemically intuitive

Molecule	Number of Off Center Charges	Atom	ESP Error (kcal/mol) Before	After	Comments
C_2H_6	0				
C_3H_8	0				
C_5H_5N	1	N	2.61	0.19	Two bonds, bisector
$C_6H_5CCH_2CH_3$	0				
$C_6H_5CF_3$	0				
$C_6H_5CH_3$	0				
C_6H_5Cl	1	Cl	0.91	0.45	One bond
C_6H_5CN	0				
$C_6H_5COCH_3$	0				
$C_6H_5CONH_2$	0				
$C_6H_5COOCH_3$	2	O (C-O-C)	1.23	0.67	Two bond, perpendicular to plane
C_6H_5F	0				
$C_6H_5N(CH_3)_2$	1	N	1.86	0.41	Three bond
$C_6H_5NH_2$	1	N	1.02	0.46	Three bond
$C_6H_5NHCH_3$	0				
$C_6H_5NO_2$	0				
$C_6H_5OCH_3$	1	O	1.20	0.51	Two bond, bisector
C_6H_5OH	0				Two bond, in plane
$C_6H_5SCH_3$	2	S	1.44	0.60	Two bond, perpendicular to plane
C_6H_5SH	2	S	4.08	0.81	Two bond, in plane
C_6H_6	0				
$(CH_3)_2CO$	0				
$(CH_3)_2NH$	1	N	2.24	1.06	Three bond
$CH_3CH_2NO_2$	0				
CH_3CHCH_2	0				
CH_3CHO	0				
CH_3Cl	2	Cl	0.95	0.10	One bond
CH_3CN	0				
$CH_3CON(CH3)_2$	0				
CH_3CONH2	0				
CH_3COOCH_3	0				
CH_3COOH	0				
CH_3NH_2	2	N	2.10	0.33	Three bond
CH_3OCH_3	1	O	1.50	0.50	Two bond, bisector
CH_3OH	0				
CH_3SCH_3	2	S	1.92	0.42	Two bond, perpendicular to plane
CH3SH	2	S	2.09	1.12	Two bond, perpendicular to plane

Molecule	Number of Off Center Charges	Atom	ESP Error(kcal/mol) Before	After	Comments
CH_3SOCH_3	1	S	4.03	0.75	Three bond
$HCOOCH_3$	2	O (C-O-C)	1.33	0.79	Two bond, perpendicular to plane

Table 5.1: The change in the error of the electrostatic potential for a set of small molecules. The atom exhibiting anisotropy is stated along with the number of charges added.

positions for the added charges, as expected from using the symmetric constraints, however the positions of the extra charges do not always correspond to those expected. For example in Fig. 5.6c the lone pairs are perpendicular to the plane but the off center point charges are in the plane.

It is interesting to note which atoms are designated as displaying anisotropy and those which are not. For example, the chlorine atoms in C_6H_5Cl and CH_3Cl both require additional charges whilst none of the fluorine atoms do. All four fluorine atoms have an ESP error below 0.60 kcal/mol. This is in agreement with the trends seen in Ref. 66 and can be explained by fluorine generally not having a σ hole present.[94] Additionally, the poor ability for atom-centered point charges to recreate the electrostatic properties for sulphur was shown in Ref. 66. In Table 5.1, the maximum error in the table (4.08 kcal/mol) is due to a sulphur atom and this error is over 50% larger than the error for any other element.

The addition of extra point charges changes the dipole moment of the molecule. Table 5.2 shows the difference in the dipole moment between the point charge model and the actual DFT values. For all the molecules present the difference is 0.20 eBohr or below. It would be preferable if this value was lower and there was better recreation of the molecule's dipole. However, the differences between the dipoles of the molecule does not significantly affect liquid properties, as shown in the following section.

The set of the small molecules tested are not representative of the types of molecules for which the anisotropy method will ultimately be used. One such application will be to improve the ESP recreation of ligands in free energy calculations of protein-ligand complexes. Therefore, our anisotropy method was also applied to two drug-like molecules. One of the ligands chosen for our test set binds to the p38 kinase protein, the second was tested as a potential inhibitor of plasmin enzymes.[101] Nitrogen, oxygen, chlorine, and sulphur atoms are present in this set of ligands. These atoms have previously been shown to exhibit anisotropy and therefore could cause inaccuracies in calculations.

Table 5.3 shows the charges and ESP errors of the two molecules, with the positions of the charges shown in Fig. 5.8. The off center point charges cause a reduction of 54.2% in the ESP error. With the exception of the sulphur atom, this reduction is achieved without the

	Overall Dipole (e bohr)			
Molecule	**Before**	**After**	**Correct**	**Error**
C_5H_5N	0.95	1.15	1.08	0.07
C_6H_5Cl	0.90	0.78	0.75	0.03
$C_6H_5COOCH_3$	0.99	0.86	0.90	0.04
$C_6H_5N(CH_3)_2$	0.94	0.95	0.90	0.04
$C_6H_5NH_2$	0.82	0.74	0.75	0.02
$C_6H_5OCH_3$	0.63	0.71	0.61	0.10
$C_6H_5SCH_3$	0.64	0.69	0.62	0.07
C_6H_5SH	0.59	0.40	0.46	0.05
$(CH_3)_2NH$	0.36	0.49	0.46	0.04
CH_3CL	1.00	0.98	0.85	0.13
CH_3NH_2	0.60	0.76	0.68	0.08
CH_3OCH_3	0.54	0.71	0.55	0.16
CH_3SCH_3	0.78	0.90	0.74	0.16
CH3SH	0.83	0.85	0.71	0.14
CH_3SOCH_3	2.02	2.00	1.81	0.19
$HCOOCH_3$	0.90	0.80	0.84	0.05

Table 5.2: The dipole moments of a set of molecules before extra charges are added, after they are included, the actual value from the DFT calculation and the difference between the two values.

need for more than one off center charge. The ESP error for atom centered point charges is particularly large for the plasmin inhibitor, with both the nitrogen and oxygen atoms having an error greater than 1.50kcal/mol. This molecule was in a test set used in Ref. 101 to compare experimental results to free energy perturbation calculations. A lack of correlation between the free energy calculations and the experimental results was seen in this work and may in part have been due to the use of atom centered charges.

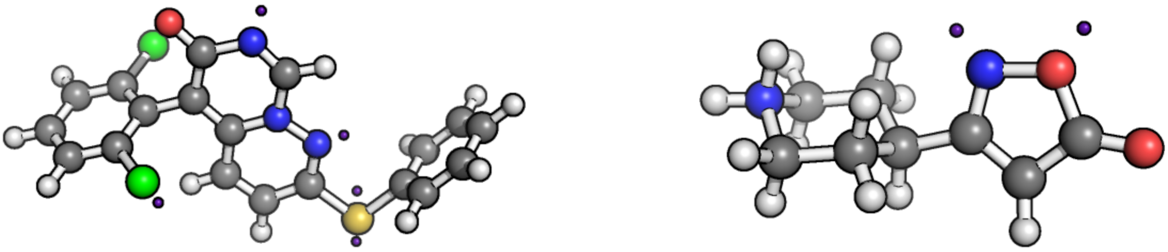


Fig. 5.7: The two drug like molecules studied with the extra charge sites shown in purple. The left hand molecule is the p38 kinase ligand and the right hand molecule is the plasmin inhibitor.

	Number of Off		ESP Energy Error (kcal/mol)	
Molecule	Center Charges	Atom	Before	After
Plasmin Inhibitor	1	N	2.16	1.10
	1	O	1.82	1.04
p38 Ligand	1	N	0.98	0.28
	1	N	1.29	0.36
	1	Cl	1.04	0.56
	3	S	1.17	0.66

Table 5.3: The change in the error in the electrostatic potential energy with addition of off center point charges, for two drug like molecules. The second chlorine in the p38 kinase ligand is symmetric to the first chlorine atom and therefore would be given identical charges to the first atom when used in simulations.

5.5 Testing Liquid Properties

The work in this section was carried out by Joshua Horton at Newcastle University.

The liquid properties of a set of small molecules was subsequently used to test the anisotropy method outlined in this chapter, as well as the general parametrization scheme that will be used for small molecules. The bond and angle parameters used in the force field were derived using the modified Seminario method, and the torsional parameters were calculated using a similar approach to that discussed in Section 4. Non-bonded terms were derived using the DDEC method, with off center point charges then added.

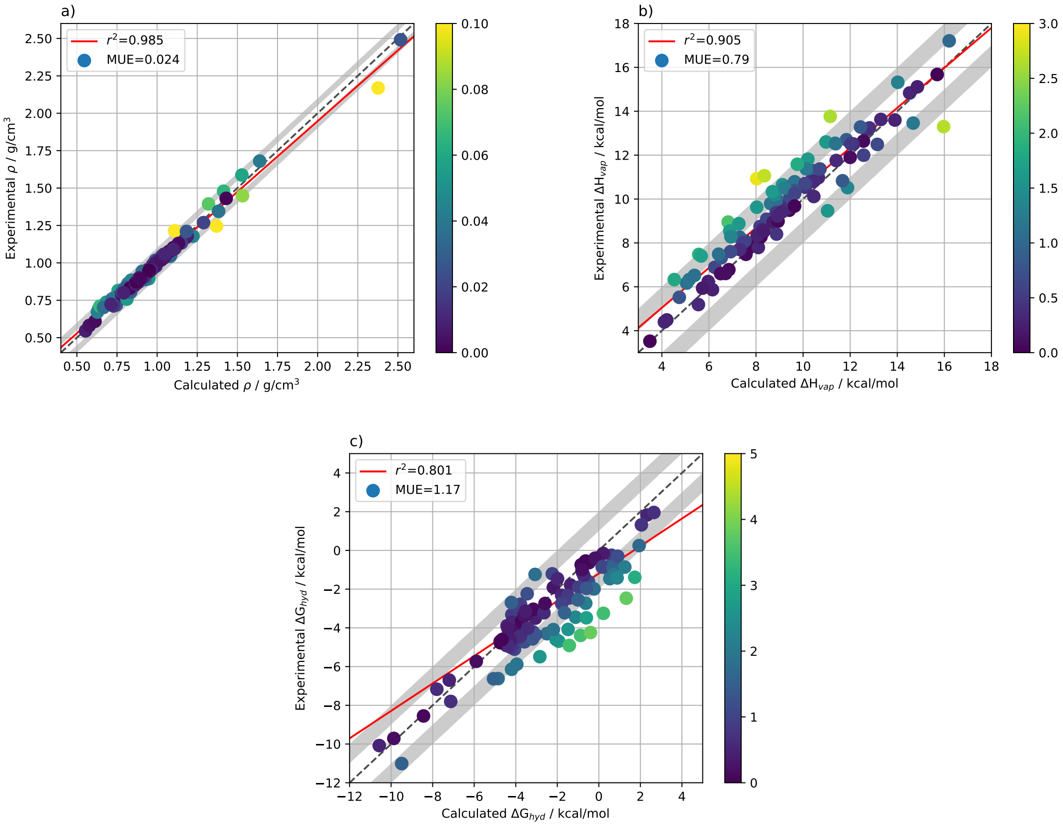


Fig. 5.8: The (a) liquid density, (b) heat of vaporization and (c) free energy of hydration calculated for the 100 organic molecule test using QUBE FF parameter and compared to experiment. The MUE and r^2 correlation are given in each figure

The accurate recreation of the free energy of hydration, heat of vaporization and liquid densities is demonstrated in Fig. 5.8. The mean unsigned errors when extra sites are not included are 0.023 g/cm^3, 0.85 kcal/mol and 1.51 kcal/mol for the liquid density, heat of vaporization and free energy of hydration respectively. This shows that the improvements made to the ESP of the molecules also lead to improvements the free energy of hydration for a set of organic molecules.

5.6 Conclusions

When point charges are only present at the center of the atom, the electrostatic potential of a molecule can fail to be accurately reproduced.[66] Using off center point charges the electrostatic potential can be recreated with increased accuracy.[22]

We have developed an approach that uses the partitioned electron density, computed with the DDEC scheme implemented in ONETEP, to find the optimal positions for extra point charges. The method assigns a vector that point charges can be placed along based on the number of bonds formed by an atom. The distance along the vector and charge of the site are chosen so that the QM ESP is recreated as accurately as possible.

Using this method for a set of small molecules, an atom's mean ESP error was reduced by 65.8%. The error in the ESP of a molecule will be lower than suggested by the atom's ESP. This is because the recreation of the ESP in certain regions, such as between atoms, is inconsequential as it is not accessible. Calculations were also performed (*by Joshua Horton at Newcastle University*) to show that when off center point charges are included our force field can accurately recreate liquid properties. Although we cannot say that the method is appropriate for all possible organic molecules, it has been applied successfully to over 100 small organic molecules.

Additionally, the method was applied to two ligands to demonstrate the improvements in ESP that could occur for drug-like molecules. Further work will investigate whether the improvements in the ESP translates to more accurate free energy calculations and explore using off center point charges to better model a protein's ESP.

Chapter 6

Conclusions

Molecular mechanics (MM) force fields are used to understand and predict a wide range of biological phenomena, including protein–ligand binding free energies,[21,120] enzyme catalysis,[77] and protein folding.[76] The majority of biomolecular force fields may be decomposed into intermolecular interactions, which describe the electrostatic and van der Waals energies, and intramolecular interactions, which describe covalent bonding.[24,54,114] The work presented in this thesis aimed to create a biomolecular force field, named QUBE, with system-specific intermolecular terms derived from quantum mechanical data. System-specific terms naturally include local polarization effects, and the use of quantum mechanical data allows for an automated parametrization workflow and a consistent parametrization methodology for small and large molecules. The work in this thesis built on that of Ref. 22, which established that using the electron density to derive intermolecular interactions was a viable parametrization approach for small molecules. However, Ref. 22 used a force field which was a combination of new intermolecular terms with OPLS intramolecular terms. The reparametrization of the intramolecular components, and testing the resulting force field, was the focus of this thesis.

The intramolecular component of a force field is split into harmonic bond and angle terms, which are used to describe vibrations of the bonds and angles around their equilibrium positions, and anharmonic torsional terms. To parametrize the bond and angle terms, we proposed a modification to the Seminario method.[103] The new method was shown to derive accurate harmonic bond and angle force field parameters directly from the quantum mechanical Hessian matrix. Using the modified Seminario method, there was a reduction in the average error in the reproduction of quantum mechanical normal mode frequencies for a benchmark set of 70 molecules. The error fell from 12.3% using the original Seminario method, to under 6.3%. A complete set of bond and angle force field parameters for the twenty naturally-occurring amino acids was then produced using the modified Seminario method. These parameters are used in our protein force field.

The torsional components of a force field are dependent on the non-bonded parameters. Therefore, their reparametrization is necessary for accurate conformational preferences. Molecular dynamics simulations of short peptides and proteins were carried out to test the accuracy of the new torsional parameters, as well as the non-bonded and bond and angle parametrization methods. The simulations of short peptides were promising, with the backbone and sidechain conformations present in the simulations compatible with experimental results. However, for the Ala$_5$ peptide the intrinsic difficulties of fitting torsional parameters to a system-specific force field were seen, with regularization required to adjust the backbone torsional parameters. In the protein simulations performed, the system-specific force field managed to retain the experimental structures well in two of the five proteins tested. In the remaining three proteins, regions with no clear secondary structure or with a turn had a particularly high RMSD. However, the majority of regions in the proteins remained close to the experimental structures and the J coupling error of GB3 and ubiquitin were only slightly higher than that of the OPLS-AA/M force field. Given that this is the first version of our force field, compared to decades of work carried out on OPLS this is a good starting point for the QUBE force field. In future work improving the functional form or fitting process of the torsional parameters may be sufficient for the QUBE force field to have RMSDs that are comparable to transferable force fields such as OPLS-AA/M or AMBER ff15ipq.

The final part of this work looked at improving the electrostatic potential of the QUBE force field by using additional off center point charges. The method developed takes as input only the electron density, is automated and is suitable for large biological molecules. The use of off center point charges greatly improved the recreation of the atoms' ESP, both for a set of small molecules and for two drug like molecules. The increased accuracy of the ESP led to improved free energies of hydration for a set of small molecules.

The addition of off center sites not only demonstrated the advantages of extra charges but also the ease that systematic improvements can be made to a force field that is derived using just quantum mechanical data. We hope that using system-specific non-bonded terms with a traditional force fields' functional form will lead to levels of accuracy that are not obtainable with transferable force fields. However, in order to achieve this, further improvements to the functional form may be required. Possible changes that could be made to the force field include the addition of a r^{-8} term for the van der Waals component or using a Born-Mayer potential to describe the short ranged component. Again, the use of quantum mechanical data, and our existing automated workflow, should help to simplify these alterations. A more radical change that could be made to our force field is to adapt it from a non-polarizable to a polarizable force field. The parametrization of the polarizable terms should be able to be determined directly from the ONETEP calculation, and therefore fit easily into our parametrization process.[96]

The development of a system-specific biomolecular force field has not been the only outcome of this work. Two new methods have also been developed that could assist with the parametrization of traditional force fields. Force field parametrization is typically only performed by experts in the field, we hope that the advances made will help to open up the process to more users. The modified Seminario method provides a fast and accurate way to obtain bond and angle parameters and, as the software implementing the method has been made freely available (*https://github.com/aa840/ModSeminario*), it can be easily used by other research groups. This could be required when studying biomolecular systems containing metals, which can lack general force field parameters, or molecules with poorly parametrized bond and angle terms.[72] The approach used in the anisotropy method may also be useful for other force fields. Schemes could be developed that use the same set of predefined vectors for off center point charges, and rely on only the ESP, but use the parametrization methodologies of traditional force fields. Therefore, although the focus of this project was the creation of a system-specific QM force field, parametrization methods that can be used with traditional force fields have also been developed.

References

[1] A. E. A. Allen, M. C. Payne, and D. J. Cole. Harmonic Force Constants for Molecular Mechanics Force Fields via Hessian Matrix Projection. *J. Chem. Theory Comput.*, 14 (1):274–281, 2018.

[2] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, USA, 1989.

[3] R. Anandakrishnan, B. C., O. A. V., and I. S. Point Charges Optimally Placed to Represent the Multipole Expansion of Charge Distributions. *PLOS ONE*, 8(7):e67715, 2013.

[4] F. Avbelj, S. G. Grdadolnik, J. Grdadolnik, and R. L. Baldwin. Intrinsic backbone preferences are fully present in blocked amino acids. *Proc. Natl. Acad. Sci. U.S.A.*, 103(5):1272–1277, 2006.

[5] E. J. Bautista and J. M. Seminario. Harmonic force field for glycine oligopeptides. *Int. J. Quantum Chem.*, 108(1):180–188, 2008.

[6] R. Best, N.-V. Buchete, and G. Hummer. Are Current Molecular Dynamics Force Fields too Helical? *Biophys. J.*, 95(1):L07–L09, 2008.

[7] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput.*, 8(9):3257–3273, 2012.

[8] D. R. Bowler, T. Miyazaki, and M. J. Gillan. Recent progress in linear scaling ab initio electronic structure techniques. *J. Phys. Condens. Matter*, 14(11):2781, 2002.

[9] M. Buck, S. Bouguet-Bonnet, R. W. Pastor, and A. D. MacKerell. Importance of the CMAP Correction to the CHARMM22 Protein Force Field: Dynamics of Hen Lysozyme. *Biophys. J.*, 90(4):L36–L38, 2005.

[10] P. Bultinck, C. Van Alsenoy, P. W. Ayers, and R. Carbó-Dorca. Critical analysis and extension of the Hirshfeld atoms in molecules. *J. Chem. Phy.*, 126(14):144111, 2007.

[11] S. K. Burger, M. Lacasse, T. Verstraelen, J. Drewry, P. Gunning, and P. W. Ayers. Automated Parametrization of AMBER Force Field Terms from Vibrational Analysis with a Focus on Functionalizing Dinuclear Zinc(II) Scaffolds. *J. Chem. Theory Comput.*, 8(2):554–562, 2012.

[12] D. Bykov, T. Petrenko, R. Izsák, S. Kossmann, U. Becker, E. Valeev, and F. Neese. Efficient implementation of the analytic second derivatives of Hartree–Fock and hybrid DFT energies: a detailed analysis of different approximations. *Mol. Phys.*, 113(13-14): 1961–1977, 2015.

[13] R. Cammack. *Oxford dictionary of biochemistry and molecular biology*. 2nd edition, 2006.

[14] D. Case, D. Cerutti, T. Cheatham, T. D. III, R. Duke, T. Giese, H. Gohlke, A. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. Roe, A. Roitberg, C. Sagui, C. Simmerling, W. Botello-Smith, J. Swails, R. Walker, J. Wang, R. Wolf, X. Wu, L. Xiao, D. York, and P. Kollman. AMBER 2017. *University of California*, 2017.

[15] D. S. Cerutti, W. C. Swope, J. E. Rice, and D. A. Case. ff14ipq: A Self-Consistent Force Field for Condensed-Phase Simulations of Proteins. *J. Chem. Theory Comput.*, 10(10):4515–4534, 2014.

[16] C. Chipot and A. Pohorille. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*. Springer, 2007.

[17] J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande. Alchemical free energy methods for drug discovery: progress and challenges. *Curr. Opin. Struct. Biol.*, 21(2):150–160, 2011.

[18] J. J. Chou, D. A. Case, and A. Bax. Insights into the Mobility of Methyl-Bearing Side Chains in Proteins from $^3J_{CC}$ and $^3J_{CN}$ Couplings. *J. Am. Chem. Soc.*, 125(29): 8959–8966, 2003.

[19] D. J. Cole and N. D. M. Hine. Applications of large-scale density functional theory in biology. *J. Phys. Condens. Matter.*, 28(39):393001, 2016.

[20] D. J. Cole, E. Rajendra, M. Roberts-Thomson, B. Hardwick, G. J. McKenzie, M. C. Payne, A. R. Venkitaraman, and C.-K. Skylaris. Interrogation of the Protein-Protein Interactions between Human BRCA2 BRC Repeats and RAD51 Reveals Atomistic Determinants of Affinity. *PLOS Comput. Biol.*, 7(7):1–16, 2011.

[21] D. J. Cole, J. Tirado-Rives, and W. L. Jorgensen. Molecular dynamics and Monte Carlo simulations for protein–ligand binding and inhibitor design. *Biochim. Biophys. Acta*, 1850(5):966–971, 2015.

[22] D. J. Cole, J. Z. Vilseck, J. Tirado-Rives, M. C. Payne, and W. L. Jorgensen. Biomolecular Force Field Parameterization via Atoms-in-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.*, 12(5):2312–2323, 2016.

[23] E. R. Davidson and S. Chakravorty. A test of the Hirshfeld definition of atomic charges and moments. *Theor. Chim. Acta*, 83(5-6):319–330, 1992.

[24] K. T. Debiec, D. S. Cerutti, L. R. Baker, A. M. Gronenborn, D. A. Case, and L. T. Chong. Further along the Road Less Traveled: AMBER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. *J. Chem. Theory Comput.*, 12 (8):3926–3947, 2016.

[25] K. Demadis, D. Dattelbaum, E. Kober, J. Concepcion, J. Paul, T. Meyer, and P. White. Vibrational and structural mapping of [Os(bpy)^3]$^{3+/2+}$ and [Os(phen)^3]$^{3+/2+}$. *Inorg. Chim. Acta*, 360(3):1143–1153, 2007.

[26] R. W. Dixon and P. A. Kollman. Advancing beyond the atom-centered model in additive and nonadditive molecular mechanics. *J. Comput. Chem.*, 18(13):1632–1646, 1997.

[27] L. S. Dodda, I. Cabeza de Vaca, J. Tirado-Rives, and W. L. Jorgensen. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.*, 45(W1):W331–W336, 2017.

[28] L. S. Dodda, J. Z. Vilseck, J. Tirado-Rives, and W. L. Jorgensen. 1.14*CM1A-LBCC: Localized Bond-Charge Corrected CM1A Charges for Condensed-Phase Simulations. *J. Phys. Chem. B*, 121(15):3864–3870, 2017.

[29] J. Dziedzic, H. H. Helal, C.-K. Skylaris, A. A. Mostofi, and M. C. Payne. Minimal parameter implicit solvent model for ab initio electronic-structure calculations. *EPL*, 95(4):43001, 2011.

[30] J. Dziedzic, S. J. Fox, T. Fox, C. S. Tautermann, and C.-K. Skylaris. Large-scale DFT calculations in implicit solvent—A case study on the T4 lysozyme L99A/M102Q protein. *Int. J. Quantum. Chem.*, 113(6):771–785, 2013.

[31] D. Frenkel and B. Smit. *Understanding Molecular Simulation*. Academic Press, San Diego, 2002.

[32] 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, J. M. 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 Revision E.01, 2009. Gaussian Inc. Wallingford CT.

[33] G. Giuliani and G. Vignale. *Introduction to the electron liquid*. Cambridge University Press, 2005.

[34] J. Graf, P. H. Nguyen, G. Stock, and H. Schwalbe. Structure and Dynamics of the Homologous Series of Alanine Peptides: A Joint Molecular Dynamics/NMR Study. *J. Am. Chem. Soc.*, 129(5):1179–1189, 2007.

[35] J. Grdadolnik, V. Mohacek-Grosev, R. L. Baldwin, and F. Avbelj. Populations of the three major backbone conformations in 19 amino acid dipeptides. *Proc. Natl. Acad. Sci. U.S.A.*, 108(5):1794–1798, 2011.

[36] S. Grimme. A General Quantum Mechanically Derived Force Field (QMDFF) for Molecules and Condensed Phase Simulations. *J. Chem. Theory Comput.*, 10(10): 4497–4514, 2014.

[37] D. Hames and N. Hooper. *Instant Notes in Biochemistry*. Instant Notes. Taylor & Francis, 2006.

[38] N. Hansen and W. F. van Gunsteren. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.*, 10(7):2632–2647, 2014.

[39] E. Harder, V. M. Anisimov, I. V. Vorobyov, P. E. M. Lopes, S. Y. Noskov, A. D. MacKerellJr., and B. Roux. Atomic Level Anisotropy in the Electrostatic Modeling of Lone Pairs for a Polarizable Force Field Based on the Classical Drude Oscillator. *J. Chem. Theory Comput.*, 2(6):1587–1597, 2006.

[40] P. D. Haynes, A. A. Mostof, C.-K. Skylaris, and M. C. Payne. ONETEP: linear-scaling density-functional theory with plane-waves. *J. Phys. Conf. Ser.*, 26(1):143, 2006.

[41] N. Hine, P. Haynes, A. Mostofi, C.-K. Skylaris, and M. Payne. Linear-scaling density-functional theory with tens of thousands of atoms: Expanding the scope and scale of calculations with ONETEP. *Comput. Phys. Commun.*, 180(7):1041–1053, 2009.

[42] N. D. M. Hine, M. Robinson, P. D. Haynes, C.-K. Skylaris, M. C. Payne, and A. A. Mostofi. Accurate ionic forces and geometry optimization in linear-scaling density-functional theory with local orbitals. *Phys. Rev. B*, 83(19):195102, 2011.

[43] S. A. Hollingsworth and P. A. Karplus. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol. Concepts*, 1(3-4):271–283, 2010.

[44] J. Horton, A. E. A. Allen, L. S. Dodda., M. C. Payne, and D. J. Cole. QUBEKit: Automating the Derivation of Force Field Parameters from Quantum Mechanics. *Paper in Preparation*.

[45] J.-S. Hu and A. Bax. Determination of ϕ and $\chi 1$ Angles in Proteins from 13C-13 Three-Bond J Couplings Measured by Three-Dimensional Heteronuclear NMR. *J. Am. Chem. Soc.*, 119(27):6360–6368, 1997.

[46] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, L. d. G. Bert, H. Grubmüller, and A. D. MacKerell. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods.*, 14(1):71–73, 2016.

[47] D. J. Huggins, A. R. Venkitaraman, and D. R. Spring. Rational Methods for the Selection of Diverse Screening Compounds. *ACS Chem. Biol.*, 6(3):208–217, 2011.

[48] J. Hughes, S. Rees, S. Kalindjian, and K. Philpott. Principles of early drug discovery. *Br. J. Pharmacol.*, 162(6):1239–1249, 2011.

[49] M. A. Ibrahim. Molecular mechanical study of halogen bonding in drug discovery. *J. Chem. Theory Comput.*, 32(12):2564–2574, 2011.

[50] K. K. Irikura, R. D. Johnson, and R. N. Kacker. Uncertainties in Scaling Factors for ab Initio Vibrational Frequencies. *J. Phys. Chem. A*, 109(37):8430–8437, 2005.

[51] R. D. Johnson III. NIST Computational Chemistry Comparison and Benchmark Database , October accessed 2016. http://cccbdb.nist.gov/.

[52] R. O. Jones and O. Gunnarsson. The density functional formalism, its applications and prospects. *Rev. Mod. Phys.*, 61(3):689–746, 1989.

[53] W. L. Jorgensen and P. Schyman. Treatment of Halogen Bonding in the OPLS-AA Force Field: Application to Potent Anti-HIV Agents. *J. Chem. Theory Comput.*, 8(10): 3895–3801, 2012.

[54] W. L. Jorgensen and J. Tirado-Rives. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110(6):1657–1666, 1988.

[55] W. L. Jorgensen and J. Tirado-Rives. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J. Comput. Chem.*, 26(16):1689–1700, 2005.

[56] W. L. Jorgensen and J. Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U.S.A.*, 102(19):6665–6670, 2005.

[57] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[58] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc*, 118(45):11225–11236, 1996.

[59] W. L. Jorgensen, J. P. Ulmschneider, and J. Tirado-Rives. Free Energies of Hydration from a Generalized Born Model and an All-Atom Force Field. *J. Phys. Chem. B*, 108 (41):16264–16270, 2004.

[60] W. L. Jorgensen, K. P. Jensen, and A. N. Alexandrova. Polarization Effects for Hydrogen-Bonded Complexes of Substituted Phenols with Water and Chloride Ion. *J .Chem. Theory Comput.*, 3(6):1987–1992, 2007.

[61] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B*, 105(28): 6474–6487, 2001.

[62] W. Kang, F. Jiang, and Y.-D. Wu. Universal Implementation of a Residue-Specific Force Field Based on CMAP Potentials and Free Energy Decomposition. *J. Chem. Theory Comput.*, 14(8):4474–4486, 2018.

[63] P. G. Karamertzanis, P. Raiteri, and A. Galindo. The Use of Anisotropic Potentials in Modeling Water and Free Energies of Hydration. *J. Chem. Theory Comput.*, 6(5): 1590–1607, 2010.

[64] N. Kodan, S. Auluck, and B. Mehta. A DFT study of the electronic and optical properties of a photovoltaic absorber material Cu2ZnGeS4 using GGA and mBJ exchange correlation potentials. *J. Alloys Compd.*, 675:236–243, 2016.

[65] J. Kohanoff. *Electronic Structure Calculations for Solids and Molecules: Theory and Computational Methods*. Cambridge University Press, 2006.

[66] C. Kramer, A. Spinn, and K. R. Liedl. Charge Anisotropy: Where Atomic Multipoles Matter Most. *J. Chem. Theory Comput.*, 10(10):4488–4496, 2014.

[67] T. J. Lane, D. Shukla, K. A. Beauchamp, and V. S. Pande. To Milliseconds and Beyond: Challenges in the Simulation of Protein Folding. *Curr. Opin. Struct. Biol.*, 23(1):58–65, 2012.

[68] L. P. Lee, D. J. Cole, C.-K. Skylaris, W. L. Jorgensen, and M. C. Payne. Polarized Protein-Specific Charges from Atoms-in-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.*, 9(7):2981–2991, 2013.

[69] L. P. Lee, N. G. Limas, D. J. Cole, M. C. Payne, C.-K. Skylaris, and T. A. Manz. Expanding the Scope of Density Derived Electrostatic and Chemical Charge Partitioning to Thousands of Atoms. *J. Chem. Theory Comput.*, 10(12):5377–5390, 2014.

[70] J. M. Leonard and W. P. Ashman. Molecular mechanics parameterization: Bond lengths and angles for nitrogen and phosphorus containing compounds. *J. Comput. Chem.*, 11(8):952–957, 1990.

[71] G. Lever, D. J. Cole, R. Lonsdale, K. E. Ranaghan, D. J. Wales, A. J. Mulholland, C.-K. Skylaris, and M. C. Payne. Large-Scale Density Functional Theory Transition State Searching in Enzymes. *J. Phys. Chem. Lett.*, 5(21):3614–3619, 2014.

[72] P. Li and K. M. Merz. MCPB.py: A Python Based Metal Center Parameter Builder. *J. Chem. Inf. Model.*, 56(4):599–604, 2016.

[73] T. C. Lillestolen and R. J. Wheatley. Atomic charge densities generated using an iterative stockholder procedure. *J. Chem. Phys.*, 131(14):144101, 2009.

[74] F. Lin and R. Wang. Systematic Derivation of AMBER Force Field Parameters Applicable to Zinc-Containing Systems. *J. Chem. Theory Comput.*, 6(6):1852–1870, 2010.

[75] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, 78(8):1950–1958, 2010.

[76] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How Fast-Folding Proteins Fold. *Science*, 334(6055):517–520, 2011.

[77] R. Lonsdale, K. E. Ranaghan, and A. J. Mulholland. Computational enzymology. *Chem. Commun.*, 46(14):2354–2372, 2010.

[78] O. Madelung and B. Taylor. *Introduction to Solid-State Theory*. Springer Series in Solid-State Sciences. Springer, 1996.

[79] M. W. Mahoney and W. L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.*, 112(20):8910–8922, 2000.

[80] T. A. Manz and D. S. Sholl. Chemically Meaningful Atomic Charges That Reproduce the Electrostatic Potential in Periodic and Nonperiodic Materials. *J. Chem. Theory Comput.*, 6(8):2455–2468, 2010.

[81] T. A. Manz and D. S. Sholl. Improved Atoms-in-Molecule Charge Partitioning Functional for Simultaneously Reproducing the Electrostatic Potential and Chemical States in Periodic and Nonperiodic Materials. *J. Chem. Theory Comput.*, 8(8):2844–2867, 2012.

[82] R. M. Martin. *Electronic structure: basic theory and practical methods*. Cambridge Univ. Press, Cambridge, 2004.

[83] C. G. Mayne. *Parameterizing Small Molecules Using the Force Field Toolkit (ffTK)*, September 2015.

[84] C. G. Mayne, J. Saam, K. Schulten, E. Tajkhorshid, and J. C. Gumbart. Rapid parameterization of small molecules using the force field toolkit. *J. Comput. Chem.*, 34(32), 2013.

[85] K. Merz, D. Ringe, and C. Reynolds. *Drug Design: Structure and Ligand-based Approaches*. Cambridge University Press, 2010.

[86] V. Minkin, B. Simkin, and R. Minyaev. *Quantum Chemistry of Organic Compounds: Mechanisms of Reactions*. Springer Berlin Heidelberg, 2012.

[87] B. Monserrat, N. D. Drummond, P. Dalladay-Simpson, R. T. Howie, P. López Ríos, E. Gregoryanz, C. J. Pickard, and R. J. Needs. Structure and Metallicity of Phase V of Hydrogen. *Phys. Rev. Lett.*, 120(25):255701–255708, 2018.

[88] P. S. Nerenberg and T. Head-Gordon. Optimizing Protein-Solvent Force Fields to Reproduce Intrinsic Conformational Preferences of Model Peptides. *J. Chem. Theory Comput.*, 7(4):1220–1230, 2011.

[89] K. Nilsson, D. Lecerof, E. Sigfridsson, and U. Ryde. An automatic method to generate force-field parameters for hetero-compounds. *Acta Crystallogr., Sect. D: Biol. Crystallogr.*, 59(2):274–289, 2003.

[90] J. K. Nørskov, F. Abild-Pedersen, F. Studt, and T. Bligaard. Density functional theory in surface chemistry and catalysis. *Proc. Natl. Acad. Sci. U.S.A.*, 108(3):937–943, 2011.

[91] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Rev. Mod. Phys.*, 64(4):1045–1097, 1992.

[92] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77(18):3865–3868, 1996.

[93] C. Pérez, F. Löhr, H. Rüterjans, and J. M. Schmidt. Self-Consistent Karplus Parametrization of 3J Couplings Depending on the Polypeptide Side-Chain Torsion χ_1. *J. Am. Chem. Soc.*, 123(29):7081–7093, 2001.

[94] P. Politzer, J. S. Murray, and T. Clark. Halogen bonding and other σ-hole interactions: a perspective. *Phys. Chem. Chem. Phys.*, 15(27):11178–11189, 2013.

[95] J. W. Ponder, D. A. Case, et al. Force fields for protein simulations. *Adv. Protein Chem.*, 66:27–85, 2003.

[96] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B*, 114(8):2549–2564, 2010.

[97] L. E. Ratcliff, S. Mohr, G. Huhs, T. Deutsch, M. Masella, and L. Genovese. Challenges in large scale quantum mechanical calculations. *WIREs Comput. Mol. Sc.*, 7(1):e1290, 2017.

[98] J.-L. Rivail, M. Ruiz-Lopez, and X. Assfeld. *Quantum Modeling of Complex Molecular Systems*, volume 21. Springer, 2015.

[99] M. J. Robertson, J. Tirado-Rives, and W. L. Jorgensen. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J. Chem. Theory Comput.*, 11 (7):3499–3509, 2015.

[100] L. Salasnich. *Quantum Physics of Light and Matter: A Modern Introduction to Photons, Atoms and Many-Body Systems*. Springer International Publishing, 2014.

[101] T. C. Schmidt, P.-O. Eriksson, D. Gustafsson, D. Cosgrove, B. Frølund, and J. Boström. Discovery and Evaluation of Anti-Fibrinolytic Plasmin Inhibitors Derived from 5-(4-Piperidyl)isoxazol-3-ol (4-PIOL). *J. Chem. Inf. Model.*, 57(7):1703–1714, 2017.

[102] A. P. Scott and L. Radom. Harmonic Vibrational Frequencies: An Evaluation of Hartree-Fock, Møller-Plesset, Quadratic Configuration Interaction, Density Functional Theory, and Semiempirical Scale Factors. *J. Phys. Chem.*, 100(41):16502–16513, 1996.

[103] J. M. Seminario. Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem*, 60(7):1271–1277., 1996.

[104] Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder, and P. Ren. Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.*, 9(9): 4046–4063, 2013.

[105] D. Sholl and J. Steckel. *Density Functional Theory: A Practical Introduction*. Wiley, 2011.

[106] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne. Introducing ONETEP: Linear-scaling density functional simulations on parallel computers. *J. Chem. Phys.*, 122(8):084119, 2005.

[107] C. K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne. Recent progress in linear-scaling density functional calculations with plane waves and pseudopotentials: the ONETEP code. *J. Phys. Condens. Matter.*, 20(6):064209–064218, 2008.

[108] A. J. Stone. Intermolecular Potentials. *Science*, 321(5890):787–789, 2008.

[109] J. Tao, J. P. Perdew, and A. Ruzsinszky. Accurate van der Waals coefficients from density functional theory. *Proc. Natl. Acad. Sci. USA*, 109(1):18–21, 2012.

[110] D. Ting, W. Guoli, S. Maxim, M. Rajib, J. M. I., and D. J. R. L. Neighbor-Dependent Ramachandran Probability Distributions of Amino Acids Developed from a Hierarchical Dirichlet Process Model. *PLOS Comput. Biol.*, 6(4):e1000763, 2010.

[111] A. Tkatchenko and M. Scheffler. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.*, 102(7):073005–073009, 2009.

[112] M. E. Tuckerman. Ab initio molecular dynamics: basic concepts, current trends and novel applications. *J. Phys. Condens. Matter.*, 14(50):1297–1355, 2002.

[113] K. Vanommeslaeghe and A. D. MacKerell. CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochim. Biophys. Acta*, 1850 (5):861–871, 2014.

[114] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. E. M. Lopes, I. Vorobyov, and A. D. M. Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, 31(4):671–690, 2010.

[115] K. Vanommeslaeghe, O. Guvench, and A. D. MacKerell. Molecular Mechanics. *Curr. Pharm. Des.*, 20(20):3281–3292, 2014.

[116] K. Vanommeslaeghe, M. Yang, and A. D. M. Jr. Robustness in the fitting of molecular mechanics parameters. *J. Comput. Chem.*, 36(14):1083–1101, 2015.

[117] J. G. Vinter. Extended electron distributions applied to the molecular mechanics of some intermolecular interactions. *J. Comput. Aided Mol. Des.*, 8(6):653–668, 1994.

[118] V. G. Vivekanand and A. Tkatchenko. Scaling laws for van der Waals interactions in nanostructured materials. *Nat. Commun.*, 4:2341, 2013.

[119] B. Vögeli, J. Ying, A. Grishaev, and A. Bax. Limits on Variations in Protein Backbone Dynamics from Precise Measurements of Scalar Couplings. *J. Am. Chem. Soc.*, 129 (30):9377–9385, 2007.

[120] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, D. L. Romero, C. Masse, J. L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D. L. Mobley, W. L. Jorgensen, B. J. Berne, R. A. Friesner, and R. Abel. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.*, 137(7):2695–2703, 2015.

[121] L.-P. Wang, T. J. Martinez, and V. S. Pande. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *J. Phys. Chem. Letters*, 5(11):1885–1891, 2014.

[122] L.-P. Wang, K. A. McKiernan, J. Gomes, K. A. Beauchamp, T. Head-Gordon, J. E. Rice, W. C. Swope, T. J. Martínez, and V. S. Pande. Building a More Predictive Protein Force Field: A Systematic and Reproducible Route to AMBER-FB15. *J. Phys. Chem. B*, 121(16):4023–4039, 2017.

[123] R. Wang, M. Ozhgibesov, and H. Hirao. Partial Hessian fitting for determining force constant parameters in molecular mechanics. *J. Comput. Chem.*, 37(26):2349–2359, 2016.

[124] T. Watanabe, T. A. Manz, and D. S. Sholl. Accurate Treatment of Electrostatics during Molecular Adsorption in Nanoporous Crystals without Assigning Point Charges to Framework Atoms. *J. Phys. Chem. C*, 115(11):4824–4836, 2011.

[125] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106(3):765–784, 1984.

[126] X. C. Yan, M. J. Robertson, J. Tirado-Rives, and W. L. Jorgensen. Improved Description of Sulfur Charge Anisotropy in OPLS Force Fields: Model Development and Parameterization. *J. Phys. Chem. B*, 121(27):6626–6636, 2017.

[127] S. Zheng, Q. Tang, J. He, S. Du, S. Xu, C. Wang, Y. Xu, and F. Lin. VFFDT: A New Software for Preparing AMBER Force Field Parameters for Metal-Containing Molecular Systems. *J. Chem. Inf. Model.*, 56(4):811–818, 2016.

Appendix A

Bond and Angle Parameters

This section contains additional information about the bond and angle parameterisation method discussed in Chapter 3.

A.1 Errors in Normal Mode Frequencies

		Original	Modified
	OPLS (%)	**Seminario** (%)	**Seminario** (%)
α**-methylstyrene**	8.63(46.95)	15.94 (133.74)	8.00 (41.18)
Acetaldehyde	5.44 (69.46)	8.08 (102.01)	4.91 (69.19)
Acetamide	15.46 (82.34)	21.76 (127.61)	13.70 (63.34)
Acetic acid	5.61 (66.40)	6.90 (79.14)	4.28 (51.12)
Acetone	6.17 (48.49)	10.85 (100.74)	7.02 (67.12)
Acetophenone	5.02 (36.80)	13.68 (131.34)	4.93 (35.44)
Aniline	6.40 (69.80)	15.70 (159.46)	4.73 (39.83)
• **Anisole**	4.65 (40.04)	12.37 (133.92)	3.90 (36.26)
Benzamide	6.32 (58.83)	15.49 (148.49)	5.19 (36.68)
Benzene	5.47 (57.94)	15.27 (172.51)	4.51 (50.38)
Chlorobenzene	5.55 (55.16)	13.52 (145.65)	3.91 (36.60)
Cyclohexane	5.68 (49.45)	9.77 (106.38)	5.77 (57.76)
Cytosine	8.61 (70.51)	16.08 (127.91)	7.65 (44.76)
Dimethylacetamide	9.71 (68.93)	11.03 (92.22)	8.06 (55.83)
Dimethylamine	6.01 (76.29)	9.83 (124.73)	5.16 (66.50)
Dimethylaniline	6.64 (42.27)	10.53 (122.23)	4.96 (43.25)
Dimethylether	8.01 (94.84)	10.04 (131.24)	5.67 (71.71)

Table A.1: The mean percentage error between the QM and MM normal mode frequencies for a set of 38 small molecules. The mean unsigned error is shown in brackets (cm^{-1}). The QM frequencies used in the calculation of the error have been scaled to better reproduce experimental frequencies.[102]

	OPLS (%)	Original Seminario (%)	Modified Seminario (%)
Dimethylsulfide	4.32 (45.40)	8.22 (67.11)	9.06 (76.35)
Dimethylsulfoxide	7.41 (60.88)	8.99 (77.43)	7.23 (60.22)
Ethane	4.47 (55.92)	5.59 (84.69)	4.61 (69.91)
Fluorobenzene	5.40 (57.44)	13.67 (153.35)	3.11 (34.32)
Methyl acetate	9.96 (78.05)	9.05 (82.40)	6.73 (53.07)
Methyl benzoate	5.15 (37.47)	12.21 (121.42)	5.06 (36.32)
Methyl formate	9.32 (94.96)	6.29 (78.81)	6.53 (77.59)
Methylacetamide	8.54 (78.79)	11.94 (130.73)	6.54 (86.30)
Methylamine	6.62 (81.13)	7.19 (94.94)	3.79 (51.64)
Methylaniline	5.90 (45.86)	13.43 (139.77)	4.22 (37.05)
Nitrobenzene	8.19 (53.22)	14.41 (130.10)	7.82 (41.66)
Nitroethane	19.93 (82.77)	19.34 (88.79)	19.5 (75.61)
Phenol	4.53 (48.38)	14.23 (154.53)	3.40 (36.60)
Phosphorine	7.26 (66.06)	13.70 (145.41)	5.02 (41.65)
Propane	4.05 (45.52)	7.05 (92.03)	4.31 (58.54)
Propene	5.41 (52.47)	12.64 (129.80)	6.13 (66.51)
Pyridine	4.00 (43.04)	13.75 (160.66)	3.83 (41.01)
Thioanisole	6.15 (39.79)	12.94 (114.28)	6.18 (42.87)
Triethylamine	6.72 (56.39)	8.04 (93.84)	4.61 (52.59)
Trifluorobenzene	7.49 (76.61)	13.02 (131.72)	5.55 (42.93)
Trifluorotoluene	18.52 (58.58)	24.74 (130.17)	16.67 (38.47)
Mean	7.34 (60.35)	12.30 (119.51)	6.38 (52.32)

	OPLS (%)	**Original Seminario** (%)	**Modified Seminario** (%)
Adenine	7.98 (75.86)	12.34 (107.04)	6.58 (44.77)
β-lactam	7.77 (67.18)	12.79 (128.79)	6.19 (49.22)
Caprolactam	7.53 (47.26)	7.67 (104.10)	6.83 (53.01)
Dibenzofuran	6.99 (68.07)	12.61 (148.94)	4.77 (34.34)
Guanine	12.90 (98.03)	14.76 (145.56)	11.07 (49.25)
Oxetane	5.74 (64.49)	7.81 (109.88)	3.63 (52.15)
Oxirane	9.72 (115.20)	9.83 (132.73)	8.01 (89.73)
Pyrrole	12.13 (135.15)	17.51 (193.96)	6.22 (64.96)
Thiazepine	6.10 (76.63)	9.82 (103.45)	6.38 (48.49)
Uracil	8.91 (77.78)	11.54 (148.37)	8.53 (41.96)
Mean	8.58 (82.56)	11.67 (132.28)	6.82 (52.79)

Table A.2: The mean percentage error between the QM and MM normal mode frequencies for a set of 10 heterocyclic molecules. The mean unsigned error is shown in brackets (cm^{-1}). The QM frequencies used in the calculation of the error have been scaled to better reproduce experimental frequencies.[102]

	OPLS (%)	Original Seminario (%)	Modified Seminario Averaged (%) [1]	Modified Seminario Specific (%) [2]
Ala	9.05 (54.74)	14.55 (107.36)	7.86 (46.21)	7.42 (44.36)
Arg	7.07 (50.76)	12.76 (109.80)	5.86 (40.22)	6.10 (42.53)
Asn	6.88 (48.83)	14.73 (109.79)	6.74 (37.79)	5.91 (35.83)
Asp	7.95 (48.64)	11.20 (96.90)	4.43 (35.68)	4.85 (35.98)
Cys	6.62 (48.67)	12.35 (100.25)	5.86 (39.48)	6.34 (37.98)
Gln	8.69 (46.95)	13.47 (105.39)	6.55 (35.82)	6.74 (34.60)
Glu	7.07 (43.97)	12.15 (96.43)	5.51 (36.56)	5.57 (39.66)
Gly	10.16 (64.98)	17.40 (114.96)	10.75 (52.09)	10.61 (49.40)
Hip	7.62 (43.78)	12.43 (110.15)	6.27 (36.20)	5.86 (36.08)
His	7.53 (43.39)	9.16 (81.65)	5.85 (38.51)	6.24 (40.80)
Ile	5.71 (43.77)	11.44 (97.37)	5.69 (39.21)	5.74 (40.55)
Leu	5.66 (41.09)	11.47 (99.16)	5.39 (39.03)	4.84 (42.35)
Lys	6.85 (50.97)	9.49 (98.73)	4.99 (43.89)	4.78 (43.50)
Met	6.30 (44.72)	11.81 (91.38)	6.79 (41.00)	6.30 (39.04)
Phe	6.31 (34.76)	13.53 (124.08)	5.00 (34.35)	4.79 (31.99)
Pro	6.00 (46.00)	12.66 (109.59)	5.51 (41.76)	5.35 (41.86)
Ser	6.78 (52.06)	10.54 (94.48)	5.52 (42.68)	5.40 (44.07)
Thr	7.58 (45.80)	10.47 (91.29)	5.80 (39.29)	5.23 (38.72)
Trp	6.49 (42.80)	13.19 (126.24)	5.13 (34.44)	4.15 (34.62)
Tyr	5.27 (35.40)	13.54 (124.01)	4.52 (34.42)	4.75 (31.72)
Val	5.49 (45.99)	12.80 (100.54)	6.97 (41.53)	6.74 (42.39)
Mean	7.00 (46.57)	12.44 (104.27)	6.05 (39.53)	5.89 (39.43)

Table A.3: The mean percentage error between the QM and MM normal mode frequencies for a set of dipeptides. The mean unsigned error is shown in brackets (cm^{-1}). The QM frequencies used in the calculation of the error have been scaled to better reproduce experimental frequencies.[102] Hip stands for protonated histidine.

[1] Bond and angle parameters that have been averaged over all 80 dipeptide structures.

[2] Bond and angle parameters are specific to the individual dipeptide.

A.2 Comparison of OPLS and Modified Seminario Method

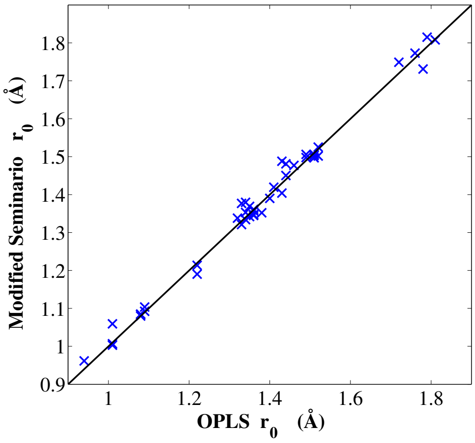


Fig. A.1: A comparison between the OPLS and modified Seminario bond lengths for a set of small molecules.

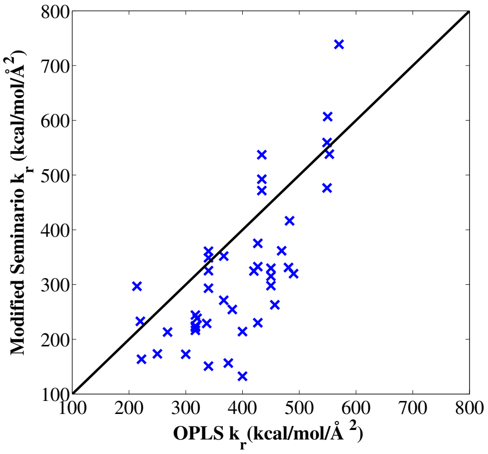


Fig. A.2: A comparison between the OPLS and modified Seminario bond force constants for a set of small molecules.

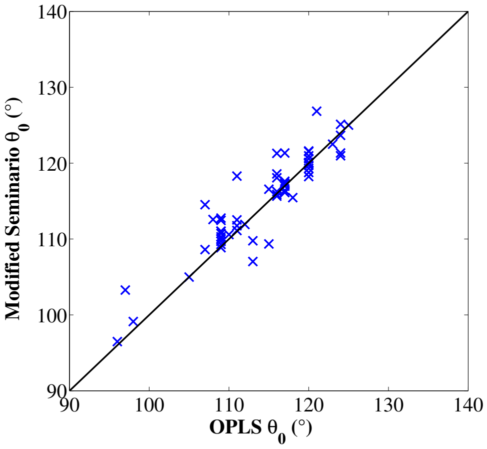


Fig. A.3: A comparison between the OPLS and modified Seminario equilibrium angles for a set of small molecules.

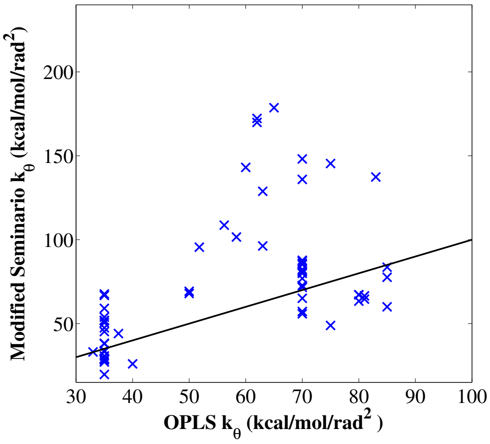


Fig. A.4: A comparison between the OPLS and modified Seminario angle force constants for a set of small molecules.

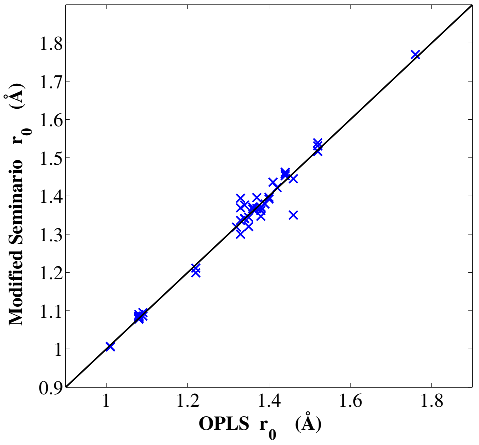


Fig. A.5: A comparison between the OPLS and modified Seminario bond lengths for a set of heterocyclic molecules.

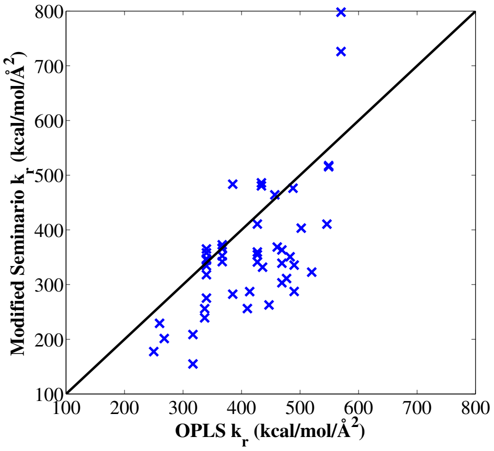


Fig. A.6: A comparison between the OPLS and modified Seminario bond force constants for a set of heterocyclic molecules.

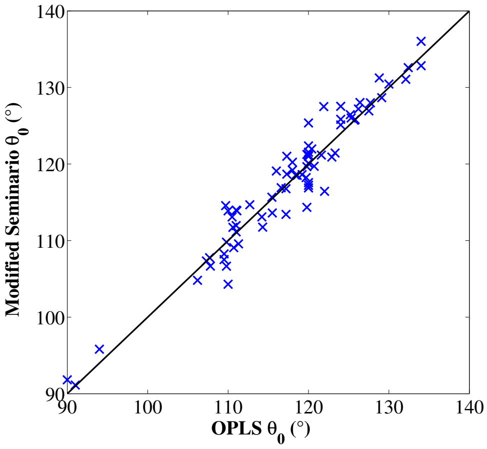


Fig. A.7: A comparison between the OPLS and modified Seminario equilibrium angles for a set of heterocyclic molecules.

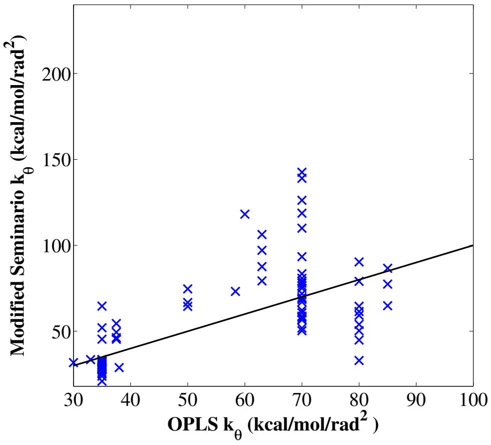


Fig. A.8: A comparison between the OPLS and modified Seminario angle force constants for a set of heterocyclic molecules.

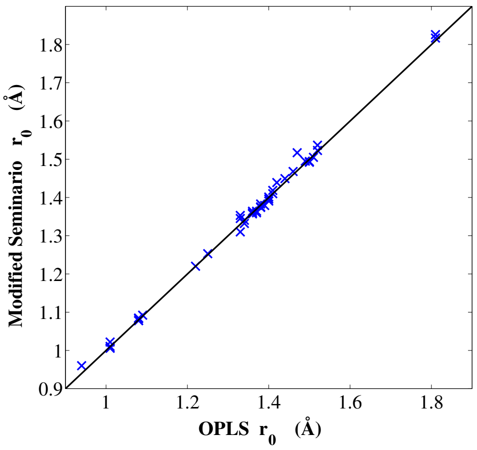


Fig. A.9: A comparison between the OPLS and modified Seminario bond lengths for a set of dipeptides.

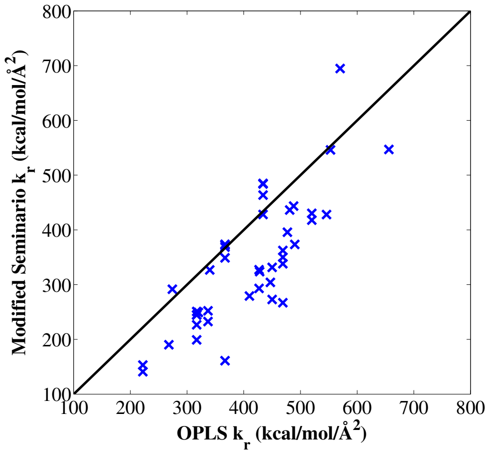


Fig. A.10: A comparison between the OPLS and modified Seminario bond force constants for a set of dipeptides.

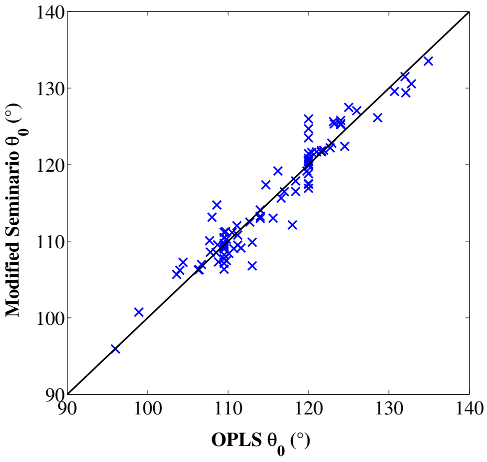


Fig. A.11: A comparison between the OPLS and modified Seminario equilibrium angles for a set of dipeptides.

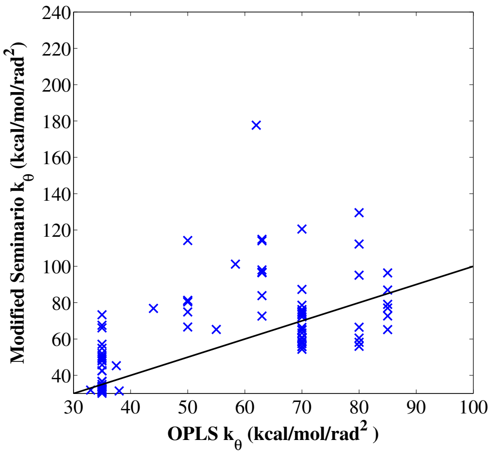


Fig. A.12: A comparison between the OPLS and modified Seminario angle force constants for a set of dipeptides.

Appendix B

Torsional Parameters

B.1 Preliminary Work

Within this section the preliminary work carried out to determine appropriate weighting and regularization values is summarized. MD simulations were then carried out using the backbone torsional parameters. A set of six dipeptides were tested.

B.1.1 Alanine/Glycine Backbone Torsional Parameters

Regularization and Weighting	Dihedral Angle	Torsional Parameters				Error QUBE
$\lambda = 0.00$	ϕ	-0.60	1.18	-2.76	0.00	1.254
No Weighting	ψ	1.26	2.88	-2.08	0.00	
	ϕ'	-3.04	-0.04	1.56	0.00	
	ψ'	1.96	-0.42	0.93	0.00	
$\lambda = 0.01$	ϕ	0.02	0.41	-0.09	0.00	2.880
No Weighting	ψ	0.09	0.43	-0.29	0.00	
	ϕ'	-0.37	-0.16	0.00	0.00	
	ψ'	-0.26	-0.32	-0.24	0.00	

Table B.1: Preliminary backbone torsional parameters tested.

Regularization and Weighting	Dihedral Angle	Torsional Parameters				Error QUBE
$\lambda = 0.00$	ϕ	-0.67	1.13	-3.02	0.00	1.231
1000K	ψ	1.21	2.70	-1.28	0.00	
	ϕ'	-3.08	-0.07	1.85	0.00	
	ψ'	1.88	-0.50	0.11	0.00	
$\lambda = 0.01$	ϕ	0.08	0.99	-0.33	0.00	1.804
1000K	ψ	0.16	1.00	-0.66	0.00	
	ϕ'	-1.00	-0.32	-0.08	0.00	
	ψ'	-0.34	-0.87	-0.58	0.00	
$\lambda = 0.00$	ϕ	0.04	0.75	-0.18	0.00	2.069
2000K	ψ	0.14	0.86	-0.51	0.00	
	ϕ'	-0.80	-0.28	-0.04	0.00	
	ψ'	-0.40	-0.64	-0.46	0.00	
$\lambda = 0.01$	ϕ	0.03	0.60	-0.12	0.00	2.360
2000K	ψ	0.12	0.64	-0.42	0.00	
	ϕ'	-0.60	-0.24	-0.02	0.00	
	ψ'	-0.37	-0.51	-0.37	0.00	

Table B.2: Preliminary backbone torsional parameters tested.

Regularization and Weighting	**J coupling Error**						
	Arg	**Asn**	**Cys**	**Gln**	**Ile**	**Phe**	**RMSE**
$\lambda = 0.00$ *No Weighting*	0.89	0.63	0.60	0.48	0.61	0.16	0.60
$\lambda = 0.01$ *No Weighting*	0.93	0.82	0.39	0.11	0.69	0.01	0.60
$\lambda = 0.00$ *1000K*	0.07	1.47	0.72	0.39	0.36	0.63	0.75
$\lambda = 0.01$ *1000K*	0.50	2.20	1.63	1.18	0.92	1.40	1.41
$\lambda = 0.00$ *2000K*	0.69	0.93	0.38	0.14	0.00	0.22	0.51
$\lambda = 0.01$ *2000K*	0.45	1.62	1.01	0.42	0.04	0.77	0.88

Table B.3: MD simulations of preliminary backbone torsional parameters.

B.1.2 Serine Backbone Torsional Parameters

Regularization was shown to be necessary for the serine backbone torsional parameters. As the backbone terms are dependent on the sidechain values, a second iteration of backbone torsional parameter fitting was performed.

Regularization	**Dihedral**	**Torsional**					**J Coup.**	**Rotamer**
and Weighting	**Angle**	**Parameters**					**Error**	**Error**
$\lambda = 0.00$	ϕ	0.94	0.21	-2.93	0.00	*Ser*	4.75	0.32
No Weighting	ψ	-1.81	2.42	2.57	0.00			
	ϕ’	-3.07	4.74	-1.61	0.00			
	ψ’	0.33	2.29	-2.34	0.00			
$\lambda = 0.05$	ϕ	0.62	-0.50	-1.03	0.00	*Ser*	0.66	0.37
No Weighting	ψ	0.09	0.43	0.18	0.00			
	ϕ’	-0.79	1.51	-0.80	0.00			
	ψ’	0.69	0.30	0.20	0.00			
$\lambda = 0.05$	ϕ	0.51	-0.52	-0.96	0.00	*Ser*	0.27	0.11
No Weighting	ψ	0.28	0.4	0.19	0.00	*Thr*	0.22	0.08
Second Iteration	ϕ'	-0.69	1.32	-0.79	0.00			
	ψ'	0.77	0.32	0.31	0.00			

Table B.4: The change in the J coupling and rotamer error with the regularization and weighting scheme used to fit the torsional parameters.

B.1.3 Alanine Backbone Torsional Parameters

Regularization λ Value	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
0.00	ϕ	-0.67	1.13	-3.02	0.00	1.231	0.927
	ψ	1.21	2.70	-1.28	0.00		
	ϕ'	-3.08	-0.07	1.85	0.00		
	ψ'	1.88	-0.50	0.11	0.00		
0.10	ϕ	0.08	0.99	-0.33	0.00	1.804	
	ψ	0.16	1.00	-0.66	0.00		
	ϕ'	-1.00	-0.32	-0.08	0.00		
	ψ'	-0.34	-0.87	-0.58	0.00		
0.20	ϕ	0.04	0.75	-0.18	0.00	2.069	
	ψ	0.14	0.86	-0.51	0.00		
	ϕ'	-0.80	-0.28	-0.04	0.00		
	ψ'	-0.40	-0.64	-0.46	0.00		
0.30	ϕ	0.03	0.60	-0.12	0.00	2.360	
	ψ	0.12	0.64	-0.42	0.00		
	ϕ'	-0.60	-0.24	-0.02	0.00		
	ψ'	-0.37	-0.51	-0.37	0.00		
0.40	ϕ	0.02	0.49	-0.09	0.00	2.589	
	ψ	0.10	0.51	-0.34	0.00		
	ϕ'	-0.47	-0.20	-0.01	0.00		
	ψ'	-0.32	-0.41	-0.31	0.00		

Regularization λ Value	Dihedral Angle	Torsional Parameters				Error QUBE	Error OPLS
0.50	ϕ	0.02	0.41	-0.07	0.00	2.761	
	ψ	0.09	0.42	-0.29	0.00		
	ϕ'	-0.38	-0.17	0.00	0.00		
	ψ'	-0.28	-0.34	-0.26	0.00		
0.60	ϕ	0.01	0.35	-0.06	0.00	2.889	
	ψ	0.08	0.36	-0.25	0.00		
	ϕ'	-0.33	-0.14	0.00	0.00		
	ψ'	-0.24	-0.29	-0.22	0.00		

Table B.5: The alanine torsional parameters as a function of the regularization strength used.

Regularization	J Coupling Error		
λ Value	Set 1	Set 2	Set 3
0.00	7.74(5.57)	7.62(7.29)	9.33(7.95)
0.10	3.67(2.64)	2.41(2.79)	3.36(3.45)
0.20	2.26(1.71)	1.96(1.68)	1.77(2.06)
0.30	1.16 (1.00)	3.03(0.87)	1.25(1.04)
0.40	0.97(0.89)	3.74(0.81)	1.38(0.90)
0.50	0.87(0.80)	3.59(0.63)	1.28(0.78)
0.60	0.93(0.90)	4.15(0.82)	1.54(0.90)

Table B.6: The change in the J coupling error for Ala$_5$ with the regularization used in the torsional parameter fitting process.

B.2 Torsional Parameter

B.2.1 Backbone Torsional Parameters

This section details the torsional parameters and corresponding errors of the terms used in the dipeptide, peptide and protein simulations. The error shown uses all the scans included in the OPLS-AA/M fitting process, even when the QUBE terms are fitted to a subset of scans. The OPLS-AA/M force field is fit using a 2000K QM weighting scheme, whilst the error used here is not weighted.

Dipeptide Name	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
Ala/Gly							
Reg = 0.00	ϕ	-0.60	1.18	-2.76	0.00	1.254	0.929
	ψ	1.26	2.88	-2.08	0.00		
	ϕ'	-3.04	-0.04	1.56	0.00		
	ψ'	1.96	-0.42	0.93	0.00		
Reg = 0.50	ϕ	0.02	0.41	-0.09	0.00	2.880	
	ψ	0.09	0.43	-0.29	0.00		
	ϕ'	-0.37	-0.16	0.00	0.00		
	ψ'	-0.26	-0.32	-0.24	0.00		
Ala							
Reg = 0.00	ϕ	-0.67	1.13	-3.02	0.00	1.231	0.927
	ψ	1.21	2.70	-1.28	0.00		
	ϕ'	-3.08	-0.07	1.85	0.00		
	ψ'	1.88	-0.50	0.11	0.00		
Reg = 0.50	ϕ	0.02	0.41	-0.07	0.00	2.761	
	ψ	0.09	0.42	-0.29	0.00		
	ϕ'	-0.38	-0.17	0.00	0.00		
	ψ'	-0.28	-0.34	-0.26	0.00		

Dipeptide Name	Dihedral Angle	Torsional Parameters				Error QUBE	Error OPLS
Gly							
Reg = 0.00	ϕ	-0.38	1.39	-2.01	0.00	1.307	
	ψ	1.33	4.32	-1.62	0.00		
Reg = 0.50	ϕ	0.07	0.42	-0.37	0.00	4.008	
	ψ	0.14	0.52	-0.33	0.00		
Pro							
Reg = 0.00	ψ	2.11	0.95	-2.36	0.00	1.128	1.116
	ψ'	1.57	-1.09	1.32	0.00		
Reg = 0.50	ψ	0.10	0.22	-0.21	0.00	1.598	
	ψ'	-0.01	-0.28	-0.17	0.00		
Ser							
Reg=0.05	ϕ	0.51	-0.52	-0.96	0.00	2.541	2.834
	ψ	0.28	0.4	0.19	0.00		
	ϕ'	-0.69	1.32	-0.79	0.00		
	ψ'	0.77	0.32	0.31	0.00		
Reg = 0.50	ϕ	0.08	-0.16	-0.13	0.00	3.109	
	ψ	-0.03	0.00	0.12	0.00		
	ϕ'	-0.02	0.19	-0.12	0.00		
	ψ'	0.17	0.12	0.12	0.00		

B.2.2 Sidechain Torsional Parameters

Dipeptide Name	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
Arg	X1	1.91	-0.29	0.56	0.00	0.606	0.662
	X1’	-1.1	-0.35	0.44	0.00		
	X2	-2.10	-0.50	0.00	0.00	2.302	0.759
Asn	X1	-7.51	0.53	-0.56	0.00	1.904	1.246
	X1’	4.79	0.60	0.21	0.00		
	X2	0.81	0.51	0.00	0.00	1.819	0.759
	X2’	1.59	-0.10	-0.10	0.00		
Asp	X1	-8.04	-0.57	0.73	0.00	1.244	1.807
β_{180} *only*	X1’	3.58	0.73	-0.12	0.00		
Set to zero values	X2	0.00	0.00	0.00	0.00	1.897	
Cys	X1	3.31	0.29	1.72	0.00	1.143	1.062
β_{60} *only*	X1’	-2.81	-0.03	-0.09	0.00		
	X2	0.93	2.90	-0.32	0.00	0.661	0.813
	X2’	0.78	1.04	2.09	0.00		

Dipeptide Name	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
Gln	X1	2.13	-0.37	1.58	0.00	0.947	0.910
	X1’	-1.62	0.24	0.08	0.00		
	X2	-1.16	1.01	-0.06	0.00	2.884	1.384
Glu	X1	-0.57	-1.45	1.33	0.00	3.120	2.132
β Scans only	X1’	-2.26	-1.87	1.04	0.00		
	X2	-8.81	-0.11	-0.80	0.00	3.105	2.819
Hid	X1	-0.70	0.19	0.50	0.00	1.238	1.177
	X1’	0.20	0.60	0.20	0.00		
	X2	2.91	-1.61	0.40	0.00	1.523	
Hie	X1	-3.12	0.39	0.92	0.00	0.803	1.113
	X1’	1.90	0.50	0.10	0.00		
	X2	-1.19	-0.50	0.03	0.00	0.749	
Hip	X1	-1.31	-0.37	0.98	0.00	1.482	1.747
β Scans only	X1’	-0.28	-0.91	0.01	0.00		
	X2	-0.38	0.50	0.19	0.00	1.135	
Ile	X1	7.51	-0.59	-0.21	0.00	0.694	1.066
	X1’	5.01	-0.19	0.99	0.00		
	X2	0.10	-0.14	0.13	0.00	0.855	0.661

Dipeptide Name	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
Leu	X1	-2.08	-0.65	0.58	0.00	0.797	1.238
	X1’	0.53	0.34	0.12	0.00		
	X2	-1.09	0.71	0.20	0.00	0.623	0.733
Lys	X1	0.00	0.00	0.00	0.00	1.683	1.024
	X1’	0.00	0.00	0.00	0.00		
OPLS-AA/M	X2	1.30	-0.20	0.20	0.00	2.689	
Met	X1	0.19	-0.83	-0.74	0.00	0.916	0.553
β_{-60} *only*	X1’	0.81	0.03	1.63	0.00		
	X2	-1.15	-0.45	-0.23	0.00	1.586	1.196
Phe	X1	-0.94	0.76	0.91	0.00		
	X1’	0.46	0.58	0.32	0.00	0.576	0.990
	X2	-5.21	0.19	0.00	0.00	0.828	0.616
Pro	X1	2.11	0.95	-2.36	0.00	1.128	1.116
	X1’	1.57	-1.09	1.32	0.00		
Ser	X1	5.32	0.75	0.23	0.00	1.526	1.388
β_{60} *only*	X1’	-6.93	0.95	1.17	0.00		
	X2	0.21	-1.21	0.41	0.00	0.961	0.929

Dipeptide Name	**Dihedral Angle**	**Torsional Parameters**				**Error QUBE**	**Error OPLS**
Thr	X1	-3.34	1.18	-1.72	0.00	1.879	1.455
β_{60} *only*	X1’	-3.87	3.12	-0.05	0.00		
	X1	3.64	2.84	3.91	0.00		
	X1’	-1.61	2.13	-0.89	0.00		
Trp	X1	1.09	0.59	0.88	0.00	0.661	0.970
	X1’	-0.40	0.90	0.12	0.00		
	X2	0.48	-0.58	-0.19	0.00	0.875	1.295
Tyr	X1	-0.12	0.70	1.07	0.00	0.764	1.014
	X1’	-0.13	1.02	-0.01	0.00		
	X2	-4.40	0.19	0.59	0.00	0.864	0.662
Val	X1	3.81	-0.55	0.79	0.00	0.540	0.550
	X1’	-1.74	-0.11	0.03	0.00		
Mean Values						1.29	1.12

B.3 Dipeptides

B.3.1 J Coupling Values

Dipeptide	**Backbone J Coupling Error**			
Name	**First Run**	**Second Run**	**Third Run**	**Average**
Arg	1.17	1.15	0.81	1.04
Asn	0.36	0.41	0.36	0.38
Asp	0.10	0.07	0.07	0.08
Cys	0.43	0.46	0.45	0.45
Gln	0.42	0.42	0.45	0.43
Glu	0.39	0.26	0.18	0.28
Hid	N/A	N/A	N/A	N/A
Hie	N/A	N/A	N/A	N/A
Hip	0.50	0.44	0.53	0.49
Ile	0.71	0.67	0.68	0.69
Leu	0.10	0.07	0.09	0.09
Lys	0.22	0.24	0.24	0.23
Met	0.38	0.45	0.39	0.41
Phe	0.09	0.06	0.04	0.06
Ser	0.27	0.38	0.31	0.32
Thr	0.22	0.12	0.09	0.14
Trp	0.14	0.16	0.14	0.15
Tyr	0.04	0.01	0.05	0.03
Val	0.47	0.44	0.41	0.44
RMSE				0.42

Table B.7: The J coupling error for the set of dipeptides.

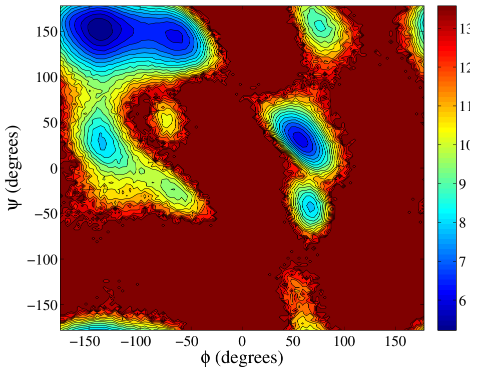
Dipeptide	Sidechain Rotamer Error			
Name	First Run	Second Run	Third Run	Average
Arg	0.08	0.09	0.10	0.09
Asn	0.17	0.16	0.17	0.17
Asp	0.28	0.29	0.28	0.28
Cys	0.08	0.11	0.13	0.11
Gln	0.16	0.15	0.15	0.15
Glu	0.31	0.17	0.10	0.19
Hid	0.10	0.16	0.10	0.12
Hie	0.13	0.13	0.06	0.11
Hip	0.26	0.24	0.23	0.24
Ile	0.13	0.11	0.13	0.12
Leu	0.10	0.11	0.10	0.10
Lys	0.14	0.13	0.17	0.15
Met	0.13	0.15	0.14	0.14
Phe	0.07	0.09	0.08	0.08
Ser	0.11	0.09	0.07	0.09
Thr	0.08	0.08	0.10	0.09
Trp	0.14	0.16	0.19	0.16
Tyr	0.08	0.09	0.08	0.08
Val	0.18	0.20	0.21	0.20
MUE				0.14

Table B.8: The error in the rotamer populations for the set of dipeptides.

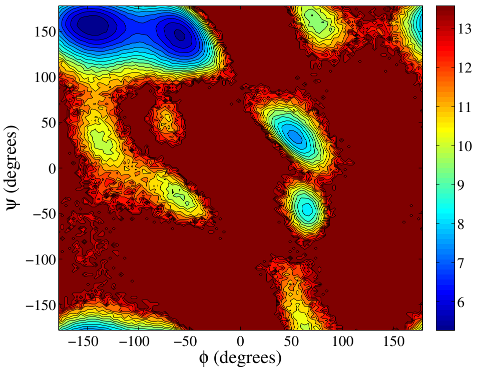
B.3.2 Dihedral Angle Distributions ϕ and ψ

The dihedral distributions are given in this section. Regions of low probability are shown in dark red as are regions that are not sampled during the simulation.

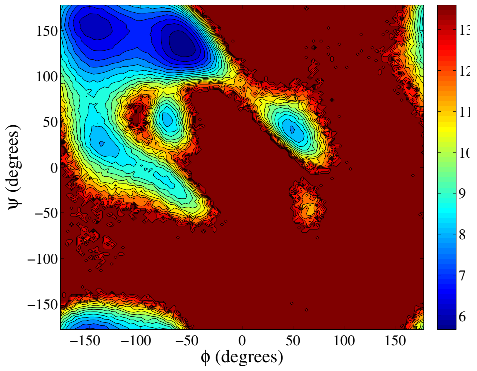
Arginine



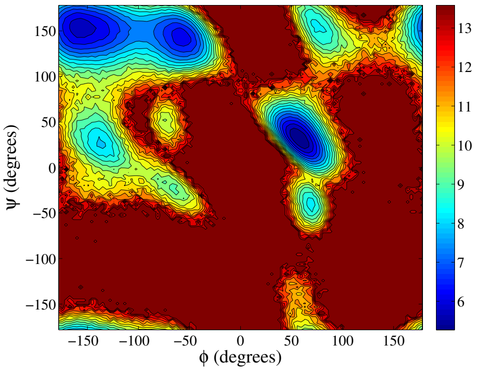
Asparagine



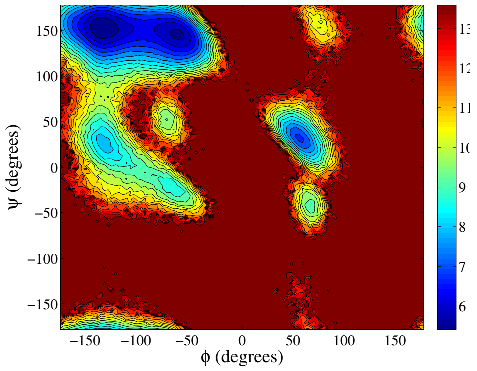
Aspartate



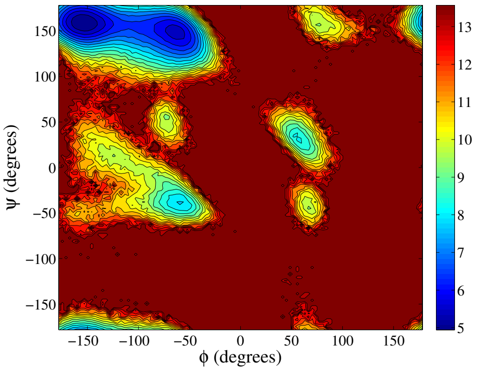
Cysteine

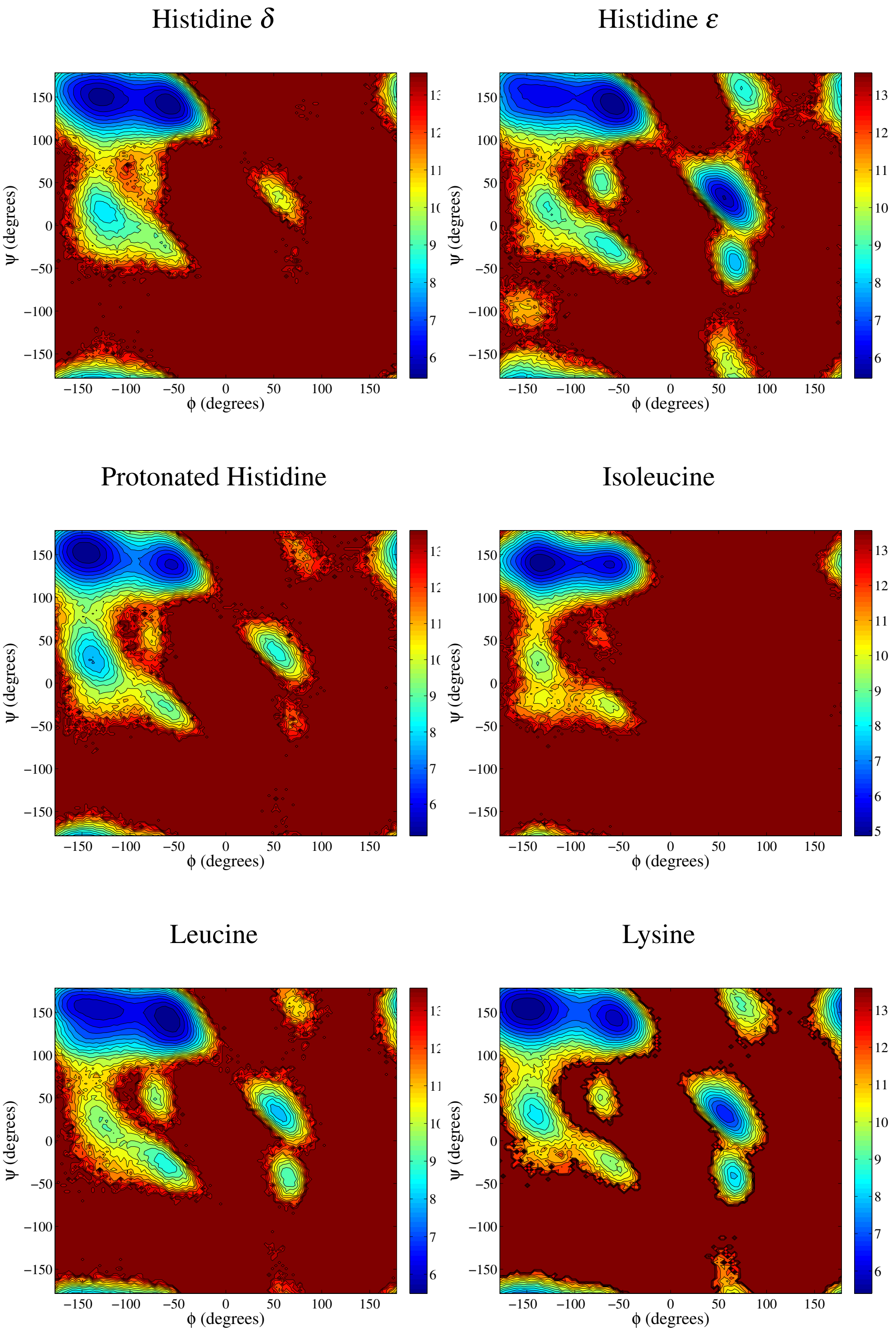


Glutamine



Glutamate





Histidine δ

Histidine ε

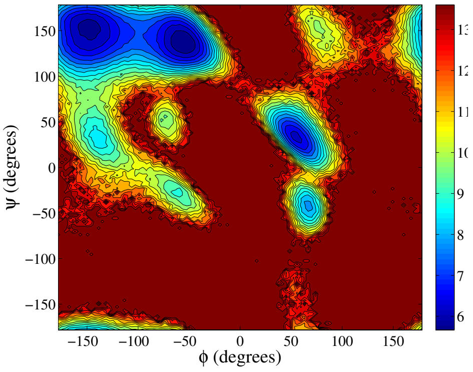
Protonated Histidine

Isoleucine

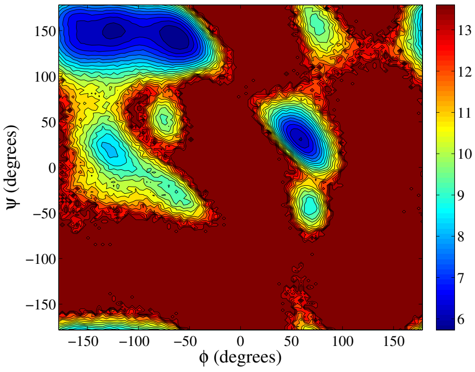
Leucine

Lysine

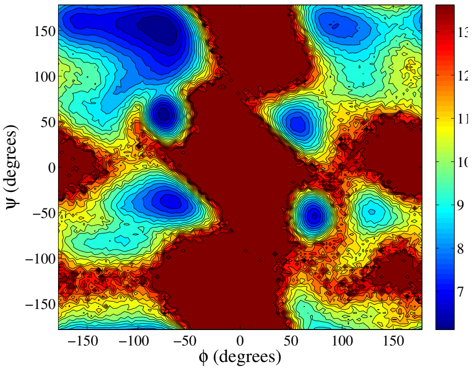
Methionine



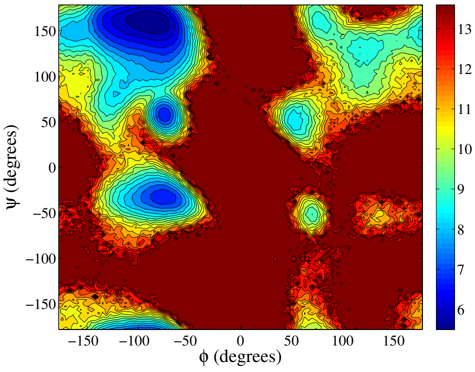
Phenylalanine



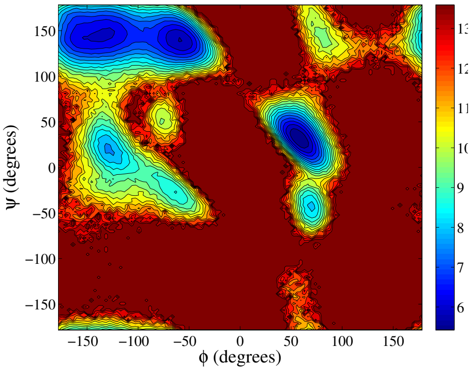
Serine



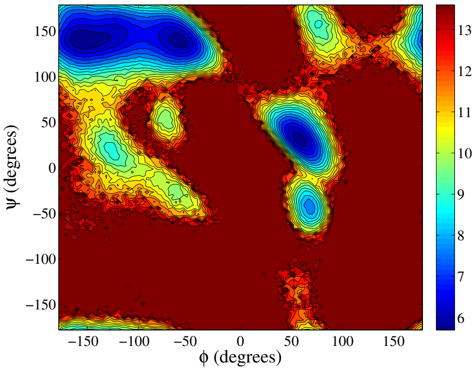
Threonine



Tryptophan



Tyrosine



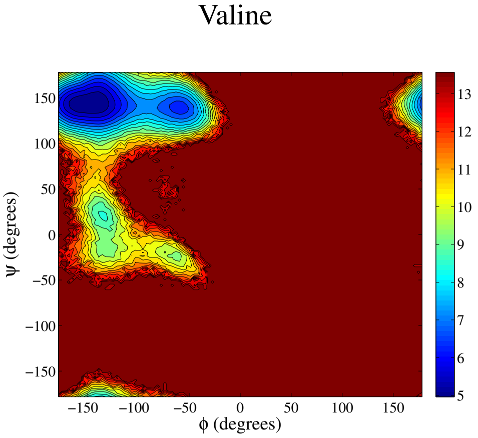


Fig. B.1: The ψ and ϕ distribution for the dipeptide simulations.

B.4 Peptides

B.4.1 J Coupling Errors

The backbone J coupling errors and conformations populated by Ala_5 are presented in this section.

Peptide	**J Coupling Error**			
Name	**Run 1**	**Run 2**	**Run 3**	**Average**
Ala_5	0.93(0.89)	0.91(0.87)	0.87(0.83)	0.90 ± 0.03 (0.86 ± 0.03)
Gly_3	5.69	5.90	5.48	5.69 ± 0.21

Table B.9: The J coupling error for Ala_5 using the first set of Karplus parameters from Ref. 6.

Peptide	**J Coupling Error**			
Name	**Run 1**	**Run 2**	**Run 3**	**Average**
Ala_5	4.16(0.84)	4.16(0.82)	4.15(0.78)	4.16 ± 0.01(0.81 ± 0.03)
Gly_3	7.13	7.25	6.96	7.11 ± 0.15

Table B.10: The J coupling error for Ala_5 using the second set of Karplus parameters from Ref. 6.

Peptide	**J Coupling Error**			
Name	**Run 1**	**Run 2**	**Run 3**	**Average**
Ala_5	1.52(0.89)	1.52(0.88)	1.49(0.83)	1.51 ± 0.02(0.87 ± 0.03)
Gly_3	6.82	6.96	6.61	4.52 ± 0.18

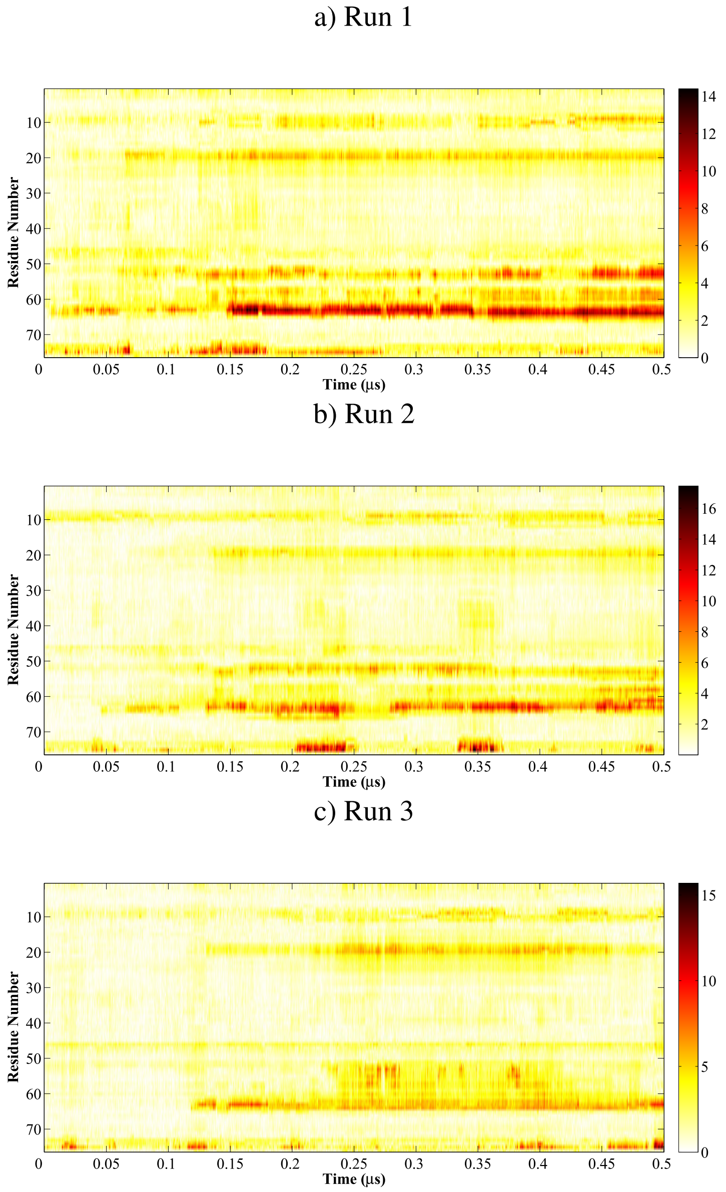
Table B.11: The J coupling error for Ala_5 using the third set of Karplus parameters from Ref. 6.

Peptide	**Conformation Percentages**		
Name	PP2	β	α
Ala_5	62.10 ± 2.23	23.5 ± 1.20	12.93 ± 2.82
Gly_3	45.93 ± 1.46	29.77 ± 7.74	24.30 ± 8.57

Table B.12: The population of each conformation present in the MD simulation of Ala_5

B.5 Protein MD Results

B.5.1 1UBQ

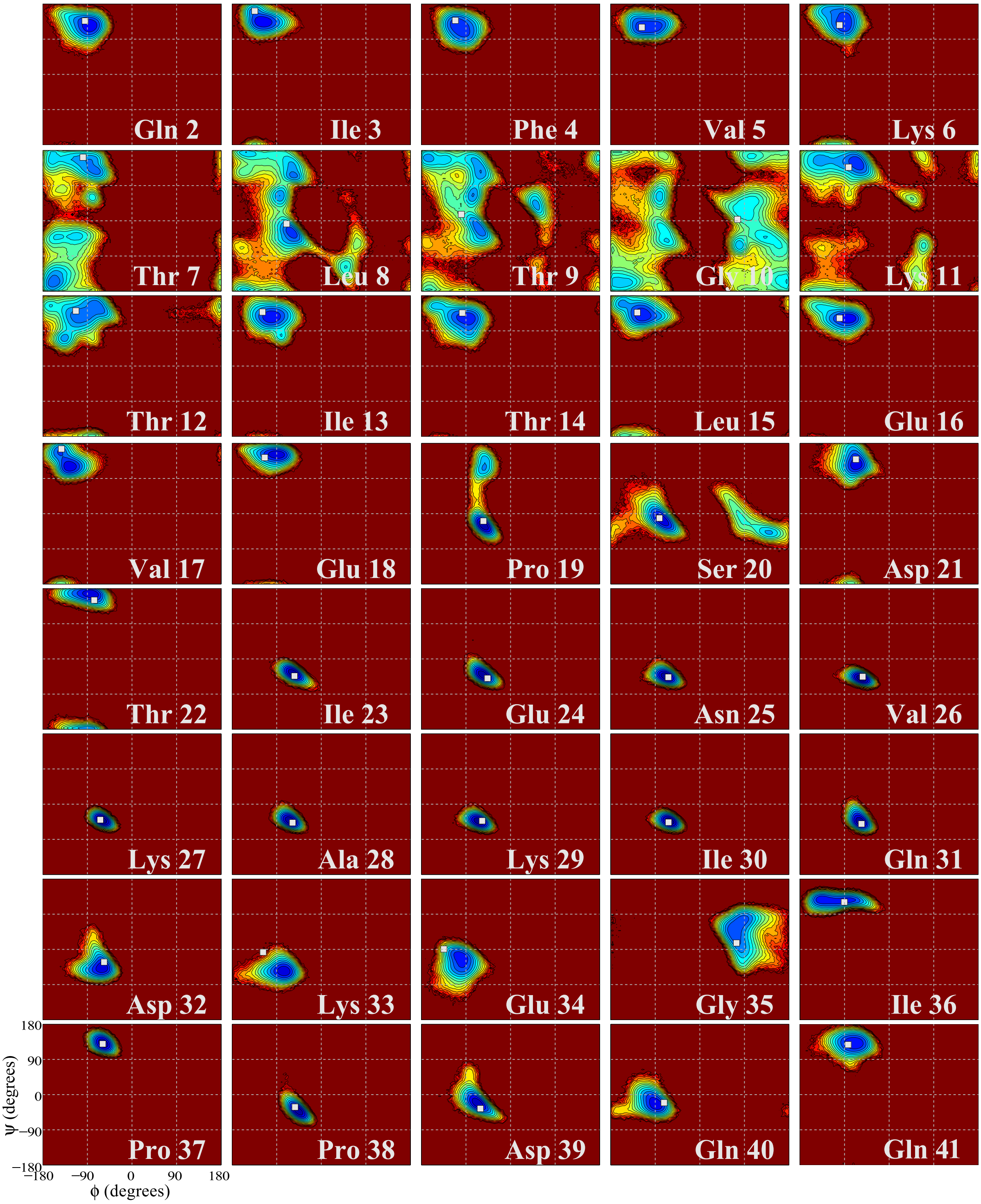


a) Run 1

b) Run 2

c) Run 3

Fig. B.2: The RMSD per residue, relative to the crystal structure with PDB code 1UBQ, of three simulations of the ubiquitin protein.


Gln 2
Ile 3
Phe 4
Val 5
Lys 6
Thr 7
Leu 8
Thr 9
Gly 10
Lys 11
Thr 12
Ile 13
Thr 14
Leu 15
Glu 16
Val 17
Glu 18
Pro 19
Ser 20
Asp 21
Thr 22
Ile 23
Glu 24
Asn 25
Val 26
Lys 27
Ala 28
Lys 29
Ile 30
Gln 31
Asp 32
Lys 33
Glu 34
Gly 35
Ile 36
Pro 37
Pro 38
Asp 39
Gln 40
Gln 41
180
90
0
−90
−180
ψ (degrees)
−180
−90
0
90
180
ϕ (degrees)

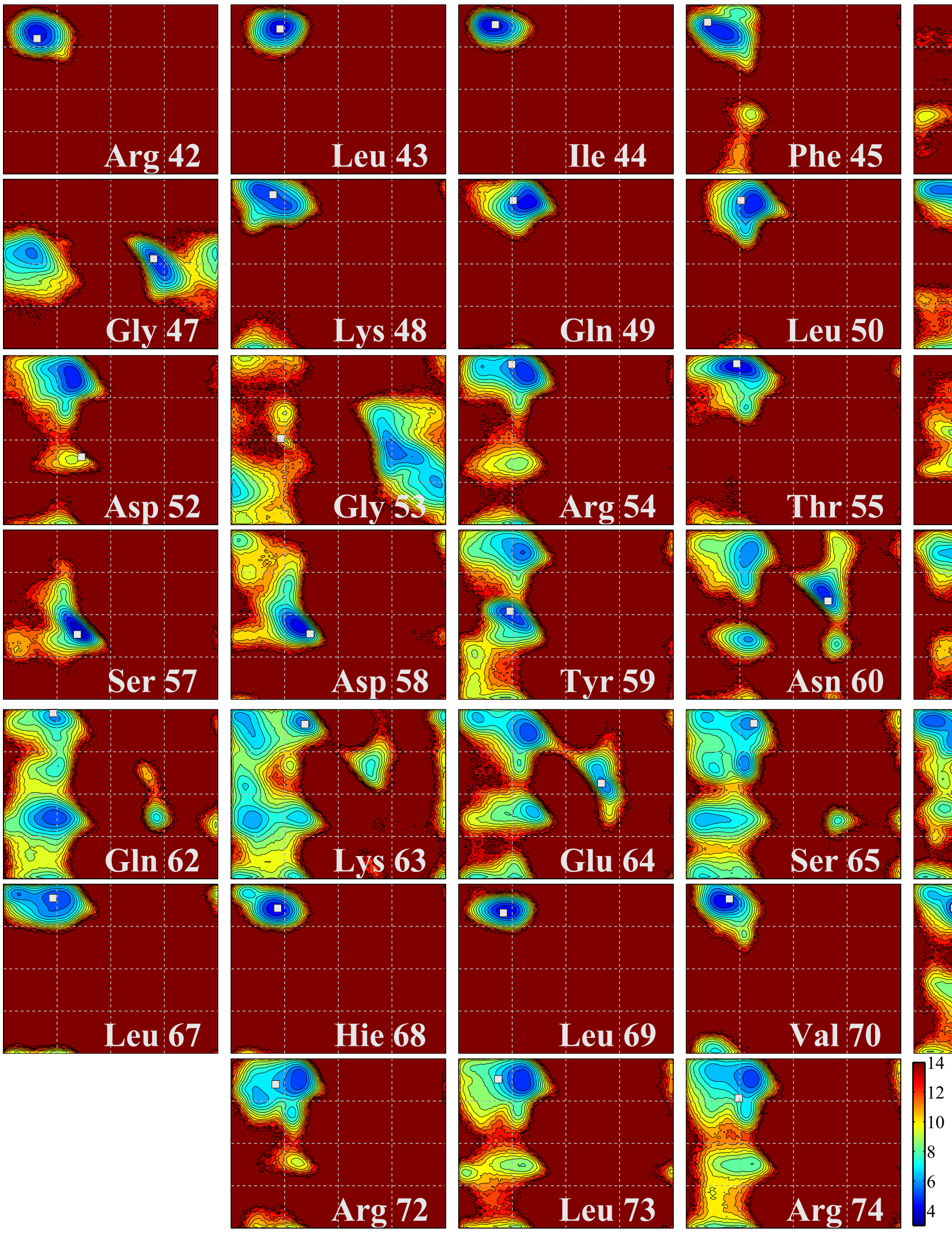
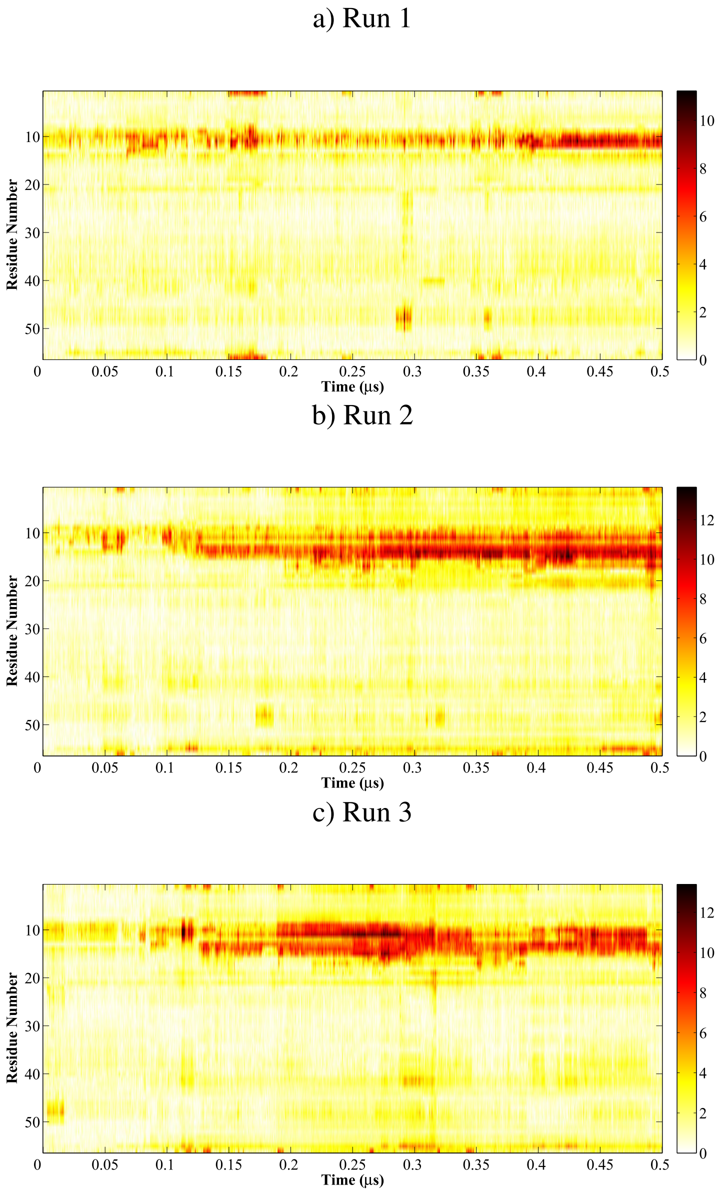


Fig. B.3: The ψ and ϕ distributions observed in the simulation of the ubiquitin protein. The white square corresponds to the experimental structure given in the PDB file 1UBQ. The dark red regions correspond to low probability regions that include conformations that are not sampled during the simulation.

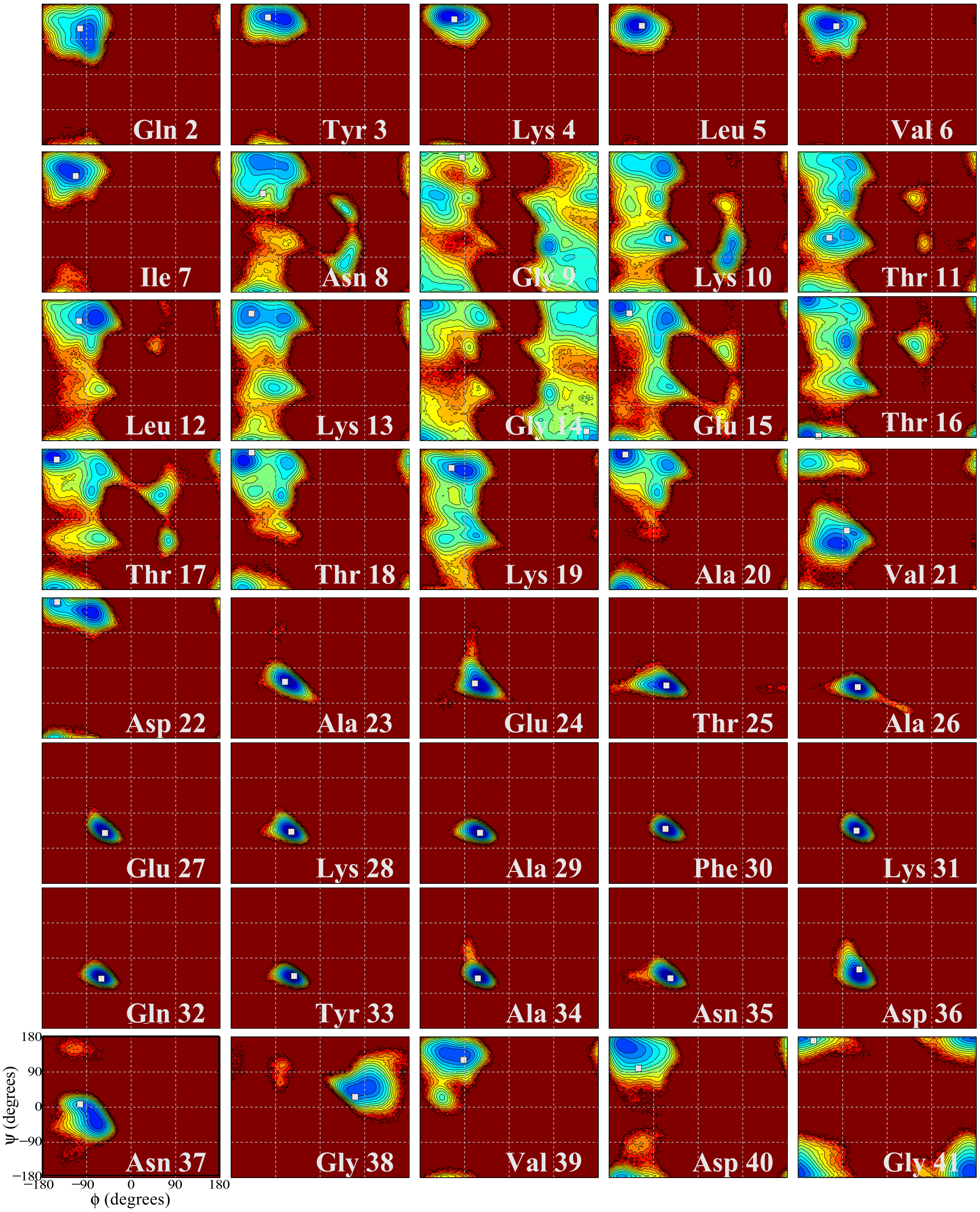
B.5.2 1P7E


a) Run 1

b) Run 2

c) Run 3

Fig. B.4: The RMSD per residue, relative to the crystal structure with PDB code 1P7E, of three simulations of GB3.


Gln 2
Tyr 3
Lys 4
Leu 5
Val 6
Ile 7
Asn 8
Gly 9
Lys 10
Thr 11
Leu 12
Lys 13
Gly 14
Glu 15
Thr 16
Thr 17
Thr 18
Lys 19
Ala 20
Val 21
Asp 22
Ala 23
Glu 24
Thr 25
Ala 26
Glu 27
Lys 28
Ala 29
Phe 30
Lys 31
Gln 32
Tyr 33
Ala 34
Asn 35
Asp 36
Asn 37
Gly 38
Val 39
Asp 40
Gly 41
180
90
0
−90
−180
ψ (degrees)
−180
−90
0
90
180
ϕ (degrees)

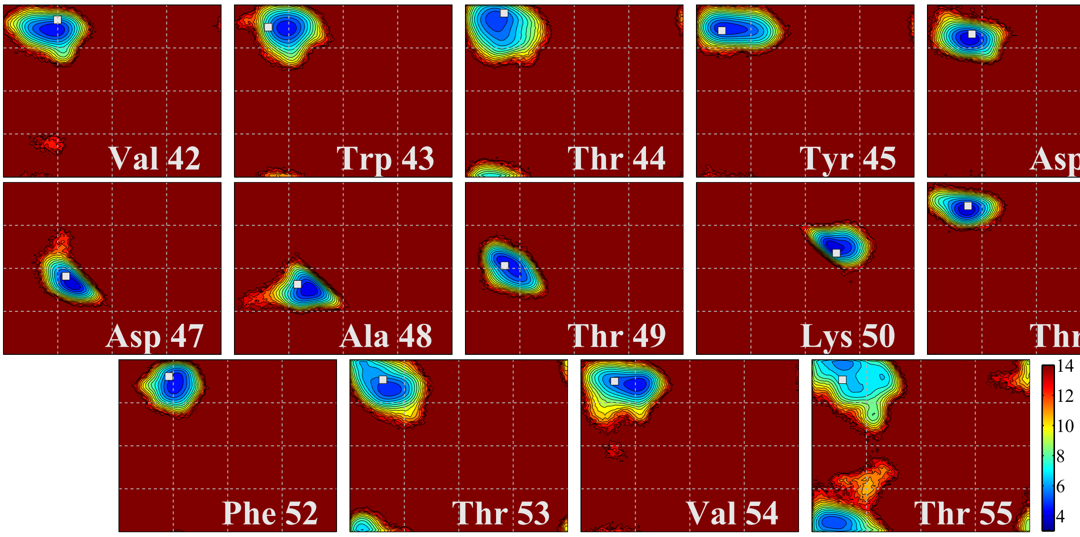
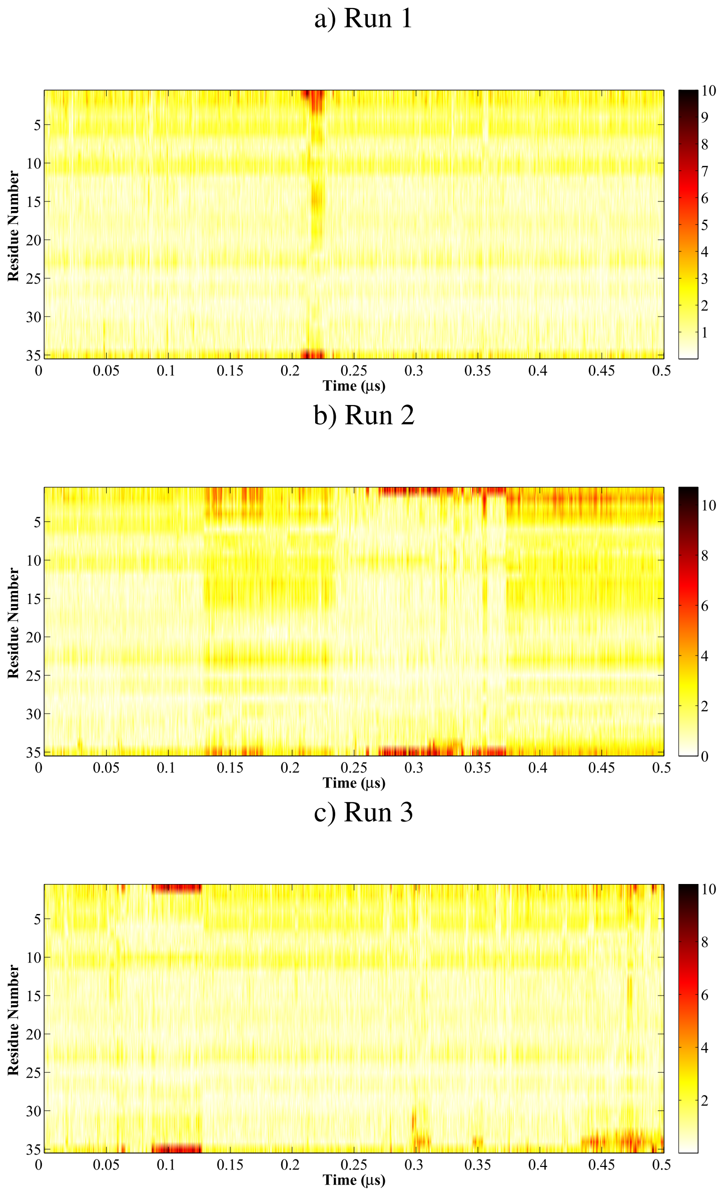


Fig. B.5: The ψ and ϕ distributions observed in the simulation of the GB3 protein. The white square corresponds to the experimental structure given in the PDB file 1P7E. The dark red regions correspond to low probability regions that include conformations that are not sampled during the simulation.

B.5.3 2F4K



a) Run 1

b) Run 2

c) Run 3

Fig. B.6: The RMSD per residue, relative to the crystal structure with PDB code 2F4K, of three simulations of the villin headpiece.

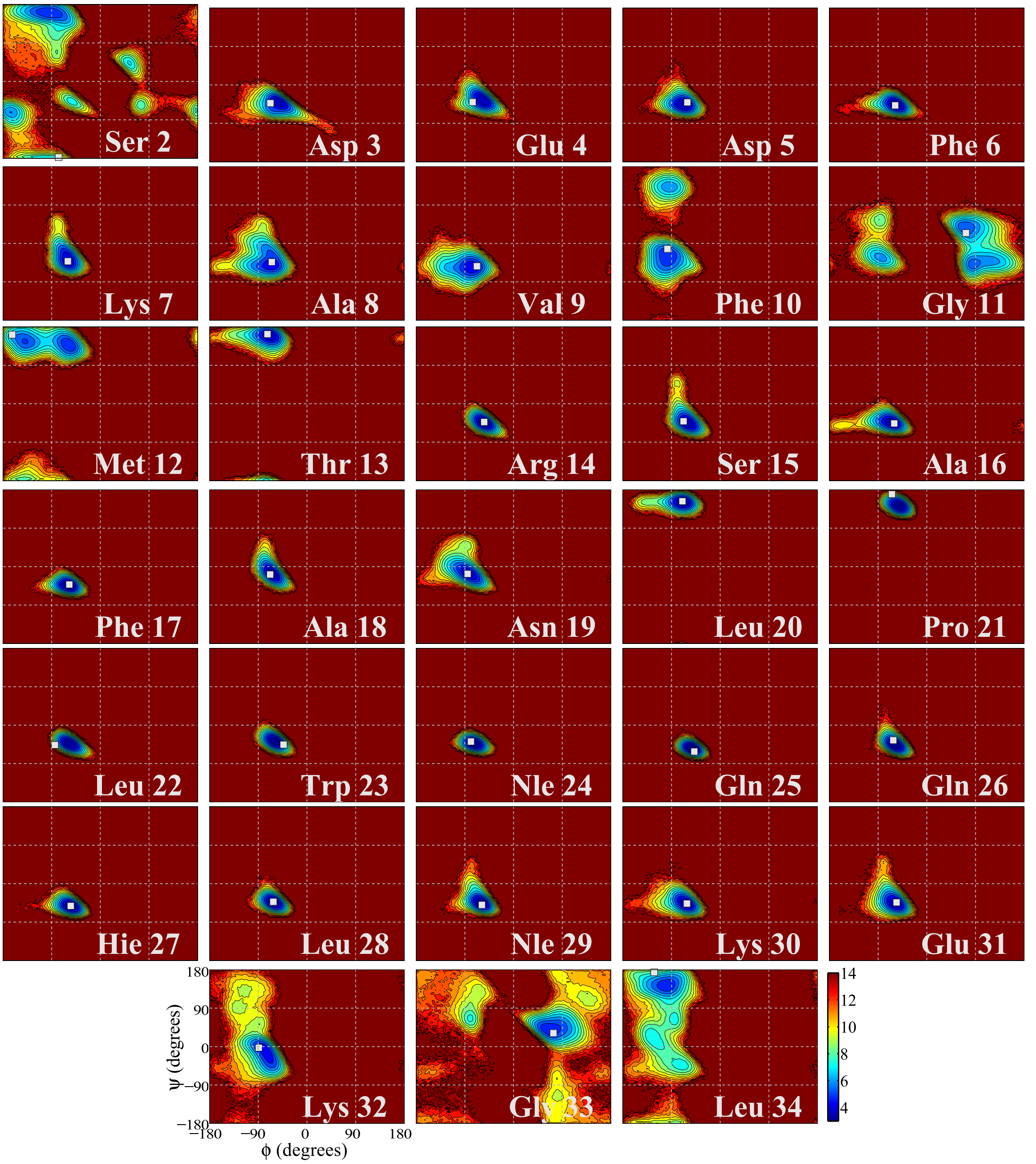
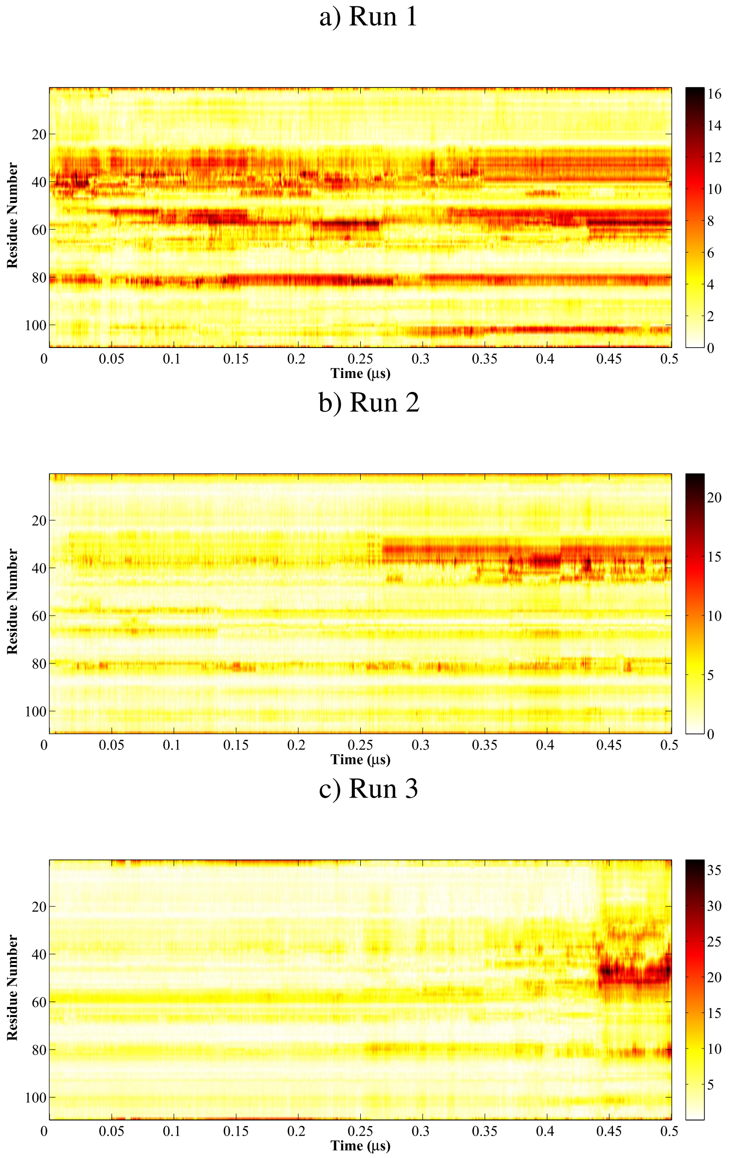


Fig. B.7: The ψ and ϕ distributions observed in the simulation of the villin headpiece. The white square corresponds to the experimental structure given in the PDB file 2F4K. The dark red regions correspond to low probability regions that include conformations that are not sampled during the simulation.

B.5.4 1BUJ

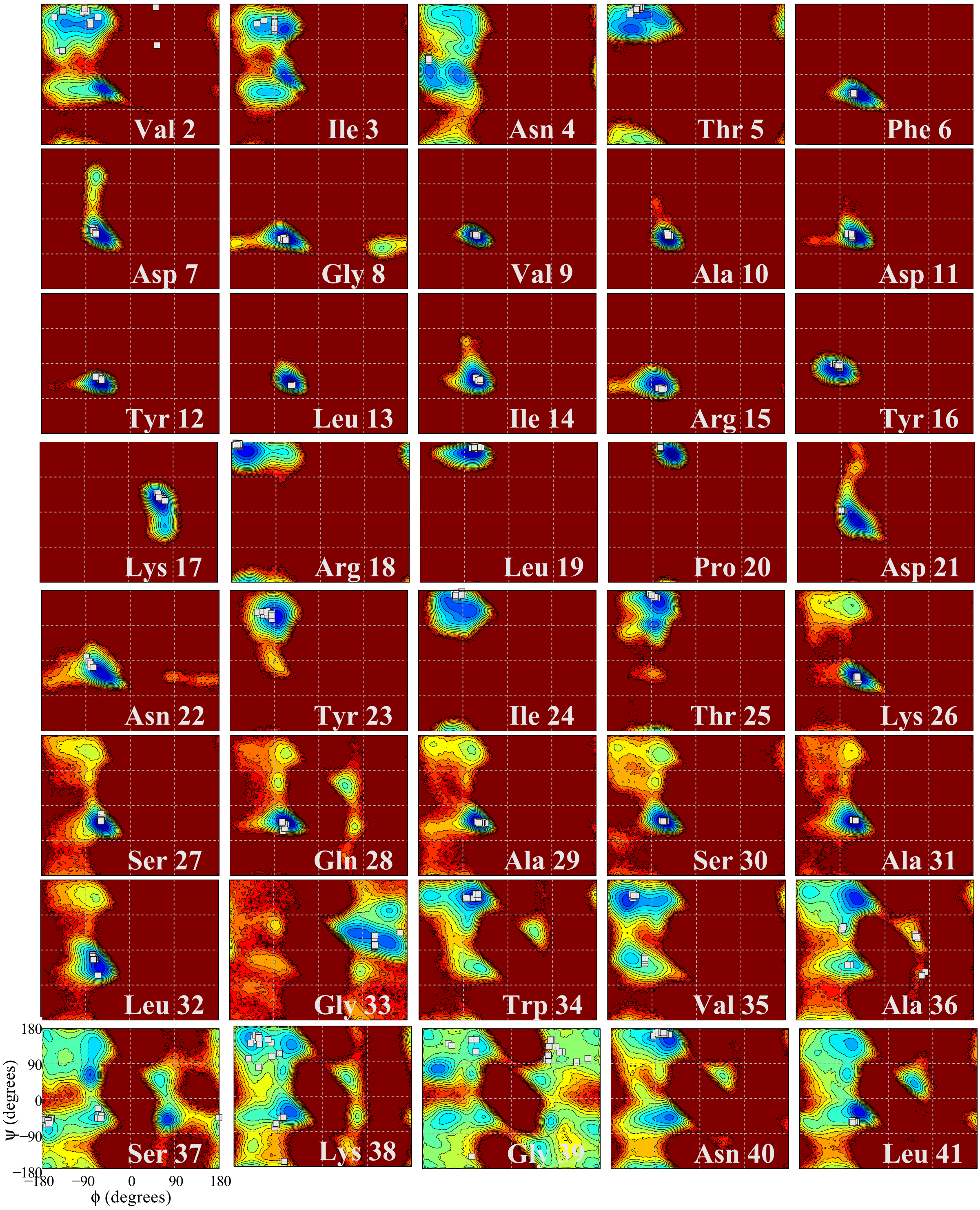


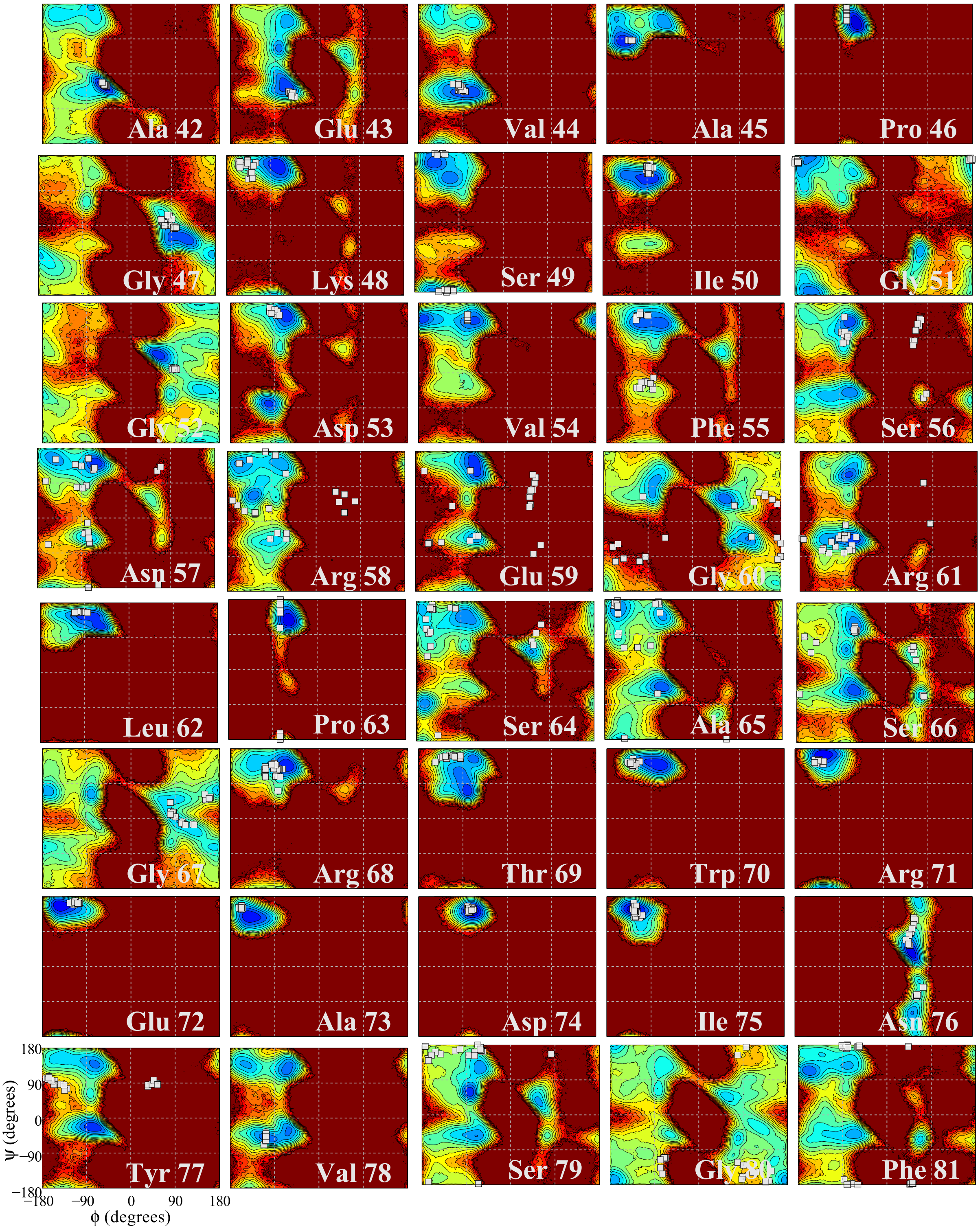
a) Run 1

b) Run 2

c) Run 3

Fig. B.8: The RMSD per residue, relative to the crystal structure with PDB code 1BUJ, of three simulations of the binase protein.


Val 2
Ile 3
Asn 4
Thr 5
Phe 6
Asp 7
Gly 8
Val 9
Ala 10
Asp 11
Tyr 12
Leu 13
Ile 14
Arg 15
Tyr 16
Lys 17
Arg 18
Leu 19
Pro 20
Asp 21
Asn 22
Tyr 23
Ile 24
Thr 25
Lys 26
Ser 27
Gln 28
Ala 29
Ser 30
Ala 31
Leu 32
Gly 33
Trp 34
Val 35
Ala 36
Ser 37
Lys 38
Gly 39
Asn 40
Leu 41
180
90
0
−90
−180
ψ (degrees)
−180
−90
0
90
180
ϕ (degrees)


Ala 42
Glu 43
Val 44
Ala 45
Pro 46
Gly 47
Lys 48
Ser 49
Ile 50
Gly 51
Gly 52
Asp 53
Val 54
Phe 55
Ser 56
Asn 57
Arg 58
Glu 59
Gly 60
Arg 61
Leu 62
Pro 63
Ser 64
Ala 65
Ser 66
Gly 67
Arg 68
Thr 69
Trp 70
Arg 71
Glu 72
Ala 73
Asp 74
Ile 75
Asn 76
Tyr 77
Val 78
Ser 79
Gly 80
Phe 81
180
90
0
−90
−180
ψ (degrees)
−180
−90
0
90
180
ϕ (degrees)

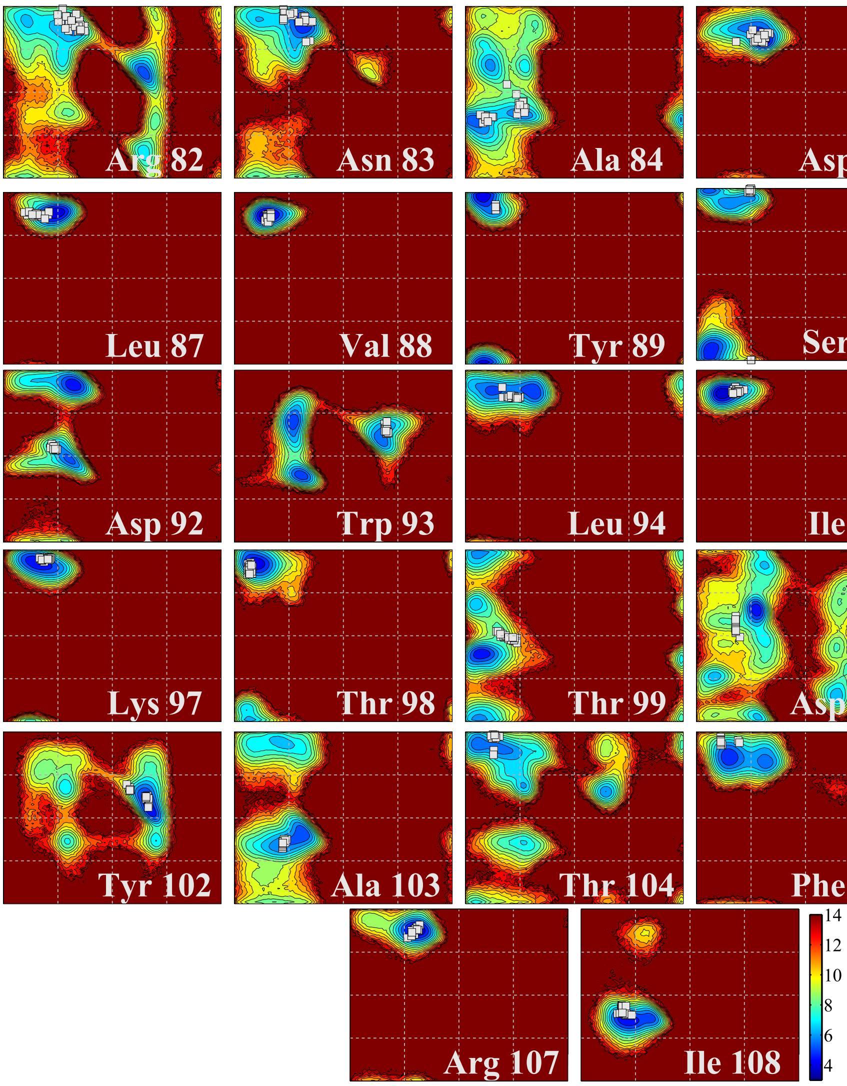
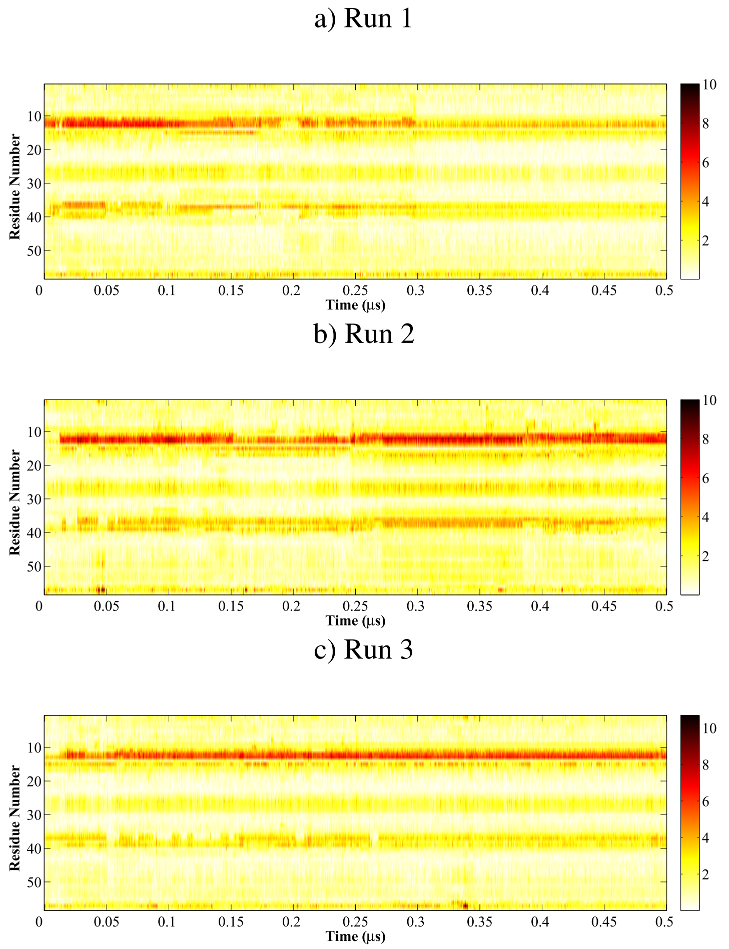


Fig. B.9: The ψ and ϕ distributions observed in the simulation of the binase protein. The white square corresponds to the experimental structure given in the PDB file 1BUJ. The dark red regions correspond to low probability regions that include conformations that are not sampled during the simulation.

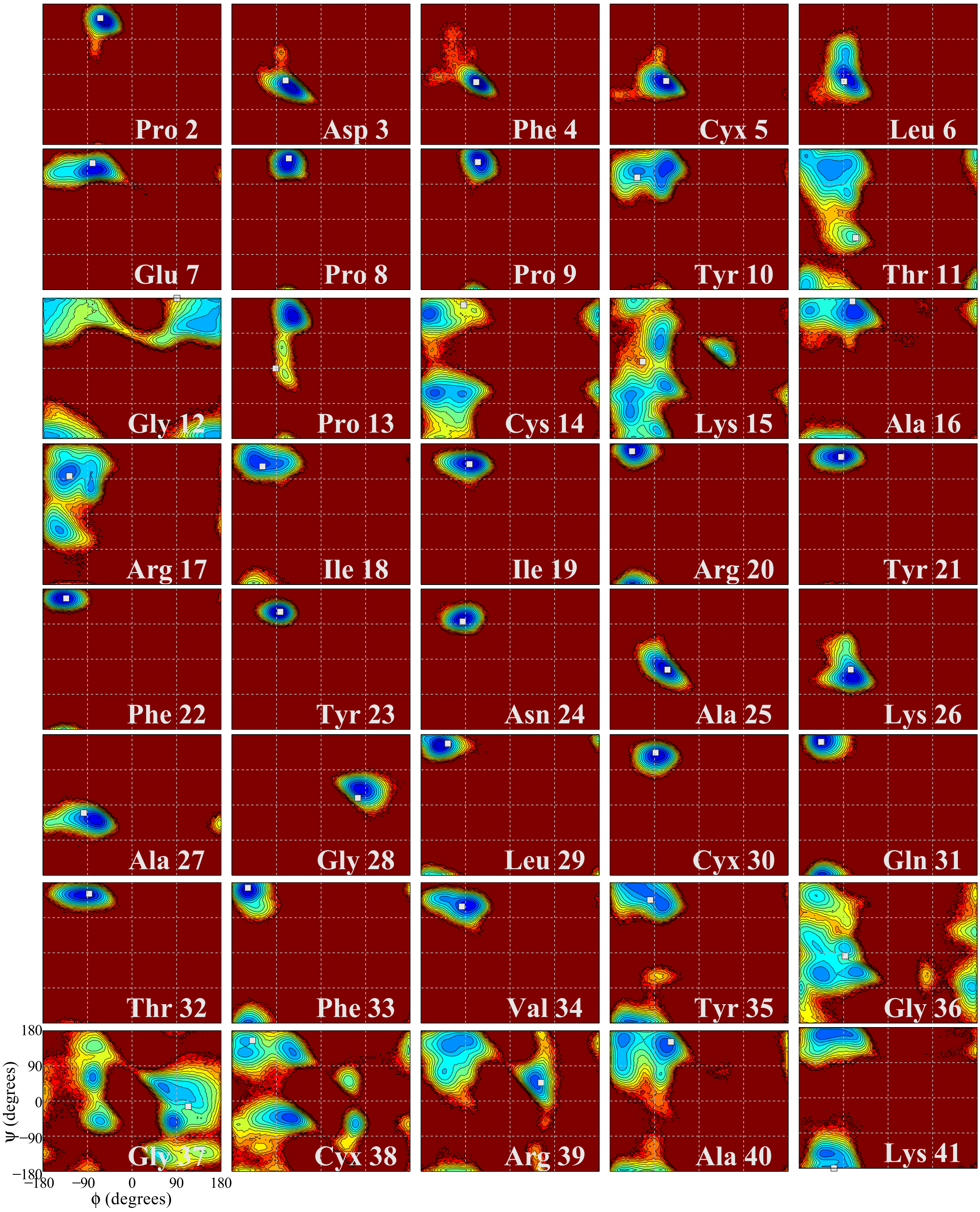
B.5.5 5PTI


a) Run 1

b) Run 2

c) Run 3

Fig. B.10: The RMSD per residue, relative to the crystal structure with PDB code 5PTI, of three simulations of the BPTI protein.


Pro 2
Asp 3
Phe 4
Cyx 5
Leu 6
Glu 7
Pro 8
Pro 9
Tyr 10
Thr 11
Gly 12
Pro 13
Cys 14
Lys 15
Ala 16
Arg 17
Ile 18
Ile 19
Arg 20
Tyr 21
Phe 22
Tyr 23
Asn 24
Ala 25
Lys 26
Ala 27
Gly 28
Leu 29
Cyx 30
Gln 31
Thr 32
Phe 33
Val 34
Tyr 35
Gly 36
Gly 37
Cyx 38
Arg 39
Ala 40
Lys 41
180
90
0
−90
−180
ψ (degrees)
−180
−90
0
90
180
ϕ (degrees)

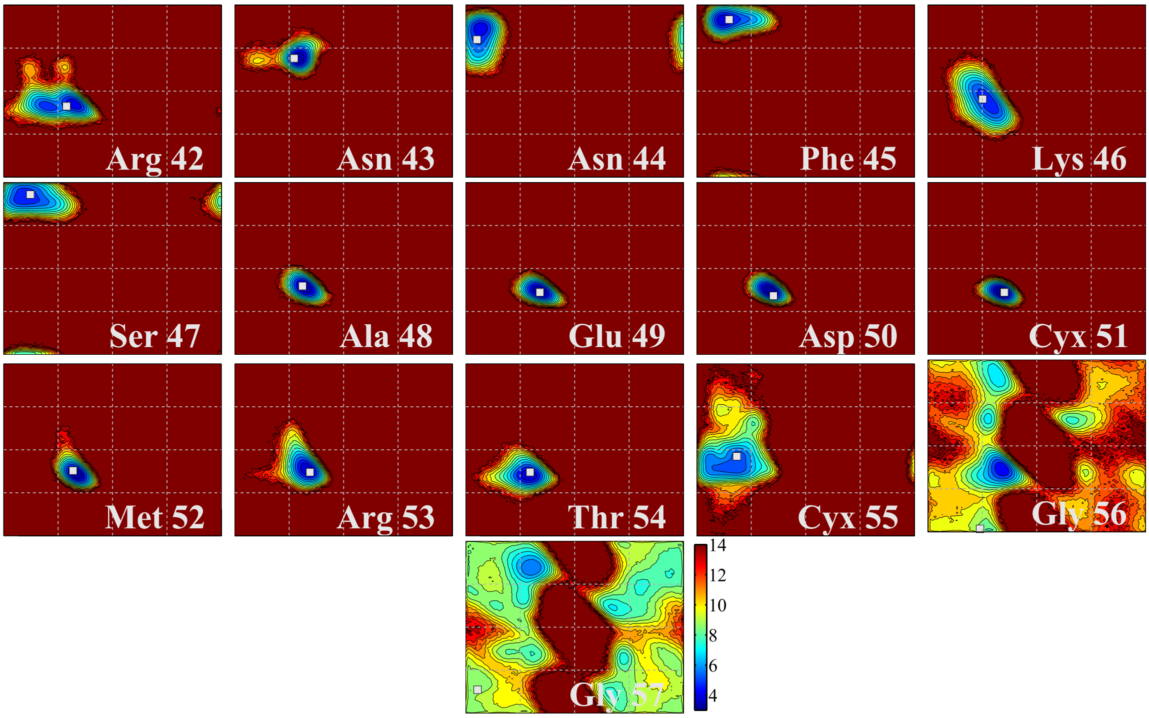


Fig. B.11: The ψ and ϕ distributions observed in the simulation of the BPTI protein. The white square corresponds to the experimental structure given in the PDB file 5PTI. The dark red regions correspond to low probability regions that include conformations that are not sampled during the simulation.