Quantum Mechanical Bespoke Force Fields for Computer-Aided Drug Design

Joshua Thomas Horton

Thesis submitted for the degree of
Doctor of Philosophy

School of Natural & Environmental Sciences
Newcastle University
Newcastle upon Tyne
United Kingdom

December 2019

Acknowledgements

I would like to thank my supervisor Dr. Daniel Cole for his dedicated mentorship and for this fantastic opportunity to work alongside him on such an exciting project. His guidance, enthusiasm and scientific insight have helped to create an exciting and fruitful research experience, which I hope to continue. I am also grateful to all of my collaborators who helped make the quantum bespoke force field a reality including Leela Dodda for his help in running the benchmark simulations in chapter 4 and Alice Allen for her continuous support and advice throughout the project. I would also like to thank Jeff Wagner for his help on implementing the skeleton force field files in chapter 6 and Lee-Ping Wang who helped extend geomeTRIC and torsiondrive to meet the projects needs. I would also like to thank Chris Ringrose for his help in developing the second version of the toolkit, Lauren Nelson who has extensively tested it and everyone in the Cole group at Newcastle University who were always there to help, working with them has been a pleasure. Lastly I would like to thank my family for their support and Sophie Needham, whom with out many late nights of proof reading this would not have been possible.

Abstract

The ability to accurately model complex biological processes such as protein-ligand binding with an atomistic level of detail is critical to their thorough understanding. Typically a molecular mechanics simulation is used, which represents the system using a force field that is a physically motivated linear combination of empirically parameterised potentials. Traditionally their parameterisation has involved the recreation of experimental and quantum mechanical data for a target set of representative structures, ranging from small molecules to peptides. This potentially limits the progress of general transferable force fields to time and labour-intensive incremental improvements. In this thesis, we aim to challenge this "parameterise once and transfer" philosophy, with that of a transferable parametrisation methodology that can be readily applied to new systems with a consistent level of accuracy. We collect together recently developed force field parameterisation techniques from the literature to develop a protocol suitable to derive virtually all required force field parameters for small molecules directly from quantum mechanics. This protocol forms the basis of the QUantum mechanical BEspoke force field (QUBE) and is delivered to users through a reliable and extensible software toolkit named QUBEKit. Here we extensively benchmark the methodology and software presented through typical force field performance metrics which involve the prediction of thermodynamical properties of small organic molecules. In this regard, we achieve very competitive accuracy with popular general transferable force fields such as OPLS which have been extensively optimised to reproduce such properties. We also demonstrate how the QUBE force field is a suitable alternative in a computer-aided drug design setting via the retrospective calculation of the relative binding free energies of 17 inhibitors of p38α MAP kinase. Again good agreement with both experiment and transferable force fields is achieved despite this being the first generation of the force field. The results of this work are then particularly important to those studying systems which are not covered or inaccurately represented by standard transferable force fields, as we present an accurate framework towards their complete parameterisation.

Contents

List of Figures

List of Tables

List of Abbreviations

AIM	atoms-in-molecule
AM1	Austin charge model 1
AMBER	assisted model building with energy refinement
B3LYP	Becke, 3-parameter, Lee-Yang-Parr hybrid functional
BCC	bond charge corrections
BFGS	Broyden-Fletcher-Goldfarb-Shanno
CADD	computer-aided drug design
CGENFF	CHARMM general force field
CI	configuration interaction
CLI	command line interface
CM1A	Cramer-Truhlar charge model
COSMO	conductor-like screening model
cpuhrs	cpu hours
DDEC	density derived electrostatic and chemical charges
DE	differential evolution global optimisation method
DFT	density functional theory
DFT-D	density functional theory with empirical dispersion corrections
ESP	electrostatic potential
FEP	free energy perturbation
FF	force field

GAFF	general AMBER force field
GGA	generalised gradient approximation
GPU	graphical processing units
GUI	graphical user interface
H-F	Hartree-Fock
H-K	Hohenberg-Kohn
HPC	high performance computing
HREM	Hamiltonian replica exchange method
IH	iterative Hirshfeld
ISA	iterative Stockholder atoms
K-S	Kohn-Sham
LDA	local density approximation
L-J	Lennard-Jones
LS-DFT	linear-scaling density functional theory
MC	Monte Carlo
MD	molecular dynamics
MM	molecular mechanics
NGWF	non-orthogonal generalised Wannier functions
N-M	Nelder-Mead
NNRTI	non-nucleoside reverse transcriptase inhibitors
NPT	isothermal-isobaric ensemble
MUE	mean unsigned error
OFF	open force field
OPLS	optimised potentials for liquid simulations, force field
PBC	periodic boundary conditions
PBE	Perdew, Burke and Erzenhoff functional

PES	potential energy surface
PME	particle-mesh-Ewald method
QM	quantum mechanics
QMDFF	quantum mechanically derived force field
QUBE	quantum mechanical bespoke
QUBEKit	quantum mechanical bespoke toolkit
REM	replica exchange method
RESP	restrained electrostatic potential
REST	replica exchange with solute tempering
RMSD	root mean square deviation
RMSE	root mean square error
SCRF	self-consistent reaction field
TDDFT	time-dependent density functional theory
TD	TorsionDrive
T-F	Thomas-Fermi
TREM	temperature replica exchange method
T-S	Tkatchenko-Scheffler
xc	exchange-correlation
k_r	bond-stretching force constant
k_θ	angle-bending force constant

Chapter 1

Introduction

Complex biological processes such as protein-ligand binding [10, 11], enzyme catalysis, and protein folding are often best understood when studied at the atomic scale which has driven an increase in the popularity of molecular mechanics (MM) and computational experiments. The ability of MM to model systems ranging in sizes from thousands to millions of atoms makes it indispensable across a wide range of sciences from biology to materials physics. The key to the general success of MM stems from the force field (FF) and its functional form, which allow the approximate description of the potential energy surface (PES) of a system as a simple function of its geometry[12]. Traditionally this energy function is composed of parametrised potentials that each represent one of the different intra/inter molecular interactions present in an atomic system:

$$U = \underbrace{V_{bond-stretching} + V_{angle-bending} + V_{torsions}}_{bonded\,interactions} + \underbrace{V_{LJ} + V_{elec}}_{non-bonded\,interactions} \tag{1.1}$$

Where U corresponds to the total potential energy of the system and the contributing potentials denoted $V_{bond-stretching}$, $V_{angle-bending}$ and $V_{torsions}$ represent the bonded (bond-stretching, angle-bending and torsional) strain energy. While V_{LJ} and V_{elec} represent the non-bonded Lennard-Jones and electrostatic contributions. The simplicity of the functional form, which has generally remained unchanged since its initial inception, is also well suited to computational implementation as the function can be evaluated at great speed via the use of modern hardware including graphical processing units (GPUs) and high performance computing (HPC) clusters making nanosecond simulations routinely achievable for hundreds of systems a week [13].

In organic/medicinal chemistry popular transferable FFs, which contain sets of parameters for eq 1.1, such as GAFF (general AMBER FF)[14], CGENFF (CHARMM general FF) [15] and OPLS-AA [16] are designed to be used in conjunction with their respective highly optimised and benchmarked biological FF counterpart. They are primarily used in simulating drug-like components of systems in, for example, computer-aided drug design (CADD), and give non-expert users the ability to parametrise highly diverse expanses of chemical space at very little computational cost. In particular, they have been crucial in advancing the lead-optimisation stages of drug design campaigns notably leading to the rational design of the most potent non-nucleoside inhibitors of HIV-1 reverse transcriptase (NNRTIs) [17, 18]. The requirement that a FF be transferable stems from two key points, 1) the parametrisation process is a complex and error-prone task that is daunting to the inexperienced user, and 2) an attempt to accurately parametrise all of chemical space would be inconceivable. It is therefore generally assumed that as long as a wide selection of chemical space is covered in the parametrisation set then these results can readily be applied to new molecules. Each of the general FFs use libraries composed of thousands of pre-tabulated parameters [19], intensively fit to experimental and quantum mechanical (QM) data for a set of small molecules that make up their training set. The parametrisation goal of these particular FFs focuses on recreating experimental data concerning the condensed phase thermodynamic properties of small organic molecules, such as liquid densities, heats of vaporisation and free energies of hydration [15]. This parametrisation philosophy follows sound logic as these properties describe the FF's ability to accurately characterise the non-bonded interactions that are also key in protein-ligand binding events. Furthermore, the accuracy and applicability of transferable FFs are aided by efforts such as ForceBalance [20] and the Open Force Field (OFF) Consortium [4], which aim to expand the areas of chemical space that can be automatically parametrised via well-documented protocols.

However, no matter how much effort is put into parameterising small molecules against experimental data, the assumption of transferability remains. That is, the assumption that parameters that are optimal for small organic molecules are also suitable for larger molecules, such as drug-like compounds or even biomolecules. It is well-established that charges polarise in response to their environment, for example the presence of electron donating or withdrawing groups has been predicted to weaken hydrogen bond strengths by up 2 kcal/mol in the case of para-subsitiuted

phenols [21]. Indeed, users of transferable FFs typically derive system-specific charges to account for this polarisation, either from semi-empirical or QM calculations [22–24]. Moreover, it is becoming increasingly apparent that van der Waals parameters themselves show interesting environment-dependent responses with carbon-carbon C_6 coefficients of graphene predicted as being 5.3 times larger than that for graphite [25, 26]. Accounting for changes in van der Waals parameters with changes in FF charges, or the atomic environment, is beyond the scope of most transferable FF protocols.

A fundamentally different approach to FF parameterisation is to instead derive the FF directly from QM simulations of the molecule under study. The potential of using such calculations to develop intermolecular FF potentials for small molecules has long been recognised [27–31]. Here, instead of assuming transferability, the user is able to derive parameters that are specific to their system using a range of automated protocols. Perhaps the most conceptually straightforward approach to QM-based intermolecular force field derivation is to generate many configurations of the system, and fit force field parameters to reproduce the QM energies and/or forces [32–35]. This approach may be applied to quite large molecules using the fragmentation reconstruction method, but extensive sampling of the intermolecular potential energy surface is required for accurate parameter derivation [36]. Alternatively, *ab initio* force fields have been developed that break down the QM interaction energy into physically motivated components using intermolecular perturbation theory [37–39]. These methods incorporate important electronic effects, allow for systematic improvement of intermolecular energies, and can potentially be derived from a very small number of high level *ab initio* calculations [40]. However, compared with more widely-used transferable force fields, *ab initio* force fields generally employ a more complex functional form, which is slower to evaluate, and due to the cost of the underlying QM calculation the majority of applications are to relatively small system sizes. In this regard, Grimme's quantum mechanically derived force field (QMDFF) has several advantages. It takes as input only the QM equilibrium structure, partial charges, Hessian matrix, covalent bond orders and semi-empirical torsion scans, and outputs a full molecule-specific force field [41]. The QM input can be relatively cheap, it is has been applied to molecules comprising more than 100 atoms, and can even be used to model bond dissociation and metals. However, it again uses a more complex functional form compared to standard, transferable force fields, and its accuracy in the condensed phase and the feasibility of extending the approach to heterogeneous problems, such as protein-ligand binding, are yet to be established.

Our goal in this thesis is to describe a QM-derived force field that has the potential to be easily extended to the types of problems usually reserved for standard, transferable FFs, such as host-guest binding in solution [42], simulation of biomolecular assemblies [43], and CADD [44]. To set up a transferable FF for a small molecule, a user typically performs a QM geometry optimisation to fit atomic charges (typically to the QM electrostatic potential), and maybe performs torsional scans for key dihedrals. In order to be competitive with transferable FFs, our FF derivation technique should i) allow users to derive all system-specific bonded and non-bonded FF parameters from these two simple QM input calculations, ii) scale up to relatively large system sizes (e.g. 50–100 atoms), iii) provide parameters suitable for use in mixed simulations (e.g. for the molecule in a solvent or in host-guest simulations), iv) retain the simple functional form of transferable FFs for implementation in the majority of classical molecular dynamics (MD) codes and for use in free energy calculations, v) retain or improve on the accuracy of transferable FFs for modelling of condensed phase properties (and hence implicitly account for many-body effects). That is, we aim to remove any FF limitations associated with parameter transferability, and instead adopt a transferable FF derivation methodology akin to the semi-empirical models routinely used for charge derivation.

Towards this goal, a range of methods for deriving FF parameters directly from QM calculations with minimal experimental fitting have previously been investigated and developed. One of the techniques employed in this study is the modified Seminario method [45, 46], which enables the derivation of bond stretching and angle bending force constants directly from the QM Hessian matrix computed at the optimised geometry. Deriving bonded FF parameters from QM data [47–50], and in particular from the Hessian matrix [46, 51–56] is a well-established concept. The recent adaptation of the original Seminario method [46] yields high quality parameters without relying on iterative fitting of the MM Hessian matrix, which avoids interdependency between force field parameters [45]. In particular, the modified Seminario method has been shown to give parameters that are able to reproduce QM vibrational frequencies with an average error of 49 cm^{-1} for a test set of 70 molecules, which is slightly lower than that achieved by OPLS-AA (59 cm^{-1}) and competitive with methods that rely on iterative fitting of the MM Hessian matrix [41, 57, 55]. The second of the methods employed here is atoms-in-molecule (AIM) analysis, which provides a means to partition the QM molecular electron density amongst the constituent atoms, and hence assign atom-

centered partial charges, even for systems comprising many thousands of atoms [58, 59]. Furthermore, the partitioned atomic electron densities can also be used in conjunction with the Tkatchenko-Scheffler (T-S) relations [25] to calculate all of the Lennard-Jones (L-J) parameters for a molecule. This method of using QM-derived non-bonded parameters has been shown to perform well in recreating liquid densities and thermodynamic properties when applied to a test set of 40 organic molecules [59]. Collectively these methods form the basis of the QUantum mechanical BEspoke (QUBE) FF [60] and to support the adoption and widespread use of the QUBE FF in computational workflows we present QUBEKit [61] as the main focus of this thesis. QUBEKit is a software toolkit that is designed to help users in developing their own QUBE FF in an automated, intuitive and reproducible way that minimises common errors. The QUBEKit workflow combines the previously described and benchmarked independent derivation methods for non-bonded, bond stretching, angle bending and torsion parameters from QM, using the same functional form as the OPLS-AA FF, into a python package.

Now that we have set out the motivation behind deriving FF parameters directly from QM we next move on to briefly cover the underlying theory behind the derivation and application of the QUBE FF. Then, we detail the development cycle of QUBEKit before using it to thoroughly benchmark the QUBE FF through the use of the standard FF metrics in a proof-of-concept automated workflow. We also expand the capabilities of the FF by including parametrisation options for bromine, boron, silicon and phosphorus-containing compounds. Then we demonstrate how the QUBE FF is well suited to a typical drug discovery application by benchmarking its performance in the calculation of relative binding free energies for a series of 17 drug-like inhibitors of the protein, p38α MAP kinase. This system is well suited to the benchmarking of FF performance as it offers a typical medicinal chemistry setting in the lead optimisation stages where we seek in improve drug potencies via well chosen substitutions around a benzene ring, with activities that span 2-3 orders of magnitude. To date this is the most extensive test of the QUBE FF, not only due to the complex sampling requirements, but also the size of the parametrisation problem which included 191 residues (2961 atoms) and 18 ligands of around 40 atoms each.

Finally, we revisit the design of QUBEKit and look to improve on the original torsion parameter optimisation strategy by integrating more open source tools while also trying to improve the quality of the parameters generated. Importantly this work aims to demonstrate that QUBE is a viable alternative to using general

transferable FFs and also provides the community with the means to generate bespoke parameters in a reproducible manner.

Chapter 2

Theory

2.1 Modern Computational Chemistry

The modern computational chemist has many well-established methods at their disposal, from accurate *ab initio* electronic structure calculations to microsecond phase-space molecular mechanics simulations, that all aid in the study of complex physical systems. The accessibility of such techniques has been aided with a surge in open-source programs such as Psi4 [62], OpenMM [63] and BioSimSpace that aim to make performing these complicated calculations as simple and routine as possible. This alongside the increase in computer hardware power makes an ever increasing range of systems accessible to computational study. In this chapter, we highlight the important theory underpinning these techniques which are essential to the construction and application of the QUBE FF starting with modern quantum chemistry calculations and the utilisation of linear-scaling and standard density functional theory (DFT) in combination with implicit solvent models. Next, we describe molecular mechanics and the force field approximation while comparing transferable and specific FF parametrisation routes, before moving on to extracting thermodynamic properties of interest such as free energies from molecular mechanics simulations.

2.2 Quantum Mechanics

Quantum mechanics underlies most of computational chemistry due to the insight into the structure, reactivity and photochemical properties of molecules that can

be gained from electronic structure calculations. Our FF relies heavily on QM calculations as we aim to extract all of the required FF parameters from a few simple widely used QM procedures such as geometry optimisations and frequency calculations. The ultimate goal of any quantum chemistry method is to calculate the electronic structure of a system of atoms which is done through the approximate solution of the time-independent, non-relativistic Schrödinger equation

$$\hat{H}\Psi = E\Psi \tag{2.1}$$

where Ψ is the wavefunction and contains all of the information that can be known about the quantum system. $\hat{H}$ is the usual differential Hamilton operator corresponding to a system of M nuclei and N electrons with no external magnetic or electric fields and is shown below in atomic units where all physical constants are set to unity.

$$\hat{H} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \frac{1}{2}\sum_{A=1}^{M}\frac{1}{M_A}\nabla_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}} \tag{2.2}$$

Here i,j and A,B run over the electrons and nuclei of the system respectively, Z_A and M_A represent the nuclear charge and mass. r_{ij} and R_{ij} represent the distance between two particles or nuclei respectively and can be expressed as $r_{ij} = |\vec{r}_i - \vec{r}_j|$. Beyond the hydrogen atom the Schrödinger equation can not be solved exactly as it becomes a complex many-body problem. Hence we are forced to make some approximations that aim to simplify this while still resulting in an accurate and physically reasonable answer. The first such simplification comes from considering the significant differences in the masses of nuclei and electrons, this indicates that the relative motion of electrons is much greater than that of the nuclei on the quantum scale. Thus, by separating the motion of the two we can reduce this complicated system, using the Born-Oppenheimer approximation, to a simpler one whereby the nuclei are fixed in place and the electrons interact with the potential field they generate. We then arrive at the electronic Hamiltonian

$$\hat{H}_{elec} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} = \hat{T} + \hat{V}_{Ne} + \hat{V}_{ee} \tag{2.3}$$

which is only a function of the N-electron wavefunction and depends solely on the 3N spatial and N spin electron coordinates. Clearly eq 2.3 does not include the

nuclear-nuclear repulsion term which is a constant of the geometry. Here we have also introduced new notations for the kinetic energy $\hat{T}$, attractive nucleus-electron $\hat{V}_{Ne}$ and repulsive electron-electron $\hat{V}_{ee}$ potentials. After separating the nuclear and electronic components of the wavefunction and providing we can set up a specific Hamilton operator for the target system, we are now in a position of solving the Schrödinger equation in order to determine the ground state energy. However it is not clear from the solution alone if we have arrived at the ground state and so we apply the variational principle. This states that the expectation value of the application of the Hamiltonian operator to some initial guess wavefunction will be an upper bound to the energy of the ground state, and thus can be used to determine the quality of our calculated wavefunction. We are then left with an optimisation problem where we need to minimise the energy functional E[Ψ] in order to find the ground state wavefunction Ψ_0, this can be written as

$$E_0 = min_{\Psi \to \Psi_0} E[\Psi] = min_{\Psi \to \Psi_0} \langle \Psi | \hat{T} + \hat{V}_{Ne} + \hat{V}_{ee} | \Psi \rangle \tag{2.4}$$

It would however be impossible to search through all valid N-electron wavefunctions in order to find the true wavefunction so we instead apply another approximation to narrow the selection. The Hartree-Fock (H-F) approximation represents the many-electron wavefunction as an antisymmetrised product of N one-electron wavefunctions or spin orbitals, known as a Slater determinant. While this is a rather simple approximation, it does capture some essential wavefunction behaviour such as antisymmetry under the exchange of two electrons, hence exchange is treated exactly in the H-F model. Now we can construct the expectation value of the Hamiltonian using the Slater determinant as our trial ansatz and by separating out the components of the Hamiltonian we eventually arrive at the H-F equations.

$$\hat{f}_i = -\frac{1}{2}\nabla_i^2 - \sum_A^M \frac{Z_A}{r_{iA}} + V_{H-F}(i) \tag{2.5}$$

$$V_{H-F}(\vec{x}_1) = \sum_j^N \left(\hat{J}_j(\vec{x}_1) - \hat{K}_j(\vec{x}_1) \right) \tag{2.6}$$

These are then used to identify the ground state spin orbitals which are defined as giving the lowest H-F energy, but most importantly, we also define the Fock operator $\hat{f}$ and crucially V_{H-F}, the H-F potential, which is the average potential

felt by an electron due to the other N-1 electrons. The H-F potential is composed of two terms as shown in equation 2.6 corresponding to the *local* Coulomb operator $\hat{J}$ and *non-local* non-classical exchange operator $\hat{K}$. The Fock operator then greatly simplifies the complicated Coulombic electron-electron repulsion interactions of the Hamiltonian to a one-electron operator V_{H-F} which includes repulsion into the system in an average way, but totally neglects Coulomb correlation effects. Consequently the Slater determinant is never the true wave function of a many-electron system, but is the exact wave function of a system of N non-interacting particles moving through the effective potential V_{H-F} and this is an essential construct of DFT which we shall discuss.

It should also be noted that we are assuming, via the use of a single Slater determinant, that the ground state is the only important factor when determining the true N-electron wavefunction, which is not the case as the excited states also have some significant contribution to the system's wavefunction. Thus to calculate a better approximation to the many-body wavefunction one must use post H-F methods like full configuration interaction (CI), in which we consider every possible combination of electron excitations described with a unique Slater determinant at a substantial computational cost that typically scales as $O(N^7)$. Despite even modern hardware, these methods are currently far too slow for any practical application to typical drug like molecules and thus these techniques are beyond the scope of this project.

2.2.1 Density Functional Theory

While the H-F approximation has simplified the problem of approximating a solution to the Schrödinger equation, we are still to deal with the many body wavefunction which is composed of 4N components where N is the number of electrons. Thus the amount of variables quickly becomes unmanageable as we move to typical system sizes of interest in chemistry and biology, rendering even computational treatment difficult. Hence an early attempt to reduce the dimensionality of the problem was made resulting in the Thomas-Fermi (T-F) model. While the model had limited application due to the very coarse approximation of the kinetic energy, Thomas and Fermi were able to construct an equation for the energy of an atom purely from the

electron density.

$$E_{T-F}[\rho(\vec{r})] = \frac{3}{10}(3\pi^2)^{2/3}\int \rho^{5/3}(\vec{r})d\vec{r} - Z\int \frac{\rho(\vec{r})}{r}d\vec{r} + \frac{1}{2}\int\int \frac{\rho(\vec{r}_1)\rho(\vec{r}_2)}{r_{12}}d\vec{r}_1 d\vec{r}_2 \quad (2.7)$$

This then led on to the pioneering work of Hohenberg and Kohn who in 1964 using only two theorems laid down the theoretical pillars which would physically justify the use of the electron density as a basic variable, and go on to create modern DFT. The first theorem proves that the many-body electron density uniquely determines the Hamiltonian operator and consequently all properties of a system or to quote Hohenberg and Kohn directly,

> "*the external potential* $V_{ext}(\vec{r})$ *is (to within a constant) a unique functional of* $\rho(\vec{r})$*; since, in turn* $V_{ext}(\vec{r})$ *fixes* $\hat{H}$ *we see that the full many particle ground state is a unique functional of* $\rho(\vec{r})$".

This then allows us to write the energy of the system as functionals, that is a function of a function, of the electron density but we can separate the terms into two categories, those that are universal and those that are system specific.

$$E_0[\rho_0] = \underbrace{\overbrace{\int \rho_o(\vec{r})V_{Ne}d\vec{r}}^{Nuclei-electron}}_{system\ dependent} + \underbrace{\overbrace{T[\rho_0]}^{kinetic\ energy} + \overbrace{E_{ee}[\rho_0]}^{electron-electron}}_{universally\ valid} \quad (2.8)$$

Then by simply collecting together all of the system independent terms we arrive at the famous universal Hohenberg-Kohn (H-K) functional $F_{H-K}[\rho_0]$, which if known would allow us to solve the Schrödinger equation exactly for any system regardless of size. The second theorem, like the variational principle, provides a way of identifying the ground state density of a system, as this density will minimise the energy functional above.

These vital theorems, however, do not indicate how we should construct the universal functional but simply realises its existence. It would take the approach of Kohn-Sham (K-S) to lay the foundations of the functional application of DFT. The K-S formulation involved mapping the reference many-body system onto an *auxiliary* system of non-interacting particles that recreate the ground state density of the original system. Now if we recall from the H-F theorem that a Slater determinant is the exact wavefunction of a fictitious system of non-interacting particles moving in

an effective potential we can utilise the exact H-F expression for the kinetic energy.

$$T_{H-F} = -\frac{1}{2}\sum_{i}^{N}\left\langle \chi_i|\nabla^2|\chi_i\right\rangle \tag{2.9}$$

While this is not equal to the exact kinetic energy of the real interacting system it is a substantial part and K-S recognised this in their subsequent separated form of the H-K general functional.

$$F_{H-K}[\rho(\vec{r})] = T_S[\rho(\vec{r})] + J[\rho(\vec{r})] + E_{XC}[\rho(\vec{r})] \tag{2.10}$$

In this form we have exact expressions for the non-interacting kinetic energy T_S and Hartree energy J, we also introduce the exchange-correlation energy E_{XC}, which collects together all of the non-classical electrostatic contributions which are unknown. The quality of the results of the DFT formulation then strongly depend on the accuracy of the approximation of E_{XC}, hence a substantial amount of research effort has been spent on the development of a wide range of functionals with differing levels of complexity and accuracy. The local density approximation (LDA) is the simplest model on which virtually all approximate functionals are built. The model assumes that the exchange and correlation energy are solely dependent on the local electron density and can be calculated via analogy to a uniform electron gas with the same density.

$$E_{XC}^{LDA}[\rho] = \int \rho(\vec{r})\epsilon_{XC}(\rho(\vec{r}))d\vec{r} \tag{2.11}$$

Where ϵ_{XC} is the exchange-correlation energy per electron of a uniform electron gas of density $\rho(\vec{r})$, which can then be further split into separate exchange and correlation contributions $\epsilon_{XC}(\rho(\vec{r})) = \epsilon_X(\rho(\vec{r})) + \epsilon_C(\rho(\vec{r}))$. The exchange term can then be calculated exactly for a homogeneous electron gas via Slaters approximation to H-F exchange.

$$\epsilon_X = -\frac{3}{4}\sqrt[3]{\frac{3\rho(\vec{r})}{\pi}} \tag{2.12}$$

While no exact formulation is known to compute the correlation energy it can be approximated via highly accurate quantum Monte Carlo simulations of the homogeneous electron gas from the work of Ceperly and Alder [64]. The LDA represents a good starting point for most functionals, however, a uniform electron gas poorly correlates with the reality of rapidly varying electron densities in atoms and molecules. To

account for this inhomogeneity we can supplement the local density with its gradient at each point which gives rise to the generalised gradient approximations (GGA).

$$E_{XC}^{GGA}[\rho] = \int \rho(\vec{r})\epsilon_{XC}(\rho(\vec{r}), \nabla\rho(\vec{r}))d\vec{r} \tag{2.13}$$

The resulting semi-local exchange-correlation functionals show dramatic improvement over the LDA class in the prediction of total energies, atomisation energies, energy barriers and structural energy differences [65]. This thesis makes use of the advantages of the non-empirical GGA parametrisation by Perdew, Burke, and Erzenhoff, the PBE functional [65] which has been shown to recreate hydrogen bonding well [66]. Trivially then we can introduce more and more complexity into the DFT functional to improve agreement with experiment and therefore create a hierarchy of functionals with varying complexity and accuracy. While more accurate than the LDA these semi-local functions still fail to accurately describe systems where the local K-S potential can not capture long-range contributions into the exchange-correlation energy. This problem is particularly prominent when considering metals and excited-state calculations and dispersion interactions. This introduces yet another class of DFT functionals known as hybrids in which some fraction of the local approximated exchange is substituted for a proportional amount of the exact noninteracting H-F exchange as is done in popular functionals such as B3LYP (Becke, 3-parameter, Lee-Yang-Parr) [67].

$$E_{xc}^{B3LYP} = E_x^{LDA} + a_0(E_x^{H-F} - E_x^{LDA}) + a_x(E_x^{GGA} - E_x^{LDA}) + E_c^{LDA} + a_c(E_c^{GGA} - E_c^{LDA}) \tag{2.14}$$

Such functionals tend to be heavily parameterised to experimental data in order to accurately balance the linear combination of exact and estimated exchange and while they are more computationally expensive than classical DFT, they have become widely popular due to their accuracy. In the case of the B3LYP functional shown in eq 2.14 a_0, a_x, a_c are the empirical parameters and E_x^{LDA}, E_c^{LDA}, E_x^{H-F}, E_x^{GGA}, Ec^{GGA} represent the LDA exchange and correlation, H-F exact exchange and GGA-type Berke88 [68] exchange and Lee-Yang-Parr [69] correlation functionals respectively.

More recently range separated hybrid functionals such as ωB97X-D [1], which is also used in this work, have become more popular. Their success stems from the approximation that exchange over a short distance is well described using approximate local functionals. While at long range, hybrid methods will be more accurate and so

these functionals are able to switch between the two methods based on separation which is controlled by a partition function [1].

2.2.2 Modelling Dispersion Interactions

In order for DFT functionals to reach the desired chemical accuracy, it has become apparent that the proper non-local description of van der Waals dispersion interactions is required. Such interactions are ubiquitous in nature and manifest in QM forces arising from electrostatic interactions between instantaneous multipoles caused by fluctuations in the electron density. Dispersion forms part of the correlation energy of a molecular system and if the exact exchange-correlation functional was known it would be accurately described. Error is however introduced by the local density based approximate DFT functionals which fail to capture its long-range effects. The van der Waals energy is most commonly modelled as a pairwise-additive interaction energy which can be described as an expansion series over the contributing multipole interactions:

$$E_{vdW} = \sum_{i>j}^{N} \left(-\frac{C_6^{ij}}{R_{ij}^6} - \frac{C_8^{ij}}{R_{ij}^8} - \frac{C_{10}^{ij}}{R_{ij}^{10}} - ... \right) \tag{2.15}$$

Where R_{ij} is the interatomic distance between atoms i and j with their corresponding dipole-dipole dispersion coefficient described by C_6^{ij}. The other higher order terms correspond to dipole-quadrupole (C_8), quadrupole-quadrupole and dipole-octupole (C_{10}) interactions. This energy is then often used to correct DFT functional energies in a post-processing procedure with essentially zero computational cost and is known as the DFT-D (density functional theory with empirical dispersion corrections) formalism which is built into the ωB97X-D functional [1] for example. Typically the corrections focus on the leading order pairwise term of the expansion C_6 although some schemes do involve higher order terms in the calculation of the correction energy [70]. Many empirical dispersion models have been suggested, differing only in how the coefficients of the correction are calculated and the choice of damping function. This function helps avoid near singularities in the dispersion energy at small atomic distances and the artificial strengthening of bonds at medium distances [70]. In the simplest models the parameters are assigned from look up tables without modification based on the systems under study. The coefficients themselves are then predetermined in a variety of ways such as iterative fitting to reduce the error differences between those calculated with a DFT functional and

some higher level reference value [71]. The alternative *ab initio* route requires the computation of the frequency-dependent atomic polarisabilities α which can be used to estimate the leading order dipole-dipole term via the Casimir–Polder formula

$$C_6^{AB} = \frac{3}{\pi} \int_0^\infty \alpha_A(i\omega)\alpha_B(i\omega)d\omega \tag{2.16}$$

as the polarisabilities describe the extent an atom's electron density can respond to fluctuations in the local electric field [72]. These coefficients are then calculated for atoms in a variety of reference states in order to be representative of the environments which might be found in the systems under study. Alternatively in an attempt to increase accuracy the reference coefficients can be adjusted via scaling relations based on the local electron density of each atom which accounts for how an atom's polarisability adjusts to its local environment [72, 73]. Overall there are a variety of ways in which the coefficients can be determined, such as well-known Tkatchenko-Scheffler scaling relations, with varying ranges of accuracy, for more details and a comprehensive comparison of the methods see refs 73, 74. While these pairwise potentials have had great success improving the accuracy of DFT functionals it is well known that they represent only part of the true many-body nature of dispersion interactions and models beyond simple pairwise-additive are needed to further improve accuracy. The simplest of which is the Axilrod–Teller–Muto three-body term which describes the triple-dipole interaction energy as a sum over all of the atom triples in the system.

Overall this is still a very active area of research that reaches into the domain of classical MM due to the relation between the leading order dispersion interaction and the attractive component of the Lennard-Jones 12-6 potential. As such some of these dispersion coefficient approximation methods have been sucessfuly applied to parameterise specific FFs directly from QM data [75, 59] (this will be discussed in more detail in section 2.3.2).

2.2.3 Linear Scaling Density Functional Theory

Despite the reduced computational cost of DFT compared to classical wavefunction based methods, it is still far too expensive to be applied to systems beyond 1000 atoms in size (due to its $O(N_e^3)$ scaling). This unfavourable scaling rules out most biological systems of interest, thus linear-scaling-DFT (LS-DFT) is employed to alleviate this burden via the reformulation of the underlying theory to achieve the

more efficient computation of the electronic structure of large systems. There are several ways in which linear scaling can be achieved in DFT, here we describe one such method implemented into the ONETEP LS-DFT code which is used throughout this thesis. The reformulation begins by using the *density matrix* as the central quantity, instead of the electron charge density, which is formally defined in terms of the eigenstates of a single determinant system.

$$\rho(r, r') = \sum_n \psi_n(r) f_n \psi_n^*(r') \tag{2.17}$$

Here f_n is the occupation (0 for unoccupied and 1 for occupied) of the nth K-S eigenstate. The total energy of the system can then be expressed as a function of this new quantity

$$E[n(r)] = -\int \nabla_r^2 \rho(r, r')|_{r=r'} dr' + \int \rho(r, r') V_{Ne}(r) dr + E_{ee}[\rho(r, r')] + E_{XC}[\rho(r, r')] \tag{2.18}$$

Now following the usual DFT convention we aim to minimise this energy with respect to the density matrix subject to some constraints, which ensure that the density matrix completely describes the system in terms of its K-S states. The constraints ensure normalisation or consistent particle count

$$\int \rho(r, r) dr = N \tag{2.19}$$

and idempotency $\rho(r, r') = \rho(r, r')^2$, which ensures integer occupation of the states. So far the use of the density matrix has removed the expensive need of diagonalisation of the Hamiltonian matrix which scales cubically with system size, however, the size of the density matrix still scales as N^2. To achieve true linear scaling performance we must make use of the theory of locality or *near-sightedness* of quantum systems. That is the observation that the properties of a system in one region are weakly correlated with changes in another spatially far away region. Thus it can be shown that for a system with a bandgap the density matrix is a diagonally dominated matrix, with exponential decay properties as a function of two-position operators $|r - r'|$ [76]. To further reduce the computational cost the density matrix can be represented as a sparse matrix of localised orbitals provided the appropriate choice of a localised basis set. In the case of ONETEP, a minimal number of localised non-orthogonal generalised Wannier functions (NGWFs) are centred on the atoms which

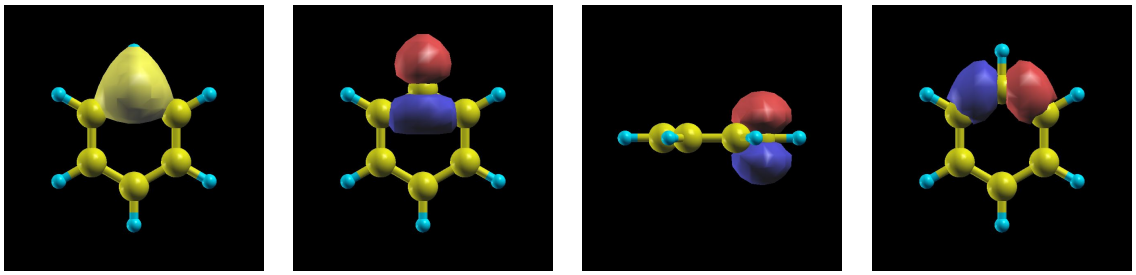

Figure 2.1: An example of how the initial NGWFs (initially placed as 2s (left) and three 2p orbitals (right)) adapt for a carbon atom in benzene during optimisation.

have a small amount of overlap with neighbouring atoms, unlike the K-S orbitals which extend over the whole system. The density can then be expressed in terms of the NGWFs

$$\rho(r, r') = \sum_{\alpha\beta} \phi^*_\alpha(r) K^{\alpha\beta} \phi_\beta(r') \tag{2.20}$$

where $K^{\alpha\beta}$ is the density kernel and ϕ^α is the NGWF basis. Here we can ensure that the density kernel is sparse by using a cut-off radius (though this cut-off is not used in this thesis) so that $K^{\alpha\beta} = 0$ for $|r_\alpha - r_\beta| > r_{cut}$. Now the energy must be minimised with respect to the density kernel and the NGWFs basis, which can be performed *in situ*, allowing the functions to adapt to their surroundings offering high accuracy with minimal size. Figure 2.1 shows an example of how the initial NGWFs adapt for a carbon atom in benzene during optimisation, this clearly demonstrates that the locality of the orbitals is also maintained during the optimisation. During optimisation, the NGWFs are expanded in a basis of periodic sinc (psinc) functions, which are truncated beyond a user-defined cut-off distance (usually around 10 Bohr), and are located on a grid spanning the simulation cell with spacing controlled by an input parameter which corresponds to a plane wave energy cut-off. Then the expanded set of NGWFs and thus the predicted electronic structure properties can be systematically improved via the control of one parameter with linear scaling effort.

2.2.4 Minimal Parameter Implicit Solvent

Biological systems such as protein-ligand complexes are naturally found in solution. QM calculations of such arrangements are usually performed in the presence of an implicit solvent to capture subsequent solvation effects. This is important in the derivation of the non-bonded terms of our FF as we want to capture the

polarisation of the electron density in response to the solvent in the fixed point charges. While it would be more accurate to use an explicit representation of the solvent molecules and average out their positional fluctuations over a large set of configurations, the computational expense is too great for even LS-DFT methods. Instead, the solvent is usually represented as an unstructured dielectric continuum surrounding the solute molecule which is embedded in a vacuum cavity. The solute is then able to polarise the solvent which produces a reactionary net electric field opposing the polarisation of the molecule which can be included into the Hamiltonian in a self-consistent fashion, giving rise to the self-consistent reaction field (SCRF) formalism [77]. Various implicit solvent implementations of differing complexity based on the SCRF method such as the popular polarisable continuum model (PCM) [78] and the conductor-like screening model (COSMO) [79] are commonly used in quantum chemistry. The construction of the vacuum cavity also varies between different implementations but should have some physical meaning, i.e. the shape of the cavity should be representative of the molecule and the majority of the solute's charge density and no solvent should be confined in its boundaries [80]. Normally the cavity is constructed by a union of overlapping atom-centred spheres with radii near the van der Waals value. This requires an extensive amount of empirical parameters to describe the radii of atoms in multiple environments which are fit to reproduce experimental solvation energies [81]. ONETEP however, uses an *ab initio* minimal parameter solvation model which defines the solute cavity using only two parameters; the isosurface value of the ground-state electron density calculated in vacuum and β which is a "*smoothness*" parameter controlling the rate at which the bulk permittivity changes when moving between cavity and solvent. The dielectric permittivity is then a smooth position-dependent potential described by eq 2.21 where $\rho(r)$ is the electron density of the solute, ϵ_∞ is the required bulk permittivity and ρ_0 is the electron density value where the permittivity drops to $\epsilon_\infty/2$ [77].

$$\epsilon(r) = 1 + \frac{\epsilon_\infty - 1}{2}\left(1 + \frac{1 - (\rho(r)/\rho_0(r))^{2\beta}}{1 + (\rho(r)/\rho_0(r))^{2\beta}}\right) \tag{2.21}$$

The total potential of the solute in the dielectric medium ϕ is then found via the solution of the inhomogeneous Poisson equation

$$\nabla[\epsilon(r)\nabla\phi(r)] = -4\pi\rho_{tot}(r) \tag{2.22}$$

where $\rho_{tot}(r)$ is the total charge density including nuclear charges. This new electrostatic potential is then used in place of the Hartree potential in the energy functional.

$$E_{solv}[\rho] = \frac{1}{2}\int \rho(r)\phi(r)dr \tag{2.23}$$

A self consistent solution is sought as the cavity will respond to changes in the ground state electron density which in turn produces a new solvation potential. While this is an accurate procedure it is computationally expensive hence we have chosen to keep the cavity fixed at the initial value calculated from the *in vacuo* ground state density thus avoiding this extra self consistency loop. It is important to stress that the results are still self consistent as the necessary terms are included into the Hamiltion, causing the electron density to respond to the medium and the error introduced due to this approximation is within a few percent of the fully self consistent solution [77].

2.3 Force Fields

The defining difference between QM and MM is the explicit inclusion of electrons in QM which even with the use of linear-scaling DFT techniques is still computationally expensive and slow for large biological systems over long time scales [82]. MM makes use of the FF approximation which allows us to describe a system's PES as a simple function of its internal geometry, meaning systems consisting of hundreds of thousands of atoms can be simulated routinely using classical methods. FFs are traditionally described using bond-stretching, angle-bending, dihedral rotation, electrostatic and L-J contributions, as exemplified by the OPLS functional form:

$$\begin{aligned} U = &\sum_{Bonds} \frac{k_r}{2}(r - r_o)^2 + \sum_{Angles} \frac{k_\theta}{2}(\theta - \theta_o)^2 \\ &+ \sum_{Dihedrals} \left[\frac{V_1}{2}(1 + cos(\phi)) + \frac{V_2}{2}(1 - cos(2\phi)) + \frac{V_3}{2}(1 + cos(3\phi)) + \frac{V_4}{2}(1 - cos(4\phi))\right] \\ &+ \sum_{Pairs} \frac{q_i q_j}{r_{ij}} + \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) \end{aligned} \tag{2.24}$$

The bond-stretching and angle-bending contributions, as depicted in figure 2.2 along with the rest of the FF potentials, require estimates of the force constants

k_r and k_θ respectively as well as reference bond lengths (r_o) and angles (θ_o). The dihedral term is described as a four component cosine series with four corresponding parameters V_1, V_2, V_3, V_4, where ϕ is the torsion angle of the dihedral being described. The OPLS FF also employs an improper dihedral term through the same potential form, using only a V_2 parameter. The final term accounts for all non-bonded interactions between pairs of atoms separated by a distance r_{ij}. The standard Coulomb potential is used to calculate the interaction between two charges q_i and q_j. Finally, the short-range repulsion and longer-range attractive van der Waals interactions are described using the L-J 12-6 potential. Here $A_{ij} = 4\epsilon_{ij}\sigma_{ij}^{12}$ and $B_{ij} = 4\epsilon_{ij}\sigma_{ij}^{6}$ where the ϵ and σ values of the L-J potential govern the energy well depth and minimum energy separation distance respectively. In the OPLS FF, non-bonded interactions are excluded for atoms separated by one or two covalent bonds, and are scaled by a factor of 0.5 for those separated by three bonds. The same set of non-bonded parameters are used to compute inter- and intra-molecular components of the FF.

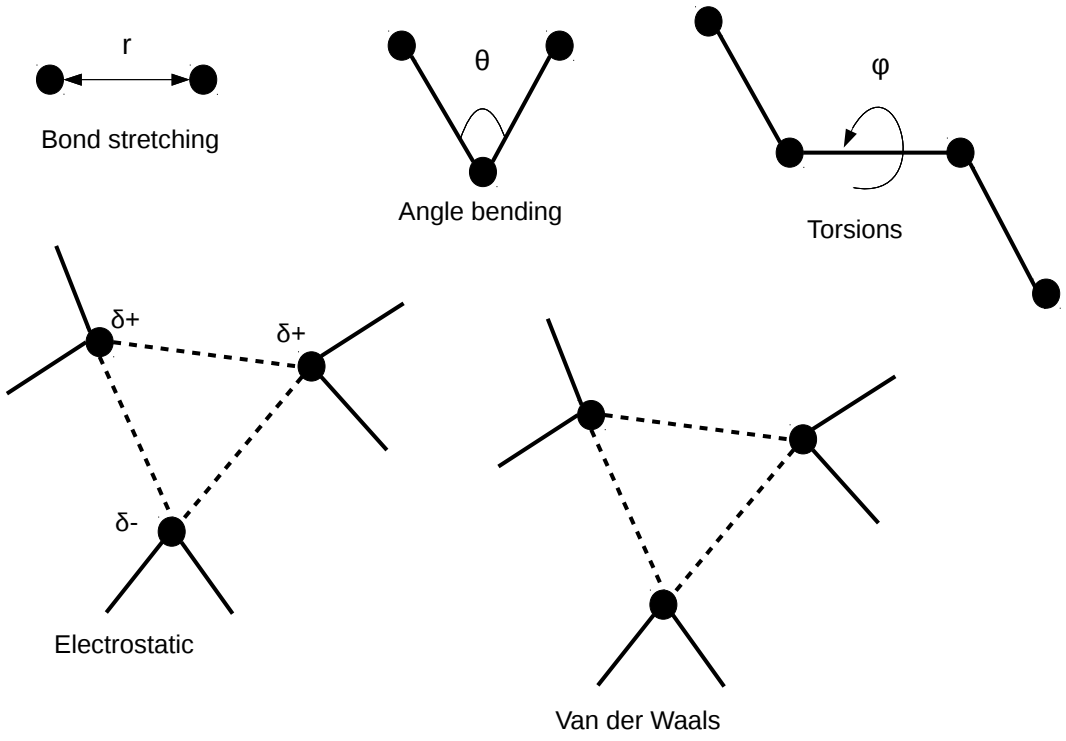

Figure 2.2: A simple illustration of the force field component terms.

A complete set of parameters for any molecule described by this FF functional form requires the derivation of all the parameters of eq 2.24. Traditionally each term has its own parameter fitting protocol and order that varies between FFs. In what follows I discuss the transferable FF philosophy before describing the parameter derivations methods used in the QUBE molecule-specific FF.

2.3.1 General Transferable Force Fields

The general transferable FF has been an essential part of molecular simulations since the 1960's, where its potential to quickly parametrise large biomolecular systems was first utilised [83]. While the defining feature of the FF, the functional form, has not changed much since its initial inception the parameter libraries have been extensively extended and optimised in a continual effort to increase their applicability and accuracy [19, 48, 84]. Currently FFs can be split into two major categories known as class 1 and 2, which correspond to the complexity of their functional form, for example the OPLS form shown in eq 2.24 is the simpler class 1 variety and is usually the choice, alongside AMBER, CHARMM and GROMOS when studying large systems. The more complex forms are exemplified by the MM FF from Allinger's group with MM4 having over 15 different contributions to the energy function. The additional components take the form of cross-terms corresponding to the interdependent motion of the atomic bonds such as torsion-bend or torsion-stretch [85]. The inclusion of such terms makes these FFs well suited to the prediction of vibrational spectra of small organic molecules where sensitivity to geometry displacements is important, however, the extra computation time of evaluating the extended functional form makes the class 1 FF a better choice for large organic simulations.

Another increasingly common type of FF are those which implicitly include charge polarisation effects via a modification to the non-bonded potentials of the functional form in order to accurately recreate molecular polarisabilities and electrostatic potentials [86]. There are many ways in which these many-body effects can be included into FFs that can be further divided into two sub-categories consisting of 1) atomic charge polarisation via the Drude oscillator [87] or induced dipole [86] models or, 2) the flow of charges between atoms via the fluctuating charge model [88]. The most notable implementation of these is in the AMOEBA FF which uses the induced dipole method to model the electronic polarisation and multipoles up to

quadropoles to describe the permanent electrostatic interactions. While AMOEBA has been shown to give higher accuracy, there is an expensive parametrisation penalty associated with the increased complexity of the FF. As in the case of the MM4 FF evaluating these non-standard potentials with extra parameters and the added process of converging the induced multipoles causes a detrimental slow down in the throughput of simulations.

Transferable FFs such as OPLS and GAFF come with vast pre-determined libraries of parameters due to the complexity and time cost of the parametrisation procedure. They are then assigned to a system instantaneously using an atom typing engine which identifies chemical similarity. Even modern transferable FFs such as those presented by the OFF follow this design pattern but use more sophisticated and chemically meaningful approaches to parameter assignment. The technique termed chemical perception is based on SMARTS substructure searching and is able to assign correct GAFF parameters while significantly reducing the redundancy of repeated parameters. Overall the transferable FF has become the most commonly used FF in CADD due to the simplicity of its use and has seen some great success in its applications to date [17, 18, 89–92]. However, there are still some challenging cases particularly in free energy calculations like those regularly performed in the SAMPL challenges where classical FF performance seems to be stagnating [42]. This is problematic as to extract high-quality results from a simulation one requires high-quality and robust parameters which are not necessary contained within a general library as the specific chemistry may not have been seen before. Hence many computational studies now start with some sort of parameter refinement involving QM calculations which provide a cheap and reliable way of generating the required fitting data. Here we aim to extend this refinement concept and challenge the transferable FF philosophy of "parametrise once, use everywhere" with the idea of molecule-specific parametrisation in order to give an accurate representation of the chemical system. In line with the modern data driven approach to transferable FF parametrisation from the OFF Initiative, molecule-specific parametrisation allows us to totally do away with the limitations of atom typing that plague other transferable FF. While the OFF's more recent parameter assignment methodologies based on hierarchical chemical perception SMIRKS patterns show great promise, they still come with a heavy parametrisation burden of associating parameters and patterns.

2.3.2 QUBE Force Field

The QUBE FF removes the assumption of parameter transferability by deriving all of the required parameters for a system from QM, instead opting for a set of transferable parameter derivation methodologies. The resulting bespoke parameter sets, from a few simple QM calculations, are always complete and easy to modify due to their minimal interdependencies. Interdependencies have always been part of general transferable FFs due to the order in which the parameters are fit. For example, imagine changing the charge derivation method in AMBER. Such a change would require a re-fit of the L-J parameter set to reproduce experimental liquid properties, which in turn would necessitate a re-fit of the torsion parameter library. Such interdependencies crucially limit the rate of progress. By adopting this QM derivation technique however, we are easily able to swap out parts of the FF and quickly derive entirely new FFs e.g. to utilise improvements to DFT exchange-correlation functionals without any code base changes. Specific parameters also have the benefit of not being restricted by atom types as the chemical proprieties of the atoms are inferred directly from the electron density at all times. While the benefits of using a bespoke FF are clear the methods which should be used to derive the parameters are not, resulting in many different groups taking unique approaches to solve this complicated optimisation problem. Here we describe one such way of deriving a full set of FF parameters directly from QM and for each component of the QUBE FF we detail and compare the methods to those used in the original creation of the general transferable FFs which we aim to replace.

Bond and Angle Parameters

For each bond and angle in our molecule, we require a force constant and equilibrium value in order to describe the internal energy contribution associated with the vibrational motion. It has been noted that, to describe all of the basic atom type combinations in GAFF, some 20,000 angle parameters would be required [14]. Such large parameter libraries are commonplace, with OPLS3 containing 15,236 angle-bending parameters, and a continuing effort to expand this list as new chemistries are encountered [19]. To generate these parameters, general FFs have to use a wide range of reference data combining experiment and QM. QM data actually already play a role in the derivation of the majority of the transferable parameters in these FFs due to the lack of experimental data available for unique chemical species and the ease

of generating accurate QM data on-the-fly. While many of the equilibrium terms are collected from x-ray crystallography and nuclear magnetic resonance studies of small molecules, some have to be determined from QM predicted minimum energy structures [14–16, 19]. Force constants are then manually fit in an iterative process which aims to recreate the QM vibrational frequencies using an initial guess for the other required parameters as described in the development of CGENFF [15] and AMBER [14]. While this method is effective, it does create interdependencies in the FF parameters as the force constants are dependent on the rest of the original parameter set, meaning that ideally all parameters should be continually updated in a self-consistent fashion until convergence is reached [15].

Instead, we have adopted the modified Seminario method for deriving bond and angle force field parameters. The standard Seminario method derives force constants directly from the QM Hessian matrix [46] and has been incorporated into specialized FF fitting tools for metal complexes such as the VFFDT plugin [93], or in the MCPB.py [94] program which is part of AmberTools. This method estimates bond force constants by projecting the decomposed forces felt by an atom due to the displacement of a neighbouring atom onto their mutual bond vector via eq 2.25 [46].

$$k_r = \sum_{i=1}^{3} \lambda_i^{AB} |\hat{u}^{AB}.\hat{v}_i^{AB}| \tag{2.25}$$

Where λ^{AB} and $\hat{v}^{AB}$ represent the eigenvalues and vectors of the 3x3 sub-Hessian matrix $[k_{AB}]$ related to the bonded atoms A and B with bond vector $\hat{u}^{AB}$.

$$[k_{AB}] = -\begin{bmatrix} \frac{\partial^2 E}{\partial x_A \partial x_B} & \frac{\partial^2 E}{\partial x_A \partial y_B} & \frac{\partial^2 E}{\partial x_A \partial z_B} \\ \frac{\partial^2 E}{\partial y_A \partial x_B} & \frac{\partial^2 E}{\partial y_A \partial y_B} & \frac{\partial^2 E}{\partial y_A \partial z_B} \\ \frac{\partial^2 E}{\partial z_A \partial x_B} & \frac{\partial^2 E}{\partial z_A \partial y_B} & \frac{\partial^2 E}{\partial z_A \partial z_B} \end{bmatrix} \tag{2.26}$$

In the case of angle force constants we can break the problem up into a linear combination of two springs, projecting the eigenvalues onto vectors perpendicular to the bonds involved in the angle. Thus for an angle composed of two bonds **AB** and **BC** and their corresponding perpendicular bond vectors $\hat{u}^{PAB}$ and $\hat{u}^{PCB}$ as shown in figure 2.3, we define the angle force constant k_θ via eq 2.27

$$\frac{1}{k_\theta} = \frac{1}{R_{AB}^2 k_{PAB}} + \frac{1}{R_{CB}^2 k_{PCB}} \tag{2.27}$$

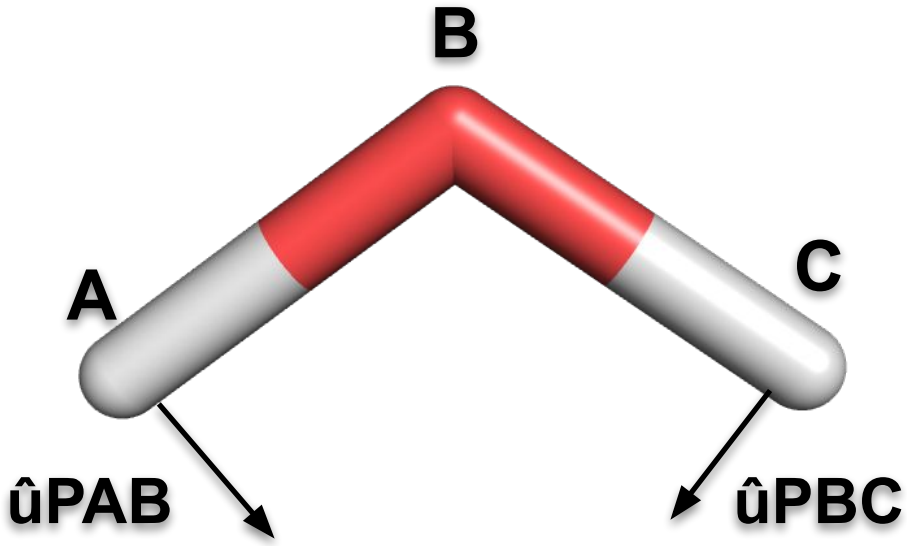

Figure 2.3: The Seminario method as applied to a water molecule.

where k_{PAB} and k_{PCB} are the individual bond components described by

$$\begin{aligned} k_{PAB} &= \sum_{i=1}^{3} \lambda_i^{AB} |\hat{u}^{PAB}.\hat{v}_i^{AB}| \\ k_{PCB} &= \sum_{i=1}^{3} \lambda_i^{CB} |\hat{u}^{PCB}.\hat{v}_i^{CB}| \end{aligned} \tag{2.28}$$

However, this method results in undesirably stiff force constants due to the double counting of angle bending contributions in larger molecules [45]. This can be best seen when looking at benzene as an example, displacing one hydrogen atom perpendicular to the C-H bond deforms two C-C-H angles. However, in the original Seminario method, the entire energy change is attributed to a single C-C-H angle (effectively leading to an over-estimation of the angle force constants). The modified method, however, accounts for an atom's chemical environment and has been shown to recreate QM vibrational frequencies with a low average error of 6.3% across all vibrational modes for a wide range of molecules [45]. This is a vast improvement on the original Seminario method and very competitive with the OPLS FF which reported average errors of 12.3% and 7.4% respectively on the same set of molecules [45]. The modification is done via the rescaling of the Seminario angle force constant by a factor which accounts for the average energy contribution of neighbouring angle changes when the target angle is displaced. As the neighbouring angles may not

necessarily be in the same plane the effect of the displacement may be reduced, hence the neighbouring bond vector is first projected along the vectors $\hat{u}^{PAB}$ and $\hat{u}^{PCB}$. We then arrive at the modified Seminario method for angle force constants.

$$\frac{1}{k_\theta} = \frac{1 + AB}{R_{AB}^2 k_{PAB}} + \frac{1 + CB}{R_{CB}^2 k_{PCB}} \tag{2.29}$$

$$AB = \begin{cases} 0 \text{ if N} = 1 \\ \frac{\sum_{i=2}^{N} |\hat{u}^{PAB_1} . \hat{u}^{PAB_i}|^2}{N-1} \text{ for N} > 1 \end{cases} \qquad CB = \begin{cases} 0 \text{ if M} = 1 \\ \frac{\sum_{i=2}^{M} |\hat{u}^{PCB_1} . \hat{u}^{PCB_i}|^2}{M-1} \text{ for M} > 1 \end{cases} \tag{2.30}$$

Where N (M) is the number of neighbouring angles with central atom B and involve the moment of the AB (CB) bond.

The ability to accurately derive the bonded parameters directly from the QM Hessian matrix without the need for initial parameter guesses simplifies the procedure for non-expert users by removing sources of human error and also speeds up the process making it suitable for automation. We shall also show that the derived force constants retain a low percentage error in recreating QM vibrational frequencies when combined with the rest of the QUBE FF.

Non-Bonded Parameters

The non-bonded interactions incorporate multiple QM effects, such as electrostatics, induction, dispersion and exchange-repulsion, through effective non-bonded Coulombic and L-J interactions. In fixed point charge models there are many methods to derive partial charges from high-level QM calculations using a mixture of population analysis techniques, but ultimately no unique solution. While *ab initio* calculations yield high-quality charges they are often disregarded as being too computationally expensive and are substituted by a variety of semi-empirical QM based methods. These methods allow the rapid assignment of charges and are heavily parametrised in order to reproduce charges observed at higher levels of theory. For example, GAFF employs Mulliken charges produced from semi-empirical Austin Model 1 (AM1) calculations [95] that are then subject to bond charge corrections (BCC) to better recreate experimental hydration free energies [22, 23]. BCC improve the atomic electrostatic potential via the redistribution of partitioned charges in specific bonds. The resulting electrostatic potential is then comparable to that calculated at the HF/6-31G* level which was used to parameterise the AMBER

restrained electrostatic potential (RESP) charges [14]. OPLS-AA, on the other hand, uses Cramer-Truhlar CM1A [24] charges, and recently also included an AM1-BCC inspired localised BCC version of the OPLS-AA/CM1A FF that is available through the LigParGen server [96–98]. It should also be noted that, as these semi-empirical QM calculations are performed in vacuum, they have to be modified to include polarisation effects to make them suitable for condensed phase modelling. This is often performed via the inclusion of the BCC mentioned in the case of GAFF and OPLS, and/or in the form of charge scaling factors all of which are only used on neutral molecules.

On the other hand, CGENFF relies heavily on *ab initio* calculations. CGENFF I charges can be first assigned by a similarity search through a library of parametrised fragments or can be derived using MP2/6-31G(d) Merz-Kollman charges [15] which are fit to reproduce the molecular electrostatic potential. With either starting parameterisation, the charges are subsequently optimised by fitting to QM-calculated scaled interaction energies at the HF/6-31G(d) level between the molecule and water in a variety of conformations. Again we note the choice of low-level theory, this an artefact from the initial derivation of the CHARMM additive FF, to ensure any new parameters are compatible with the biological CHARMM terms. Importantly this means the overall charge description is compatible between systems that require a mix of transferable and biological FFs.

Computational cost is also kept to a minimum in standard transferable FFs by assigning the L-J parameters from a library of pre-fit parameters. This has become standard practice across transferable FFs, with OPLS3 containing 124 different atom types so far, and many general FFs borrowing terms from their biological counterparts [14, 19]. The L-J potential parameters are often tuned to accurately recreate experimental liquid properties [15, 16, 50, 99]. While this technique works very well for atoms covered in the original parameterisation, more atom types often have to be introduced to account for new chemical environments. Infact the poor or miss typing of an atom has been shown to limit the performance of GAFF for some molecules [4]. During the optimisation of the GAMMP/GAFF-LJ* parameters, for example, it was found that for a test set of 430 compounds the 41 standard atom types of GAFF were restricting the maximum achievable accuracy of the FF. The performance was then substantially increased with the addition of 11 new atom types, reducing the average unsigned relative error in the heat of vaporisation from 17.9% to 5.9% [50]. Clearly increasing the number of atom types will help increase

the overall accuracy of a FF as new exceptions to current atom types arise. Logically this implies that system-specific FF parameters have the potential to lead to an overall more accurate FF.

The QUBE FF follows this QM-based philosophy by deriving both L-J parameters and charges from a single ground state QM electron density. An AIM partitioning method divides the total molecular electron density ($\mathrm{n}(\mathbf{r})$) into approximately spherical, uniform overlapping atomic densities ($\mathrm{n}_i(\mathbf{r})$) via:

$$n_i(\mathbf{r}) = \frac{w_i(\mathbf{r})}{\sum_k w_k(\mathbf{r})} n(\mathbf{r}) \tag{2.31}$$

The weighting factor $w_i(\mathbf{r})$ is determined by the choice of AIM partitioning method, in our case the density derived electrostatic and chemical charges (DDEC) [100, 101] scheme is employed. This method iteratively optimises the weighting factor to resemble the spherical average of $n_i(\mathbf{r})$ and the density of a similar reference ion using a mixture of iterative Hirshfeld (IH) and iterative Stockholder atoms (ISA) [59, 100] AIM population analysis techniques. The charges are then found by integrating the atomic electron density over all space:

$$q_i = z_i - N_i = z_i - \int n_i(\mathbf{r}) d^3\mathbf{r} \tag{2.32}$$

Where N_i is the number of electrons associated with atom i and z_i is the nuclear charge. The electron density is calculated as the direct solution of the inhomogeneous Poisson equation in a medium with a dielectric constant $\epsilon = 4$ [59]. It was found that "half-polarising" the molecule with a low dielectric constant resulted in non-bonded terms that are suitable for condensed phase modelling. Including polarisation in this manner allows us to avoid parametrising any BCC or charge scaling factors as employed by CGENFF, OPLS/CM1A and OPLS/CM5 [102].

Additionally, the QUBE FF employs the T-S method to derive the A_{ij} and B_{ij} terms of the FF in equation 2.24 by rescaling reference free atom data, proportionally to AIM electron densities [25]. The dispersion coefficient B_i is estimated as:

$$B_i = \left(\frac{V_i^{AIM}}{V_i^{free}} \right)^2 B_i^{free} \tag{2.33}$$

The atomic volume is readily calculated from the same AIM partitioned electron

density as used in charge assignment via:

$$V_i^{AIM} = \int r^3 n_i(\mathbf{r}) d^3\mathbf{r} \tag{2.34}$$

The B_i^{free} coefficients are computed using time-dependent density functional theory (TDDFT) calculations on free atoms in vacuum [6]. Specifically TDDFT is used to calculate the dynamic polarisabilities, α of the isolated atoms which can then be used in the Casimir-Polder integral (eq 2.16) evaluated over all imaginary frequencies, $i\omega$ to find the C_6^{AA} dispersion coefficient between two isolated atoms. V_i^{free} is the reference volume of the atom calculated using the MP4(SDQ)/aug-cc-pVQZ method in Gaussian 09 [2] and the chargemol code [5] for each of the elements in our model (table 4.5). To ensure that the dispersion and repulsion coefficients result in a minimum in the L-J potential close to the van der Waals radius of the atom, it can be shown that the A_i coefficient can be approximated by:

$$A_i = \frac{1}{2} B_i (2R_i^{AIM})^6 \tag{2.35}$$

The AIM effective radius R^{AIM} of each atom is found by rescaling the reference free atom radius using the T-S method:

$$R_i^{AIM} = \left(\frac{V_i^{AIM}}{V_i^{free}} \right)^{1/3} R_i^{free} \tag{2.36}$$

After the partitioning of the electron density and the calculation of the L-J terms we do see some slightly undesirable asymmetries that must be accounted for by evenly distributing charge and L-J terms across identical atoms defined by their local environment. We also then adjust the L-J terms on any polar hydrogen atoms, that is those bonded to O, N or S by transferring the hydrogen L-J contributions to the parent heavy atom via

$$\sqrt{B'_x} = \sqrt{B_x} + n_H \sqrt{B_H} \tag{2.37}$$

Where B'_x and B_x are new and old dispersion terms of the heavy atom X respectively, n_H corresponds to the number of bonded hydrogen atoms and B_H is their pre-zeroed dispersion component. A full description of the non-bonded parameter derivation methods can be found in Ref. 59.

At simulation runtime we then have to combine the A_i and B_i parameters between two interacting atoms using the geometric combination rules of the OPLS FF:

$$A_{ij} = \sqrt{A_i A_j} \quad \text{and} \quad B_{ij} = \sqrt{B_i B_j} \tag{2.38}$$

It is this feature that also makes the mixing of different transferable FFs undesirable as while AMBER and OPLS share a common functional form AMBER instead uses the Lorentz-Berthelot combination rules which are compared in figure 2.4.

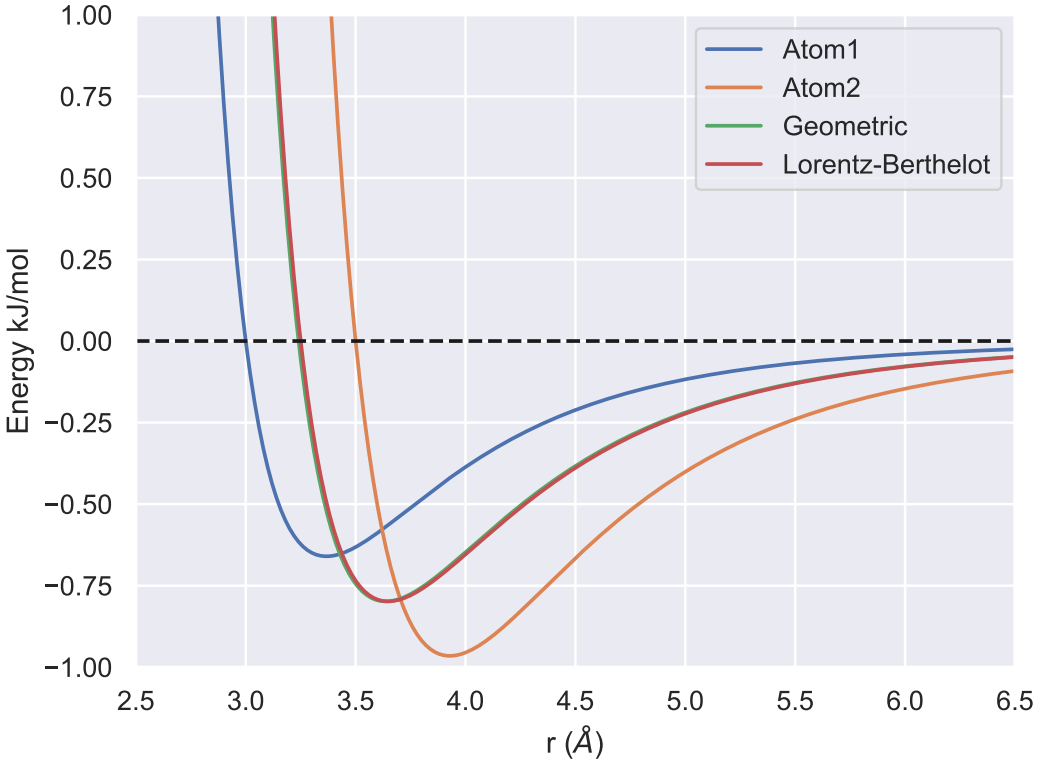

Figure 2.4: An example plot of the Lennard-Jones potential calculated using the Lorentz-Berthelot and geometric combination rules.

Anisotropy

This method was developed by Alice Allen at the University of Cambridge as part of a collaborative project [61].

While atom-centered point charges provide a good representation of the QM electrostatic potential (ESP) if the partitioned atomic electron density is spherical,

in many cases this simple representation is inadequate[59]. This situation occurs when there is significant anisotropy in the underlying electron distribution, and is common in molecules containing nitrogen, sulfur or halogens [103]. Here, to model electron anisotropy, we employ off-center, "virtual" sites, which have been shown to be competitive with the use of more computationally expensive higher-order multipole electrostatics [104]. Virtual sites are commonly used in water models, such as TIP4P [105], and various force fields for modelling lone pairs and σ-holes [106], but the positions and charges of the virtual sites require fitting to experiment. On the other hand, it has recently been shown that virtual site positions may be derived directly from localized QM molecular orbitals [107, 108], but currently the magnitudes of the charges are derived by fitting to the molecular dipole moment, which may be problematic for extension to larger molecules that contain multiple sites. In keeping with our goal of avoiding fitting FF parameters to experiment and developing methods that scale to biological molecules, a method was proposed that relied on the dipole and quadrupole moments of the partitioned atomic electron density, to optimize the charges and locations of virtual sites [59]. However, the method employed did not consistently converge and resulted in a large number of off-center point charges. Modifications were required to correct these issues and improve the usability of the method in an automated high-throughput scenario.

Here, we propose a method for the derivation of virtual site positions and charges directly from the QM electron density in which the virtual sites are positioned so as to reproduce as closely as possible the QM ESP of the partitioned atomic electron density. By determining the virtual site parameters only using atomic properties, the method scales trivially to macromolecules such as proteins. In order to reduce the search space we limit the virtual site positions to those dictated by the symmetry of the atom's bonding environment. Together these improvements allow us to define virtual sites that improve the electrostatic properties of the simulated molecule in an automated manner.

The QM ESP (Φ_i^{ref}) is calculated from the partitioned atomic electron density ($\mathrm{n}_i(\mathbf{r})$). This is advantageous as the method may be applied equally well to both surface and buried atoms. 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 $F(\Phi, \Phi^{ref})$ is given by:

$$F(\Phi, \Phi^{ref}) = \sum_{i=1}^{M} \frac{|\Phi_i - \Phi_i^{ref}|}{M} \tag{2.39}$$

where M is the number of sampling points. The MM ESP (Φ_i) is calculated as:

$$\Phi_i = \sum_{j=1}^{N} \frac{q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.40}$$

where N is the number of sites on an atom, r_{ij} is the distance from the site to the sampling point and q_j is the charge on site j. An additional threshold parameter (F_{thresh}) was required to distinguish between atoms that required extra sites and those that did not. Above this threshold the anisotropy method is used, below the threshold, no off-center charges are added. As well as this, extra charges are only added when there is a reduction in error which is controlled by a second parameter (F_{change}).

One Additional Off Center Charge

For atoms with ESP error above the threshold, we begin by attempting to model the anisotropy using a single off-center charge. The vectors for one additional off-center point charge that preserve symmetry are shown in Fig. 2.5. The vector direction is governed by the number of atoms bonded to the atom exhibiting anisotropy:

1. *One bond.* The atom A (which exhibits anisotropy) has one neighbor, atom B. The vector along which the extra charge is positioned 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.

2. *Two bond.* The atom A has two neighbors, atoms B and 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.

3. *Three bond.* The atom A has three neighbors, atoms B, C and 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.

After the vector is assigned, the optimal position along the vector and the charge of the off-center point is determined. This is carried out using a grid search of parameters to find the values which best recreate the QM ESP. Assigning a symmetry-derived search direction 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

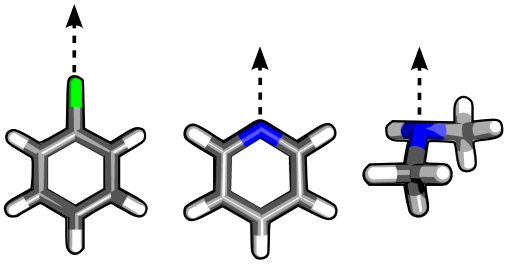

Figure 2.5: The directions along which off-center point charges are placed for an atom with one, two or three bonds.

multiple off-center point charges are added, as described in the following section. The atom-centered point charge is assigned a value such that the net charge of the atom is unchanged.

Multiple Off-Center Charges

In Ref. 59, 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 atoms bonded to the atom exhibiting anisotropy:

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

2. *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, $\mathbf{r_{1,2}} = \lambda_{\parallel}(\mathbf{r_{AB}} + \mathbf{r_{AC}}) \pm \lambda_{\perp}(\mathbf{r_{AB}} + \mathbf{r_{AC}}) \times (\mathbf{r_{AB}} \times \mathbf{r_{AC}})$, or perpendicular to this plane, $\mathbf{r_{1,2}} = \lambda_{\parallel}(\mathbf{r_{AB}} + \mathbf{r_{AC}}) \pm \lambda_{\perp}(\mathbf{r_{AB}} \times \mathbf{r_{AC}})$. An example is shown in Fig. 2.6. A third extra charge can also be added and is placed along the bisector $\mathbf{r_3} = \lambda_3(\mathbf{r_{AB}} + \mathbf{r_{AC}})$.

3. *Three bonds.* A second off-center charge is placed along the same 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 NH_2 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.

A disadvantage of using the partitioned electron density to calculate the QM ESP is that it includes regions that are not accessible during MM simulations, such

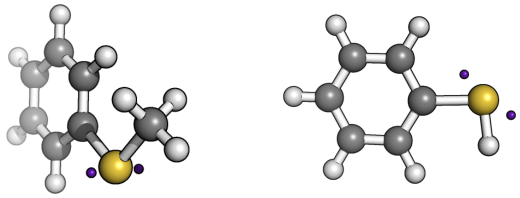

Figure 2.6: An example of off-center charge placement for the case (left) perpendicular to the plane of the bond vectors or (right) in the plane of the bond vectors.

as between bonds. This is the case for the amine group and results in other regions of the QM ESP not being adequately reproduced. The addition of an off-center site between the nitrogen and hydrogen atoms helps to overcome this issue.

Torsional Parameters

The final stage in the fitting procedure is the optimisation of torsional parameters. Torsional rotation is an important factor controlling the conformational preference of a molecule due to its association with QM stereoelectronic effects, and the parameters are therefore often a target for re-optimisation [19, 57, 84, 109–113]. In this work, we follow a standard procedure of fitting the parameters to minimise the difference between MM and QM constrained one dimensional torsional scans. During fitting, we aim to optimise the four V_n parameters of the OPLS FF torsion potential shown in eq 2.24 by automating the scheme outlined in Ref. 84 with some additional considerations. The steepest descent algorithm is employed to find the torsional parameters that minimise the regularised Boltzmann weighted error function:

$$\Omega = \sqrt{\frac{\sum_{i=1}^{n}(\Delta E_{MM}^{i} - \Delta E_{QM}^{i})^2 e^{-\Delta E_{QM}^{i}/k_B T}}{n}} + \lambda \sum_{torsions} \sum_{j=1}^{4} |V_j^{ref} - V_j| \qquad (2.41)$$

where k_B is the Boltzmann constant, T is a temperature weighting factor, n is the number of sampling points and V_j^{ref} is a reference torsional parameter. ΔE_{QM} and ΔE_{MM} are the QM and MM optimised energies at each sampled torsional angle relative to the lowest QM or MM energy. MM scans allow all other degrees of freedom to optimise, and so the structures are similar but not identical to the QM optimised structures. Overfitting is often a concern at this point in the fitting process. Here, we introduce a regularisation function controlled by a variable parameter λ, which constrains the fitted torsional parameters to be close to the reference values, V_j^{ref}. In this work, V_j^{ref} were taken from the OPLS force field, but could also be set to

zero [114]. It is also important to note that it is not possible to always perfectly recreate the entire QM PES hence users should concentrate on relatively low energy regions as these are most likely to be sampled during room temperature simulations. The weighting temperature T can be adjusted to preferentially weight the low-energy regions of the QM PES.

In molecules containing multiple flexible dihedral angles, it was found that torsional parameters were best fit in an order that started with rotations that would involve the movement of the fewest number of atoms. For example, a long chain molecule with no repeated dihedral types would be best fit by starting at the ends and working inwards. Larger molecules could also be fragmented during fitting to reduce the computational cost of the fitting procedure. It should also be noted that we do not derive any improper torsion parameters in this workflow, instead taking them from the OPLS-AA FF.

2.4 Molecular Mechanics

MM provides researchers with a means to simulate the intricate atomic motion of biomolecular systems in order to extract thermodynamic properties of interest or validate hypotheses concerning structural changes. MM has a wide range of applications across biology, chemistry and materials science and has become a vital part of CADD as we will show due to time and cost efficiency improvements. In order to achieve accurate results it is imperative that we accurately sample the phase space of our system using an appropriate technique. There are two such ways to generate system configurations in MM which use the potential energy of a system as described by a FF, MD and Monte Carlo (MC). While both methods are valid ways to generate system ensembles as demonstrated by their long running success in MM applications [17, 18] they are actually very different processes. Importantly MD represents a system's evolution over time with particle trajectories while there is no timescale associated with the randomly generated states in MC. Within this work we employ both MD and MC sampling techniques during the validation and application stages depending on the different quantities we aim to calculate along with software compatibilities. To this end we give a brief overview of the two different sampling techniques.

2.4.1 Molecular Dynamics

The overall goal of MD is to propagate a system according to Newton's second law of motion:

$$\boldsymbol{f}_i(t) = m_i \boldsymbol{a}_i(t) = -\frac{\partial V(\boldsymbol{x}(t))}{\partial \boldsymbol{x}_i(t)} \tag{2.42}$$

Where f_i, m_i and a_i are the force, mass and acceleration of the ith atom of the system at time t. The vector $\boldsymbol{x}(t)$ represents the system's configuration in Cartesian space and the associated potential $V(\boldsymbol{x})$ which is computed via the FF in eq 2.24. However, this equation can only be solved exactly for systems of limited size, hence MM relies on the repeated use of integrators that aid with numerical solutions to the problem. Integrators advance a system's state by discontinuous intervals known as time-steps, δt, resulting in a series of new system configurations corresponding to the evolution of the atomic positions. One of the most widely used examples of an integrator that demonstrates this technique is the velocity-Verlet method which is derived as the Taylor expansion of the atomic positions after some time-step.

$$\boldsymbol{x}_i(t + \delta t) = \boldsymbol{x}_i(t) + \boldsymbol{v}_i(t)\delta t + \frac{1}{2}\boldsymbol{a}_i(t)\delta t^2 \tag{2.43}$$

In this method the next set of atomic positions depends on the current positions, velocities and accelerations, we must also then update the velocities of the next configuration via

$$\boldsymbol{v}_i(t + \delta t) = \boldsymbol{v}_i(t) + \frac{1}{2}[\boldsymbol{a}_i(t) + \boldsymbol{a}_i(t + \delta t)]\delta t \tag{2.44}$$

Then by combining these steps together we have the general procedure used by MD engines such as OpenMM or GROMACS to advance a system over time, this routine can also be outlined as a flow chart as shown in fig 2.7.

From eq 2.44 we see that we have to advance the system's positions ahead of calculating the velocity at time $(t + \delta t)$ which creates a lag between the positions and velocities. Thus in order to check that our integrator is propagating the system as we expect we must check it retains physical properties of the equations of motion and hence the system. For example the total energy of a closed system should remain constant throughout simulation. To ensure this is the case a time step of 1-2 fs is typically used. On the other hand additional considerations must be made if we want to include more physical macroscopic fluctuations such as temperature

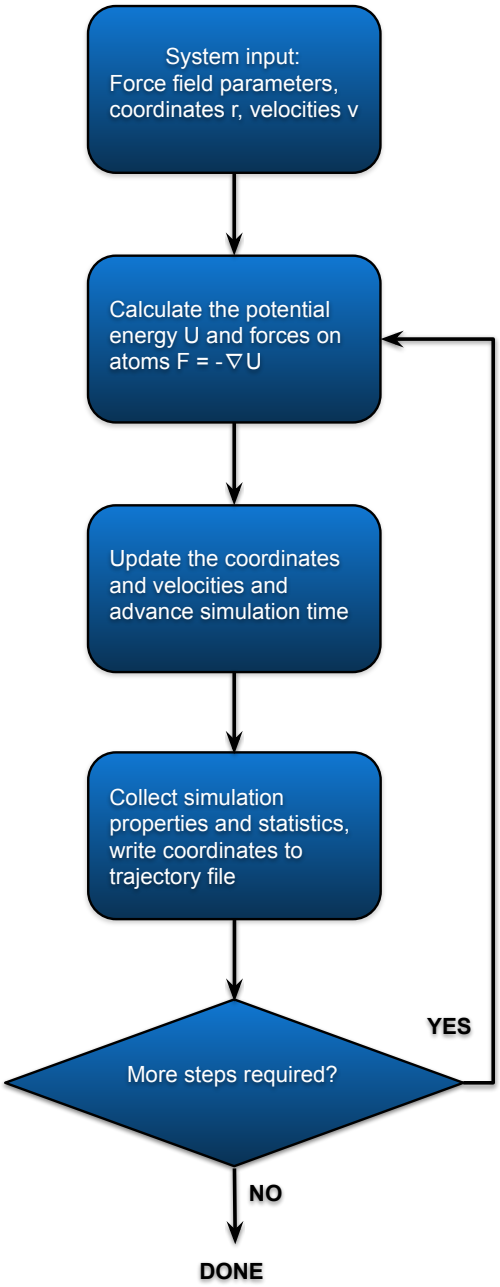

Figure 2.7: Shows the procedure followed to generate new configurations in a MD simulation.

and pressure in our simulations. Constant temperature can be achieved via the use of thermostat algorithms such as the Andersen method which rescales a random particle's momentum from a Boltzmann distribution at the desired temperature [115]. One way to achieve this is via the introduction of a stochastic collision term into the equations of motion which controls the frequency with which particles have their

momentum rescaled [12, 115]. Alternative deterministic methods are also available which modify the equations of motion to rescale the velocities, the simplest example of this would be to use a factor of $\sqrt{T/\tau}$ where τ and T are the current and desired temperatures respectively [116]. The isothermal-isobaric (NPT) ensemble can be obtained by the proper control of the system's pressure. One potential way of doing this is via the random isotropic scaling of a system's volume using a Monte Carlo barostat. Finally to allow the accurate simulation of bulk liquid properties using reasonable system sizes we often employ periodic boundary conditions (PBC), whereby a unit cell of molecules is repeated in all directions to form an infinite lattice. Long-range non-bonded interactions are then computed up to a chosen cut-off distance within the periodic cell and are smoothed towards this truncation through the use of a switching function. This allows us to recreate liquid bulk properties to high accuracy using only a small number of molecules in our pure organic liquid benchmark calculations.

2.4.2 Monte Carlo

MC sampling can be applied to the study of static, thermodynamic or equilibrium quantities that can be calculated as an ensemble average, or expectation value, of some mechanical system property such as internal energy [117]. The use of ensemble instead of dynamical time averages to calculate system properties is valid providing the states generated by MC are representative of the appropriate distribution. The goal of using MC is then to produce a Boltzmann distribution of system states via a random walk through phase space following a Markov chain in order to reach thermodynamic equilibrium. A Markov chain is built from the repeated application of a Markov process, which in this regard refers to the stochastic transformation from one system state to the next via a set of transformation probabilities which must meet the following criteria: 1) the transformation probabilities must not change over time, 2) the probability of generating the next configuration should depend on the present state only and not the previous history, 3) for any initial state the sum over all probability transitions to some final state must be one. The actual system transformations correspond to a set of small possible perturbations which are used at random during each MC step of a simulation. The allowed perturbations also depend on which system component is chosen to be moved, for example, in a protein-ligand complex the changes can be made to the ligand, protein or the surrounding

solvent. Translations and rotations are often applied to ligands and solvent molecules while backbone and side chains of proteins undergo rotations about randomly chosen flexible bonds which surround the binding site [118]. In the case of the software used in this project (MCPRO), a MC move of a fully flexible molecule involves relocating the molecule in three-dimensional space and reorientating it, before making changes to a selection of the available hard degrees of freedom such as bonds, angles and torsions [119]. A new set of Cartesian coordinates corresponding to the translation of a molecule could easily be produced via the following set of example equations:

$$x_{new} = x_{old} + (2\eta - 1)\delta r_{max} \tag{2.45}$$

$$y_{new} = y_{old} + (2\eta - 1)\delta r_{max} \tag{2.46}$$

$$z_{new} = z_{old} + (2\eta - 1)\delta r_{max} \tag{2.47}$$

Here δr_{max} is the maximum allowed step size and η is a randomly generated number between 0 and 1, fig 2.8 shows the range of possible locations that atom i can be moved to in order to generate a new system state. When designing the move

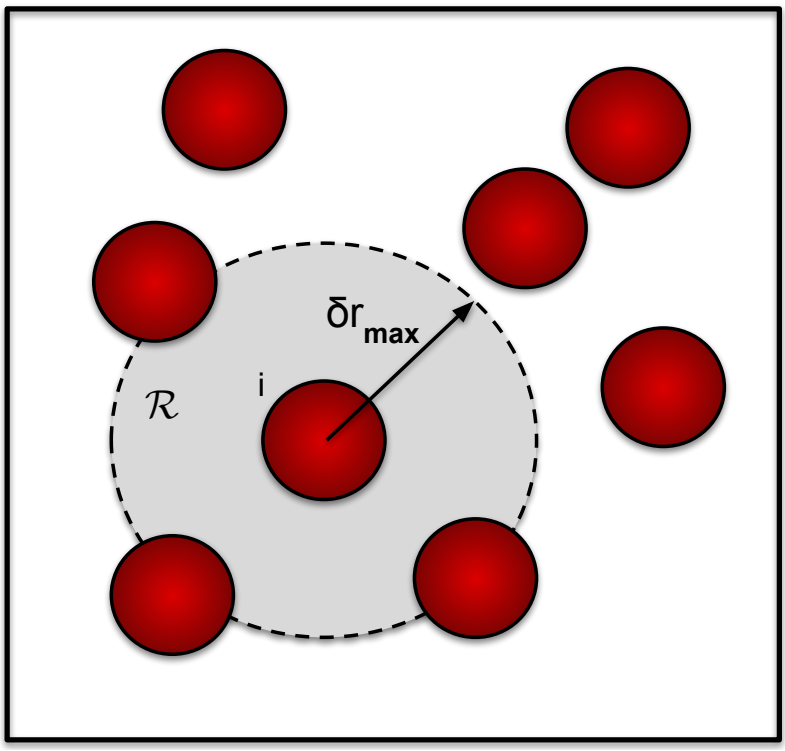

Figure 2.8: A new system state is generated by moving atom i with equal probability to any point in the region R defined by the maximum displacement distance δr_{max}.

set we must also keep in mind that in order to reach equilibrium our Markov process should be ergodic, that is there is some chance of attaining any possible final state from any current state. This gives rise to the principle of detailed balance, which requires that the probability of observing the system transition $\alpha \rightarrow \alpha'$ is the same as the reverse transition of $\alpha' \rightarrow \alpha$. Now that we can generate new system configurations we need a way to ensure that the states are representative of the desired Boltzmann distribution, hence MC is always used with the Metropolis criterion [120]. The Metropolis criterion generates an acceptance probability related to the change in a system's energy after a proposed move, which then determines if the new configuration should be accepted as a valid state. If we take the system in fig 2.8 for example and move atom i to a new position and then evaluate the potential energy difference $\Delta U = U_{new} - U_{old}$ using the FF we can check if we have decreased or increased the system's energy. The Metropolis criterion then states that we should always accept a new configuration which lowers the potential energy of the system. However, if the move increased the energy then we should only accept the state if $rand(0,1) \leqslant e^{-\Delta U/k_b T}$, where $rand(0,1)$, k_b and T represent the generation of a random number between 0 and 1, the Boltzmann constant and the temperature respectively. From these steps we now have the basic method of generating a Metropolis MC move which is repeated in order to generate an ensemble average for the system. This routine, employed by simulation engines such as BOSS and MCPRO, is outlined by a flow chart in fig 2.9.

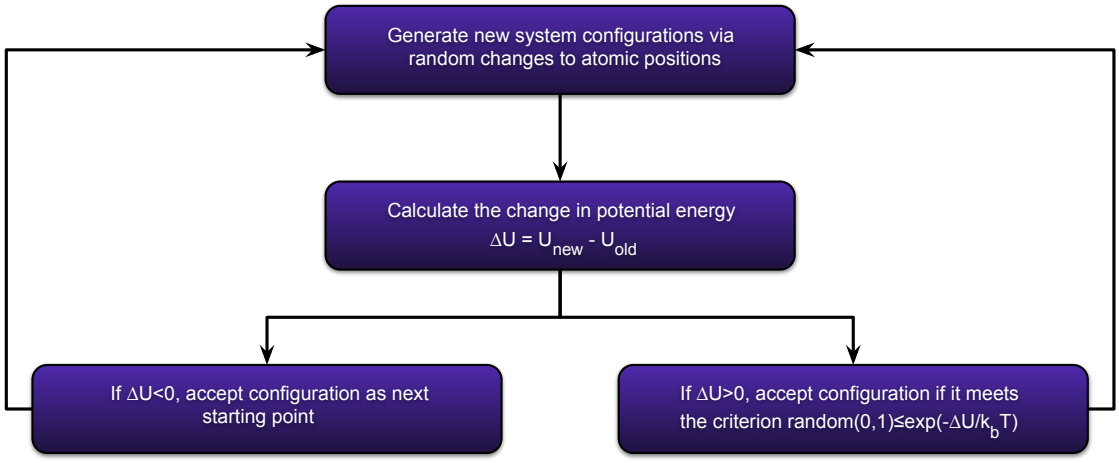

Figure 2.9: The procedure followed to generate new configurations in a MC simulation.

Clearly the acceptance rate of any new configuration is related to the system's change in energy and hence the maximum atomic displacement, making the fine

tuning of this parameter important when considering the convergence of a simulation. A small displacement is very likely to be accepted, but progress through conformational space is slow and requires many more steps to explore new areas, whereas large displacements are unlikely to be accepted. Hence we often design MC algorithms which aim to generate moves with a 50% acceptance rate to ensure we are efficiently sampling conformational space.

2.4.3 Enhanced Sampling with REST

Accurate conformational space sampling of biophysical systems with many degrees of freedom such as protein-ligand complexes remains challenging, due to large energy barriers in the PES [121]. Energy barriers with heights greater than k_bT often cause the ligand, in particular, to become trapped in local energy minima for long periods resulting in quasi-ergodic sampling [122] (that is, simulations may appear to converge but be very sensitive to the starting conditions). This incomplete sampling can have a disastrous effect in CADD where the proper prediction of ligand binding modes involves fully sampling all energetically accessible configurations to ensure accurate affinity predictions in free energy perturbation calculations. Figure 2.10 shows an example of a simple 1 dimensional PES with many local minima in which the system can become trapped making the results of the simulation highly dependent on the starting configuration. One solution might be to start the simulation from a

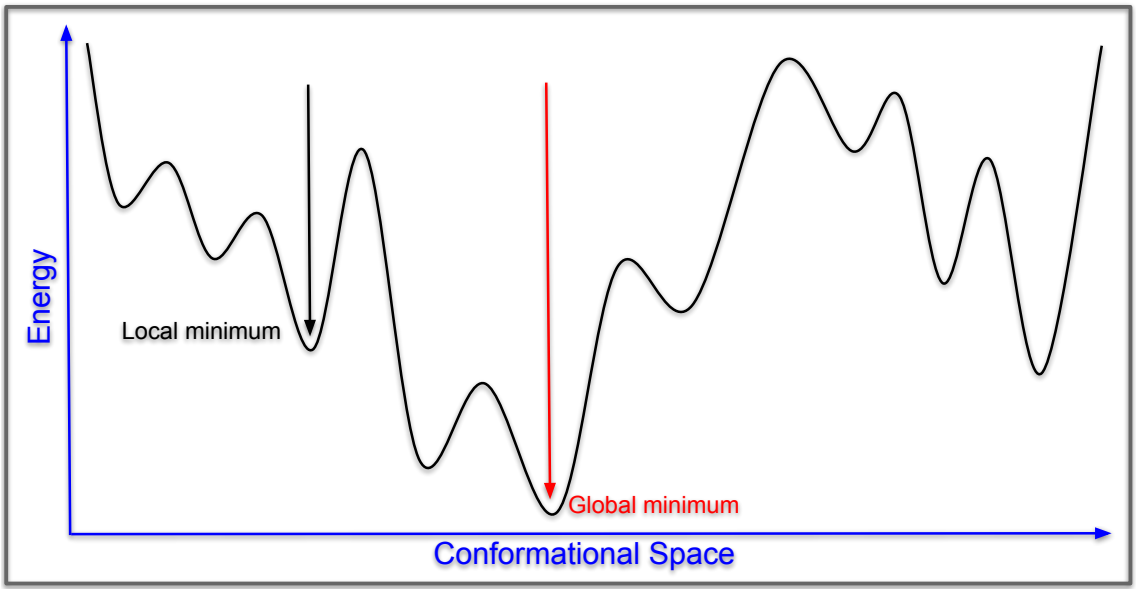

Figure 2.10: A simple example of a rough one dimensional PES with many local minima.

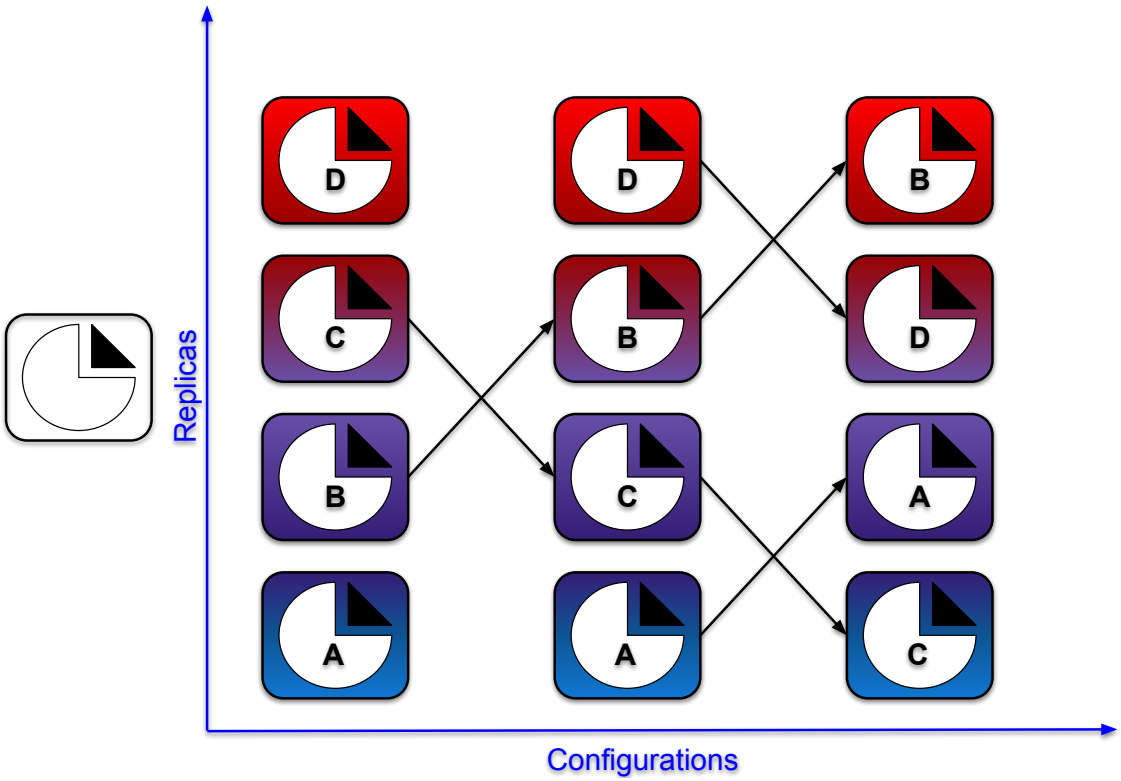

Figure 2.11: Schematic of the temperature REM in action. The colours of the replicas represent the distribution of temperatures from the target (blue) to the highest simulated temperature (red). At regular intervals the configurations are exchanged between replicas resulting in an ensemble for the target temperature composed of configurations A,A,C.

series of different conformations that are separated by high energy barriers, but this can be very computationally expensive and complicated for molecules with many fully rotatable dihedral angles. Parallel tempering, or the replica exchange method (REM), is one of the most popular schemes in CADD to avoid quasi-ergodicity and involves simulating multiple replicas of the system in parallel under different simulation conditions such as temperature (TREM). Then at regular intervals, an exchange of system configuration is attempted with a higher-temperature replica which facilitates the frequent crossing of high barriers in the PES, the acceptance of which is controlled by the Metropolis criterion [12]. Figure 2.11 shows an example of TREM in action for a distribution of temperatures ranging from the target (blue) to the highest chosen (red). The effective use of this method requires the number of replicas to be simulated to scale as $N^{1/2}$ where N is the number of degrees of freedom of the whole system, therefore limiting the applicability of TREM for large systems due to computational cost. A much more efficient alternative is to use

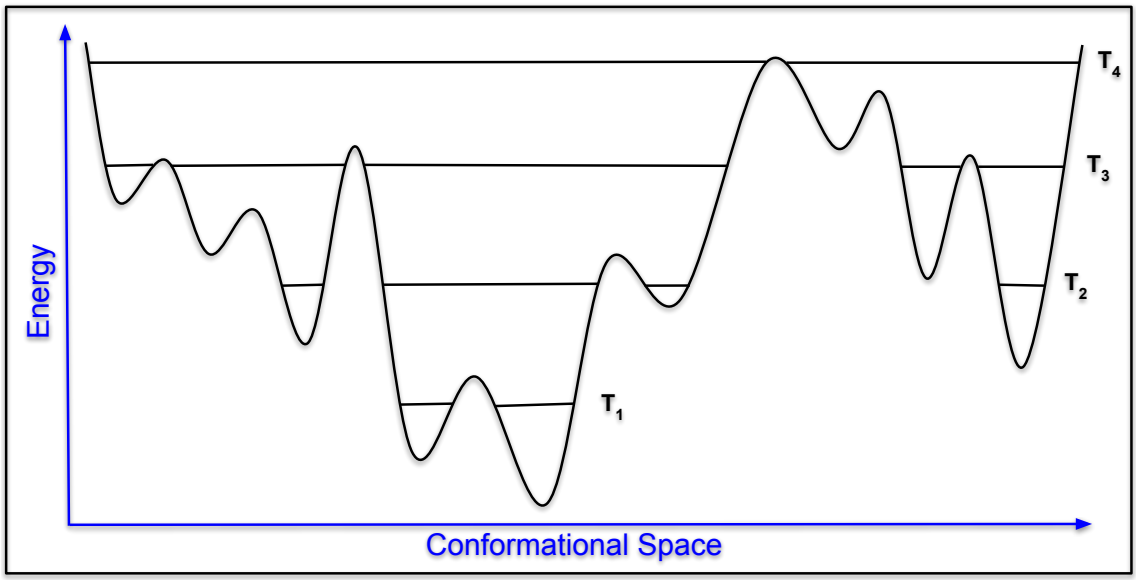

Figure 2.12: An example of how solute tempering allows the exploration of high energy states during a simulation.

Hamiltonian REM (HREM) whereby the PES of the system is incrementally scaled down in the replicas, rather than the temperature. This enables the targeting of specific relevant degrees of freedom like those of the ligand, greatly reducing the number of replicas required. Replica exchange with solute tempering (REST) is one variation of this method which has recently been implemented into MCPRO [122] and has been shown to be accurate and robust when used with just four replicas of the system at an exponentially distributed range of temperatures. Similar to REM, replica m is then simulated at temperature T_m, but the form of eq 2.48 ensures that effectively only the ligand degrees of freedom are heated. A standard procedure in the application of REST is to first breakdown the potential energy of an interacting protein-ligand system into its component parts which normally include, the ligand's intramolecular interactions (E_L) , the surrounding water and protein interaction energy (E_P) and lastly the energy from the protein-ligand interactions (E_{PL}) [122]. The scaling of the Hamiltonian is then described as

$$E_m(X_m) = E_L(X_m) + \frac{\beta_0}{\beta_m} E_P(X_m) + \sqrt{\frac{\beta_0}{\beta_m}} E_{PL}(X_m) \tag{2.48}$$

where the total energy E_m is a function of the system configuration X_m and scaling factors $\beta_m = 1/(k_b T_m)$ which are temperature T_m dependent. At the temperature

of interest T_0 eq 2.48 reduces to the normal unscaled form the potential energy expression. The effect of the targeted heating on our simple example can be seen in figure 2.12, where the chosen range of temperatures allows larger and larger regions of conformational space to become energetically accessible as the ligand can now rapidly cross energy barriers in the PES. The REST method employed in MCPRO is further enhanced with a "flip" protocol which periodically attempts large rotations (bigger than typical MC moves) around user selected dihedrals. This ensures that the chosen diherals have sufficient opportunity to transition between separated energy wells within a reasonable amount of MC moves. In combination with REST this flip protocol has been shown to efficiently enhance confirmational space sampling and successfully improve the consistency between results starting from configurations separated by large energy barriers [122, 123].

2.5 Free energy perturbation

The ability to computationally rank a congeneric series of ligands based on binding affinity to a target receptor is an invaluable technique in CADD. Conventionally this is done using free energy techniques based on MM simulations and results in either the absolute binding free energy of a single ligand-receptor complex [123] or the relative binding free energy between two similar ligands bound to the same receptor [122]. Here we focus on alchemical or relative binding free energy calculations using free energy perturbation (FEP) theory as these are generally regarded as more efficient [10]. This is due to a presumed cancellation of errors introduced by the FF or incomplete sampling of the systems during the simulation. This assumption is reasonable provided that the ligands are structurally similar enough to share binding modes or if different, they should be separated by barriers that can be easily crossed during a simulation. Formally the Zwanzig equation [124] can be used to calculate the relative free energy $\Delta G_{A\rightarrow B}$ between two system states A and B as

$$\Delta G_{A\rightarrow B} = -k_b T ln \left\langle exp\left(\frac{-\Delta E_{AB}}{k_b T}\right)\right\rangle_A \tag{2.49}$$

Here the free energy difference is calculated as the ensemble average (denoted by <..>) of the exponential difference of the state energies ΔE_{AB} with k_b and T referring to the Boltzmann constant and temperature respectively. In a CADD context, these end states would correspond to two structurally similar ligands differing only by

Reference(A)		Perturbed(B)
F	→	H
Cl	→	F
CH_3	→	H
Br	→	F
OH	→	F
CH_3	→	OH
NH_2	→	F
NH_2	→	OH

Table 2.1: A typical example set of small FEP transformations.

simple atomic substitutions which could potentially boost binding affinity (alchemical perturbation). A range of typical alchemical transformations are listed in table 2.1 although larger perturbations are possible [125]. The Zwanzig method then requires the generation of an ensemble corresponding to the reference system (state A), the energy difference is then subsequently calculated using the perturbed system (state B) on the same ensemble. However, due to the exponential average, the proper convergence of this ensemble average is only possible if the two end states are similar and there is a reasonable overlap between them. To ensure this FEP calculations often include a series of unphysical intermediate states spanning between the desired end states which facilitate the gradual alchemical transformation from the reference to the perturbed ligand. A reaction coordinate (λ) is then introduced which is coupled to the FF parameters (X) of the systems and is used to linearly scale between states A and B.

$$X_i = \lambda_i X_B + (1 - \lambda_i) X_A \tag{2.50}$$

A series of simulations are then performed at the intermediate values of λ spanning between the reference ($\lambda = 0$) and perturbed ($\lambda = 1$) states known as λ-windows. Increasing the amount of λ-windows used in FEP improves the convergence of the results as $\Delta\lambda$ is decreased meaning the neighbouring windows have greater overlap. However, the number of simulations required increases linearly with the number of windows [126]. Hence the choice of $\Delta\lambda$ is important to balance these trade-offs. It is also important to consider the type of λ-window sampling employed. For example, the simplest form is known as "direct sampling" and involves perturbing state A to B in the forward or backward direction i.e. $\lambda = 0 \rightarrow 0.25 \rightarrow 0.5 \rightarrow 0.75 \rightarrow 1$. Then if the calculation is performed in both directions any hysteresis in the results

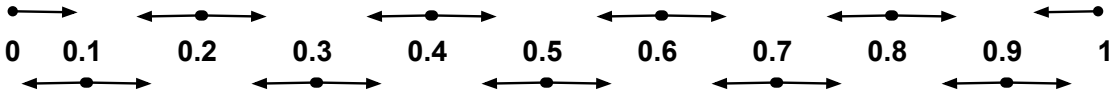

Figure 2.13: An example of the simple overlap sampling scheme used in MCPRO where dots represent simulations run at each λ window and the arrows correspond to the evaluation of the Zwanzig equation.

would be an indication of any error in the simulation resulting from poor overlap or incomplete sampling. Trivially this can be reduced by taking the average energy difference between each window and is known as "double-ended sampling". This, however, requires twice as many simulations and quickly becomes very expensive when combined with REST. Hence "simple overlap sampling" is a computationally efficient choice as one simulation per λ-window is required but the forward and backward energy changes are calculated simultaneously, as shown in figure 2.13, and averaged accordingly.

Now that we have the means to calculate the free energy change between to separate states we can construct a thermodynamic cycle composed of four legs as shown in figure 2.14 to extract the relative binding free energy. The cycle involves two physical legs (horizontal) which represent the unbinding of ligands A and B (or the absolute binding free energy) and two unphysical simulated legs (vertical) corresponding to the alchemical transformation between the ligands in pure solvent and in complex with the receptor [10]. As free energy is a function of state that only depends on the end states and is the same regardless of the intermediate path taken through the cycle, we can determine that $\Delta G_{bB} = \Delta G^{p} + \Delta G_{bA} - \Delta G^{w}$. Thus the relative free energy between ligands A and B is defined as

$$\Delta\Delta G_{A\to B} = \Delta G_{bB} - \Delta G_{bA} = \Delta G^{p} - \Delta G^{w} \tag{2.51}$$

There are also two ways in which the standard thermodynamic cycle can be implemented in a simulation referred to as *single topology* and *dual topology* methods. Here we concentrate on the *single topology* type in which one structure representing ligand A is transformed into the second B throughout the simulation. This method can often involve the use of extra non-interacting virtual sites as shown in Figure 2.15 which

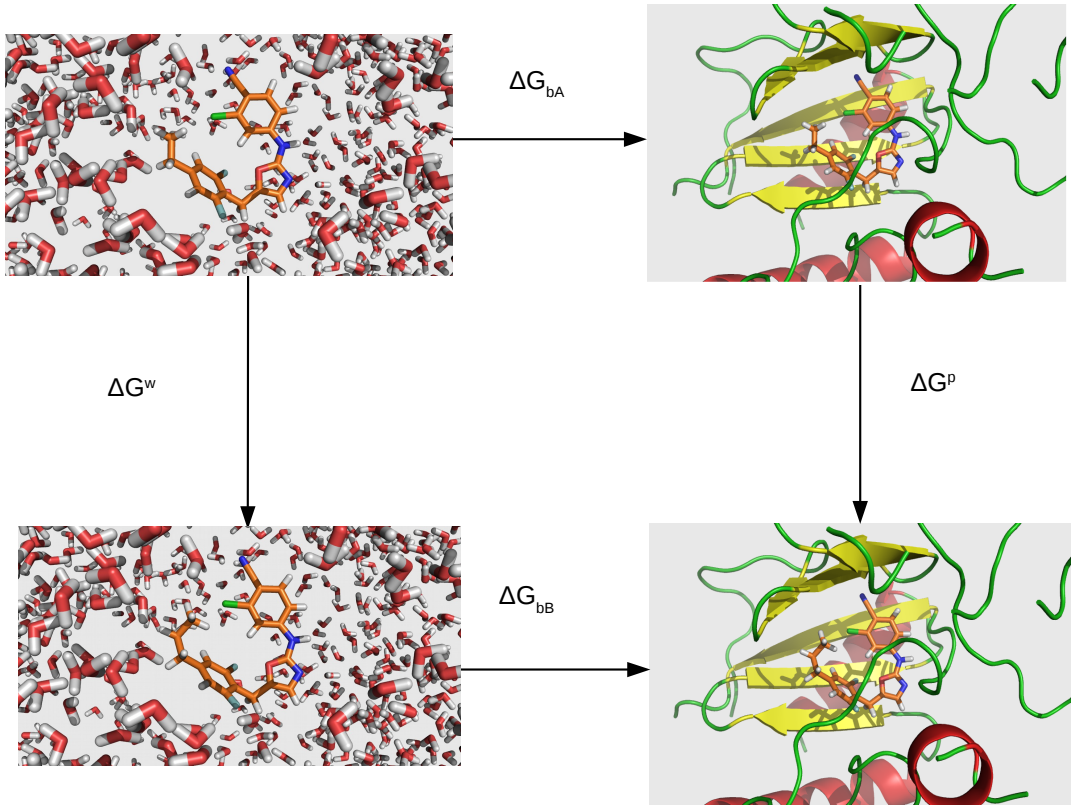

Figure 2.14: An example thermodynamic cycle used to calculate the relative binding free energy between two non-nucleoside inhibitors of HIV-1 reverse transcriptase. Where ΔG_{bA} and ΔG_{bB} are the binding free energies of the ligands to the receptor, ΔG^{w} is the free energy difference between the ligands in solution and ΔG^{p} is the free energy difference in the bound system.

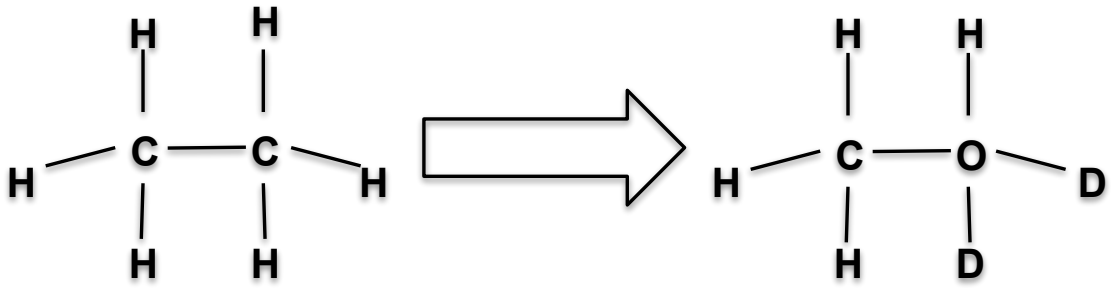

Figure 2.15: An example of a *single topology* FEP transformation between ethane and methanol, where D corresponds to dummy or virtual non-interacting particles used to aid the transition.

shows the end states of the first FEP calculation performed by Jorgensen et al. [127] where ethane was transformed to methanol in water. Note that similar ideas can be used to compute the hydration free energy which is used in this thesis for FF validation.

Advances in FEP now see its routine application in prospective CADD campaigns [17, 18], and its accuracy is only limited by finite sampling time and the accuracy of the underlying FF. This is then the motivation behind our goal of improving FF accuracy by deriving environment-specific bespoke parameters directly from QM. Now that we have identified and detailed the methods available to do this, our aim in this thesis is to bring together these techniques in an automated pipeline and validate their accuracy, ultimately testing their suitability in FEP simulations.

Chapter 3

The development of QUBEKit

3.1 The need for automated software

FF development and parametrisation has long been regarded as a '*black art*' due to the complexity and scale of systematically tuning thousands of parameters in a data driven way to result in a comprehensive biological FF [20]. Thus the process was typically only carried out by large and experienced groups which drove the community to adopt general transferable FFs. Over many years we have seen the periodic increase in the size of the required parameter libraries along side their refinement. However, as application of the FF starts to diverge from its initial conception, weaknesses have started to emerge. For example, in the case of modelling intrinsic disordered proteins a new water model TIP4P-D [128] and biological FF a99SB-disp [9] had to be developed in order to accurately model such systems. Thus users are now aware of typical FF shortcomings and parameter modifications and optimisations are now becoming standard practice. For example, many users are opting to recalculate the atomic charges [129] and optimise problematic torsion parameters based on QM calculations. Hence the need for automated parametrisation protocols has become apparent, with many becoming a regular part of the standard modelling workflow. The development of physically meaningful, specific and robust parameter derivation techniques based on abundant QM calculations, have gained a lot of traction in the community, due to their ability to rapidly parametrise molecules. However, this has also brought with it a pressure to release and maintain corresponding software that is able to facilitate the accurate application of such methods. This is critical in order for others to use these techniques in a reliable, systematic and reproducible manor. Hence the development of QUBEKit was

essential to offer users another avenue alongside the wide range of options currently available to the comunity [20, 35, 41, 55, 130–132].

In this section we detail the development of QUBEKit and its ability to automate the error-prone task of parameter derivation and also highlight features which can be integrated into other workflows.

3.2 QUBEKit design and development cycle

QUBEKit was designed to automate and reduce the complexity of the bespoke parametrisation of small molecules for CADD by combining and interfacing with a collection of recently developed parameter derivation techniques. QUBEKit does this via the automatic file writing/reading of various input/output files from a variety of external programs and then deriving parameters using the QUBE methodology described above. The software's original design was that of an interactive script (*QUBEKit.py*) that could be called from the command line with a set of intuitive flags (operating in a similar way to many bash programs to increase familiarity), such as *bonds charges dihedrals*, which controls the action to be performed. The parametrisation sequence was then broken up into an order of best practice starting with a fully relaxed geometry optimisation and frequency calculation which is needed to derive the bond-stretching and angle-bending parameters using the modified Seminario method. Initially only the Gaussian09 QM software was supported for this operation, and input files were created from an initial BOSS z-matrix and corresponding pdb file (which can be generated via the LigParGen web server [96]) using the *QUBEKit.py -f bonds -t write -p filename* command. The job file will instruct Gaussian to run a full optimisation to the tight convergence criteria before starting the subsequent Hessian calculation, bond and angle terms are then derived and inserted into a BOSS style parameter file using the *-f bonds -t fit* command. QUBEKit also produces an xyz formatted file at the optimised geometry during this step which is then to be used in ONETEP in order to derive the non-bonded parameters from a single point calculation in implicit solvent. Once that calculation has finished QUBEKit can then be called with the *-f charge -t fit* tag which will extract the AIM partitioned charges and volumes from a ONETEP output file, required to derive the specific L-J terms. During this step we also extract any virtual-site positions and charges which significantly reduce the electrostatic potential error around the target atom and automatically merge them into a MM simulation topology

file. Along side this, users are able to write a series of constrained QM optimisation input files which form the QM reference data for the torsion optimisation step. The original implementation of this method followed the standard practice of scanning the dihedral in the forward direction from 0° to 360° in user defined increments. Once the QM calculations are finished the results are automatically gathered ready for fitting. The dihedral parameter optimisation could then be selected using the *-dihedrals fit* command tag from within the QM reference scan folder, regularisation can then be easily changed using the *-l* flag.

Once fitting is complete the final BOSS style z-matrix and parameter file are created ready for simulation along with plots of the fitting results and a detailed log file of the procedure. Users also have the option to convert these BOSS style simulation files into an OpenMM (XML) or GROMACS input files using the *-X yes* operation. As QUBE, like the OPLS FF, uses the geometric combination rules during simulation we also provide a comparison of the single point energies calculated using BOSS and OpenMM to ensure they matched and the resulting XML parameter file accurately represented the FF. This sanity check function could be performed at the end of parametrisation along with the normal mode comparison, which uses BOSS to calculate the MM normal modes and compares them to their QM counterparts which are calculated during the frequency simulation, using the *-SP* and *-FR* flags respectively.

A flowchart summarising the proof-of-concept workflow used throughout this thesis is also shown in figure 3.1. While QUBEKit was designed to be simple to use we also give users advanced control over a series of important runtime parameters which can significantly alter the FF, such as the level of quantum chemistry and basis set, and the temperature weighting in the torsional fitting. These parameters are controlled through the use of an INI style configuration file which is commonly used in Microsoft operating systems and is thus easy to understand and parse through the standard python library. This allows the creation of multiple configuration files outlining specific combinations of parameters, which may refer to different projects, and can be easily selected at runtime using the *-config filename* flag. Any of the preferences can be subsequently overwritten while running using the wide range of quick command line controls such as *-charge, -multiplicity, -basis, -theory* each corresponding to the equivalent parameter in the configuration file. Overall this gives users control over almost all running settings, however, QUBEKit will default to the original settings described in chapter 4 should black box behaviour be requested.

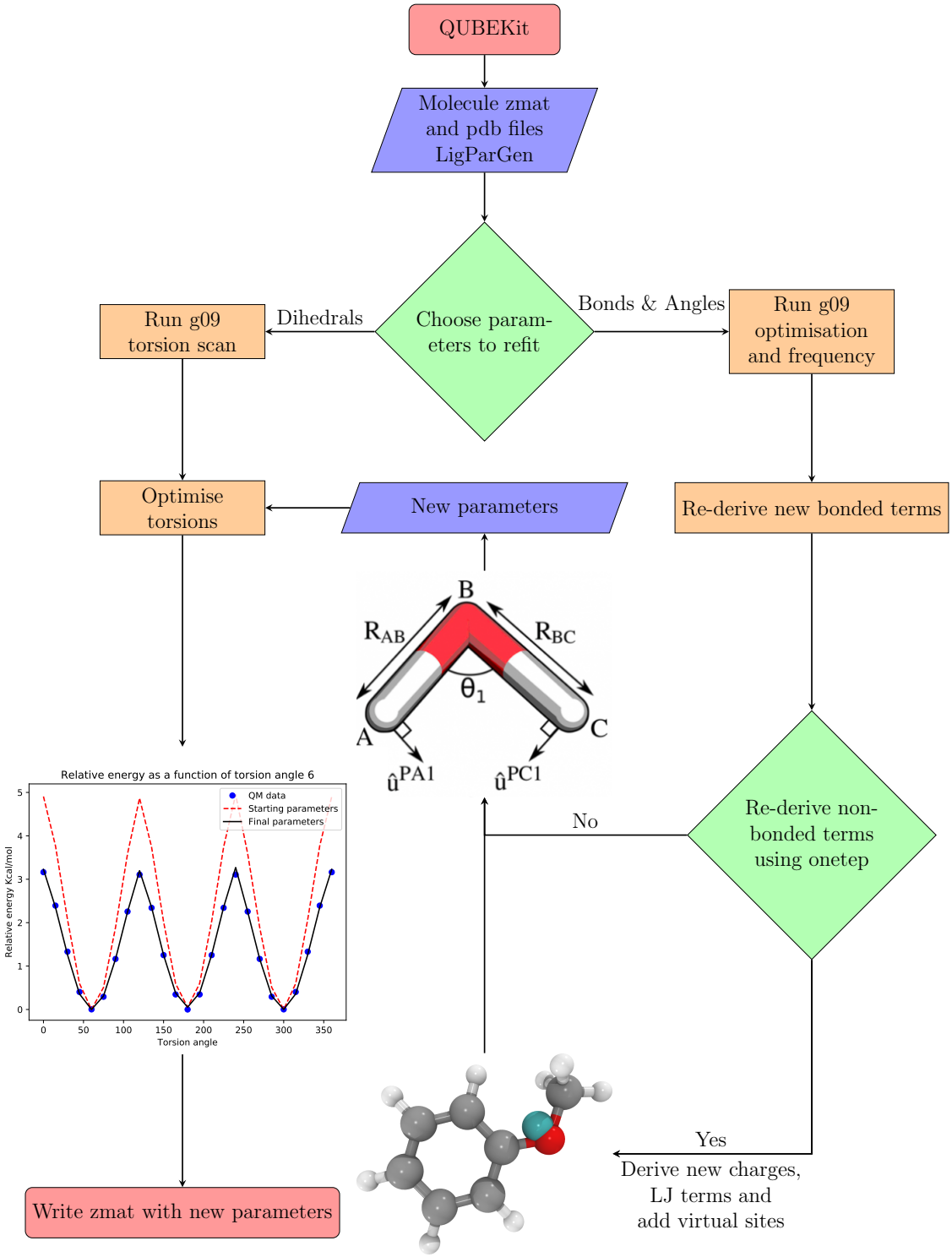

Figure 3.1: QUBEKit example workflow used throughout this thesis.

While this initial design sufficiently demonstrated how QUBEKit could be used to automate the parameter derivation process for around 150 molecules studied in development and testing phases discussed in chapters 4 and 5, the underlying software dependencies potentially limited its widespread adoption by the community. The initial software dependencies were inherited from the parameter derivation techniques combined into QUBEKit as they were developed to be compatible with specific software in mind which is often the case in proof-of-concept applications. Following this we began work on QUBEKit-v2 (co-developed by Chris Ringrose, Newcaslte University) which aimed to correct this by integrating a wider range of software options which most importantly were open-source, greatly reducing the barrier to entry. This would also see the re-writing of QUBEKit to make use of the powerful object-orientated capabilities of python, allowing us to modularise each task in the workflow and generalise the inputs, making for the easier extension and modification of QUBEKit. This also gave us the opportunity to expand on some other limiting design choices in the first iteration such as input file format and initial parametrisation method.

Figure 3.2 shows how QUBEKit-v2 was modularised such that each block would represent a python class which would do one unit of work during the operation of the program. Modularisation of the tasks during parametrisation vitally allows the trivial swapping of one class for another at any point in the workflow allowing us to easily add different software dependencies. For example, the initial parametrisation can now be performed by any of the three corresponding classes which will apply parameters from antechamber (*Antechamber()*), the OFF toolkit (*OpenFF()*) or an OpenMM style XML file (*XML()*). Each of these has been designed to work with our internal data structure class called *Ligand()* which stores any properties or information which are associated with or describe the molecule being parametrised. This also reduces the complexity of using the application interface (API) built into QUBEKit when employing the classes in other workflows such as in the example script extract shown in figure 3.3. The script extract demonstrates how the modularisation and generalisation of the modified Seminario method, allows us to derive force constants from QM reference data taken from the QCArchive [133], or any other source from which we can extract the required input data (optimised geometry and Hessian matrix). In fact this script extract is very similar in structure to our automated workflow script *run.py* which is run when QUBEKit is called via the command line interface (CLI). This main script gives a through demonstration of

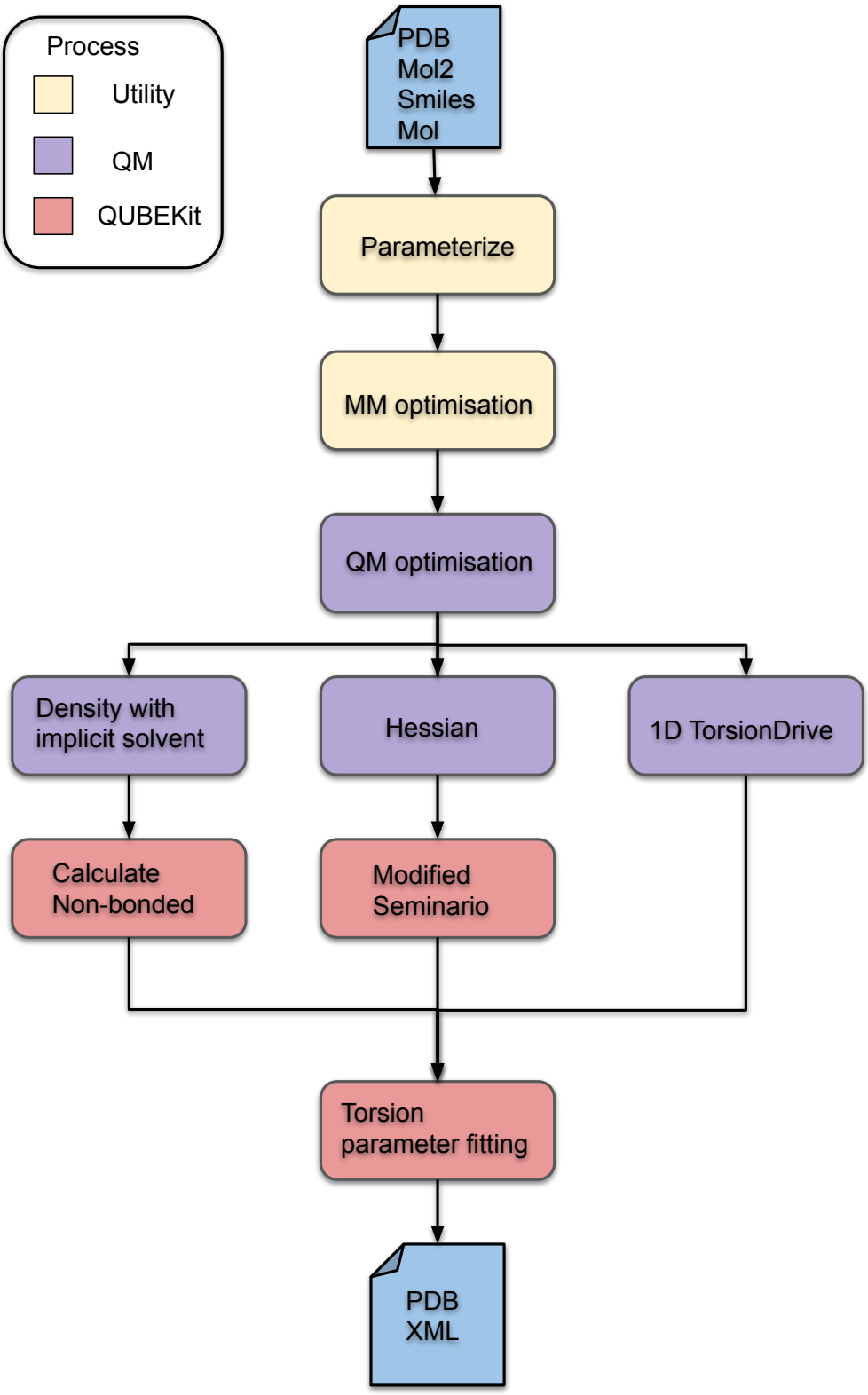

Figure 3.2: QUBEKit-v2 modularised flowchart highlighting each python class involved in the workflow and their primary function (utility, QM or QUBEKit).

```
result=client.query results(molecule=optimisation.final_molecule,
                            driver="hessian")[0]
hessian = result.return_result

# Reshape hessian
conversion = constants.HA_TO_KCAL_P_MOL / (constants.BOHR_TO_ANGS
** 2)
hessian = np.array(hessian).reshape(int(len(hessian) ** 0.5), -1)
* conversion

# Extract optimised structure
opt struct =
client.query_procedures(id=opt_record)[0].get_final_molecule()

# Initialise Ligand object using the json dict from qcengine
mol = Ligand(opt_struct.json_dict(), name='initial_test')

# Set the qm coords to the input coords from qcengine
mol.coords['qm'] = mol.coords['input']

# Insert hessian
mol.hessian = hessian

# Get Mod Sem angle and bond params
ModSeminario(mol).modified_seminario_method()
```

Figure 3.3: Part of an example work flow using the QUBEKit API to derive modified Seminario method predicted force constants for QM data extracted from a shared QM calculation database called QCArchive.

how to build a fully automated workflow using the QUBEKit API and adds extra functionality to remove some of the user intervention that was required in the first iteration to run and check the results of QM calculations. Now with proper error handling and checks we can identify exceptional situations during the execution of the QM calculation and try to take the appropriate action to minimise human intervention where possible. However, as we have now created a fully functioning python library, users are able to quickly build their own custom workflows from a few simple building blocks, which through layers of abstraction can handle complicated tasks in just a few lines of code.

In fact it is possible to link the actions required during parametrisation to a graphical user interface (GUI) to allow for a more intuitive interaction with the software, as has been done with other MM parametersiation tools based on QM calculations [55]. A prototype example of how this might look is shown in figure 3.4,

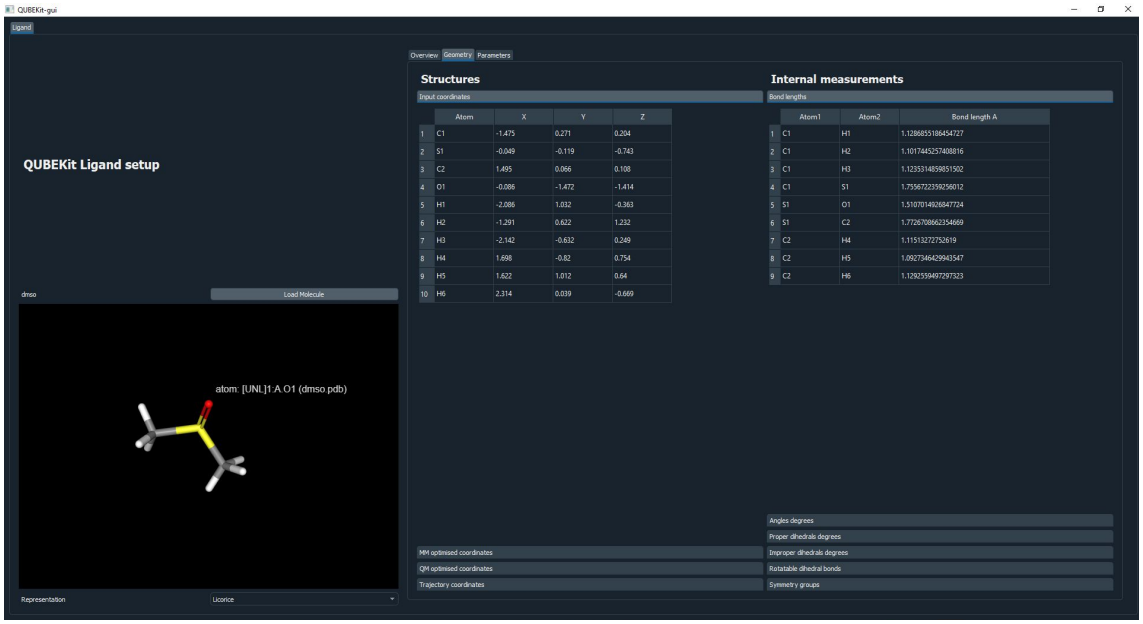

Figure 3.4: A prototype GUI that could be used to control the QUBEKit library.

where the loaded molecule can be viewed in an interactive viewing window powered by the NGL viewer [134], while internal measurements of bonds and angles are shown on the right along with parametrisation settings and steps. Future work could also be devoted to the development of this to further increase the usability of the software and give increased control when concerned with a specific parametrisation rather than a batch style automated use.

Furthermore by interfacing with software such as *QCengine*, which seeks to create a universal input format between quantum chemistry packages, we can vastly expand the range of underlying QM codes that can be used automatically with little effort. Similarly we now use RDKit to handle the initial input before loading the information into the internal data structure which allows for a wider range of input file formates including: SMILES, PDB, Mol/SDF and Mol2 which are all widely used within CADD. In fact all steps in the workflow shown in figure 3.2 that are labelled as utility or QM now have multiple running options corresponding to different software which can be used to carry out that specific task. This has also allowed us to integrate an OpenMM XML style version of the general QUBE protein bonded parameter library [60] into QUBEKit. QUBEKit can then be used to parametrise a protein with the QUBE bonded terms and can even calculate and apply non-bonded terms to the system from a ONETEP output file using the same method as applied to small molecules. The entire system can then be simulated in

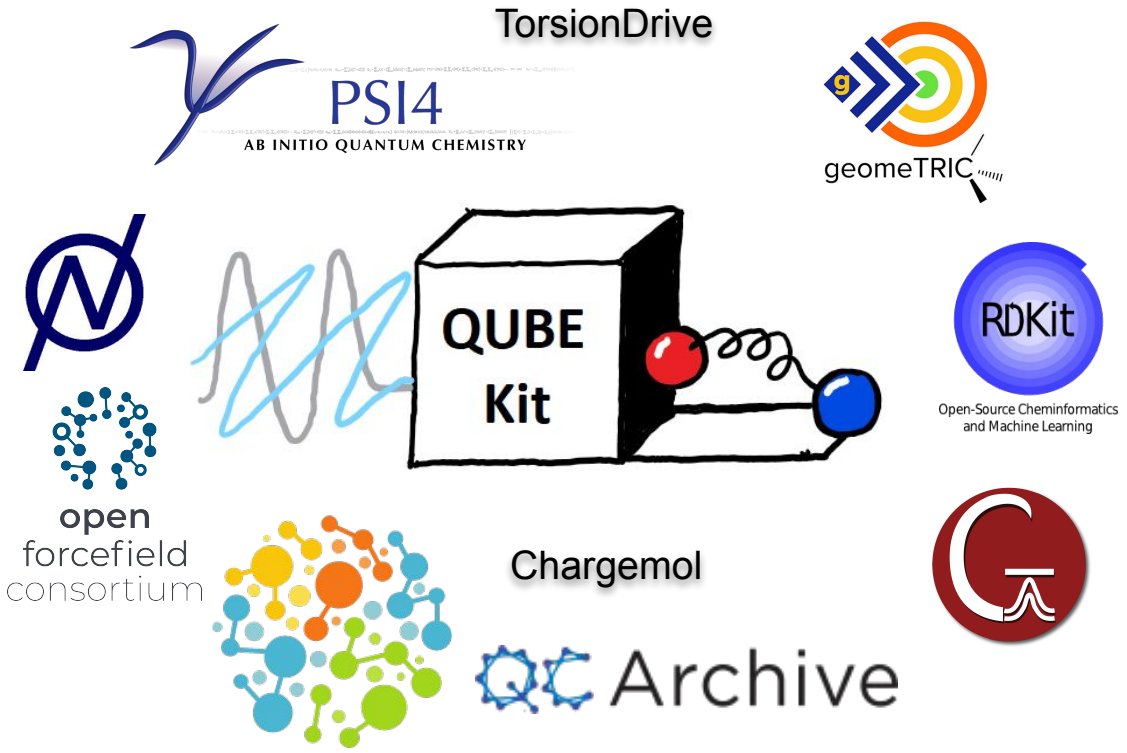

Figure 3.5: The QUBEKit ecosystem is shown to demonstrate the wide range of software with which it can interface.

OpenMM or further serialised into a specific XML style FF which can be converted for use in other software. Demonstration and application of this workflow to HIV reverse transcriptase is in progress with the University of Edinburgh.

Overall the work described in this chapter has vastly increased the size of the QUBEKit software ecosystem which is shown in figure 3.5. However, thanks to the use of python, which has become the community standard programming language, we can also easily distribute QUBEKit via the Python Package Index (PyPI) and Conda (open source package and environment management system) along with the majority of its dependencies making installation straightforward. This follows from a newly recognised need within the community to meet software development best practices which aim to ensure that software is reusable and maintainable. This also includes adhering to coding style recommendations such as PEP8, which aim to improve readability through the uniform appearance of code, unit testing and continuous integration which help ensure results are repeatable and updates are automatically rolled out. Following these best practices also allows for the easy extension of the code base by others, and as an example of this we have contributed

to the development of other open-source software used by the community to ensure long term compatibility with QUBEKit. This included adding the geometric non-bonded combining rules and custom convergence criteria to the geomeTRIC package (PR #83) as well as an interface to the Gaussian software package for TorsionDrive (PR #53) details of which can be found in the associated pull requests on github.

3.3 Conclusion

With the rise in popularity of MM simulations, automated workflow tools to aid the set up, parametrisation and execution of calculations have become central to their wide spread success and reproducible results. As such there is now a movement within the community towards professional software development practices for any critical software to ensure that reproducibility can be maintained for the long term. While this may add extra overhead during the development and implementation of any new tools, or in this case FF parametrisation methods, it can help aid adoption if users can easily install, use and recreate published results. Following this QUBEKit has gone through an intensive design cycle starting with a proof-of-concept implementation script to a full python library that follows these coding best practices. Furthermore QUBEKit is open-source and developed so users can see any changes and issues live which helps maintain transparency and allows others to easily extend the project. To aid the widespread use of QUBEKit, based on community feedback, we have also tried to use open-source dependencies where possible to lower the barrier to entry, resulting in a range of choices in software that can be used at most stages during parametrisation. This is possible due to the modularisation of the functions in the automated workflow which, as we have shown in figure 3.3, allows parts of the QUBEKit library to be used in other external pipelines. Overall we have created an extensible QM parameter derivation library that currently contains all of the methods need to derive the QUBE FF [61]. In future we envision that this could easily be extended to add other parametrisation techniques such as internal Hessian fitting, so users can easily and routinely build their own bespoke FFs from QM through combinations of different parameter derivation methods. Future work should also aim to add open-source alternatives at every stage of the pipeline. Currently the specific implicit solvent model and virtual site derivation method devloped for QUBEKit [61] are only available in ONETEP. This would require extensive testing of other implicit solvent models and the extraction of the virtual

site derivation method into an external package.

During this work will employ QUBEKit at various stages during its development meaning that the original script is used for the studies presented in chapters 4 and 5, before switching the QUBEKit-V2 in chapter 6 to demonstrate its extension and use.

Chapter 4

Benchmarking the QUBE FF

4.1 Introduction

A common measure of the quality of FF parameters for use in biomolecular simulations is a comparison of the predicted condensed phase properties of molecules simulated using the FF with experiment. These properties, such as liquid density, the heat of vaporization and free energy of hydration, can be calculated routinely due to low sampling requirements, thus making FF inaccuracies the main contributor to any differences between the computed data and experiment. Therefore we have chosen a benchmark dataset comprising 109 small organic molecules, which are representative of the key functional groups commonly observed in biology and drug design. This then necessitated the derivation of a new R_{free} fitting term to include bromine into the covered elements of the QUBE FF, which increases the potential application range. Importantly most of the molecules used in the set are also part of the training data used during the parametrisation of many of the general transferable FFs mentioned, including the OPLS/1.14*CM1A-LBCC FF (see chapter 2.3.2), which allows for direct comparison of the FFs. In this section we start by outlining the optimisation of the new fitting parameter before taking a detailed examination of the FF parameters and consequently properties predicted by the QUBE FF for our benchmark set. We also compare to other transferable FFs wherever possible to assess the overall level of accuracy achievable with the FF.

OPLS bond type	k_r (kcal/mol/$\mathring{A}^2$)	equilibrium bond length ($\mathring{A}$)
CT-CT	212.4 / 214.9	1.525 / 1.527
CT-HC	312.0 / 315.2	1.097 / 1.097
CT-NT	248.9 / 240.1	1.453 / 1.461
NT-H	439.1 / 428.6	1.014 / 1.018

Table 4.1: Comparison of the Modified Seminario Method derived bond stretching parameters and DFT/MP2 predicted equilibrium bond lengths for N-butyl-1-butanamine.

OPLS angle type	k_θ (kcal/mol/rad^2)	equilibrium angle (degrees)
CT-CT-CT	115.8 / 96.8	112.9 / 112.7
CT-CT-HC	48.6 / 47.8	109.8 / 109.7
HC-CT-HC	32.7 / 33.4	107.0 / 107.4
CT-CT-NT	126.9 / 116.4	111.6 / 111.2
NT-CT-HC	55.0 / 57.0	110.0 / 109.9
CT-NT-CT	119.4 / 121.4	113.6 / 112.7
CT-NT-H	47.3 / 48.0	109.4 / 108.4

Table 4.2: Comparison of the Modified Seminario Method derived angle bending parameters and DFT/MP2 predicted equilibrium angles for N-butyl-1-butanamine.

4.2 Computational details

Quantum Mechanical Calculations

All Gaussian09 input files were prepared using QUBEKit, which takes PDB files and the corresponding BOSS/MCPRO style z-matrices generated using the LigParGen web server as input. All optimization routines and frequency calculations used for the bond stretching and angle bending terms were performed with the ωB97X-D [1] functional using the 6-311++G(d,p) basis set and a vibrational scaling factor of 0.957 [45]. Users of QUBEKit are free to choose their own QM methods based on required accuracy and computational expense. For comparison, Tables 4.1 and 4.2 show the derived bond and angle parameters of N-butyl-1-butanamine computed using ωB97X-D/6-311++G(d,p) and MP2/6-311++G(d,p).

Torsional constrained optimizations were performed in Gaussian09 [2] with the

same functional and basis set so as to be consistent with the other bonded terms. The torsional scan optimizations were performed in 15° increments from 0° to 360°. The majority of the dihedral parameter fitting was done using no Boltzmann weighting (corresponding to T=∞) and regularization against OPLS reference values was applied with $\lambda = 0.1$, see equation 2.41. This was only changed in rare cases where it was particularly difficult to recreate the QM energy landscape, in which case $\lambda = 0$ and $T = 2000K$ were used as previously suggested [84].

Ground-state electron density calculations for non-bonded parameter derivation were performed using the linear-scaling DFT code ONETEP [135]. Four nonorthogonal generalized Wannier functions (NGWFs), with radii of 10 Bohr, were used for all atoms with the exception of hydrogen, which used one. NGWFs were expanded in a periodic sine (psinc) basis, with a grid size ($0.45a_o$), corresponding to a plane wave cut-off energy of 1020 eV. The PBE exchange-correlation functional was used with PBE OPIUM norm-conserving pseudopotentials [136]. The calculation was carried out in an implicit solvent using a dielectric of 4 to model induction effects [137, 59, 77]. The DDEC module implemented in ONETEP was used to partition the electron density and assign atom-centered point charges and atomic volumes [138, 58]. The charges were assigned with a IH to ISA ratio of 0.02. The ESP error threshold, F_{thresh}, was set to 0.9025 kcal/mol. The additional charges are only added if the decrease in ESP error is larger than $F_{change} = 0.0625$ kcal/mol. The locations of the virtual sites were restricted using maximum distance cut-offs chosen by element, as virtual sites near the van der Waals radius can be detrimental. The cut-offs were defined as follows: 0.8 Å for N, 1 Å for O, S and F, and 1.5 Å for Cl and Br.

Pure Liquid Simulations

Pure liquid simulations were performed using OpenMM [63] with a custom non-bonded potential to describe the mixing rules and 1-4 interactions employed by the OPLS (and QUBE) FF. The required .xml files were generated using QUBEKit with extra sites included automatically by QUBEKit using the local coordinate site construction function in OpenMM. All extra sites were modelled as virtual particles, and do not contribute bond and angle force field terms. For the construction of neighbor lists for 1-4 interactions, their only connection is made to the parent atom.

Simulations were performed in the isothermal-isobaric (NPT) ensemble at 1 atm and comprised 267 molecules in a periodic cubic box. Long-range electrostatic interactions were calculated using the Particle-Mesh-Ewald (PME) method [139],

Molecule	ρ (g/cm^3)	ΔH$_{vap}$ $(kcal/mol)$
dmso	1.111 / 1.112	12.563 / 12.644
N-methylaniline	0.97 / 0.971	11.841 / 12.035
chloroform	1.414 / 1.413	5.569 / 5.546

Table 4.3: Comparison of the pure liquid property predictions sensitivity to the choice of time step (1/0.5-fs) for a small set of molecules.

with a 0.0005 tolerance error while also applying a long-range correction to the system energy. As in previous studies [59, 97], non-bonded interactions were truncated at distances based on molecular size (15 Å for molecules with 5 or more heavy atoms, 13 Å for 3–5 and 11 Å for fewer than 3) and smoothed over the last 0.5 Å. No long-range corrections to the Lennard-Jones energy were applied. Following minimization of the initial configuration, 3 ns simulations were run for each molecule using a 1 fs time step. The first nanosecond was treated as equilibration. Data showing the insensitivity of the computed liquid data to the choice of time step are shown in Table 4.3. The liquid and corresponding gas-phase simulations were run at 25°C or the molecule's boiling point if it was lower. The resulting densities and heats of vaporization were averaged over 2000 data points collected in the production part of the run. The heats of vaporization were computed using eq 4.1 taken from Ref. 140.

$$\Delta H_{vap}(T) = E_{gas}^{potential}(T) - E_{liquid}^{potential}(T) + \frac{1}{2}R\Delta T(3N_{atoms} - 6 - N_{cons}) + RT \quad (4.1)$$

where $E_{gas}^{potential}$ and $E_{liquid}^{potential}$ are energies of the molecules in the gas and liquid phases respectively while ΔT is the difference between the simulated temperatures in the liquid and gas phases. N_{atoms} is the number of atoms in each of the molecules and N_{cons} is the number of restrained degrees of freedom which is zero in this case as the molecules are fully flexible during the simulations. Following their recommended protocol, we employed Langevin dynamics temperature regulation with a collision frequency of 5 ps^{-1}. The pressure was regulated using a Monte Carlo barostat as implemented in OpenMM. The uncertainties were found to be less than 0.003 g/cm^3 and 0.02 kcal/mol for densities and heats of vaporization respectively. Graphs showing the convergence of the properties with simulation time can also be found in Figures 4.1 and 4.2.

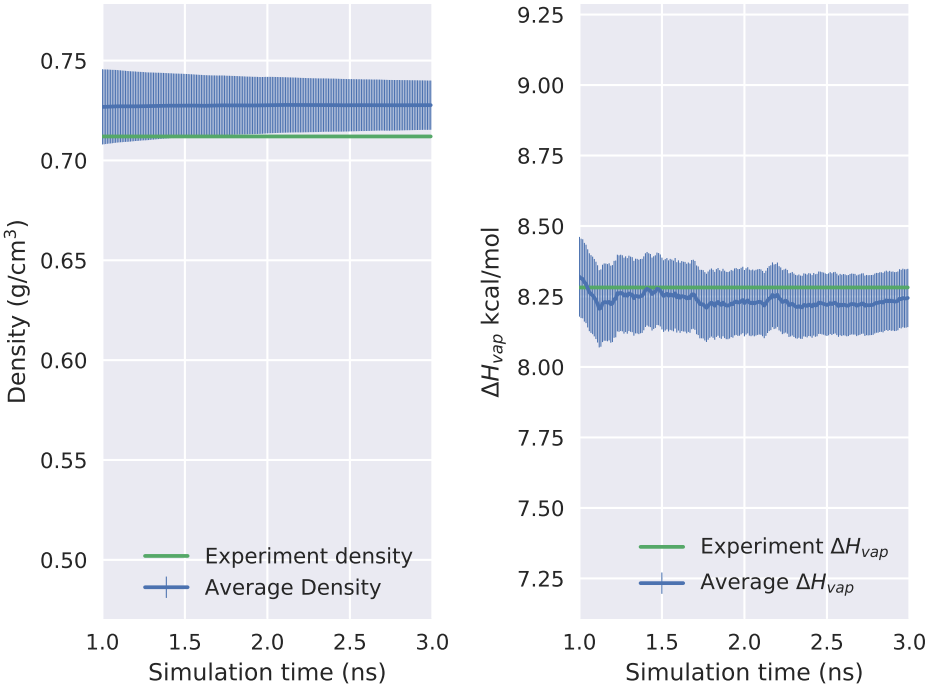

Figure 4.1: The convergence of the density and heat of vaporization is shown for N-(1-methylethyl)-2-propanamine over the course of the 2ns simulations used to measure each property. The experimental value is also shown along with the running average of the measured property and its standard deviation at each frame.

Free Energies of Hydration

Free energies of hydration were calculated using GROMACS [141] due to its ability to include extra sites during alchemical perturbation. All input files were generated using QUBEKit which writes OPLS FF style GROMACS .top and .gro files. The virtual sites were all constructed by hand using the simplest method available for each molecule, with a connection being added between the site and parent to again make the 1-4 interaction lists consistent with OpenMM and BOSS. Each molecule of the test set was annihilated from a cubic box containing approximately 1500 TIP4P water molecules using a two-step approach over 21 λ-windows, first turning off the charges followed by the L-J terms. The solute-solvent non-bonded interactions were switched off via coupling to the λ reaction parameter using soft-core potentials with settings $\alpha = 0.5$, $p = 1$ and $\sigma = 0.3$ [142]. The charges were decoupled using λ values of (0.00 0.25 0.50 0.75 1.00) and van der Waals using λ values of (0.00 0.05 0.10 0.20 0.30 0.40 0.50 0.55 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00). The

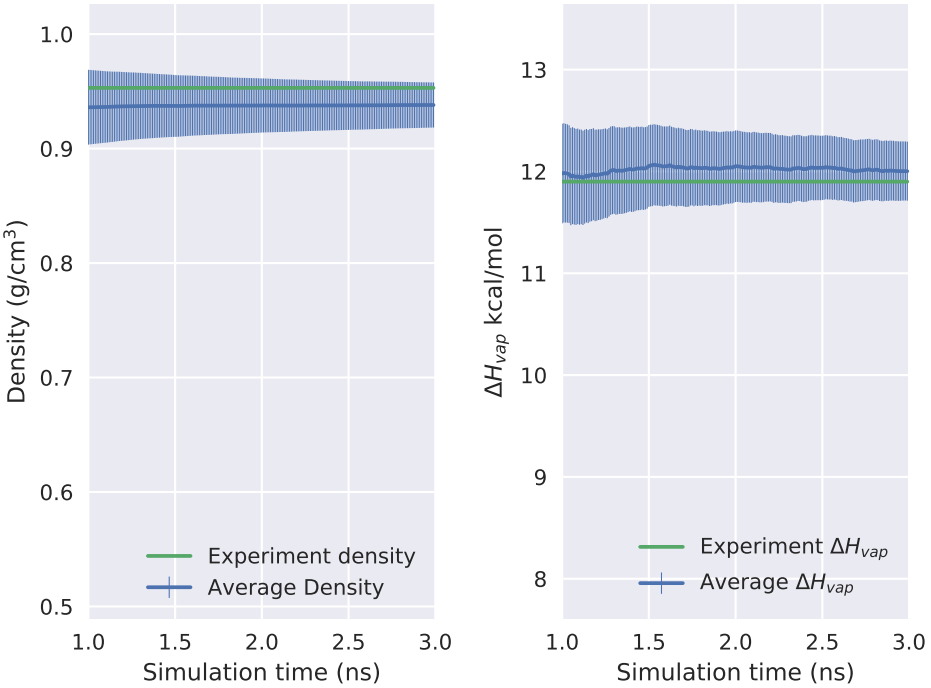

Figure 4.2: The convergence of the density and heat of vaporization is shown for N,N-dimethylaniline over the course of the 2ns simulations used to measure each property. The experimental value is also shown along with the running average of the measured property and its standard deviation at each frame.

simulations were again run in the NPT ensemble at 1 atm and 25°C. All solvent-solute and solvent-solvent non-bonded interactions were truncated at 10 Å and smoothed over the last 0.5 Å. PME was used with a long-range correction applied to the total energy and pressure. Each λ-window was run using Langevin dynamics and a two femtosecond time step with bonds involving hydrogen constrained using the LINCS algorithm [143]. The starting configurations at each λ-window were first minimized before being equilibrated twice. The first was a 100 ps run in the canonical ensemble (NVT) followed by a 200 ps run in the NPT ensemble. Finally, the production stage was run for 1 ns and the free energy of hydration was calculated using Bennett's acceptance ratio as implemented in the GROMACS BAR module [144]. All uncertainties for the calculations were found to be less than 0.3 kcal/mol.

4.3 Results and Discussion

Extending the QUBE FF

The extension of the QUBE FF is quite a trivial task due to its limited number of interdependent fitted parameters (one for each element). Here we aim to develop an R_{free} parameter for bromine following the same method as that used to parameterise the other elements in the model which so far includes H, C, N, O, S, F and Cl [59]. The R_{free} term, as described earlier, controls the scaling of the partitioned ground state electron density used to calculate the short-ranged repulsive interactions of the L-J potential shown in equation 2.24. Pure liquid property simulations are then a perfect target to guide the optimisation of these parameters as their sampling requirements are low, allowing the identification of FF inaccuracy specifically. Thus from the literature benchmark data, we created two sets of bromine-containing molecules to act as the fitting and testing data. For the three molecules (bromobenzene, 1,2-dibromoethane, bromoethane) of the fitting set, we then calculated their predicted density and heat of vapourisation at a range of R_{free} values to identify that which minimised the MUE as shown in Table 4.4. Here we found that a value of 1.96 Å gave a good compromise in accuracy between the two metrics, note that the heat of vapourisation was included in this fitting due to the abnormally large errors introduced when only fitting to the density. The complete set of parameters for the elements included in this benchmark can be found in Table 4.5.

One point to consider about this optimisation strategy is that the parameters derived here are dependent on those of the existing model. This then creates some degree of interdependency between them. A more robust and accurate optimisation may be achievable via the co-optimisation of all of the parameters of the model simultaneously. On the other hand, it is not obvious that we would arrive at a substantially different level of accuracy although this would need future investigation. Instead, we limit our self to the simple FF extension described above in this case.

ρ (g/cm^3)						
Molecule	R_{free}=2.02	R_{free}=2.00	R_{free}=1.98	R_{free}=1.96	R_{free}=1.95	Experimental
bromobenzene	1.455	1.486	1.482	1.495	1.501	N/A
1,2-dibromoethane	2.211	2.272	2.328	2.377	2.483	2.169
bromoethane	1.45	1.475	1.500	1.532	1.546	1.449
MUE	0.0215	0.0645	0.105	0.1455	0.2055	
ΔH_{vap} $(kcal/mol)$						
bromobenzene	8.601	8.841	9.026	9.124	9.195	10.65
1,2-dibromoethane	7.736	8.14	8.518	8.899	11.547	9.974
bromoethane	6.015	6.094	6.328	6.496	6.613	6.601
MUE	1.6243	1.3833	1.1177	0.902	1.0133	

Table 4.4: Bromine R_{free} fitting data.

Element	V^{free}(Bohr^3)	B^{free}(Ha.Bohr^6)	R^{free}
H	7.6	6.5	1.64
C	34.4	46.6	2.08
N	25.9	24.2	1.72
O	22.1	15.6	1.60
F	18.2	9.5	1.58
S	75.2	134	2.00
Cl	65.1	94.6	1.88
Br	95.7	162.0	1.96

Table 4.5: The free atom data used with the TS method to derive all L-J terms. The V^{free} term was calculated using the MP4(SDQ)/aug-cc- pVQZ method in Gaussian09[2] and the chargemol[5] code. B^{free} was taken from ref 6 with R^{free} being fit to experimental densities and heats of vaporization.

4.3.1 Condensed Phase Properties

Figure 4.3 shows the results of the condensed phase property calculations for the test set where experimental data are available, along with the correlations and MUE, while Table 4.6 compares the latter with some examples of widely-used transferable FFs for the same test set [97, 19, 59]. The average errors in the density and heat

Force field	ρ (g/cm^3)	ΔH_{vap} $(kcal/mol)$	ΔG_{hyd} $(kcal/mol)$
OPLS/1.14*CM1A [97]	0.024	1.40	1.26
GAFF/AM1-BCC [97]	0.039	1.31	0.94
OPLS/CM5 [97]	0.024	1.06	0.94
OPLS/1.14*CM1A-LBCC [97]	0.024	1.40	0.61
DDEC/OPLS [59]	0.014	0.65	1.03
QUBE (this work)	0.024	0.79	1.17

Table 4.6: Mean unsigned errors between calculated liquid properties and experiment for various FFs. Note that different parameter sets were also used in each of the benchmarks.

of vaporization (0.024 g/cm^3 and 0.79 kcal/mol, respectively) indicate that QUBE performs extremely well in the prediction of pure liquid properties, that is despite only using eight fitting parameters in the derivation of non-bonded parameters (the van der Waals radii of the elements H, C, N, O, S, F, Cl, Br used in this benchmark). Table 4.7 lists the thirteen molecules used for fitting and shows that removing them from the validation set has negligible effect on the analysis. Some of the outliers in the heat of vaporization predictions include interactions between aromatic rings,

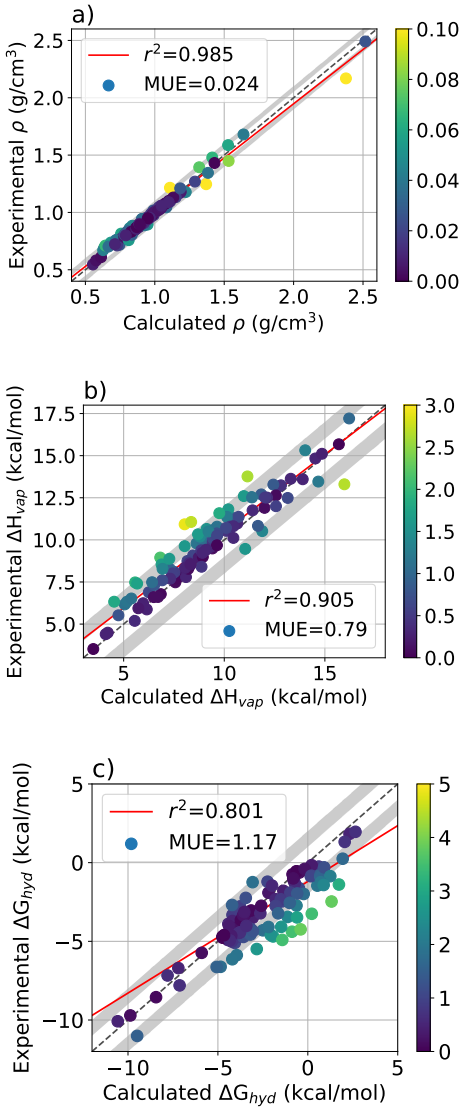

Figure 4.3: Force field liquid property metrics (a) liquid density, (b) heat of vaporization (c) free energy of hydration. Calculated for the organic molecule test set using QUBE FF parameters. MUE compared to experiment and r^2 correlation are also included.

which may be due to the difficulty of describing van der Waals interactions using a simple r^{-6} interaction, which neglects higher-order dispersion and many-body effects. The general transferable force fields are of similar accuracy to QUBE, despite being extensively parametrized against data sets similar to these.

As has been found previously [59], hydration free energies are more difficult to predict (MUE 1.17 kcal/mol). This could be due to limitations in the functional form, particularly the neglect of an explicit polarization term, in describing the transfer of a molecule between low dielectric (vacuum) and high dielectric (water)

media, or the mixing rules used to compute L-J interactions. Though it should be noted that a MUE of 0.72 kcal/mol is reported using the OPLS3 FF on an expanded 239 molecule test set, which indicates that there is room for further improvement within the current FF functional form [19]. The largest outliers in Figure 4.3(c) are for apolar molecules with a low (less negative) free energy of hydration, for which QUBE under-estimates their solubility. This is particularly problematic again for molecules containing aromatic rings, and may indicate an imbalance between dispersive and electrostatic contributions to hydration when QUBE is used in combination with a standard transferable water model (TIP4P).

Another potentially problematic group of compounds are aliphatic alcohols as we found the ten in our test set to have a relatively high MUE (1.27 kcal/mol) in hydration free energy. The poor description of alcohol groups was also previously found to be a trait of the OPLS/CM1A FF [97, 145]. The charges assigned to the head group of 1-octanol by OPLS/CM1A are shown in Table 4.8. It has been suggested that scaled CM1A charges are too positive, resulting in the poor prediction of densities and heats of vaporization as shown in Table 4.9 [145]. To tackle problematic groups such as these, the OPLS/1.14*CM1A-LBCC parametrization was developed which adds a systematic bond charge correction to various functional groups and was fit to better reproduce experimental free energies of hydration [97]. In the case of the aliphatic alcohols, the correction transfers a $0.1e^{-}$ charge to the oxygen of the head group from the neighboring carbon atom as can be seen in Table 4.8. Thus with the same L-J parameters, the density, heat of vaporization and free energy of hydration are subsequently improved for 1-octanol, as shown in Table 4.9 along with the values obtained by the QUBE FF. This same BCC was also found to reduce the MUE for the hydration free energy from 1.95 to 0.43 kcal/mol for 32 aliphatic alcohols in the development of the LBCC parameters[97]. Importantly the fitted correction scheme gives roughly the same charge as our AIM partitioning method which demonstrates the successful inclusion of polarization into our charges

Force field	ρ (g/cm^3)	$\Delta \mathrm{H}_{vap}$ $(kcal/mol)$	$\Delta \mathrm{G}_{hyd}$ $(kcal/mol)$
QUBE(this work)	0.024	0.83	1.19

Table 4.7: The calculated FF accuracy metrics adjusted for the training set data is shown. Training set molecules: Ethane, benzene, acetone, methanol, acetamide, chlorobenzene, dimethylsulfide, methanethiol, fluorobenzene, trifluorobenzene, bromobenzene, 1-2-dibromoethane, bromoethane.

Force field	charge	σ	ϵ
OPLS/1.14*CM1A	-0.588	3.120	0.170
OPLS/1.14*CM1A-LBCC	-0.687	3.120	0.170
GAFF/AM1-BCC	-0.598	1.721	0.210
QUBE	-0.673	3.129	0.127

Table 4.8: The non-bonded parameters for the head group oxygen in 1-octanol are shown for a variety of FF and charge combinations. The LigParGen server was used to parameterize the OPLS variants, and Antechamber for GAFF with QUBE coming from this work.

Force field	ρ (g/cm^3)	ΔH$_{vap}$ $(kcal/mol)$	ΔG$_{hyd}$ $(kcal/mol)$
OPLS/1.14*CM1A	0.807	15.201	-1.26
OPLS/1.14*CM1A-LBCC	0.809	16.038	-3.12
GAFF/AM1-BCC	0.834	20.354	-3.12
QUBE	0.793	16.206	-2.19
Experiment [145, 7]	0.822	17.208	-4.09

Table 4.9: The liquid properties of 1-octanol predicted using different FF and charge parametrization methods are displayed and compared with experiment.

at the point of derivation rather than via subsequent corrections. We also observe similar σ parameters between the QUBE FF and OPLS, which is reassuring considering OPLS is extensively fit to reproduce liquid properties. While the ϵ values do differ noticeably, it has been found that liquid property predictions can be greatly improved with the systematic tuning of this parameter [140]. However, this would not be compatible with the philosophy of a QM derived FF, and future work will instead investigate modifications to the FF functional form.

Finally, it should be noted that there is an increase in the MUE of each of the properties computed using the QUBE FF compared with an original benchmark study (Table 4.6), which used AIM-derived non-bonded parameters in combination with OPLS bonded parameters (DDEC/OPLS) [59]. This is likely the result of the expanded test set used here as on further inspection of the data concerning only the same molecules that were included in the original benchmark we find the MUEs to be 0.017 g/cm^3, 0.59 kcal/mol and 1.08 kcal/mol for the density, heat

of vaporization and free energy of hydration respectively, which are very similar to the original values. This is promising considering the original properties were computed using MC simulations of the liquids under the same conditions for which the R_{free} values were fit. This demonstrates transferability in the parameters between computational protocols, however accuracy maybe slightly improved with the refitting of the parameters specifically for each of the protocols. Overall we conclude that bonded parameters, while crucial to the conformational preferences of larger molecules, are not too important in the description of the liquid properties of small molecules.

With the inclusion of larger molecules and molecules that contain multiple functional groups, the increase in overall error of the liquid properties is to be expected if we consider the accuracy on a per functional group basis. This effect is exemplified by the case of o-chloroaniline, which has unsigned errors in ΔH_{vap} of 2.61 kcal/mol and in ΔG_{hyd} of 3.49 kcal/mol. By way of comparison, the smaller molecules aniline and chlorobenzene showed unsigned errors in ΔH_{vap} of 1.63 and 1.17 kcal/mol and in ΔG_{hyd} of 2.66 and 1.67 kcal/mol, respectively. This should be kept in mind when applying QUBE (and other force fields) to the study of, for example, absolute protein-ligand binding free energies for larger organic molecules containing multiple functional groups.

4.3.2 Bond, Angle and Dihedral Parameters

As discussed in the previous section, it appears that the bonded parameters have little effect on the accuracy of liquid properties. However, given the importance of torsional parameters in determining conformational preferences of larger molecules, and bond and angle parameters in modelling molecular vibrations, which are important for example in photochemistry applications, we examine the properties of the derived parameters here in more detail.

The first point to note is that by deriving bond and angle parameters directly from the QM Hessian matrix, there is no possibility of missing parameters in the QUBE FF. In contrast, even for this small test set, we found one missing bond parameter and six missing angle parameters using a standard transferable FF. The QUBE predicted values for these terms along with the OPLS atom types are shown in Tables 8.1 and 8.2 of the appendix. In practice, these parameters would be inferred from similar atom types or re-parameterized by the user, which may introduce

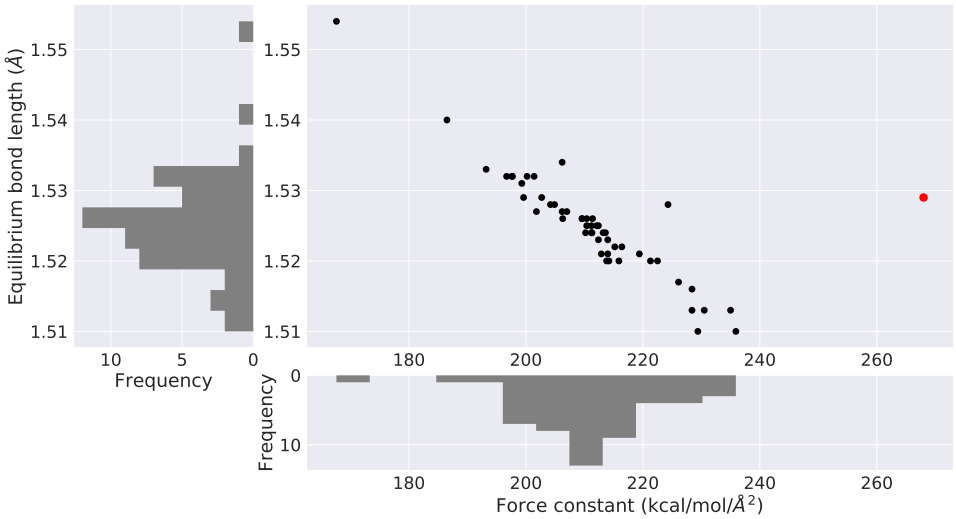

Figure 4.4: A common bond type is analyzed by comparing the QM predicted equilibrium bond length to the associated derived force constant of each molecule they appear in for the CT-CT bond type. The OPLS parameters are shown in red.

inaccuracy. QUBE allows the user to rapidly and automatically derive all necessary parameters with no compromise in accuracy. In this benchmark, the QUBE FF maintains a low mean percentage error in MM vibrational frequencies of 6.5% (MUE of 54 cm^{-1}), which is very similar to the values initially reported reaffirming the wide-scale applicability of the method [45]. We note that the modified Seminario method derives the force constants directly from the QM Hessian matrix with no information required about the torsional and non-bonded parameters. In practice, these components of the FF will also contribute to molecular vibrations. It appears that slight improvements in accuracy are achievable by fitting the full MM Hessian matrix to the QM Hessian [55, 41]. For example, a MUE of 44 cm^{-1} is reported using the QMDFF on a set of 22 molecules [41]. Where high accuracy in molecular vibrations is key, for example in spectroscopic applications, it may be desirable to include coupling FF terms which account for off-diagonal terms in the Hessian matrix [146]. However, for our intended applications in computer-aided drug design, we favor the relative simplicity of the modified Seminario method.

Given the widespread use of transferable bond and angle parameters, it is worth analyzing to what extent these parameters vary in our benchmark test set. Figure 4.4 plots the range of QUBE bond lengths and force constants for all atoms defined with

CT-CT bond types in the test set, and compares them with the OPLS parameters. Further plots like this for all bonds and angles that are present in at least ten of the molecules in the test set can be found in the appendix Figures 8.1-8.22. As reported previously [45], the modified Seminario method gives bond-stretching force constants that are on average lower than their OPLS counterpart. The QUBE parameters typically span a range of around 0.05 Å and 100 kcal/mol/Å^2 for the bond length and force constant respectively, indicating that use of a single average, transferable value should not introduce significant error. Interestingly, there is a negative correlation between force constant and equilibrium bond length, supporting the use of bond length to infer force constants in early studies [147]. These results indicate that it may be possible to derive more explicit algorithms for 'learning' force field parameters directly from the molecular geometry. We also envisage QUBE parameters as providing a reasonable starting point for optimization if further fitting to QM potential energy surfaces is desired [20].

Torsional parameters, like the bond and angle parameters, were derived separately for each molecule. Due to the use of virtual sites, we found that parameters were often not transferable between similar molecules, and those that were such as methyl group rotations remained close to the initial OPLS parameters. The overall accuracy of the torsional scan fitting was very good when regularization was used and only a handful of molecules with poor predicted energy surfaces required the setting to be switched off. A sample of torsion fitting data taken directly from the QUBEKit output is shown in Figure 4.5, along with the overall error and regularization error bias where appropriate. Other examples can also be found in the appendix Figures 8.23-8.24.

4.3.3 Extra sites

To test the effect of the additional off-center point charges, the liquid properties for the benchmark test set were also calculated in the absence of extra sites. This led to a general worsening of the results with the MUEs becoming 0.023 g/cm^3, 0.85 kcal/mol and 1.51 kcal/mol in the density, the heat of vaporization and free energy of hydration respectively (Figure 8.25). As expected, since it is governed mostly by Lennard-Jones interactions, the error in the density remained approximately constant. However, the decline in accuracy of the other properties indicates that modelling of anisotropy in electron density is required to accurately describe intermolecular

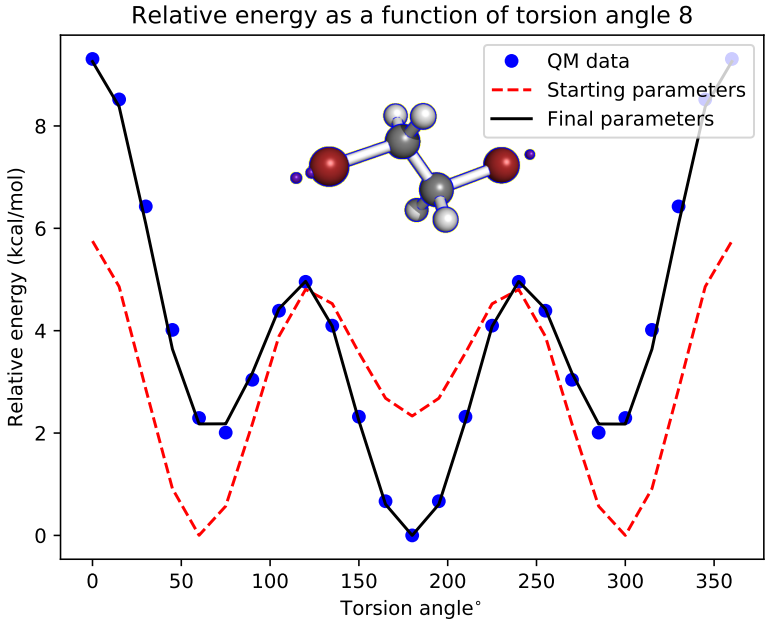

Figure 4.5: The QUBEKit generated torsional scan is shown for 1,2-dibromoethane. Where the QM data is calculated using the ωB97X-D[1] DFT functional and 6-311++G(d,p) basis set in Gaussian09 [2], the starting parameters are taken from OPLS and the final parameters are found using QUBEKit. Final error = 0.893 kcal/mol, bias = 0.797 kcal/mol.

interactions. This is consistent with the increasing use of virtual sites in multiple FFs [19, 148].

While there is no unique way to derive virtual site parameters, it would seem that deriving the parameters to minimize the error in ESP for an individual atom is effective. Figure 4.7 compares the ESP error around atoms before and after the addition of virtual sites. While some residual error is to be expected given the simplicity of the FF functional form, the errors on these atoms displaying highly anisotropic electron density is now much closer to, and in many cases below, the average ESP error across every atom in the benchmark set. Figure 4.6 shows a selection of molecules from the test set that required virtual sites. Here we can see that the derived positions are chemically intuitive, with σ-holes and lone-pairs well-represented. A more detailed analysis of the virtual site positions and charges is also shown for three molecules (morpholine, anisole and DMSO) of the test set in Figure 4.8. Interestingly we see two different virtual site positions identified as lowering the ESP of an oxygen atom in similar environments. In the case of morphline the site

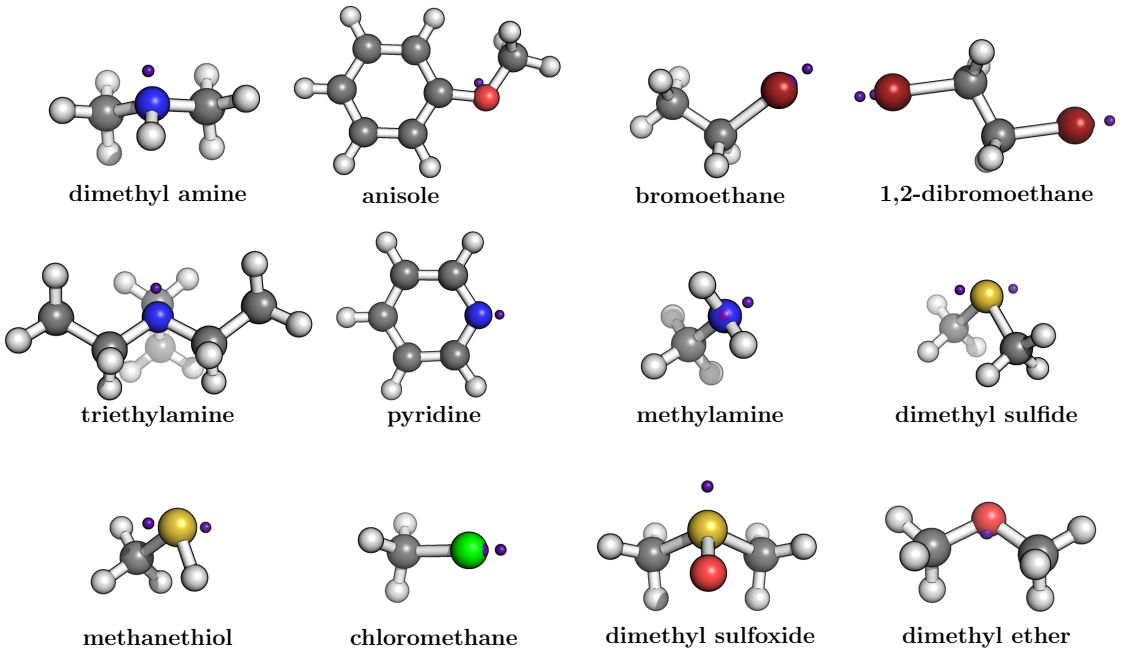

Figure 4.6: A selection of 12 molecules from the benchmark test set with their extra sites depicted as purple spheres. Charges and positions of the extra sites were derived from the partitioned atomic electron density.

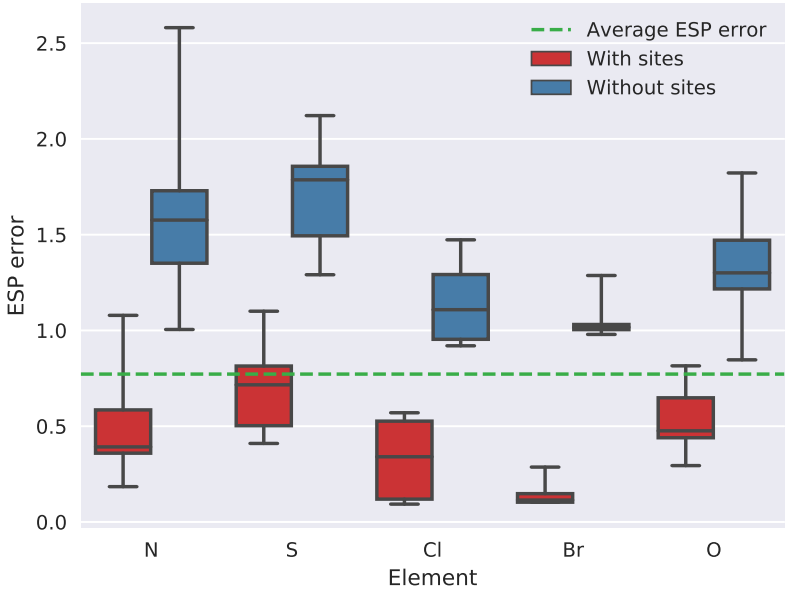

Figure 4.7: The average and range of the ESP error around each element for molecules in the test set before and after the addition of virtual sites. The dashed line represents the average error across all atoms in the benchmak set.

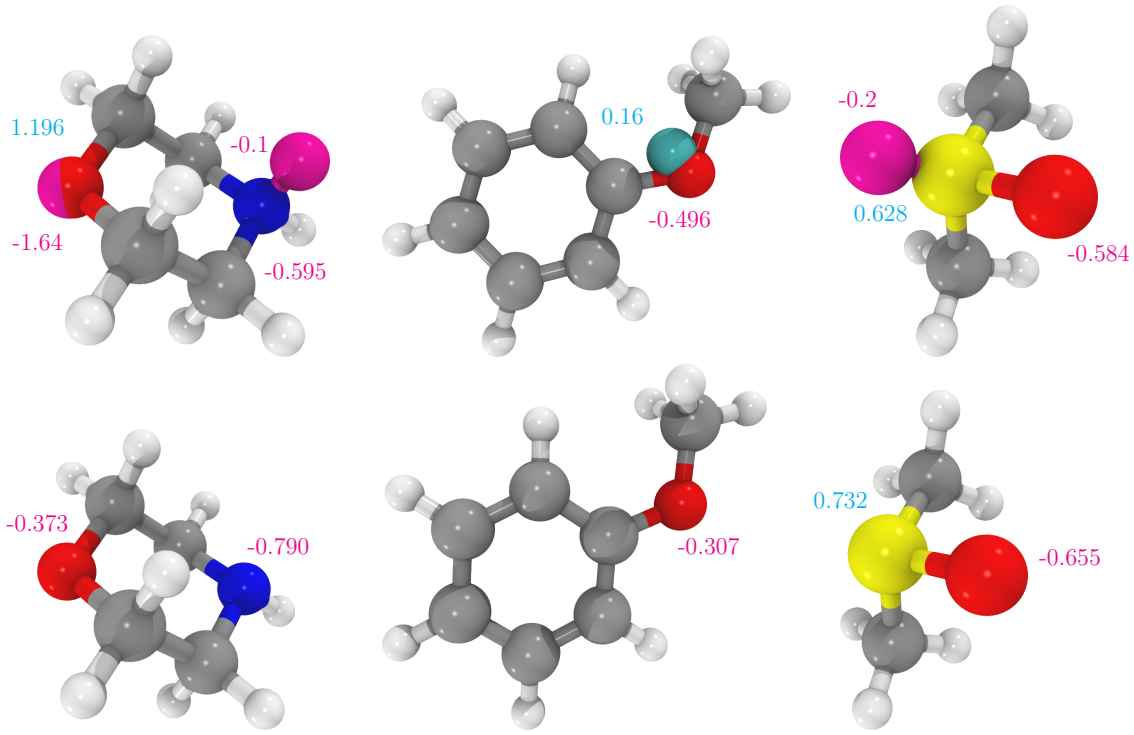

Figure 4.8: The virtual site positions and charges derived using QUBEKit for three molecules (morpholine, anisole and DMSO) are shown in comparison to the semi-empirical charges predicted using the 1.14*CM1A OPLS FF. Here all positive (negative) charges and virtual sites are shown in cyan (magenta).

is placed where we expect to find a lone pair with a large negative charge however, for anisole the site is placed opposing the lone pair position with a small positive charge resembling that of the TIP4P water model. Overall both oxygen atoms' ESP error is substantially reduced to very similar values, as shown in table 4.10, which are well within the error threshold, demonstrating that the positions are both valid. In total 50 of the 109 molecules in the test set required at least one virtual site, and on average a molecule whose functional group ESP error is initially above the chosen threshold requires 2.1 virtual sites. While this is more than is typical in molecular mechanics simulations, the computational cost of virtual sites in an MD simulation is small [104]. Furthermore, QUBEKit substantially simplifies the process for the user by deriving the virtual site parameters from QM and writing them to simulation-ready input files.

Some molecules with large ESP errors were not assigned off-center virtual sites. Chlorobenzene, for example, was found to have a large ESP error on the Cl atom just below the set threshold of 0.90 kcal/mol. However, the resulting liquid property

Molecule	ESP before	ESP after
Morphline (N)	1.607	0.906
Morphline (O)	1.350	0.692
Anisole (O)	1.101	0.688
DMSO (S)	2.005	0.871
Chloromethane (Cl)	0.978	0.318
Pyridine (N)	1.245	0.516

Table 4.10: Compares the ESP error of the DDEC fixed charges before and after the addition of virtual sites around the parent atom, which is shown in brackets.

predictions were not significantly affected (Table 8.3). Methanol was another example of a molecule that was not assigned virtual sites despite having an ESP error of 1.50 kcal/mol, which is above the threshold. After performing the grid search it was found that the addition of virtual sites did not substantially reduce the ESP error of the oxygen atom by the required amount F_{change}. This was the case for all aliphatic and aromatic alcohols in the test set which could also contribute to the poor performance of alcohols overall.

4.3.4 Test cases

While the molecules in the validation set represent many of the functional groups often used in drug design, they contain many fewer rotatable dihedral bonds and functional groups than a typical drug-like molecule. Thus, following previous work investigating the use of QM derived FF parameters we have used QUBEKit to derive a QUBE FF for 3-hydroxypropionic acid (3-HA) [110]. The molecule shown in Figure 4.10 incorporates carboxyl and hydroxyl functional groups, has been identified as a potentially useful agent for organic synthesis and is also a surrogate for a typical fragment scaffold. QM-based fitting techniques have previously been used to derive the bonded parameters for the molecule from a series of single point energy calculations, with the L-J terms being taken from AMBER and the partial charges assigned according to the CHelpG scheme [149]. In addition, we have selected two further molecules from the FreeSolv database [7], which allows us to compare computed hydration free energies with experiment for more challenging small drug-

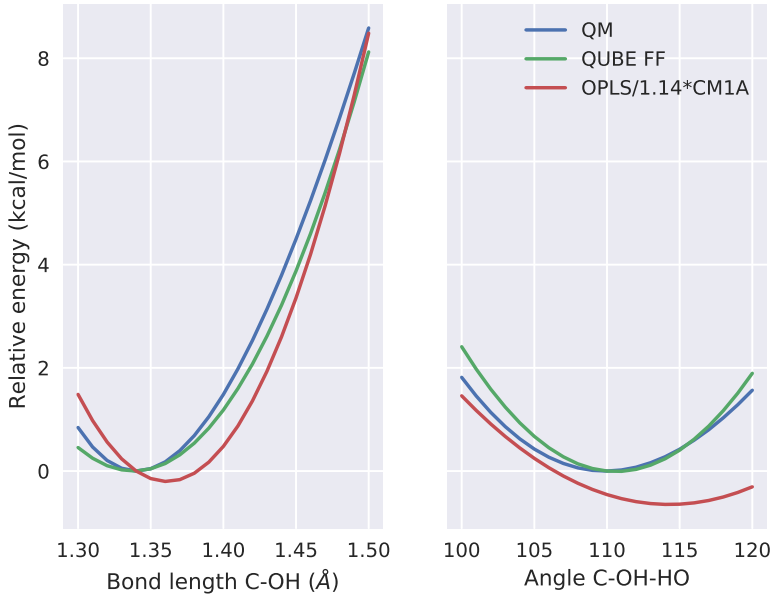

Figure 4.9: Comparison of the calculated relative single point energies using QM, OPLS and QUBE for C-OH bond-stretching and C-OH-HO angle-bending motions in 3-HA.

like molecules [4]. The two molecules, captan and bromacil (Figure 4.10), were selected due to the presence of halogens, and they therefore provide an additional test of the virtual site assignment procedure in QUBE.

Starting with the molecule 3-HA, Figure 4.9 compares the QUBE and OPLS force fields with QM single point calculations for a range of molecular geometries. Since we compute the bond and angle force constants in a one-off calculation directly from the QM Hessian matrix, with no iterative fitting, it is not obvious how accurate they will be in reproducing QM conformational energetics when combined with the rest of the QUBE FF parameters. However, Figure 4.9 reveals that the QUBE FF reproduces extremely well, not only the QM minimum energy conformations, but also describes small changes in these same bond lengths and angles. This is also well replicated across all calculated vibrational modes for the molecule with an average percentage error of 6.7% compared to the QM vibrational frequencies.

Next, with the goal of evaluating the ability of QUBE to recreate intramolecular energetics including torsional rotations, separate liquid simulations of 3-HA, captan and bromacil solvated in boxes containing 1000 TIP4P water molecules were performed. We then extracted 500 conformations from each simulation and computed the relative single point energies of each snapshot of the molecule using OPLS, QUBE and

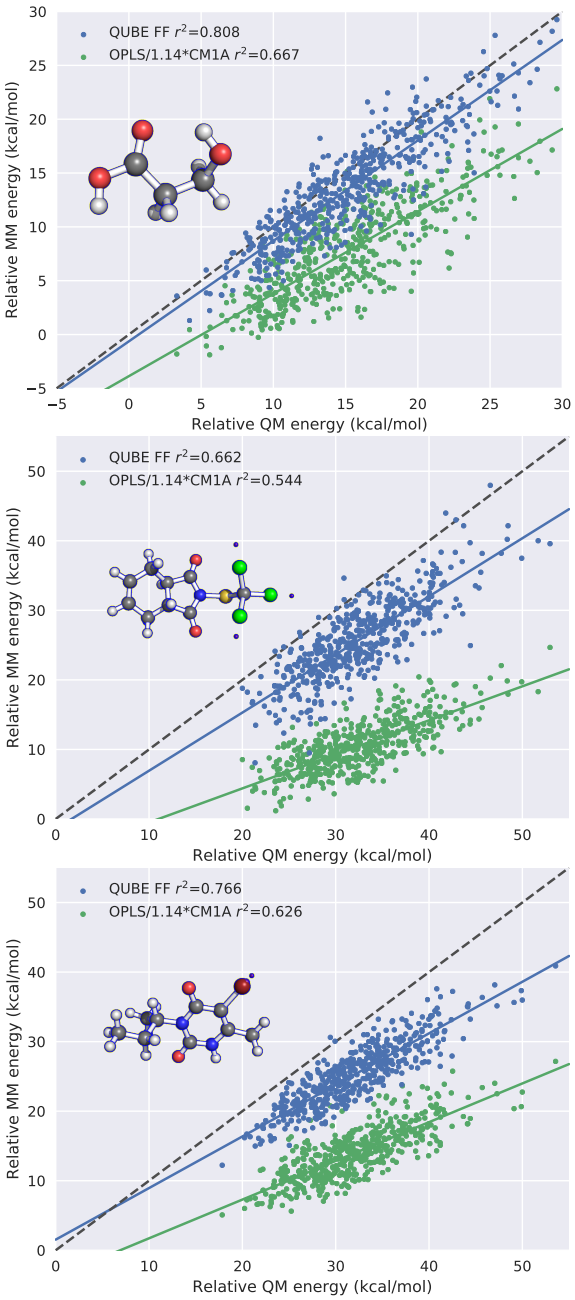

Figure 4.10: Comparison between the relative QM and MM energies using the QUBE FF and OPLS for 500 conformations extracted from a MD simulation of 3-HA (top), captan (center) and bromacil (bottom) which are shown as insets.

QM (with the same DFT functional and basis sets as used for the parameter derivation). Figure 4.10 shows the correlation between the relative MM and QM energies for each of the three molecules. We note in making this comparison that, unlike QUBE, the OPLS FF was not parametrized against this QM model chemistry. Compared to OPLS, the correlation between MM and QM energetics is improved, and significantly QUBE does not sample any configurations that are lower in energy than the optimized QM structures. Figure 4.11 shows in more detail the fitting of QUBE torsion parameters to QM potential energy scans, as well as the dihedral angles sampled during MM dynamics in water. Encouragingly,

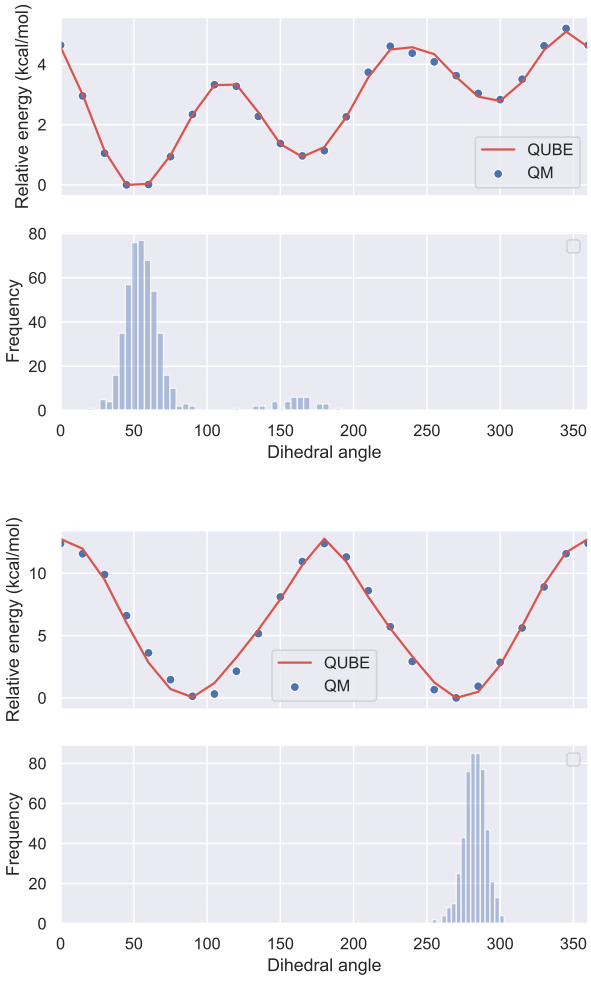

Figure 4.11: The gas phase QM and QUBE predicted potential energy surfaces during the dihedral fitting (top panel) are shown with the frequency of the dihedral angle sampled during the water simulation (bottom panel) for bromacil and captan respectively.

despite the simple MM functional form used, and the fact that it is optimized

	ΔG_{hyd} $(kcal/mol)$		
	QUBE	GAFF	Experiment
Captan	-5.48	-8.72	-9.01
Bromacil	-14.05	-14.50	-9.73

Table 4.11: The free energy of hydration predicted for two molecules from the FreeSolv database using the QUBE FF, compared to GAFF and experiment [7].

for reproducing condensed phase properties, QUBE is not only able to reproduce the minimum energy structures, but also sample physically reasonable structures in liquid simulations, which is encouraging for future use in computer-aided drug design.

Finally, the free energies of hydration of captan and bromacil were calculated using the same protocol described earlier, and the results are shown in Table 4.11 alongside the experimental data and those computed using a GAFF parametrization [7]. The errors of around 4 kcal/mol in the QUBE FF are higher than those reported for the small molecule benchmark set, but consistent with expected cumulative errors in hydration free energy prediction. Nevertheless, improvements in accuracy are required, particularly for hydration free energy calculations, if QUBE is to be used in predictive computer-aided drug design. Future strategies along these lines are discussed in the next two sections.

TIP4PD

The free energy of hydration is widely considered as the most important (simple) FF performance metric in regards to the application to computer-aided drug design due to its clear links to binding free energy calculations. Thus any improvements that can be made to the performance of QUBE in this respect are vital. It was noted above that generally the QUBE FF tended to underestimate the solubility of the molecules in the benchmark. Which could be related to the reduced dispersion parameters derived for the molecules compared to that of the iteratively fit OPLS values. It has been speculated that this would require the addition of higher-order terms in the L-J potential to correct this. While these parameters can be calculated from an AIM partitioned electron density [75, 150] the extension of the functional form is beyond the scope of this thesis. To this end, it was investigated whether an appropriate choice of water model could also account for this deficiency within

	TIP4P	TIP4PD
hydrogen charge	0.52	0.58
virtual charge	-1.04	-1.16
σ(Å)	3.15365	3.165
ϵ(kcal/mol)	0.155	0.223841
C_6(kcal/mol Å^6)	610	900
C_{12}(kcal/mol Å^{12})	600000	904657

Table 4.12: The FF parameters and estimates of the corresponding C_6 and C_{12} terms for the TIP4P [8] and TIP4PD [9] water models.

the current physics-based model. TIP4PD was selected as a viable candidate due to its accentuated dispersion parameter on the oxygen atom of the 4-site model as shown in table 4.12 which compares the parameters of the TIP4PD and TIP4P water models. The model was developed in response to the poor performance of standard water models when simulating intrinsically disordered proteins whose ensembles were found to be too compact [128]. The whole benchmark test set was then re-run under the same conditions as those described above with the TIP4PD water model to calculate the free energy of hydration, the results of which are shown in Figure 4.12. The use of the TIP4PD water model has improved the correlation

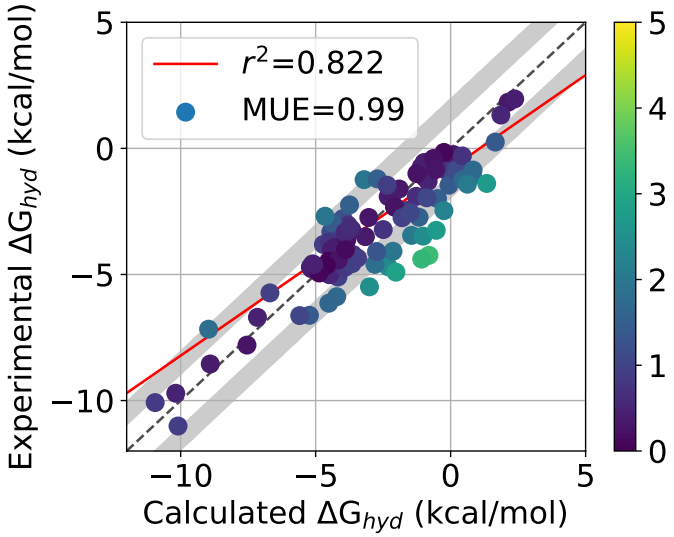

Figure 4.12: The free energy of hydration calculated for the organic molecule test set using QUBE FF parameters and the TIP4PD water model. MUE compared to experiment and r^2 correlation are also included.

	ΔG_{hyd} $(kcal/mol)$		
molecule	TIP4P	TIP4PD	Experiment
DMSO	-14.83	-12.43	-10.11
dimethyl sulfide	-0.5	-1.90	-1.61
thiophenol	-1.01	-1.66	-2.55

Table 4.13: The free energy of hydration predicted for two molecules from the FreeSolv database using the QUBE FF, compared to GAFF and experiment [7].

and MUE to 0.822 and 0.99 kcal/mol respectively, indicating that this water model does go some way to accounting for the underestimated dispersion terms of the QUBE FF within the standard functional form. Importantly this does not affect the other reported metrics in section 4.3.1 as they do not depend on the water model, but are pure liquid properties. Furthermore, on analysis of the results we recognise that most molecules see a systematic decrease in their predicted hydration free energy of between 0.2-0.6 kcal/mol. However, some specific groups show a more significant response which could be an indication that their dispersion is dramatically underestimated, and is detrimental to the accurate description of their interaction. In particular, sulphur-containing molecules, such as thiols and sulfoxides showed the biggest improvements with the unsigned error of DMSO reducing by 2.4 kcal/mol as shown in table 4.13. This is even more surprising when we consider that the free energy of hydration became more positive in the case of DMSO opposing the general trend in the response of the molecules to TIP4PD.

Within the current construction of the QUBE FF, the MUE of the predicted free energies of hydration can be improved via the use of the TIP4PD water model. While the model does show a balanced and accurate recreation of many pure water properties, it is not clear how compatible this model will be with the QUBE protein FF in terms of protein dynamics which would have to be validated. In theory, due to the derived nature of the QUBE parameters, they are compatible with any water model which is optimised using the OPLS non-bonded combination rules. Unlike general transferable FFs, which are often optimised against a specific water model, thus improved water models can be quickly utilised with the QUBE FF to improve its performance. Future work could also take advantage of this by co-optimising a water model specifically for use with the QUBE FF, correcting any deficiencies in the parameterisation process within the current functional form. The modular

construction of QUBE will enable rapid investigation of such water models and new functional forms.

4.4 Conclusions

With the spread of low-cost computing and access to automated software, it is becoming increasingly common for users to perform parameter set optimization prior to running molecular mechanics simulations. However, this optimization is typically used to supplement existing transferable force fields and is limited to the charge and torsional parameters, for which well-established protocols for fitting to QM data exist. On the other hand, QM derived force fields allow the user to obtain all (or most) of the force field parameters directly from *ab initio* calculations, but for these methods scaling to large molecules is problematic and there is no clear route to the simulation of, for example, biomolecular complexes. In this benchmark we have demonstrated how the QUBEKit software can be used in an automated fashion to derive virtually all force field parameters required to model the dynamics of small organic molecules.

Overall, we achieve mean unsigned errors of 0.024 g/cm^3, 0.79 kcal/mol and 1.17 kcal/mol in the prediction of liquid densities, heats of vaporization and free energies of hydration for a benchmark set of 109 molecules, compared to experiment. This accuracy is particularly impressive compared to standard, transferable force fields when considering heats of vaporization and liquid densities. While competitive with many transferable FFs, there is substantial room for improvement in the prediction of hydration free energies. This is particularly highlighted when comparing the QUBE data in Table 4.6 with OPLS3, or when considering the larger molecules, captan and bromacil, in Table 4.11. Importantly, however, we emphasize that to describe all molecules in the benchmark data set, we have *only fit 8* parameters (the van der Waals radii of eight elements in vacuum) to experimental data (Table 4.5). This reduction in empiricism has two key advantages. Firstly, it has the potential to substantially simplify the FF fitting process, since the parameters come directly from QM and do not rely on extensive collection of experimental fitting data, which is time-consuming for small molecules, and is rarely done for larger molecules. Secondly, the ease of FF design presents the opportunity to derive new protocols, and move beyond the standard functional form of the FF whilst retaining the ability to derive non-bonded parameters for large molecules.

Opportunities for FF improvement include: i) update of the atoms-in-molecule partitioning scheme [151–155], ii) the introduction of more rigorous descriptions of van der Waals interactions [156–158], iii) inclusion of explicit polarization, iv) a more accurate functional form for the short-range repulsion [156, 39], v) investigation of a QUBE-compatible water model [128] and vi) the investigation of Lennard-Jones combination rules. Such efforts would typically require significant re-fitting of the parameter libraries for transferable FFs. However, with the software infrastructure provided by QUBEKit, iterative improvements in the accuracy of the FF metrics presented here, particularly the hydration free energy, are envisaged.

One example of the update of our FF design protocol, is the addition of a new method for off-center virtual site parameter derivation for the modelling of anisotropic electron density. Compared to previous methods used in conjunction with the DDEC AIM partitioned charges [59], the parameter derivation process is faster and more user-friendly. By deriving the virtual site charges and positions from the molecular symmetry and partitioned atomic electron density, we do not require any experimental data for fitting. Furthermore, since the bond, angle and Lennard-Jones parameter derivation methods are independent of the charge derivation, we can trivially add extra sites without substantially altering the force field. Notably, the mean unsigned error in the free energies of hydration of our benchmark set increases to 1.51 kcal/mol if virtual sites are not included. QUBEKit writes the virtual site positions in OpenMM .xml file format for ease-of-use and easy automation of derivation and testing pipelines.

In agreement with previous work we again find that while the derived C_6 dispersion coefficients using the T-S scaling relations react to their atomic environment, they are substantially lower than their empirically fit counter parts of the OPLS FF. In fact recent work by Mohebifar et al has highlighted that this is systematic across a wide range of commonly used small molecule and protein FFs [159, 157] when compared with dispersion coefficients calculated using the exchange-hole dipole model (XDM)[160]. XDM provides a nonempirical way to calculate the dispersion coefficients of atoms and molecules to an arbitrarily high order directly from DFT calculations. The method relies on the use of a reference electron and its exchange hole as the source of dispersion and produces coefficients which respond to the local environment. Overall Mohebifar et al found that XDM approximated C_6 coefficients are on average 50% smaller than empirically fit values [159, 157]. This can be attributed to the fact that the simple functional form of the MM L-J

potential neglects higher order dispersion interactions, modelling only the leading term of the expansion (see section 2.2.2) resulting in overly large C_6 coefficients to compensate the functional form deficiency. Logically this then explains the underestimation of attractive dispersion interactions within the QUBE model which can be seen in the benchmark results. In light of this the XDM method has been used to successfully parametrise several polarisable small molecule FFs directly from DFT calculations which include higher order dispersion terms up to C_8 in an adapted functional form [75, 161]. These FFs have also been shown to achieve very competitive performance in standard pure liquid benchmarks and emphasise the ease of generating a QM derived FFs to take advantage of more complex functional forms.

In contrast to previous work [59], we have supplemented the atoms-in-molecule non-bonded parameters with molecule-specific bonded parameters derived from the QM Hessian matrix and torsional scans. In agreement with previous studies [45], we showed that the so-called modified Seminario method is able to reproduce QM normal mode vibrational frequencies to high accuracy (6.5% here). Closer examination of bond and angle force field parameters for widely used atom types reveals that these parameters are reasonably transferable between closely-related molecules. Such analyses of more complex molecules could be used to identify problems with standard force fields where bonded parameters may require re-fitting or the inclusion of more atom types. In addition, we have shown that for three molecules, QM relative energies of an ensemble of structures are modelled reasonably well with the QUBE FF when combined with torsional fitting. It should be noted that torsional fitting is the major computational expense in QUBE (since it requires a constrained QM optimization at each torsion angle), and methods to reduce this expense will be discussed in chapter 6. Improper torsional parameters are not derived in this study, and we have used those from the OPLS FF here, however, with a small modification to QUBEKit, future versions could easily extend the torsion optimization procedure to include such dihedrals scanned over a limited range [162]. Potential future improvements include support for 2D torsion scans [48], and the use of direct fitting to the Hessian matrix to allow derivation of stiff, harmonic torsional parameters and cross-terms to account for coupling between internal coordinates [54, 55, 146, 41]. Such improvements are especially important in, for example, spectroscopic applications where a faithful representation of the QM intramolecular potential energy surface is crucial [163, 164]. Additional validation of QUBE against metrics

such as condensed phase dielectric constants [145], host-guest binding [42] and many more are envisaged, and QUBEKit will facilitate this process.

Chapter 5

Retrospective study of p38α MAP Kinase using the QUBE FF

5.1 Introduction

The ability to prospectively rank order a congeneric series of inhibitors based on their predicted binding affinity is crucial to the speedy delivery of new medicines in the pharmaceutical sciences. FEP based on MM simulations can be an effective guide during the hit-to-lead stages of a drug design campaign as it provides a formally rigorous means to compute protein-ligand binding free energies [165, 166, 43]. In practice, the predictive ability of such simulations is effectively limited by two major factors, 1) the accuracy of the underlying MM FF that is used for the rapid calculation of the system energy, and 2) finite simulation times that can limit the conformational space explored [167]. In the expectation of making such calculations routinely reliable, the development of enhanced sampling methods is an active area of research [168, 169], yet virtually all FEP simulations employ transferable biological FFs, such as AMBER, OPLS, GROMOS and CHARMM, all with quite similar functional forms and parameter fitting strategies [170]. These biological FFs, alongside their small molecule counterparts, have had wide success to-date thanks to meticulous fitting of parameters to reproduce QM and experimental properties of sets of small organic molecules. However, there is room for improvement [171, 42, 172]. It is widely acknowledged that atomic point charges are sensitive to their (local and long-ranged) environment, which is why small molecule FFs typically employ atomic charges that are fit to the molecular electrostatic properties (e.g. ESP or

CMx charges [97]), on a case-by-case basis. Interestingly, this leads to a disconnect between protein and small molecule FFs, in which the former sets of atomic charges are read from a transferable library, and the latter are derived using methods that are not always consistent with the underlying biological FF. Also, standard libraries of parameters describing torsional rotation about flexible bonds are often blamed for observation of unphysical conformations in MM simulations, and these parameters are often re-derived specifically for the molecule under study [132, 20, 173, 130].

With regards to these issues, there has been recent interest in molecule-specific, or bespoke, FFs in which the parameters that govern the dynamics of the system are not assigned from a library based on predetermined atom types, but instead are inferred directly from QM calculations specifically for the molecule of interest [41, 174, 130, 61]. One such example is the QUBE FF [61], which has a particular focus on scalability to large system sizes and applications in the condensed phase [59, 60]. The QUBE FF shares its functional form with OPLS, so that it retains the favorable computational efficiency of transferable FFs, but differs in that as many parameters as feasible are derived directly from routine, molecule-specific QM calculations. The ground state electron density of the molecule under study is first computed in a weak implicit solvent to simulate the effect of environmental polarization [59]. The density is then partitioned into a set of approximately spherical atom-centered basins via the DDEC AIM approach [100, 101], from which we compute the environment-specific non-bonded parameters, including (atom-centered and off site) atomic charges and Lennard-Jones parameters [59, 61] (see section 2.3.2). Since the DDEC method is implemented in the linear-scaling density functional theory code, ONETEP [135], we can derive these parameters consistently for both small molecules and also systems comprising thousands of atoms, such as proteins [58, 138]. QUBE bond and angle FF parameters are derived directly from the QM Hessian matrix of small molecules, as described previously [45], and flexible torsions may be parametrized by fitting to constrained one-dimensional QM dihedral scans [61]. Parameter assignment is automated by the QUBEKit software package [61] presented in this thesis.

To date, the first generation of the QUBE force field has undergone extensive benchmarking against established performance metrics, such as the prediction of the condensed phase thermodynamic properties (density, heat of vaporization and free energy of hydration (see chapter 4)) of over 100 small organic molecules [61]. A custom library of bonded parameters for protein simulations has been developed and validated via the comparison of molecular dynamics trajectories with NMR

observables [60]. In all of these cases, QUBE performed to a similar standard as established and optimized transferable force fields. In the context of FEP calculations, QUBE has been applied to the study of the benchmark L99A mutant of T4 lysozyme, achieving a MUE of 0.85 kcal/mol in the prediction of the absolute binding free energies of six benzene derivatives. However, typical hit-to-lead studies in drug discovery scenarios are significantly more complex than the above study in terms of the sizes of the ligands, the nature of their interactions, and their conformational flexibility [175, 176, 166]. In this section we therefore retrospectively calculate the relative binding free energies of a series of 17 drug-like inhibitors of p38α MAP kinase (Figure 5.1). This represents a typical optimization scenario involving both polar and non-polar substitutions around a benzene ring, with activities that span 2–3 orders of magnitude (Table 5.1). As we shall discuss, the binding pose is determined to a large extent by two flexible dihedral angles (ϕ_1 and ϕ_2, Figure 5.1), which impose complex sampling requirements on the simulations. This set of transformations has been the target for a range of activity prediction methods including FEP calculations, which were used to demonstrate the importance of the initial water placement during MC simulations using the OPLS force field [3].

5.2 Computational Methods

5.2.1 System Preparation

Input structures for the complexes between p38α MAP kinase and the 18 inhibitors were prepared starting from the crystal structure (PDB: 1OUY [177]) as described below using the MCPRO 3.2 [119] and BOMB [178] software packages. The x-ray crystal structure contained an inhibitor structurally similar to ligand **17** which was extracted and truncated to serve as the common core substructure used to generate all other compounds via the molecule growing program BOMB [178]. Crystallographic water molecules were removed and the protein and ligand z-matrices were prepared using the chop and pepz utilities of MCPRO 3.2. Any residues within 20 Å of the ligand were retained and a fully flexible region was defined within this region with a cut-off distance of 10 Å. It was confirmed that an increase in the radius of the flexible region to 12.5 Å changed the computed relative binding free energy by less than 0.2 kcal/mol for the transformation of **2** to **1** (0.36 to 0.2 kcal/mol). The net charge of the system was set to zero via neutralization of distant, titratable residues

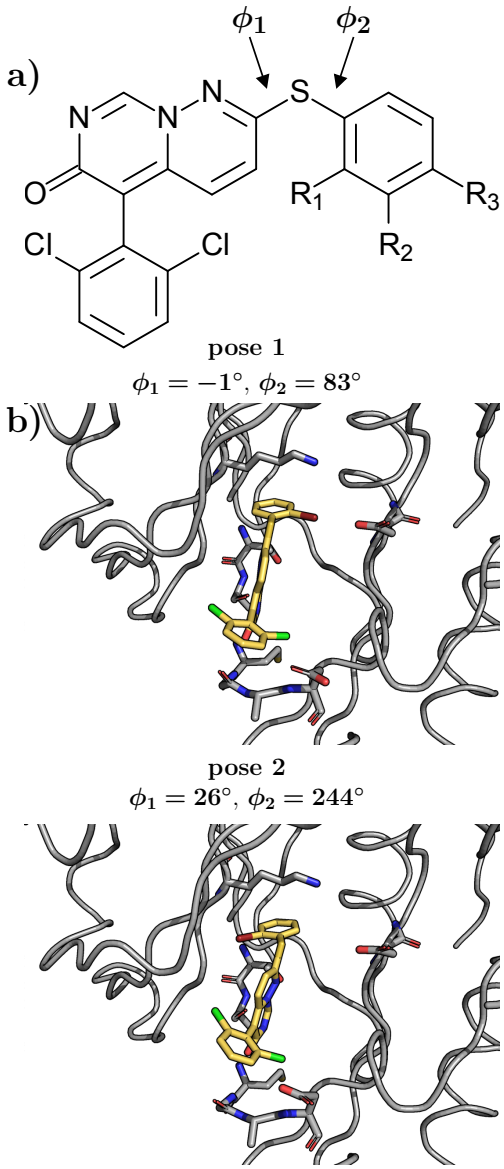

Figure 5.1: (a) Core structure of the p38α MAP kinase inhibitors studied here. Key flexible dihedrals (ϕ_1 and ϕ_2) are labelled. (b) Snapshots from FEP MC simulations of ligand **12** (yellow) highlighting binding poses 1 and 2.

and non-bonded energy terms used a 10 Å cutoff. Ligand and key host degrees of freedom were optimized using BOMB. Each protein-ligand complex was solvated in a water cap with radius 25 Å using the JAWS hydration protocol described in detail elsewhere [179].

Compound	R_1	R_2	R_3	pIC_{50}
1	H	H	H	6.6
2	H	H	F	7.0
3	H	H	CH_3	5.9
4	H	Cl	Cl	6.1
5	H	CH_3	H	5.9
6	H	CH_3	CH_3	5.7
7	H	F	H	6.3
8	CH_3	H	H	6.7
9	H	Cl	F	6.3
10	H	Cl	H	6.6
11	CH_3	H	Cl	6.7
12	Br	H	H	6.6
13	CH_3	H	CH_3	6.6
14	OH	H	H	6.4
15	NH_2	H	F	6.7
16	Cl	H	F	7.4
17	F	F	F	8.0
18	F	H	H	N/A

Table 5.1: List of p38α MAP kinase inhibitors with their pIC_{50} values. The pIC_{50} is the negative log of the experimentally measured IC_{50} activities [3] which correspond to the concentration of an inhibitory substance required to inhibit a biological process or component by 50%.

5.2.2 QUBE FF Parametrisation

Ligand force fields were parametrized using the QUBEKit software package [61]. Quantum chemistry geometry optimizations and frequency calculations were performed in Gaussian09 [2] using the ωB97XD functional and 6-311++G(d,p) basis set. Equilibrium bond lengths and angles were extracted from the QM optimized geometry, and the bond-stretching and angle-bending force constants were derived from the QM Hessian matrix via the modified Seminario method with a vibrational scaling factor of 0.957 [45]. Constrained one-dimensional torsional optimizations were also performed using Gaussian09, with the same level of theory and basis set, in 15° increments from 0° to 360°. Torsion parameter optimizations of dihedrals ϕ_1 and ϕ_2 were performed for each ligand separately using QUBEKit with no Boltzmann weighting or regularization [61]. OPLS atom types were retained during torsion fitting to reduce the parameter search space, while all remaining small molecule torsion parameters were taken from the OPLS force field. Non-bonded parameter assignment was

performed for both small molecules and the protein (2961 atoms) using the ONETEP linear-scaling density functional theory code and DDEC AIM analysis (see below). All bonded parameters of the protein were assigned from a transferable library that has been specifically designed to be compatible with the QUBE FF [60]. Water molecules were described using the TIP4P water model.

5.2.3 ONETEP Calculations

All ground state electron densities used to derive the non-bonded parameters of both the 18 ligands and the p38 kinase protein (2961 atoms) were computed using the linear-scaling density functional theory code, ONETEP [135]. ONETEP uses a basis set of spatially-truncated nonorthogonal generalized Wannier functions (NGWFs) localized on each atom. Four (NGWFs), with radii of 10 Bohr, were used for all atoms with the exception of hydrogen, which used one. NGWFs were expanded in a periodic cardinal sine (psinc) basis, with a grid size ($0.45a_o$), corresponding to a plane wave cutoff energy of 1020 eV. The PBE exchange-correlation functional was used with OPIUM norm-conserving pseudopotentials. The calculation was carried out in an implicit solvent using a dielectric of 4 to model induction effects in the ligands, and 10 in the protein. For several test cases, ligand charges were also computed using a dielectric of 10, but the RMS/maximum differences between the charge set are just 0.01/0.03 e. The DDEC module implemented in ONETEP was used to partition the electron density and assign atom-centered point charges and atomic volumes, no off center charges were used in this study [61]. The electron density partitioning was assigned using an IH to ISA ratio of 0.02 [59]. Lennard-Jones parameters were calculated using the Tkatchenko-Scheffler relations [25], and protocols described previously [59].

5.2.4 Free Energy Calculations

FEP/REST calculations (see chapter 2.4.3) were performed using the MCPRO software, version 3.2, which includes recent improvements to the efficiency of protein MC moves [180]. The free energy calculations were performed using the single topology approach for both the bound (protein-ligand complex in water) and unbound (ligand in water) simulations as part of a standard thermodynamic cycle. Ligands were transformed over the course of 11 equally spaced λ windows. Simple overlap sampling was employed, with each window comprising 10 million (M) (20M) configurations

of equilibration and 30 M (40 M) configurations of averaging for the bound (unbound) simulations. All computed free energy changes (including those presented from previous studies) were computed by aligning the mean energies of the experimental and computed distributions. REST was used during each λ window to effectively rescale the non-bonded and dihedral parameters of the ligand, thereby reducing potential energy barriers in "high temperature" replicas of the system [181, 182]. Four replicas were run in parallel with REST scaling factors exponentially distributed in the range from 25°C to 250°C (chosen to allow reasonable replica exchange). Exchange attempts between pairs of neighboring replicas were attempted every 10 000 MC steps, and the resulting free energy changes were computed from the room temperature ensemble. The "flip" MC dihedral move modification [122] was also used to encourage crossing between energetically separated poses 1 and 2, with random move sizes that ranged from 60° to 180°. Protein conformal sampling employed new protocols, which generate more efficient MC moves specifically targeted at the backbone and side-chains [180]. These moves have been shown to be in good agreement with MD for the calculation of protein conformational ensembles [180] and protein-ligand binding [183].

5.3 Results

5.3.1 Assessing parameter quality

First, we begin by analysing the parametrisation of the molecule-specific FFs for the 17 p38 kinase inhibitors (plus compound **18**, which does not have experimental data for comparison but is a useful FEP intermediate) using the QUBEKit software [61]. Non-bonded (charge and Lennard-Jones) parameters are derived using AIM partitioning of the ground state electron density as described previously [59, 61]. Parametrization of the protein non-bonded parameters is performed using the same protocols, while bonded parameters are read in from a custom library [60]. Bond and angle parameters of the small molecules (**1**–**18**) are derived using the modified Seminario method computed using the QM Hessian matrix at the optimized geometries [45]. Finally, parameters describing rotation about the two flexible dihedral angles ϕ_1 and ϕ_2 are fit to constrained QM potential energy scans. Figure 5.2 shows the results of the torsion parameter optimization for ligand **1**. The fit to the underlying QM data is very good with an average root mean square deviation between sampled QM and

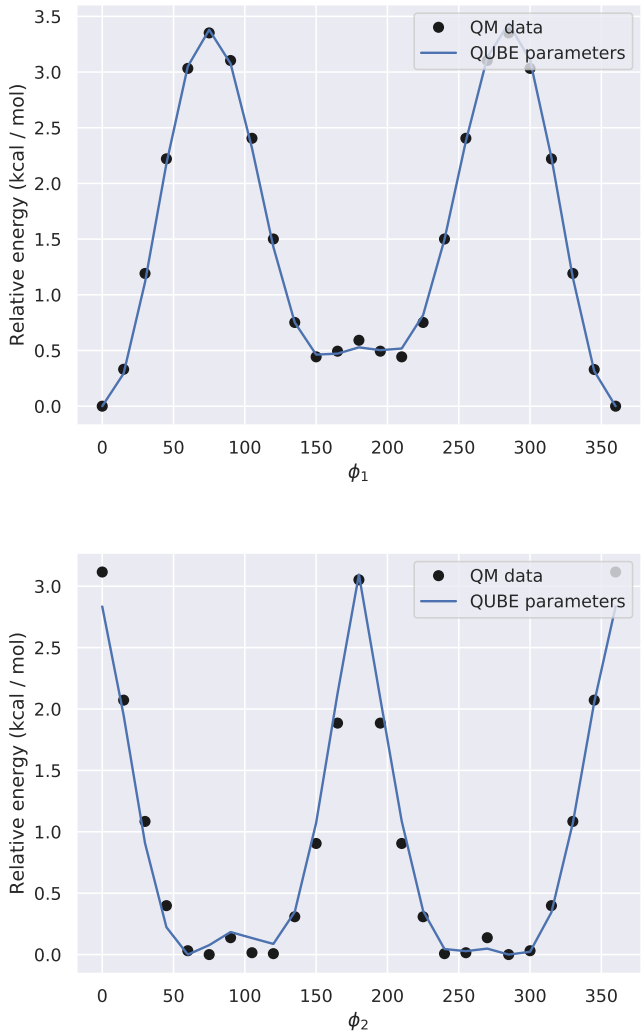

Figure 5.2: QUBE and QM PES for ligand **1** upon rotation of flexible dihedrals ϕ_1 and ϕ_2. Potential energy surfaces prior to optimization (using OPLS torsional parameters) are shown for comparison.

QUBE torsional scans of 0.07 kcal/mol. For comparison, typical errors in excess of 1.5 kcal/mol are observed using small molecule transferable force fields [184]. By deriving the QUBE FF directly from QM, our goal is to provide accurate and automated molecule-specific parameters that reproduce as closely as possible the full QM potential energy surface. Figure 5.3 shows the correlation between QUBE and QM relative energies of structures **3** and **10** extracted from Monte Carlo simulations (see later). The correlation between QUBE and QM energetics is similar to that previously reported [61], and significantly QUBE does not predict any physically unreasonable structures (either bound to the protein or in water) whilst retaining the

	rmse (kcal/mol)	
Compound	ϕ_1	ϕ_2
1	0.038	0.107
2	0.061	0.063
3	0.057	0.061
4	0.132	0.230
5	0.046	0.259
6	0.073	0.192
7	0.284	0.188
8	0.372	0.287
9	0.147	0.141
10	0.116	0.158
11	0.215	0.363
12	0.407	0.453
13	0.381	0.140
14	0.341	0.053
15	0.475	0.303
16	0.161	0.144
17	0.103	0.438
18	0.112	0.539
Average	**0.196**	**0.229**

Table 5.2: Root mean square deviation between QM and QUBE torsional energy profiles for rotation of ϕ_1 and ϕ_2 for each of the 18 molecules.

fixed MM functional form that provides us with a practical method for deployment in free energy predictions. Additional analysis of torsion scans and correlations between QM and QUBE energetics for the remaining ligands may be found in tables 5.2 and 5.3.

5.3.2 Predicted binding poses

Having parametrized the 18 inhibitors, we turn now to the computation of their relative binding free energies to p38 kinase. Free energy calculations were performed using the MCPRO software [119]. The ligand binding site is expected to be hydrated, and so the JAWS water placement algorithm [179] was used to optimize the initial solvent distribution. As reported previously [3], the majority of the ligands **1-17** are expected to bind in pose 1 (Figure 5.1). Hence, we set up the ligands initially in pose 1, but employed the REST enhanced sampling method with the goal of reducing the dependence of the computed binding free energies on the starting conditions.

	correlation (r^2)		rmse (kcal/mol)	
Compound	Bound	Unbound	Bound	Unbound
1	0.665	0.507	3.08	3.61
2	0.694	0.643	3.26	3.89
3	0.822	0.705	3.60	3.41
4	0.682	0.735	3.02	3.60
5	0.860	0.681	2.82	3.80
6	0.714	0.697	2.97	3.63
7	0.480	0.571	3.06	3.90
8	0.460	0.574	4.50	3.31
9	0.651	0.713	4.10	3.49
10	0.693	0.659	3.11	3.70
11	0.593	0.651	3.51	3.40
12	0.660	0.615	3.74	4.06
13	0.684	0.722	3.94	3.58
14	0.220	0.643	3.63	3.88
15	0.570	0.675	4.37	3.33
16	0.591	0.623	3.11	3.73
17	0.583	0.622	2.79	3.51
18	0.631	0.759	3.22	3.61
Average	**0.625**	**0.655**	**3.44**	**3.64**

Table 5.3: The correlation between the single point energies calculated using the QUBE FF and QM on structures extracted from MC simulations in the bound (protein-ligand complex in water) and unbound (ligand in water) states. Note that the correlation is relatively low for **14** in the bound state, but this appears to be due to the limited variability of structures, and hence energies, sampled.

Importantly, in MCPRO, the REST algorithm may be employed alongside the 'flip' protocol, in which selected dihedral angles (here, ϕ_1 and ϕ_2) undergo Monte Carlo moves that are much larger than typical. For example ligand **1** is symmetric under 180° flips in ϕ_2, and indeed approximately equal distributions of the two conformers are observed at $\phi_2 = 40°$ (pose 1) and $\phi_2 = 220°$ (pose 2). Figure 5.4 further illustrates the effects of this sampling procedure. Interestingly, despite starting in pose 1, **17** shows a single peak at $\phi_2 = 250°$, indicating a strong preference for pose 2. This agrees with previous observations using the OPLS force field [3], and x-ray crystal structures of similar ligands [177] (Figure 5.5). Of note, in that former study, MC simulations were required starting from both poses 1 and 2 since interconversion between the two is not expected during these simulations using either standard MC or molecular dynamics. In contrast, the use of the REST/flip

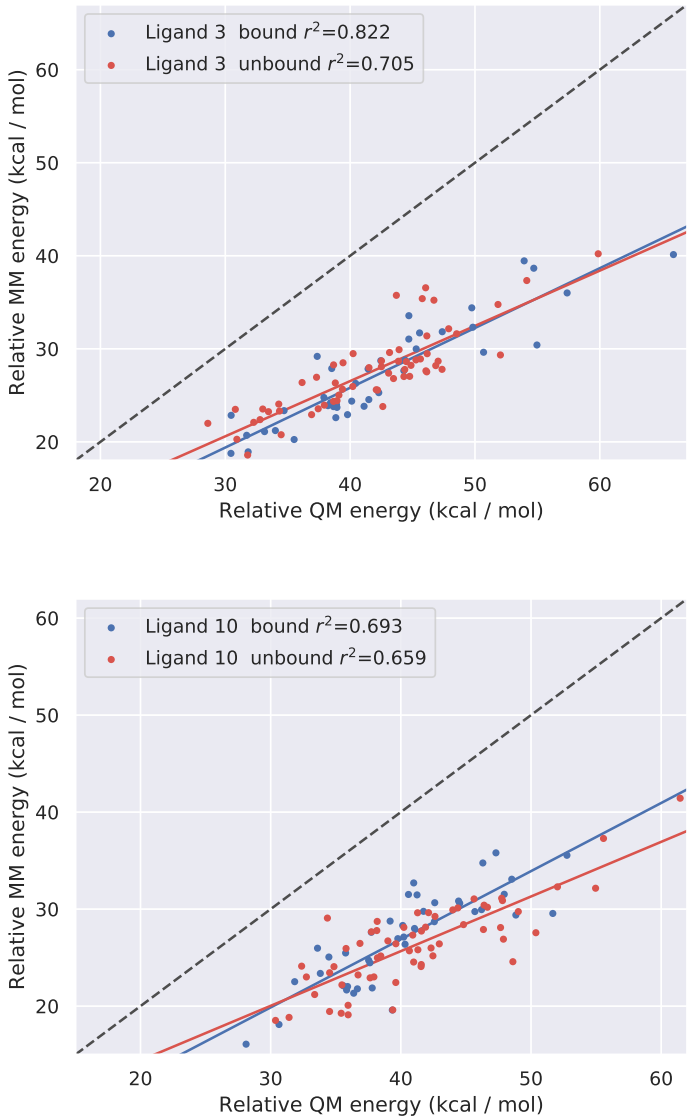

Figure 5.3: Comparison between QUBE and QM single point energies of structures of **3** (top) and **10** (bottom) extracted from bound and unbound (in water) MC simulations. The mean energies of each distribution have been shifted to zero. Also shown are the correlation (r^2) and root mean square errors (rmse, kcal/mol) between the two distributions.

algorithm facilitates binding mode determination and free energy prediction from a single MC run. Despite being asymmetric, **12** shows similar behavior to **1**, with peaks around $\phi_2 = 30°$ and $\phi_2 = 210°$ (Figure 5.4). This is perhaps reasonable, since **12** is similar in chemistry to **17**, but the bulkier Br atom may hinder full inclusion into the pose 2 binding pocket. Overall, we conclude that using the QUBE force field and REST enhanced sampling algorithm described here, the asymmetric

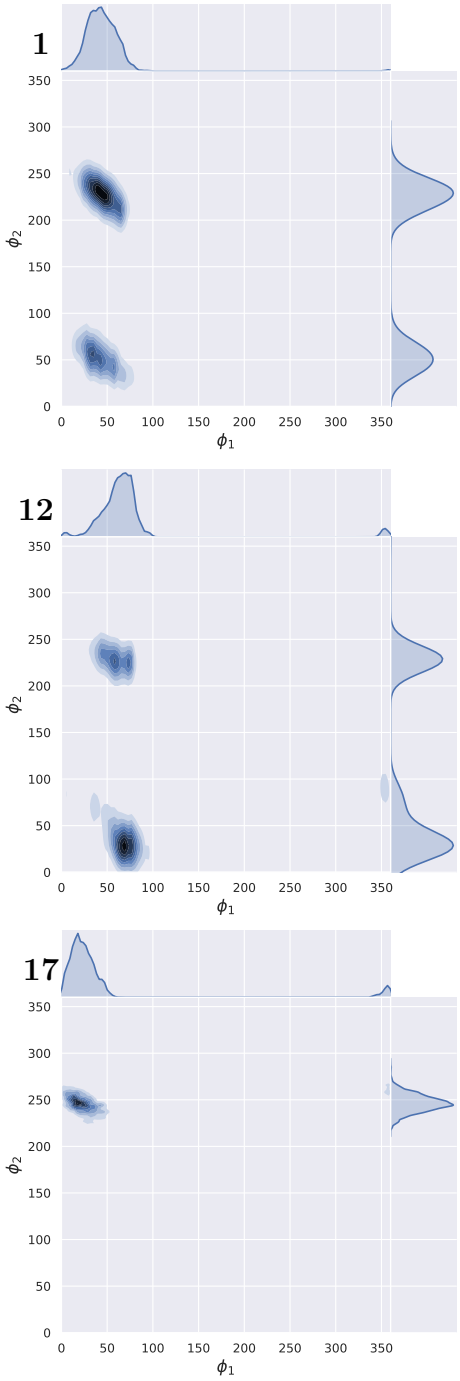

Figure 5.4: Two-dimensional dihedral distributions observed during the protein-ligand complex simulations of ligands **1**, **12** and **17**. See also Figure 5.1 for indicative poses.

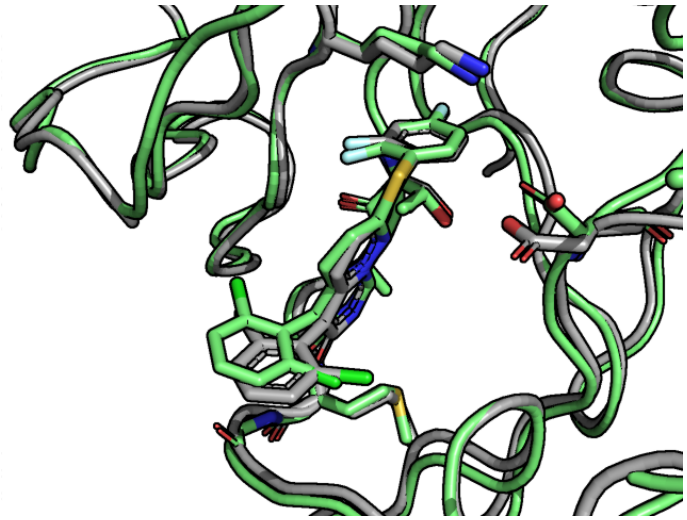

Figure 5.5: Overlay of the crystal structure (PDBID: 3FC1, gray) with the last snapshot (green) of the MC simulation of **17** bound to p38α MAP kinase.

ligands **4**–**11** and **13**–**15** bind in pose 1, ligands **16**–**18** bind in pose 2, and ligand **12** is intermediate between the two.

5.3.3 Relative binding free energies

Having elucidated the preferred binding poses of the 17 inhibitors, we turn now to the prediction of protein-ligand relative binding free energies. Figure 5.6 compares the errors in the relative binding free energies computed using the QUBE protein/ligand force field with experiment. For comparison, the corresponding quantities are also displayed for the OPLS force field from previous work [3]. Full details of the transformations used in this study are given in the appendix 8.4. Overall, the MUE using QUBE is 0.98 kcal/mol, which is competitive with the generally accepted accuracy of standard biological FFs for transformations of this type [166] and, in particular, with previous calculations using the OPLS FF on this system (0.88 kcal/mol). The largest errors, using the QUBE FF are for ligands **12**–**15**, which all include bulky and/or polar substituents at the R_1 position (Figure 5.1), as well as ligand **7**. The torsional profiles of these ligands are all reasonable, and so it seems likely that non-bonded interactions and/or sampling errors are to blame. We have found previously that QUBE can underestimate hydration free energies of some molecules containing bulky hydrophobic and hydroxyl functional groups by up to around 2 kcal/mol [61]. Although the relatively high accuracy of the **8** and **11**, for example, indicates

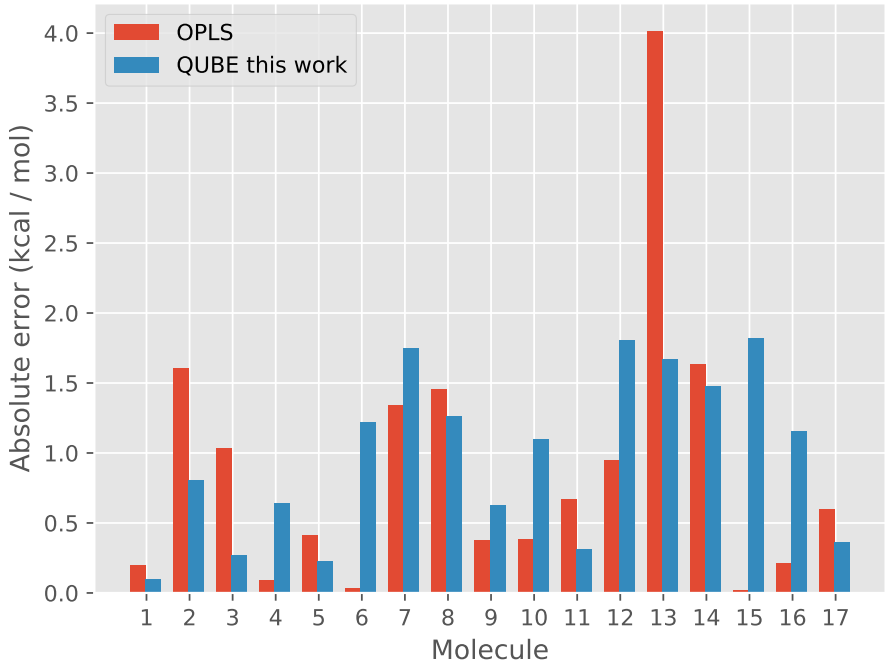

Figure 5.6: Absolute errors in predicted relative binding free energies computed using the QUBE and OPLS [3] force fields, compared to experiment.

that the presence of these functional groups at R_1 is not the only factor affecting accuracy. Figure 5.7 shows the correlation between the QUBE and OPLS predictions of the binding free energies of the 17 inhibitors to p38α MAP kinase. Although both FFs have similar errors relative to experiment, as demonstrated by several statistical measures (table 5.4) there are some quite large differences in individual predictions. For example, there are differences between QUBE and OPLS in excess of 2 kcal/mol in the computed binding free energies for compounds **2**, **12** and **13**. The latter two are perhaps not surprising given the sampling and FF difficulties discussed. Compound **2** has a F substituent at the R_3 position with QUBE non-bonded parameters: q = -0.21 e, σ = 2.89 Å, ϵ = 0.066 kcal/mol. The corresponding OPLS/CM1A parameters are: q = -0.08 e, σ = 2.90 Å, ϵ = 0.060 kcal/mol. The difference in the charge sets here may be sufficient to explain the difference in binding prediction for compound **2**, but larger datasets involving fluorinated compounds will be required to investigate further. Other possible sources of inaccuracy, highlighted by Luccarelli et al. [3], are that changes in solvent distribution in the binding pocket and/or protein side chain conformational changes are not properly sampled

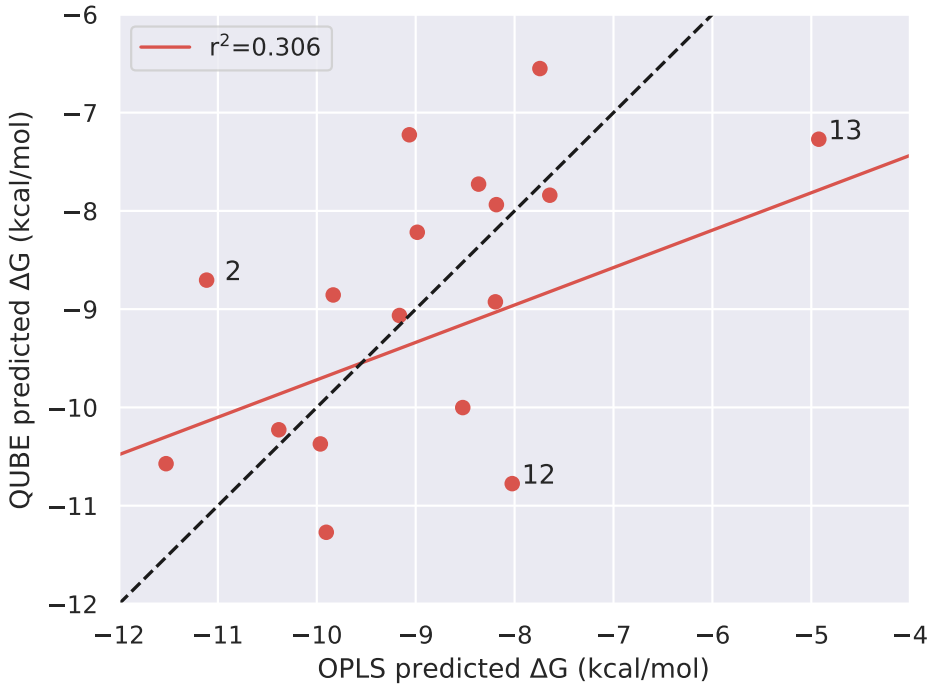

Figure 5.7: Correlation between QUBE and OPLS predictions of the binding free energy for the 17 inhibitors.

Force Field	MUE	RMSE	Spearman's rho
OPLS	0.88	1.30	0.46
QUBE	0.98	1.14	0.40

Table 5.4: Comparison between FF methods and experiment. Mean unsigned error (MUE, kcal/mol), root mean square error (RMSE, kcal/mol) and Spearman's rank correlation coefficient for each theoretical method are shown. OPLS data are taken from the previous literature [3].

during alchemical perturbation. To investigate the adequacy of the REST method for sampling the complex binding mode of **12**, we re-ran the **12**→**18** and **18**→**1** transformations starting with the ligand in pose 2. However, the error in the relative binding free energy of **12** fell only from 1.8 kcal/mol to 1.5 kcal/mol, indicating that the binding mode is sufficiently sampled during our simulations.

5.4 Conclusions

In summary, we have benchmarked the accuracy of the QUBE FF against relative

binding free energies of 17 drug-like inhibitors of p38α MAP kinase. The selected protein-ligand complex includes challenges due to sampling of protein-ligand binding modes and binding site hydration, and is therefore representative of typical hit-to-lead optimization projects. The mean unsigned error of 0.98 kcal/mol of the first generation of QUBE is competitive with widely-used biological FFs, and encouragingly the crystallographic binding pose of **17** was obtained despite starting from an alternative structure. More generally, the FEP/REST enhanced sampling protocol employed here allowed us to obtain all predictions starting from a single binding pose, in contrast to previous studies that required two [3]. One current disadvantage of QUBE is parameterization time, which can be of the same order of magnitude as the free energy calculation itself. Derivation of bond, angle and non-bonded parameters for these molecules typically require a taotal of 150 cpuhrs, while calculation of QM torsion profiles requires up to 2000 cpuhrs. However, there is future scope for the use of, for example, fragmentation schemes for reducing the computational expense of torsion scans and machine learning methods for non-bonded parameter assignment [185], especially when employed in congeneric series of ligands such as this one. Meanwhile, a wide range of accuracy improvements are envisaged, from the use of off-site charges in relative binding free energy calculations to improve the description of electron density anisotropy [61], to improved descriptions of polarization, van der Waals and short-range repulsion using advanced force field functional forms [75, 150]. Future work should then continue to improve the accuracy and throughput of the QUBE FF for binding free energy applications in prospective medicinal chemistry efforts.

Chapter 6

The future of QUBEKit-V2

6.1 Introduction

Thus far the QUBE FF has been shown to offer competitive performance against typical transferable FFs such as OPLS, GAFF and CHARMM which are commonly employed in computational drug design. This has been validated through the standard FF performance metrics presented in chapter 4 and the FF's benchmark application to a drug design setting involving the retrospective calculation of relative (chapter 5) and absolute binding free energies [123] for two different systems. Now that this first generation of the QUBE FF has been shown to reach sufficient accuracy to guide a medicinal chemistry effort we turn our attention to its future ahead of the second major release. Specifically, we focus on areas of the FF that could be improved and currently limit the adoption of the QUBE FF, such as chemical space coverage, the robustness of the torsion optimisation procedure and compatibility with open source software where possible, as discussed in chapter 3. Here we discuss how QUBEKit-V2 aims to incorporate these improvements starting with our improved initial parametrisation protocol. This feature allows users to assign and derive parameters for molecules containing elements not included in QUBE or other transferable FFs and is demonstrated for a range of boron and silicon-containing molecules some of which are then used to optimise their respective R_{free} parameters. We then move on to our new open-source implementation of an improved torsion parameter optimisation scheme which resolves some of the limitations of the previous approach such as hysteresis during the constrained optimisations. To thoroughly test this new optimisation procedure in its routine application we refit the dihedral parameters for a collection of molecules taken from eMolecules with potential energy

surfaces that are traditionally difficult to model.

6.2 Initial parametrisation

As the QUBE FF does not have the means to derive improper torsion parameters yet and many simple proper dihedral terms like those for methyl groups are already well described by a transferable FF, it relies on the use of initial parameters borrowed from another FF which typically has been OPLS. However, to increase our application range we are now able to allow users to derive parameters starting from OPLS, GAFF or OFF parametrised molecules. While this step does streamline the parametrisation procedure by avoiding many time-consuming QM calculations to fully parametrise a molecule, the range of molecules that can be studied is limited by the element coverage of these FFs. QUBE, like other transferable FFs, currently covers a wide range of commonly occurring elements typically found in biology and drug design, however, it does have the potential to have a substantially larger coverage. As the QUBE FF solely depends on QM calculations to derive almost all FF parameters the method could potentially be applied to any molecules for which we can perform an accurate *ab initio* calculation and fit the required R_{free} parameters (assuming the T-S relationships hold), which are trivial to derive as we have shown in the case of bromine (see section 4.3).

In order to be able to quickly process new molecules with missing parameters it would be ideal if we can build a "skeleton" FF, transferring reliable parameters from existing FFs, and using QUBEKit to fill in the missing terms. Such a scheme would have widespread uses in e.g. organometallic simulations. To achieve this we can take advantage of the hierarchical SMIRKS based parameter assignment method used in the parsley and Smirnoff FFs [4]. First, we check if the molecule can be parametrised using the underlying transferable FF, and if not we add generic terms which are set to zero for each parameter type such as bond-stretching, angle-bending etc. to ensure that the molecule does not cause parametrisation to fail. Thus the FF will apply any known transferable parameters to a molecule and any unknown parameters will match the generic SMIRKS patterns resulting in a semi-parametrised FF template that can be used to guide parametrisation by identifying all of the required bonded and non-bonded terms. An example of this initial parametrisation method is shown for the case of triethylborane in figure 6.1 which replicates how the FF analyses the molecule, highlighting terms that were identified and missing from the FF. From

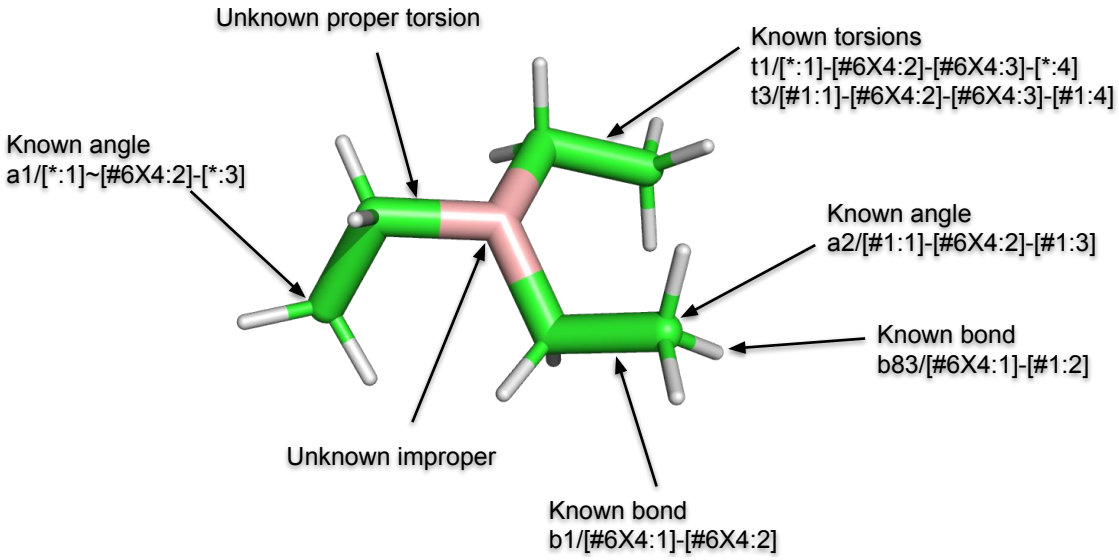

Figure 6.1: Triethylborane is shown after being processed by the OFF toolkit (which uses SMIRKS patterns explained in ref 4 to apply FF parameters), with found and missing FF terms highlighted.

this point QUBEKit can then be used routinely to replace all bond-stretching and angle-bending terms before fitting some new torsion parameters for the unknown place holder generic terms. Even if the overall goal is not to use QUBEKit to totally parametrise the molecule the FF template file creates a very useful starting point which can be easily converted between different simulation packages. This new initial parametrisation method then greatly expands the range of molecules that can be processed by QUBEKit and as we will show in the next section reduces the complexity of extending the FF to new elements.

6.2.1 Boron, silicon and phosphorus

As the non-bonded terms in transferable FFs are optimised to recreate experimental properties, their extension to cover new elements and atom types is often too complex and time consuming. The QUBE FF benefits from only having one non-bonded fitting parameter (R_{free}) per element which significantly reduces the amount of parameter space to be searched and gives competitive performance when fit to simple pure liquid simulations [59]. However until the new initial parametrisation method discussed in section 6.2 was introduced there was no reliable and automated way to process molecules with elements not covered by standard FFs through QUBEKit to fit the corresponding R_{free} parameters. Here we demonstrate how QUBEKit can

now be used to parametrise the bonded FF terms of molecules containing boron, silicon and phosphorous before using a small sample of the molecules to derive some initial R_{free} parameters.

Such parameters are of significant importance to the computational simulation community with phosphorous parameters being vital in biological simulations of DNA and RNA. Silicon and boron on the other hand have recently garnered significant interest in drug design applications due to the advantages of sila-substitution [186, 187] (the strategic replacement of a carbon with a silicon atom) and BN/CC isosterism [188–190] (the replacement of a carbon-carbon (CC) unit with a boron-nitrogen (BN) unit) to increase the chemical space of biologically active compounds. Such substitutions have also been shown to significantly increase potency [187] and alter the local electrostatic properties of motifs [191] giving rise to unique chemical and photoelectronic properties of the molecules compared to their all carbon counterparts. These simple atomic substitutions are also well suited to CADD and the relative free energy calculations employed in chapter 5, however without access to accurate and robust parameters practitioners are unable to use such techniques to guide synthesis.

The need for these parameters is then clear and work towards their accurate parametrisation has begun with the OFF initiative, in particular aiming to add transferable boron terms. As we now have the means to process molecules containing boron, silicon and phosphorus using QUBEKit and "skeleton" FF files we set out to demonstrate how QUBEKit could be used in an automated fashion with the modified Seminario method to derive the bonded parameters for a large selection of example molecules. To this end we have successfully analysed over 200 molecules ranging from fragment to drug-like, in an aim to infer the required atom types and a minimal set of transferable parameters that might be used in a transferable FF. Figures 6.2 and 6.3 show the clustering of the predicted equilibrium bond lengths and angles along with their associated force constants predicted by the modified Seminario method. Interestingly our test set of molecules indicates that the **P-S** and **P-O** bonds have at least two distinct types which could not be represented by single parameters, which coincides with the SMIRKS types found in the parsley FF which are compared in table 6.1. Example molecules from the phosphorus test showing these distinct bond types can also be found in figure 6.4. Figures 6.2 and 6.3 suggest that in the case of boron, transferable FFs would require at least two different bond-stretching parameters corresponding to the **B-N** bond.

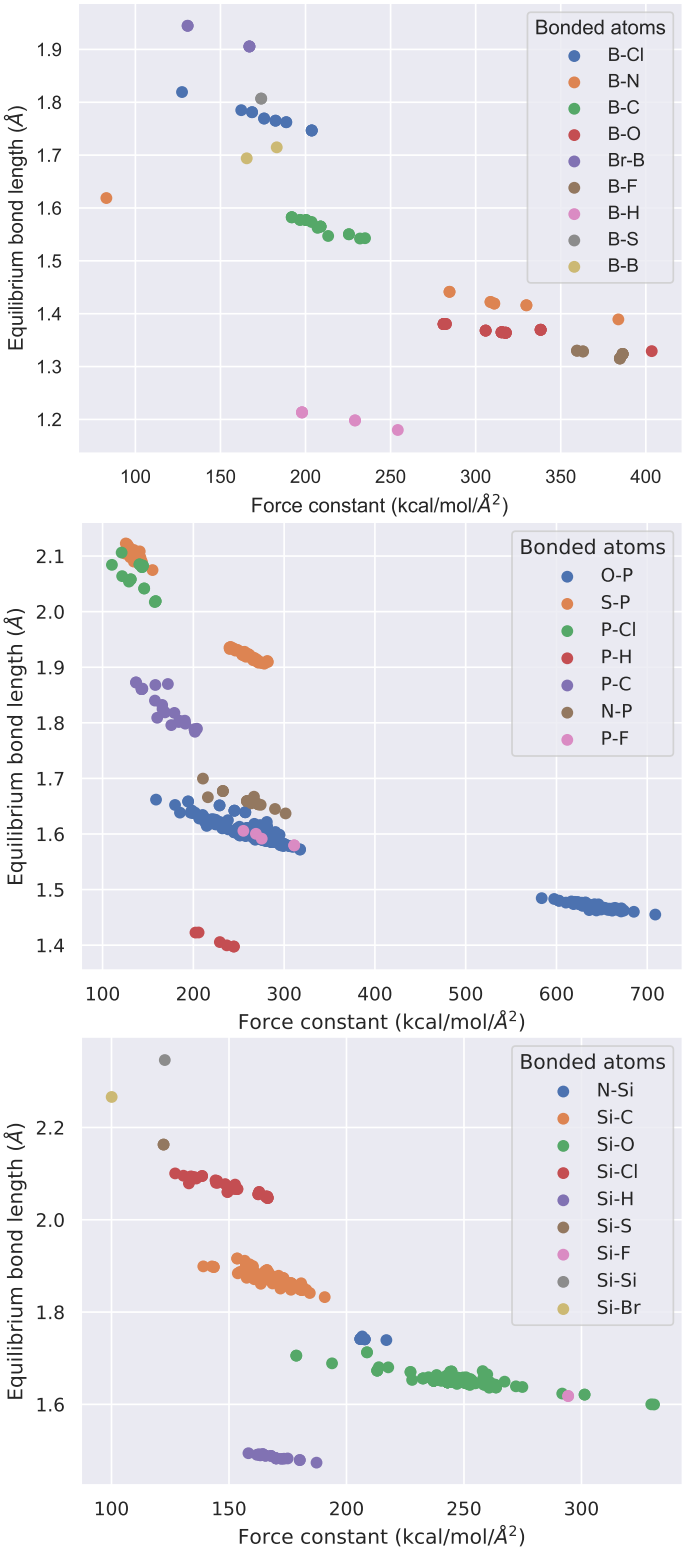

Figure 6.2: The predicted bond-streaching FF parameters taken from the QM optimised geometry and modified Seminario method for molecules containing boron (top), phosphorus (middle) and silicon (bottom).

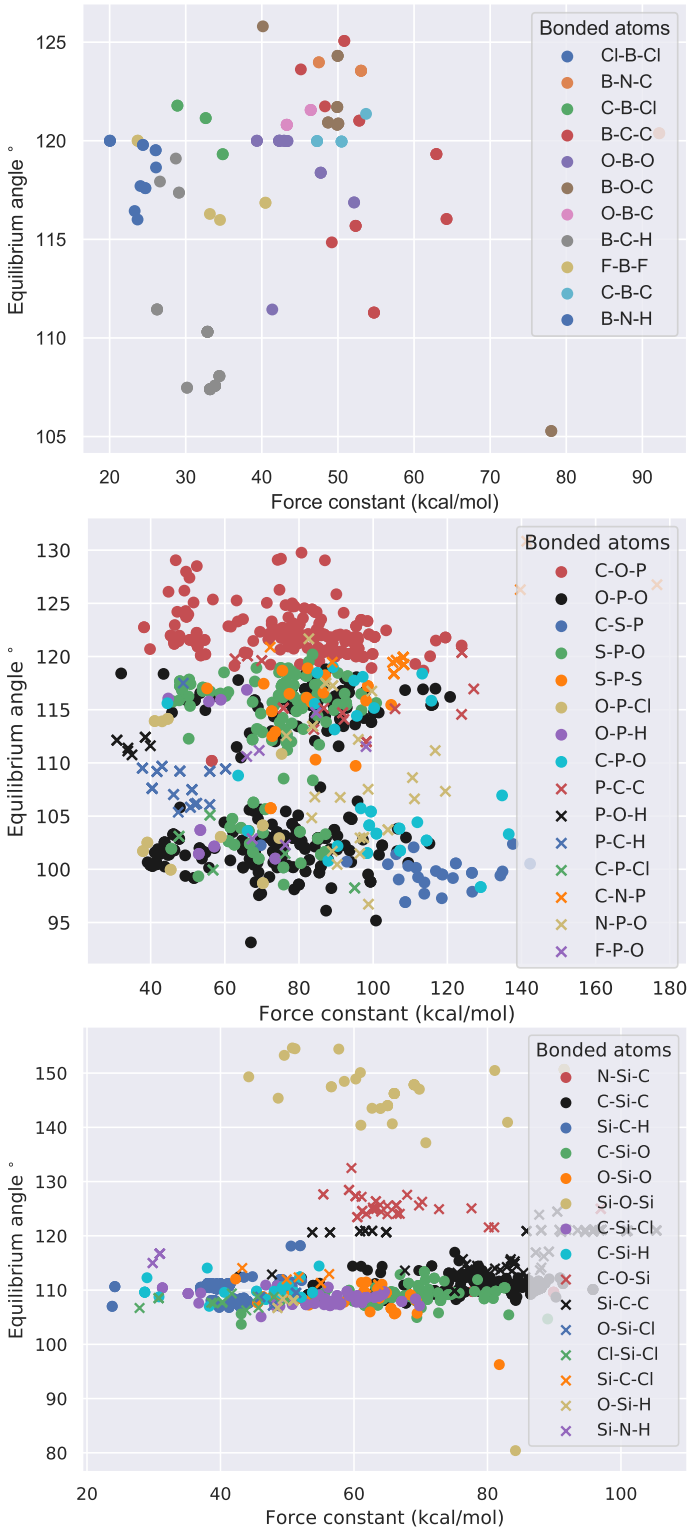

Figure 6.3: The predicted angle-bending FF parameters taken from the QM optimised geometry and modified Seminario method for molecules containing boron (top), phosphorus (middle) and silicon (bottom).

Bond	Parsley SMIRKS	Example molecule
P-S	[#15:1]-[#16:2]	acephate (a)
P-S	[#15:1]=[#16X1:2]	isofenphos (b)
P-O	[#15:1]~[#8X1:2]	acephate (c)
P-O	[#15:1]~[#8X2:2]	isofenphos (d)

Table 6.1: The parsley SMIRKS pattern is shown for each of the QUBEKit bond types identified along with the name of an example molecule whose structure can be found in figure 6.4.

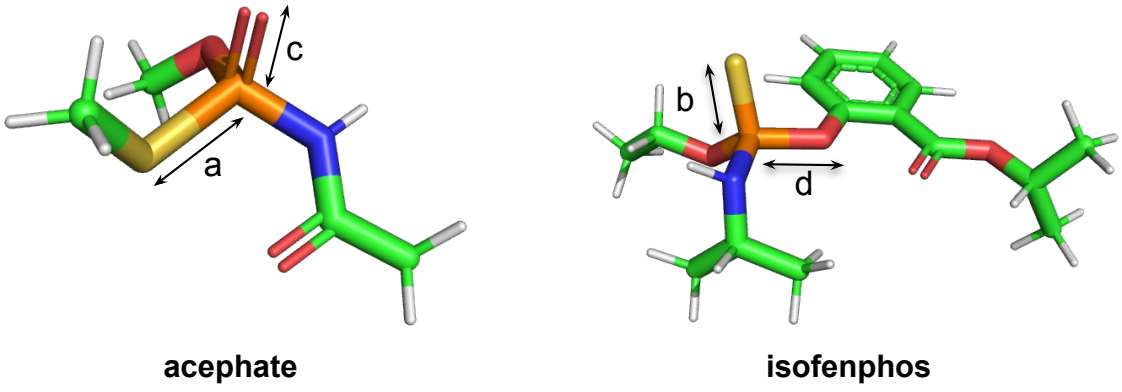

Figure 6.4: Example molecules taken from the phosphorous parametrisation set showing the two types of **P-S** and **P-O** bonds identified by QUBEKit, the corresponding parsley SMIRKS are shown in table 6.1.

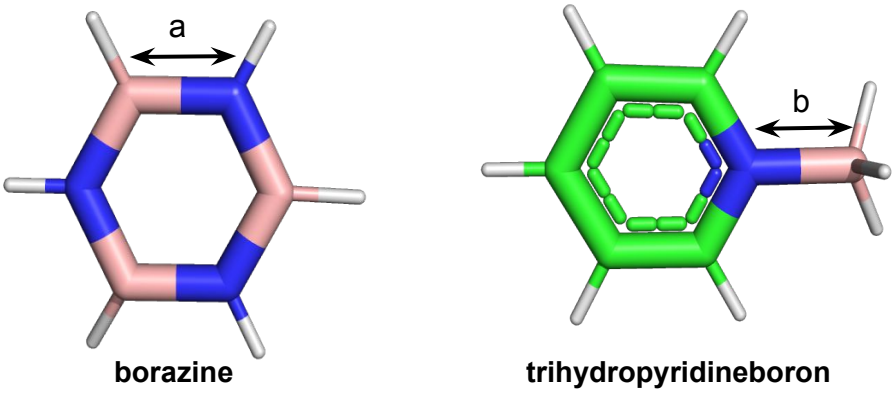

Figure 6.5: Example molecules taken from the boron parametrisation set showing the two types of **B-N** bonds identified by QUBEKit.

Example bond	equilibrium bond length (Å)	k_r (kcal/mol/Å^2)
a	1.421	316.889
b	1.619	82.949

Table 6.2: The predicted equilibrium bond length and modified Seminario force constant for the two types of **B-N** bonds found in the boron test set, example structures can be found in figure 6.5.

Example molecules showing these two categories of **B-N** bond are also shown in figure 6.5, with the corresponding predicted equilibrium bond lengths and modified Seminario force constants in table 6.2. It is also thought that carbon parameters may serve as a good starting point to derive boron parameters and in this case we see that the generic **C-N** bond in the parsley FF has a equilibrium bond length of around 1.466 Å which is similar to the **a** type bond shown in table 6.2.

Silicon, however, shows little variation within similar bond and angle types over this test set which should help ease the creation of accurate transferable parameters to cover this set of molecules. While the modified Seminario predicted values of the force constants have been shown to more flexible than their OPLS equivalents [45], they are thought to be a good starting point for iterative fitting techniques and could also help speed up the development of new parameters.

To allow the full parametrisation of molecules containing these missing elements with the QUBE FF however, we require the corresponding R_{free} values. Thus in order to determine the suitability of the T-S relations for modelling such elements within the QUBE FF a small selection of representative molecules for which experimental pure liquid data points are available were selected and are shown in figure 6.6. The

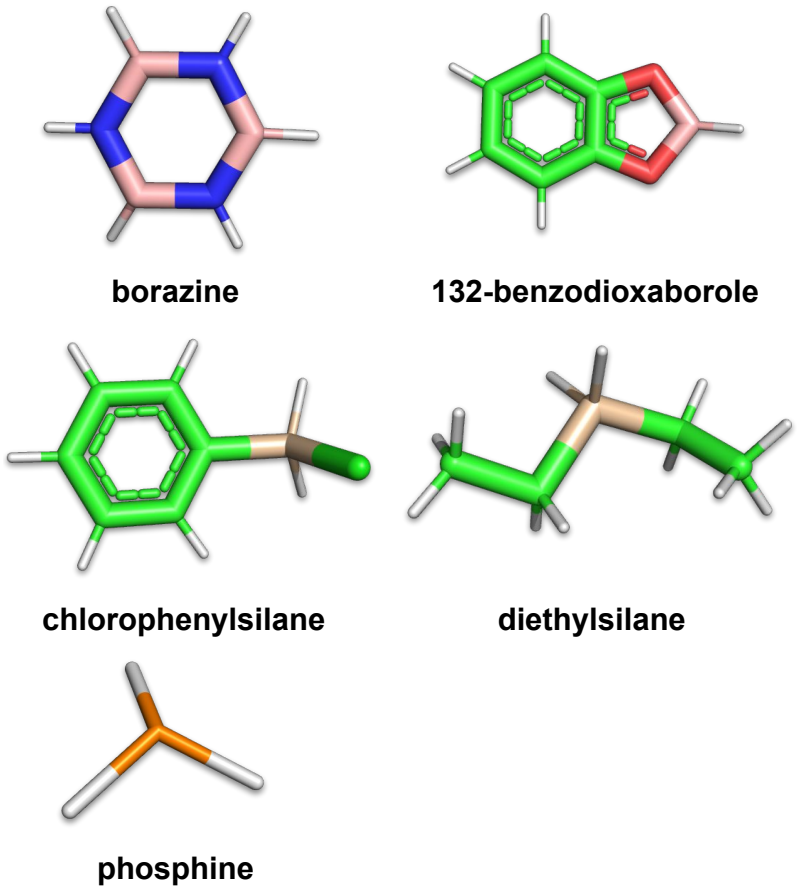

Figure 6.6: The example molecules containing boron, silicon and phosphorus.

fitting set is composed of borazine, 132-benzodioxaborole, diethylsilane, chlorophenylsilane and phosphine all of which were fully parametrised using QUBEKit with the same DFT functional and basis set employed in the benchmark application described in chapter 4 to ensure consistency. A linear parameter space search was then performed by hand via modification of the QUBEKit source files incrementing the R_{free} parameters in regular intervals over a suitable range of physically motived values following the trends in the other elements. The results of the parameter search are shown in figures 6.7, 6.8 and 6.9 for each of the elements with the predicted densities (top) and heats of vaporisation (bottom). As was found with the bromine parameters derived in chapter 4 fitting solely based on the density in the case of boron and silicon would lead to substantially worse performance in regards to the heat of vaporisation. For the two silicon containing molecules picked for fitting we see that they have different optimal R_{free} values in regards to the predicted densities. Overly large ϵ values are required to minimise the error in the

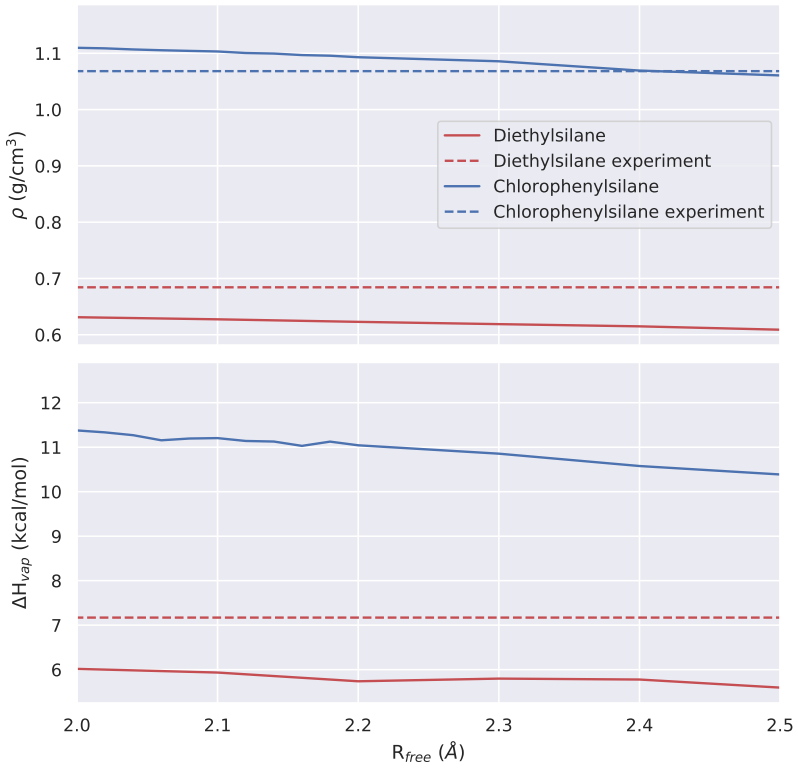

Figure 6.7: The predicted and experimental values for the density (top) and heat of vaporisation (bottom) are shown for diethylsilane and chlorophenylsilane.

heat of vaporisation for diethylsilane. For silicon the very large ϵ parameters of the L-J potential could be an indication that the element may require alternative potentials such as the Tersoff potential [192], which has been used to model silicon oxygen interactions, to be accurately described. On the other hand, this could be an indication of over fitting as some bespoke parametrisations of silicon, which have been used to model silicon-water surfaces in nano-devices with the CHARMM FF, found an ϵ value of 0.3 kcal/mol to be accurate [193]. This would correspond to a R_{free} value of around 2.2Å which is much closer to the value suggested by fitting to the density of chlorophenylsilane.

For boron, a similar disagreement between the molecules picked for fitting is observed with borazine requiring a much larger R_{free} value compared to that of 132-benzodioxaborole. To validate the large over-prediction in density for borazine the heat of vaporisation was also checked as it was available for this molecule and

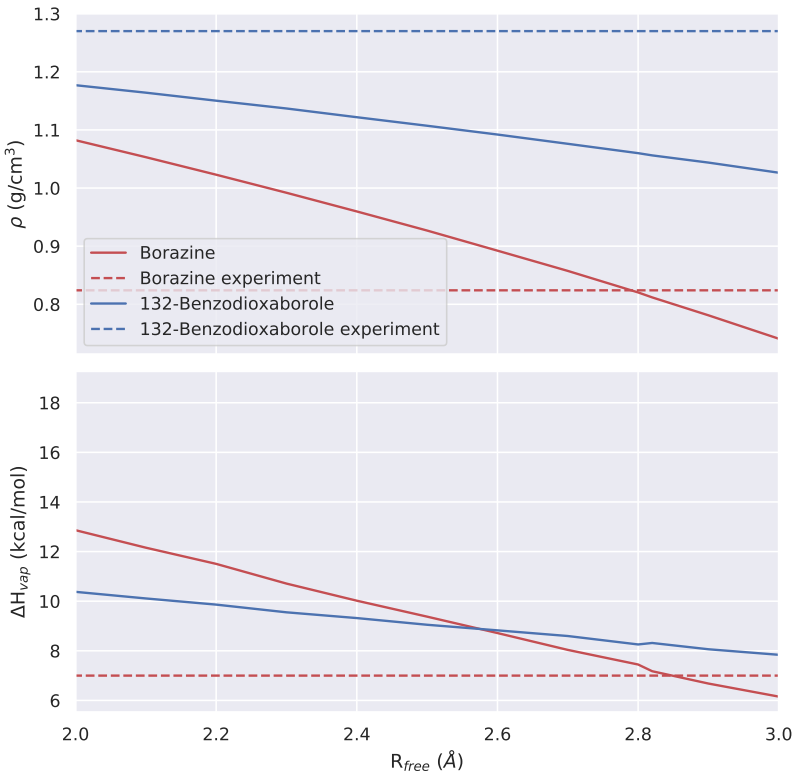

Figure 6.8: The predicted and experimental values for the density (top) and heat of vaporisation (bottom) are shown for borazine and 132-benzodioxaborole.

its dependence on R_{free} is also shown in figure 6.8. From this we can see that the dependence on R_{free} agrees with the density predictions shown in figure 6.8 leading to very small ϵ values. Interestingly there have been other bespoke parameterisations of boron in the context of boron nitride nanotubes where the L-J terms were fit to reproduce water-nanotube interaction energies calculated using QM at the B3LYP level [194]. Here two different parametrisations of ϵ were derived with values of 0.095 and 0.453 kcal/mol [194]. The local environment of the boron atoms in borazine should correspond closely to that of the nanotube meaning the ϵ values should be comparable. Here we find that fitting to the liquid density and heat of vaporisation properties of borazine leads to an ϵ value of $\approx$ 0.021 kcal/mol whereas only fitting the density data of 132-benzodioxaborole leads to a value of $\approx$ 0.638 kcal/mol. While both of these values are of the same order as those reported previously the local environment of the boron atom in 132-benzodioxaborole is significantly different to

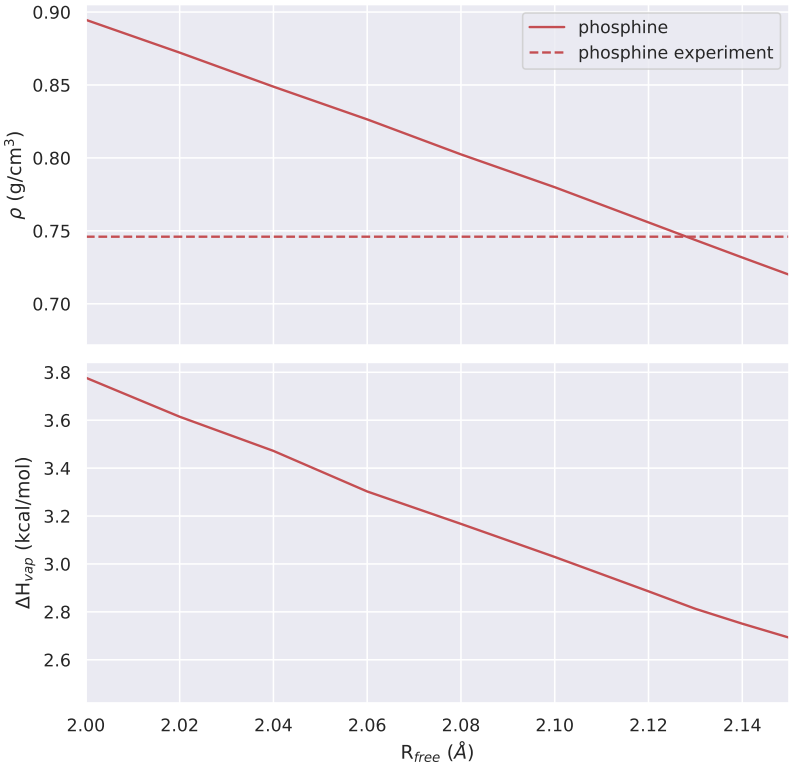

Figure 6.9: The predicted and experimental values for the density (top) and heat of vaporisation (bottom) are shown for phosphine.

that of the boron nitride nanotube. We also note the different charging methods used in previous studies which can effect the optimum choice of L-J parameters [195]. Furthermore we find that in the case of the example molecules used in fitting, any hydrogens directly bonded to silicon or boron obtain a substantial negative charge of around -0.22 e. As they were also not considered to be polar (that is their L-J terms were not transferred to the parent atom to allow for hydrogen bonding) this resulted in a large and potentially unphysical σ value of around 2.99 Å. Such surroundings were not sampled when fitting R_{free} for hydrogen. Thus the ability of the T-S relations to accurately model elements in a range of diverse environments may be limited and could be improved with the use of more complex functional forms or environment-specific element mappings. Such a study is beyond the scope of this thesis, but with the implementation of the described "skeleton" FF files, the groundwork is in place to begin this work.

In terms of the phosphorus only one target (phosphine) was used for fitting as the simple structure of the molecule should mean that the non-bonded parameters of the phosphorus atom contribute significantly to the accurate calculation of the density. The heat of vaporisation however was not available for this molecule to validate the density prediction but the estimated values are shown in figure 6.9.

6.3 Torsions

Dihedral FF terms act as a fine-tuning parameter helping to align MM predicted relative conformer energies with more expensive and accurate *ab intio* calculated values. The accuracy of a PES predicted for a flexible molecule using MM depends on the quality of the torsional parameters which often struggle with transferability within the class 1 functional form [113]. As the proper prediction of the conformational preferences of a molecule are particularly important in CADD, where we are concerned with estimating binding poses, torsional parameters are often the target for optimisation. This is particularly true when the non-bonded terms of the FF have been altered. As such the first version of QUBEKit provided a simple and automated torsion optimisation method that utilised the dihedral driver functionality of the BOSS MM package. However, this package is not widely available to the community which potentially limits adoption. Dihedral optimisation was also found to be one of the most time-consuming steps during parametrisation which affected the throughput of the parametrisation method. Thus to address these points and some other limitations of the method we have designed a new optimisation protocol using a range of open-source software including geomeTRIC [196], TorsionDrive (TD), psi4 [197], scipy and OpenMM [63] as our MM engine.

6.3.1 Computational implementation

We aimed to design a dihedral parameter optimisation method which fulfilled the following criteria: 1) hysteresis should be avoided during torsion driving where possible, 2) the method should allow greater user control over scan features such as changing the range of angles scanned, 3) it should account for geometry differences between predicted MM and reference QM structures, 4) it should offer similar or greater accuracy and speed compared with version 1 of QUBEKit. In regards to the first point, hysteresis can arise in both the QM and MM constrained torsion

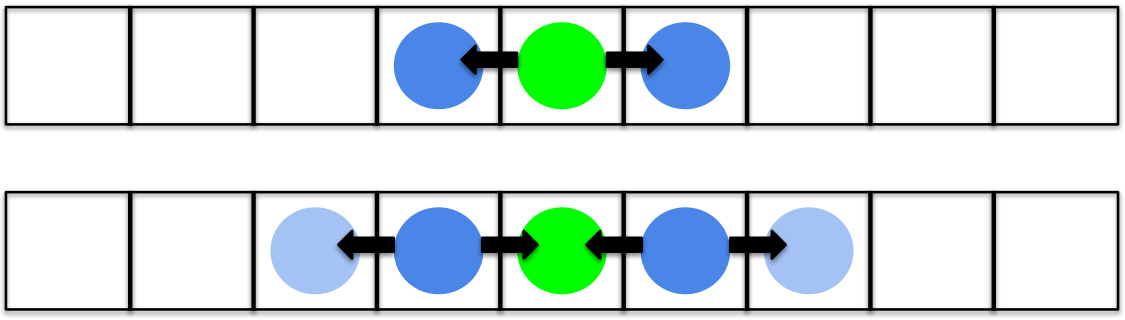

Figure 6.10: The initial grid point optimisation is performed in green which then activates neighbouring points shown in dark blue which then activate the next set of points show in light blue as well as repeating the original optimisation (green).

optimisations meaning the quality of the reference data and subsequent MM parameters can be diminished. This is often caused by the optimisation becoming trapped in local minima and is related to the choice of starting conformer which can cause asymmetries in the PES due to frustration. Due to the smaller size of the molecules studied in the benchmark testing in chapter 4, this was not an issue as it is more commonly associated with a complicated PES due to a molecule having multiple flexible bonds. However, the larger drug-like compounds studied in chapter 5 could show signs of hysteresis while scanning the ϕ_2 angle, which due to the symmetry of the molecule, should produce a symmetrical PES. A common way to account for possible hysteresis in these torsion profiles is to repeat a scan in the opposite direction and identify differences in the resulting geometries and PES [114]. To facilitate this we have chosen to use the TD software (https://torsiondrive.readthedocs.io/en/latest/) which allows users to perform multidimensional scans with a range of QM and MM engines. TD also uses a technique termed "*wavefront propagation*" which aims to avoid hysteresis by propagating the lowest energy conformers throughout the dihedral scan. In the standard 1 dimensional case, which is the focus of this work, the set of constrained optimisations to be done by TD are broken up into a grid as shown in figure 6.10. The initial input structure is then assigned to the closest grid point and optimisation is started. Once complete the energy is saved and the neighbouring grid points become active and start their respective optimisations from the optimised structure. Throughout the optimisation, new lower energy conformers are then found which reactivate neighbouring grid points causing them to start an optimisation again from the newest low energy structure, to reduce starting conformer dependence and importantly, any hysteresis. The effects

of "*wavefront propagation*" can be seen in figure 6.11 which shows the optimised structures obtained when scanning the ϕ_2 dihedral angle of compound 1 presented in chapter 5 using the BOSS and TD methods along with their corresponding PES. Here we see that the ϕ_1 angle gets stuck in a higher energy conformation in the second half of the scan when using the BOSS scanning method compared to using TD which is further shown in the PES plots in figure 6.11. Using TD also allows us to handle point 2 of our requirements as the range of the dihedral angles to be scanned is more easily controlled through an input file keyword which could allow users to also derive improper torsion parameters using a limited scan range.

After computing a reference PES surface around the chosen flexible bond, QUBEKit will automatically attempt to optimise the corresponding parameters of the dihedral. This can typically involve up to six independent dihedral parameter sets each composed of four Vn coefficients (see equation 2.24) all fit simultaneously. Before fitting we first determine the minimal number of parameters which are to be optimised, as some of the dihedral terms involved in the torsion may be composed of equivalent atoms as defined by the element and local environments. To cluster the torsions we begin by assigning each torsion an identifier pattern which is based on the graph symmetry of the molecule much like the conventional atom typing used by transferable FFs. However, as the types are created specifically for the molecule we have the freedom to assign an arbitrary pattern as we are not limited by a predefined combination of types. While in the majority of cases this symmetry approach reproduces the expected torsion clusters, such as those shown for ethanol in figure 6.12 and table 6.3, it can also identify when more types should be introduced to offer greater parameter freedom. We then optimise the torsional parameters by minimising the same objective function shown in equation 2.41. The MM single point energies are rapidly calculated using OpenMM at the reference QM optimised geometries (themselves computed using psi4 and TD) which ensures that the PES is being aligned to the same geometries. Once this first step has converged we then perform a full relaxed MM torsion scan of the molecules using TD with the same convergence criteria as those used for the QM scan. This allows us to fully assess the quality of the optimised parameters in their ability to recreate the QM predicted PES and optimised geometries. An overall iteration error is then produced composed of the energy error which is the objective function calculated on the new relaxed MM surface (RMSE) supplemented by the root-mean-square deviation (RMSD) between the atomic coordinates of the structures, $E = (RMSE/kcal/mol) + (RMSD/\mathring{A})$.

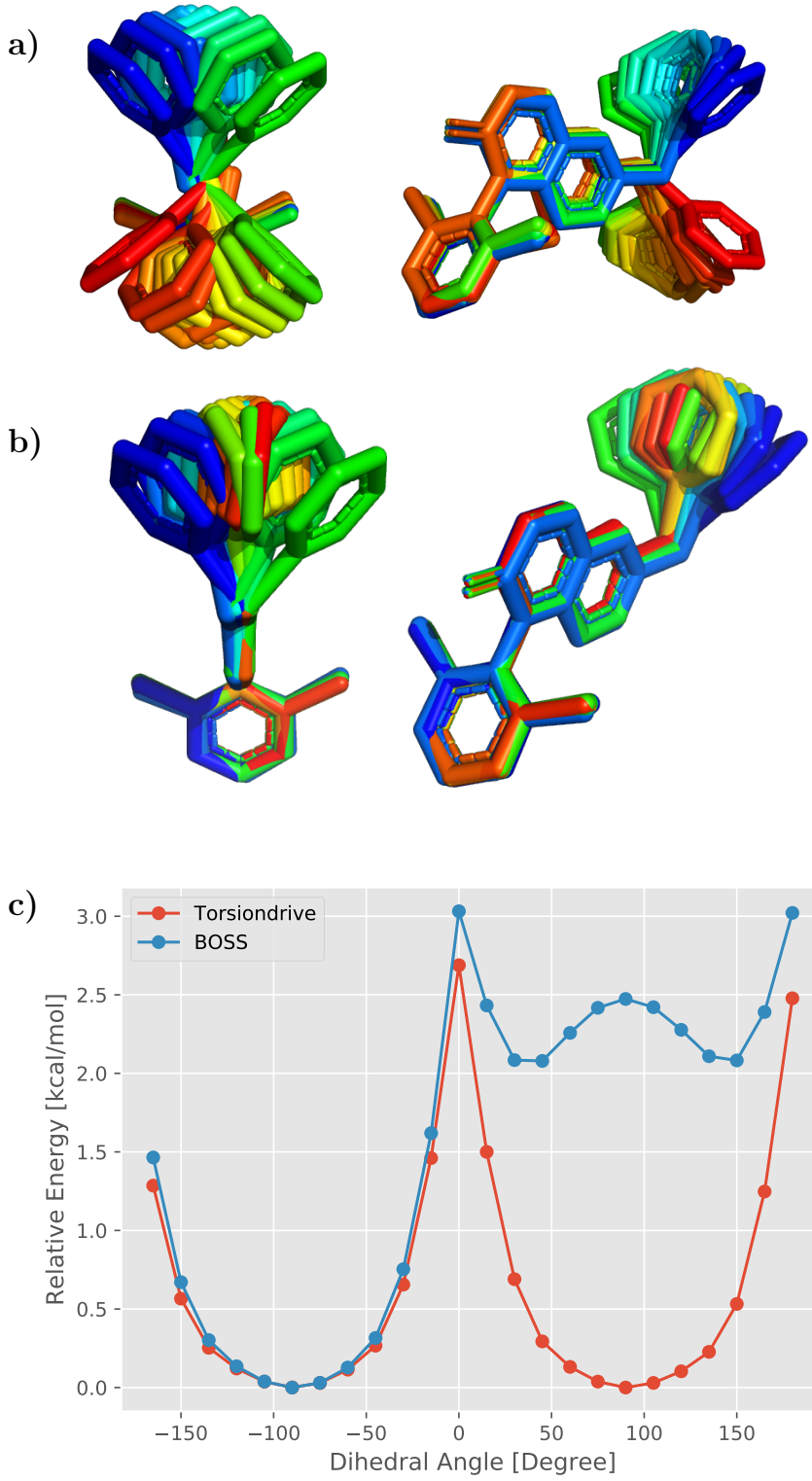

Figure 6.11: Optimised structures of ligand 1 (chapter 5) obtained using the OPLS FF and a) BOSS, b) torsiondrive constrained dihedral optimisation software. The structures are aligned to the first grid point and coloured from initial optimisation (blue) to final (red). c) The corresponding PES of the scans is also shown.

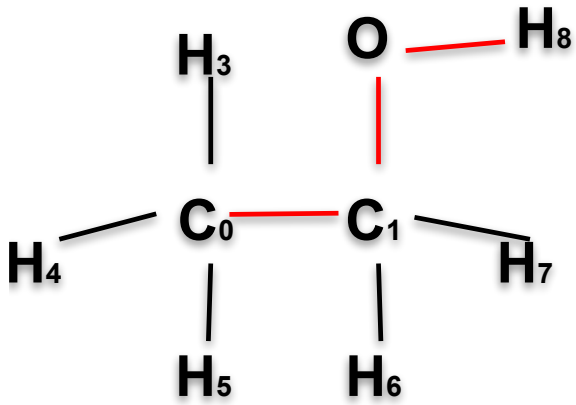

Figure 6.12: Ethanol with the torsion angle to be fit highlighted in red.

Torsion	QUBE identifier	Parsley type
O-C_1-C_0-H_3	1	t9/smirks [#1:1]-[#6X4:2]-[#6X4:3]-[#8X2:4]
O-C_1-C_0-H_4	1	t9/smirks [#1:1]-[#6X4:2]-[#6X4:3]-[#8X2:4]
O-C_1-C_0-H_5	1	t9/smirks [#1:1]-[#6X4:2]-[#6X4:3]-[#8X2:4]
C_0-C_1-O-H_8	2	t85/smirks [#6X4:1]-[#6X4:2]-[#8X2H1:3]-[#1:4]
H_6-C_1-O-H_8	3	t84/smirks [*:1]-[#6X4:2]-[#8X2:3]-[#1:4]
H_7-C_1-O-H_8	3	t84/smirks [*:1]-[#6X4:2]-[#8X2:3]-[#1:4]
H_3-C_0-C_1-H_6	4	t3/smirks [#1:1]-[#6X4:2]-[#6X4:3]-[#1:4]
H_3-C_0-C_1-H_7	4	t3/smirks [#1:1]-[#6X4:2]-[#6X4:3]-[#1:4]
H_4-C_0-C_1-H_6	4	t3/smirks [#1:1]-[#6X4:2]-[#6X4:3]-[#1:4]
H_4-C_0-C_1-H_7	4	t3/smirks [#1:1]-[#6X4:2]-[#6X4:3]-[#1:4]
H_5-C_0-C_1-H_6	4	t3/smirks [#1:1]-[#6X4:2]-[#6X4:3]-[#1:4]
H_5-C_0-C_1-H_7	4	t3/smirks [#1:1]-[#6X4:2]-[#6X4:3]-[#1:4]

Table 6.3: All dihedral parameter sets describing the highlighted main flexible bond of ethanol, figure 6.12, are shown along with their QUBE and parsley assigned types.

While the instantaneous effect of changing the parameters on the RMSD is not directly calculated during optimisation due to the time cost associated with performing a torsion scan thousands of times, the effect does guide the choice of the optimal parameters due to its inclusion in the overall error. The next iteration then starts at the resulting set of geometries computed using the previous set of optimised torsion parameters. The parameters are then optimised again to align MM energies of these structures to the original QM PES with a small regularisation penalty of 0.15. While the geometries will be slightly different from the QM reference structures due to the simplified approximation of the MM FF and the quality of the parameters, it is these geometries which will be sampled in a simulation and therefore they must

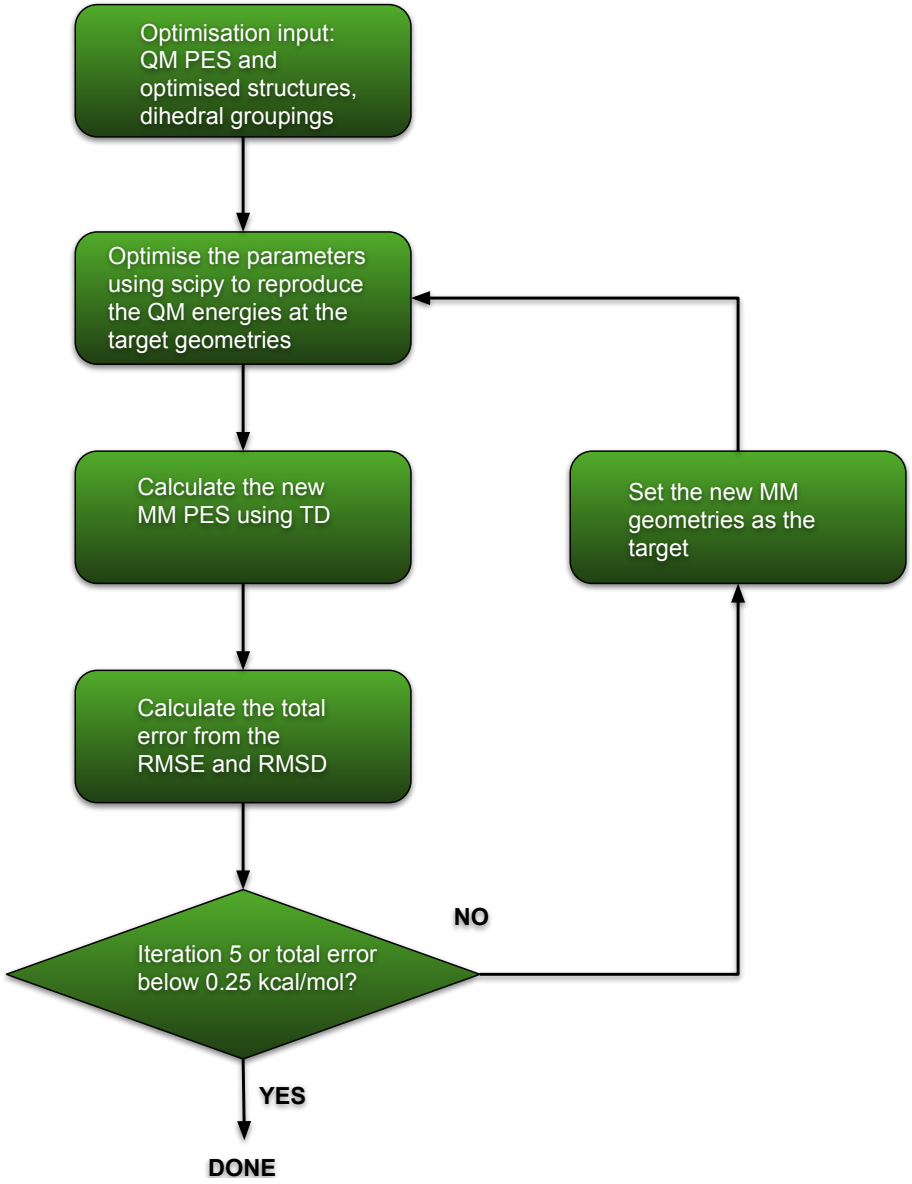

Figure 6.13: A summary of the new torsion parameter optimisation routine.

produce an accurate PES. The regularisation parameter is chosen to allow some refinement of the parameters between iterations but is restrictive enough to prevent the parameters causing a large RMSD which is measured only once per iteration. This cycle is then repeated for a maximum of five iterations or until the error reaches the threshold which is set to 0.25. If after five iterations the error has not reached the threshold the parameters which produced the lowest overall error are saved and a new FF XML is written for the molecule. This new optimisation routine is also summarised in figure 6.13.

Torsion	QUBE identifier	Parsley type
Br-C-O-H	1	t84/smirks [*:1]-[#6X4:2]-[#8X2:3]-[#1:4]
H_1-C-O-H	2	t84/smirks [*:1]-[#6X4:2]-[#8X2:3]-[#1:4]
H_2-C-O-H	2	t84/smirks [*:1]-[#6X4:2]-[#8X2:3]-[#1:4]

Table 6.4: The three dihedral parameter sets describing the main flexible bond of bromomethanol, figure 6.14, are shown along with their QUBE and parsley assigned types.

As the torsion identifiers are assigned to every rotatable bond of a molecule we are also able to simultaneously fit all symmetry equivalent torsions to those in the targeted flexible bond. This is achieved by creating a parameter vector initialised to [0,0,0,0] for each QUBE assigned type of torsion that is to be optimised corresponding to the four Vn components of the truncated Fourier series in equation 2.24. Furthermore due to the careful computational design of the problem, we are also able to take advantage of a wide range of fast and efficient parameter optimisation routines such as those found in the scipy package. So far we have implemented the Nelder-Mead simplex method [198] (N-M), Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm and the differential evolution (DE) global optimisation method [199]. Typically these methods can involve hundreds of objective function evaluations, which limited our ability to use them in the previous iteration of QUBEKit in conjunction with the BOSS software. However, thanks to the ability of OpenMM to rapidly calculate the single point energies of the molecules we can do thousands of evaluations during optimisation at a much reduced computational cost allowing for a greater search over parameter space and potentially better quality parameters.

6.3.2 Results

Test case: Bromomethanol

While being a relatively simple molecule with only one rotatable bond bromomethanol (figure 6.14) has been identified as having a PES which is poorly predicted when parameterised using GAFF [184]. Thus we aim to see if the new torsion parameter optimisation method described above can be used to improve the accuracy of the MM predicted energy surface. During parameterisation, the symmetry-based torsion pattern clustering used in QUBEKit identifies two different dihedral types, as expected. Whereas the GAFF based SMIRKS FF (parsley) of the OFF assigns the same parameters to all of the torsions as shown in table 6.4, which in this

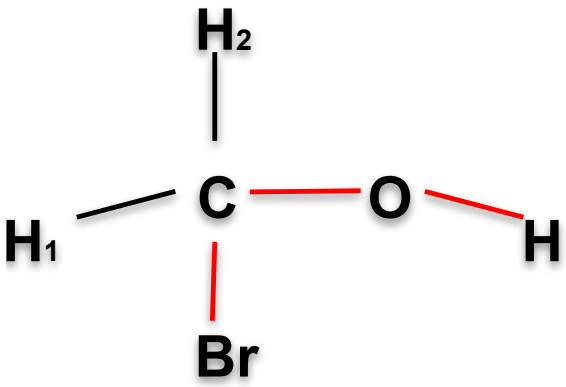

Figure 6.14: Bromomethanol with the torsion angle highlighted in red.

case is the cause for the poor FF performance and can also be shown to limit the maximum achievable accuracy. To demonstrate this we have optimised the torsion parameters corresponding to the main flexible bond in bromomethanol for three possible scenarios and the final results are shown in figure 6.15. The graph overlays the reference QM data (dots) along with the PES surface predicted by the original GAFF parameters (dashed line) which matches the observation in ref 184. The blue line corresponds to the case where all torsions are deemed to be equivalent (that is torsion parameters are optimised using the GAFF atom types). On the other hand the orange line uses the QUBE internal torsion clustering to separate the parameters into the groups shown in table 6.4, but only optimises the parameters corresponding to the Br-C-O-H term, replicating the case where a more specific FF term has been added. In the last case the green line represents the normal execution of the new procedure in QUBEKit whereby the torsions are again clustered into two groups but all of the parameters are allowed to optimise simultaneously. Thus the lacking specific torsional parameter in the GAFF FF seems to be one cause of the poor performance and even with a completely unrestrained optimisation this limits the maximum achievable accuracy.

This test case then highlights the importance of clustering the torsions into logical groups which lead to the improved accuracy of the FF due to the increased freedom of the parameters. We also see that under normal application with default parameters QUBEKit can significantly improve the quality of the torsion parameters using this new optimisation routine and that it can even be used to fit a single specific FF parameter resulting in vastly improved torsional parameters. It should also be noted that the choice of optimiser can have a significant effect on the results as

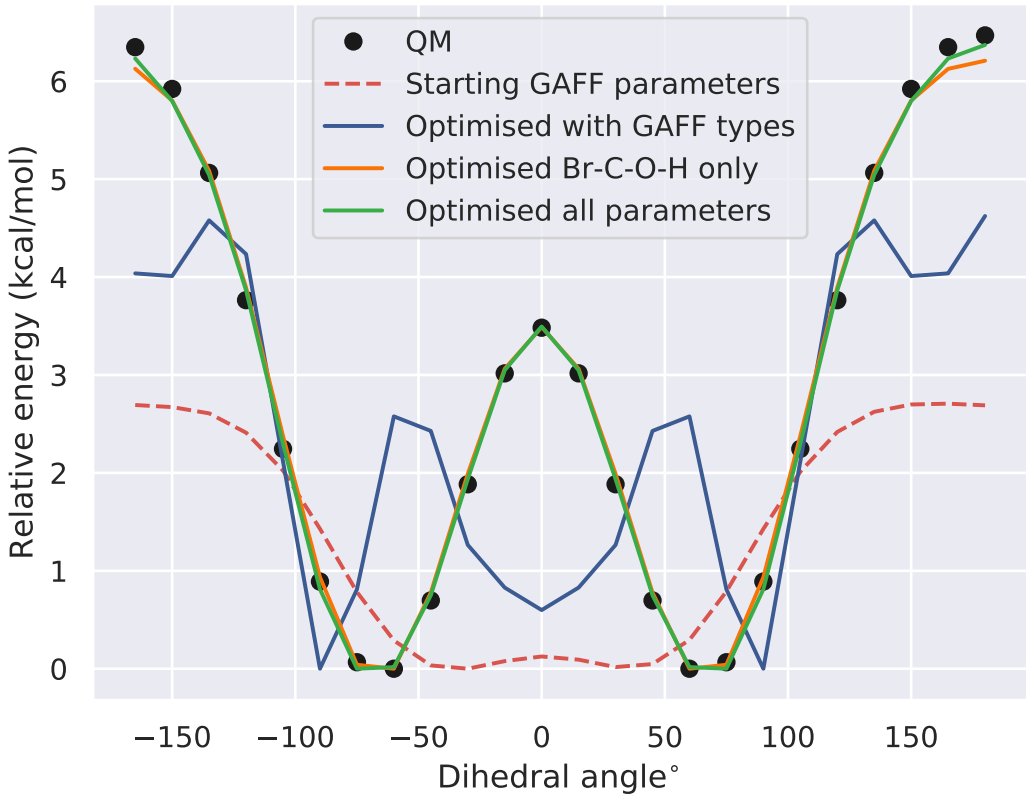

Figure 6.15: Optimisation of the bromomethanol main torsion for three different combinations of parameter clustering shown with the QM reference data and the starting PES. The blue line refers to GAFF clustering where all types are the same, orange represents the case of QUBE symmetry based clustering but only optimises the Br-C-O-H parameters. The green line corresponds to QUBE clustering but with both types of parameters being allowed to optimise simultaneously.

some can become stuck in local minima. In this example, the N-M method was used and performed a sufficiently thorough search resulting in low RMSEs of 0.157 and 0.137 kcal/mol for the case of the single parameter and full QUBE optimisations respectively. However, it was found that even with the use of the QUBEKit defined torsion clustering the BFGS optimisation algorithm could become stuck resulting in less satisfactory fitting as shown in figure 6.16 which compares the breakdown of the final errors of the three different optimisation methods. Clearly, the choice of parameter optimisation algorithm can significantly affect the performance of the method, hence QUBEKit was designed to allow users to quickly change and repeat fittings with minimal effort as well as easily incorporating new optimisation algorithms.

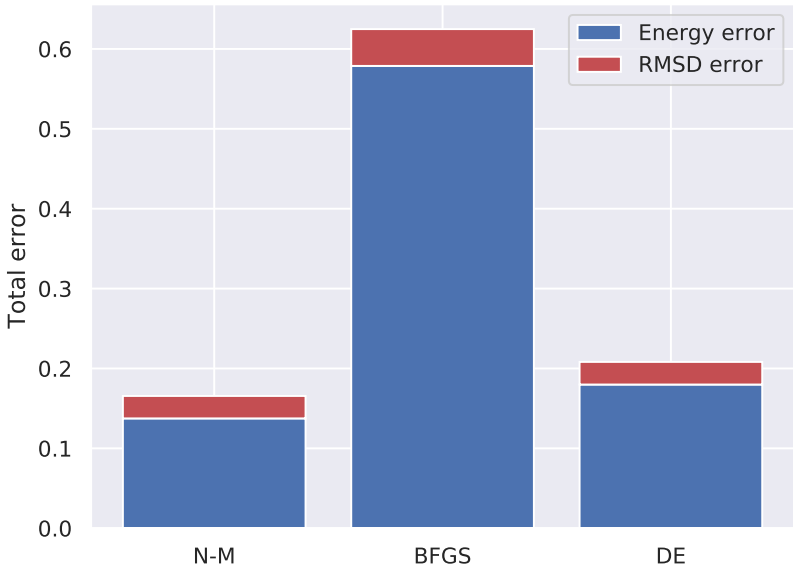

Figure 6.16: The finial optimisation error is shown in terms of its component parts for the three different optimisation methods when used on bromomethanol.

Test case: eMolecules

To fully investigate the robustness of the new parameter optimisation protocol, we decided to look at a collection of 43 molecules taken from eMolecules that have been identified by Pfizer as having a substantially different PES when calculated with OPLS3e compared to a QM reference [200]. In some cases this is due to significant nitrogen geometry rearrangement during the rotation of a flexible bond. The molecules are shown in figures 6.17 and 6.18 with the main flexible bond targeted for optimisation highlighted in red. Overall this test set represents a routine parametrisation challenge in the early stages of a drug discovery campaign as the molecules are fragment-like in terms of size and complexity and contain at least 1 rotatable bond. Following this example scenario, all molecules were entered into QUBEKit as SMILES strings and initially parametrised using the OFF toolkit before having their bond-stretching, angle-bending and all non-bonded terms replaced using the methods described in section 2.3.2. All QM geometry optimisations were performed with Psi4 [197] at the B3LYP level using the 6-31G** basis set in order to reduce computational cost, as we are less concerned with the accuracy of the reference calculations, but more in our ability to reproduce them

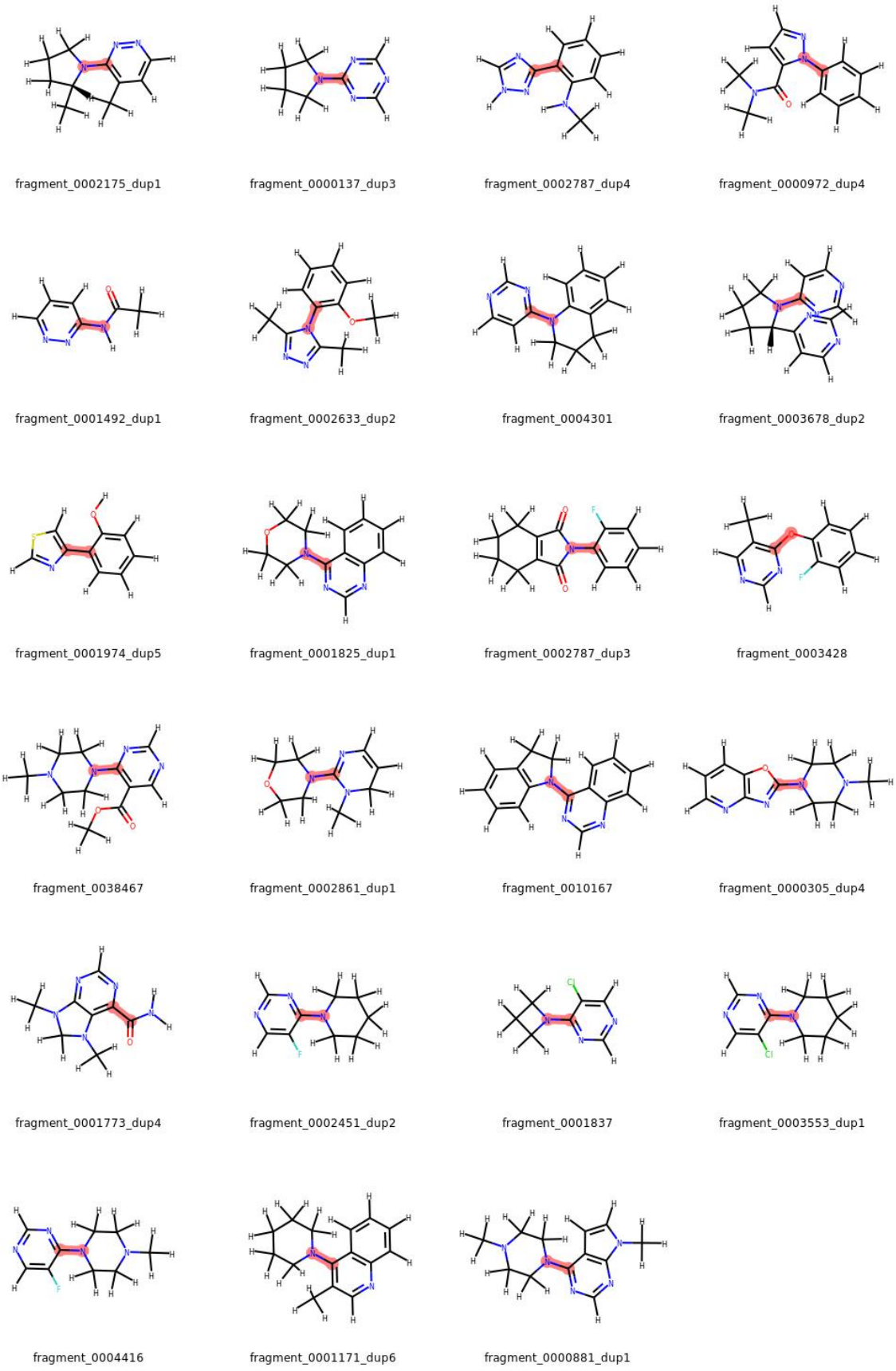

Figure 6.17: The first 23 molecules from the test set with the flexible bond targeted for optimisation highlighted in red.

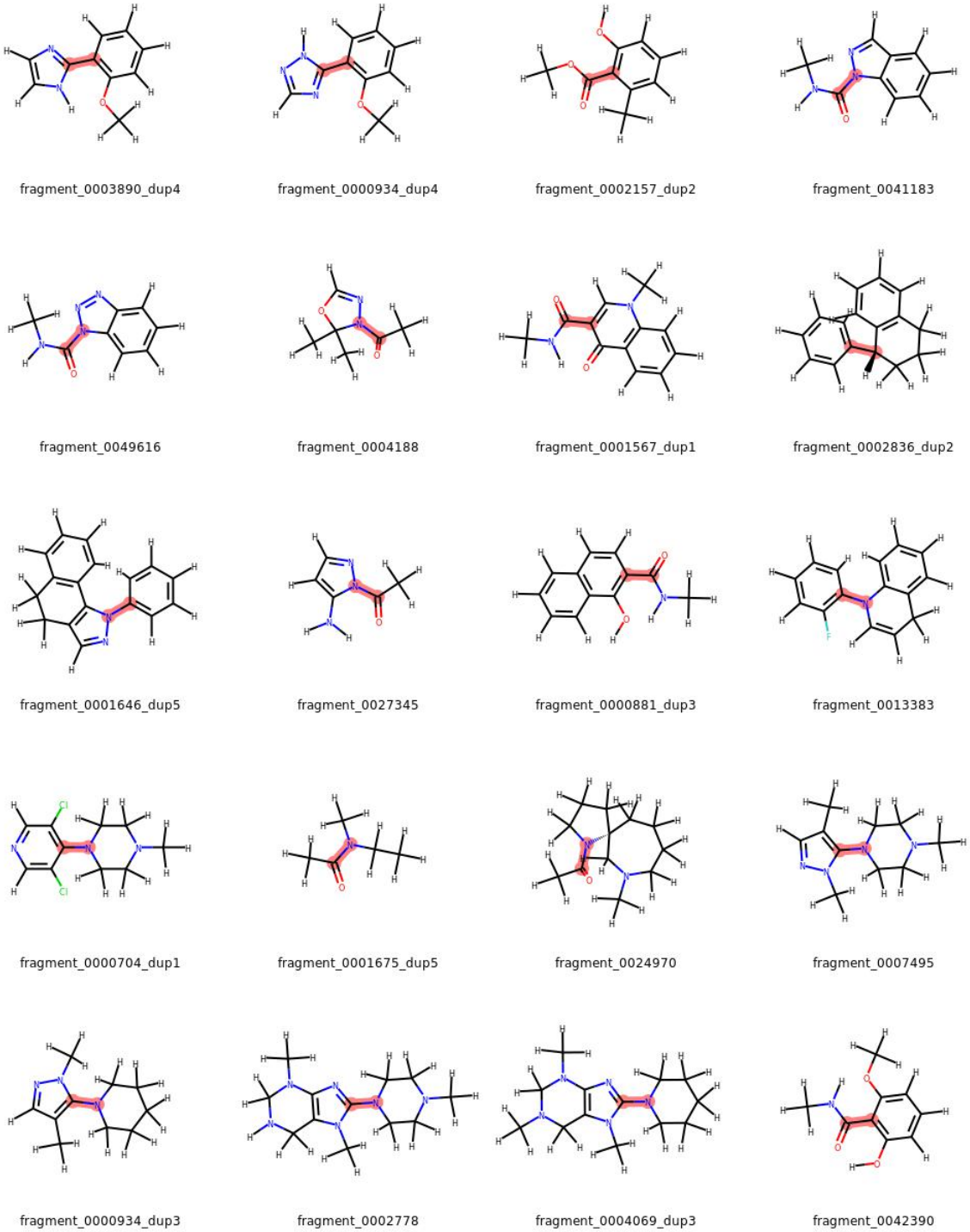

Figure 6.18: The last 20 molecules from the test set with the flexible bond targeted for optimisation highlighted in red.

using MM. The highlighted rotatable torsions were driven from -165° to 180° in 15° increments using TD and were optimised with no temperature weighting (T= ∞), no initial regularisation and a torsion parameter absolute value limit of 20 kcal/mol to reduce the parameter search space. The optimisation was performed a total of three times, each using one of the currently available parameter optimisation algorithms to establish a benchmark level of performance that can be expected during routine application. Figures 6.19, 6.20 and 6.21 show a breakdown of the final error composed of the RMSE and RMSD contributions after optimisation using the N-M, BFGS and DE methods respectively.

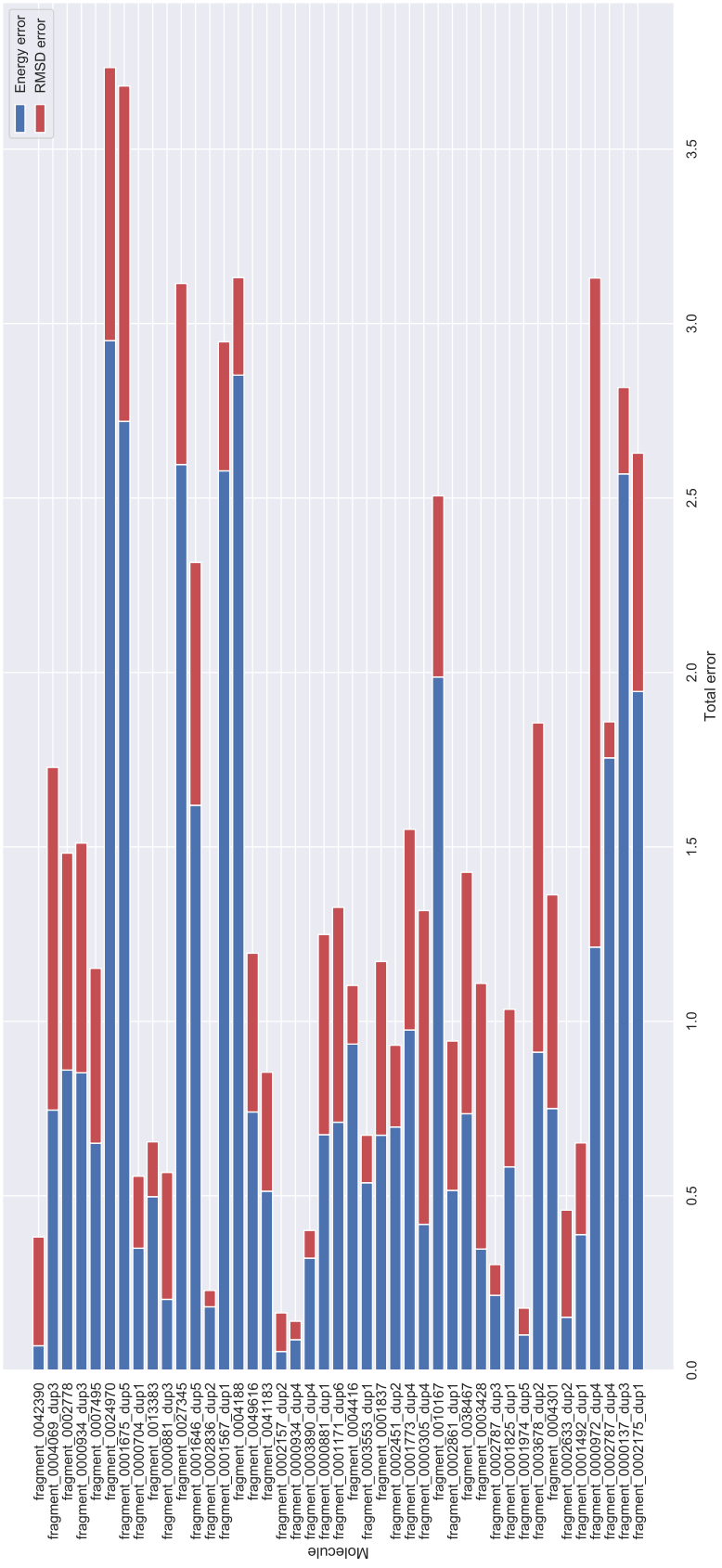

Figure 6.19: The final optimisation errors obtained using the N-M algorithm on the eMolecules test set.

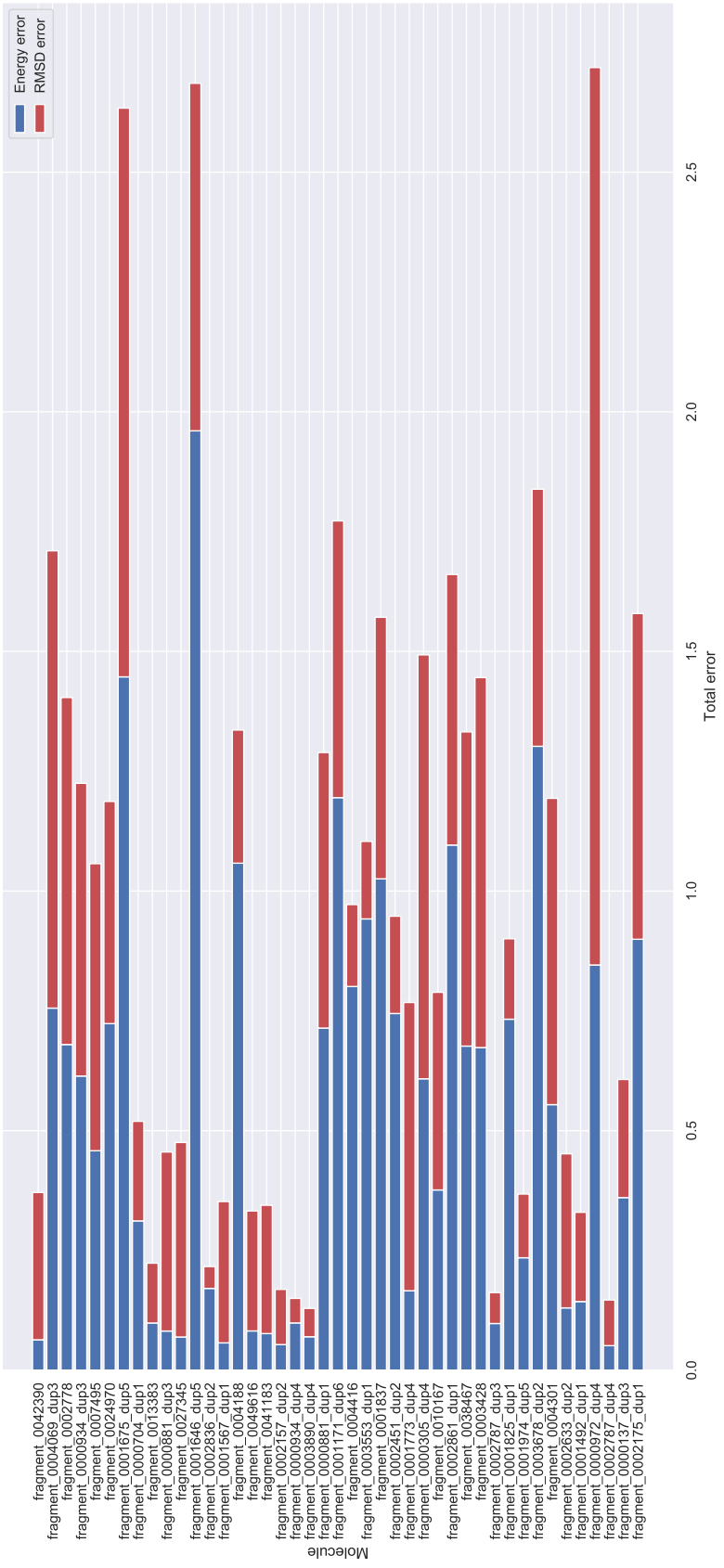

Figure 6.20: The final optimisation errors obtained using the BFGS algorithm on the eMolecules test set.

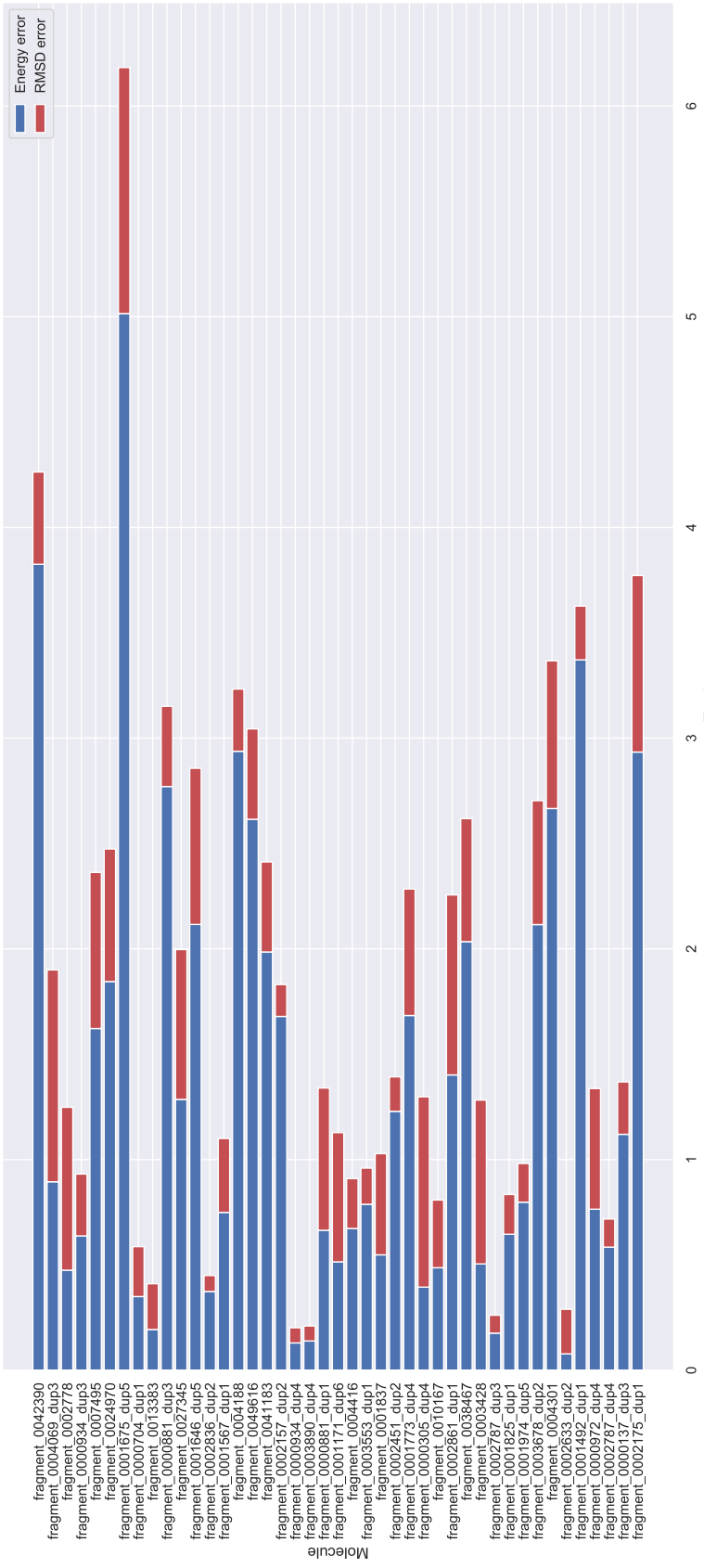

Figure 6.21: The final optimisation errors obtained using the DE algorithm on the eMolecules test set.

	Optimisation method		
	N-M	BFGS	DE
Mean RMSE (kcal/mol)	0.960	0.542	1.345
Mean RMSD (Å)	0.472	0.444	0.455
Mean timing (s)	5062	11894	21288

Table 6.5: The mean RMSE, RMSD and optimisation timings for each optimiser on the eMolecules test set.

From the graphs, we can see that the choice of optimiser largely affects the performance of the optimisation resulting in significantly different final errors for the same molecule, this is further summarised in table 6.5 which compares the mean final errors and timings for the optimisations. Despite each of the methods achieving a roughly similar RMSD of around 0.45 Å, which is measured between the MM predicted structures and the reference QM calculations, we see a larger variation in the PES error ranging from 1.35 to 0.55 kcal/mol. However, all optimisers show a large improvement over the mean initial RMSE in the PES which was found to be 2.28 kcal/mol using the transferred torsion parameters. Overall we find that BFGS optimiser results in the lowest combined errors in 25 of the 43 cases and therefore the lowest average error of the three methods and would be a logical first choice when using the new routine. The average timings also show a large variation between methods with BFGS taking roughly twice as long as the N-M, and the DE method taking twice as long again (table 6.5). In some cases, however, the choice of optimiser was irrelevant as they all attained a very similar final error despite the very different combinations of parameters. Molecules 0003890_dup4 and 0000934_dup4, in particular, had very low optimised errors which can be attributed to their less flexible simpler structure and corresponding PES, meaning that the fastest optimiser would be the best case for such molecules.

Molecules containing multiple flexible bonds were also included within the test set and were found to be poorly fit when only optimising one of the dihedrals which was mainly due to large RMSD contributions involving the other flexible bonds. One of the most notable examples of this is molecule 0001675_dup5 for which the BFGS optimiser achieved a final objective function value of 2.634 with an RMSD contribution of around 45% which remained consistent throughout optimisation. Figure 6.22 shows the extent of these geometry deviations via the alignment of the QM reference structures (grey) at torsion angles of a) -90° and b) 0° with their

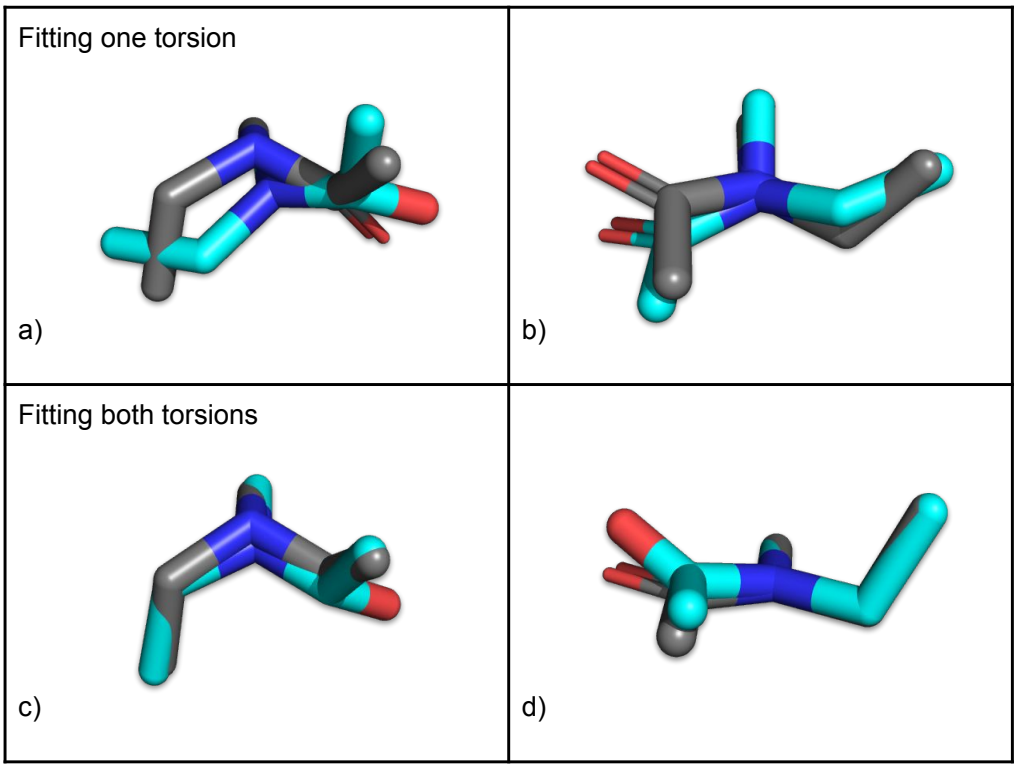

Figure 6.22: Overlay of the QM predicted optimised structures (grey) of 0001675_dup5 with MM analogues using BFGS (blue) to optimise one of the flexible bonds extracted from a TD simulation at angles of a) -90° and b) 0°. c) and d) correspond to angles -90° and 0° but after using BFGS to optimise both flexible bonds in series.

MM predicted analogues (blue). On analysis the structures in figure 6.22 it is clear that the fitting may improve if the second torsion was also optimised, hence the molecule was re-optimised using QUBEKit to fit both flexible bonds sequentially. This resulted in an improved final combined error of 0.663 which is actually less than the average across the other less flexible molecules in the test set, and example structures from the sequential optimisation are also shown in panels c and d of figure 6.22. Now we can see that the reference and MM structures align much more closely with a final mean RMSD of 0.551 Å and remaining large differences in geometry arise from pyramidalisation of the nitrogen and an out-of-plane bending of the oxygen atom which are both influenced by improper torsions in the FF. The parameters used for these terms were transferred from the Parsley FF and apply a low barrier potential of around 1 kcal/mol to both instances which may be one cause of the remaining error. Improper torsions in general could be a culprit for structural errors predicted by FFs as there are a very limited number of types (only 4 currently

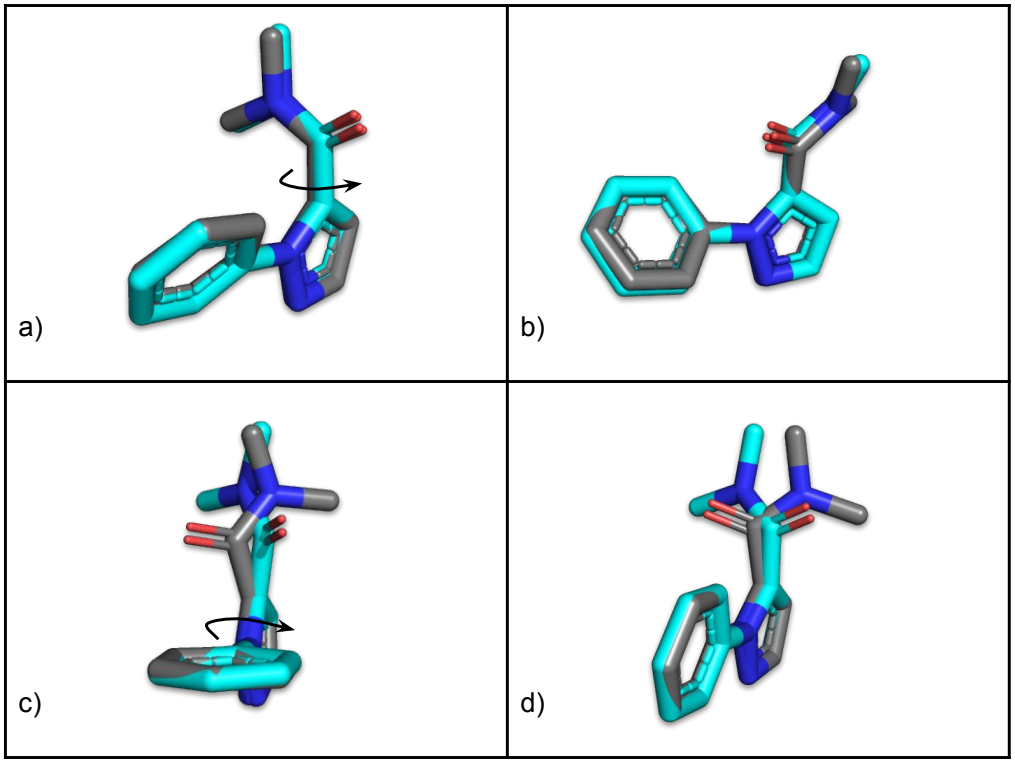

Figure 6.23: Overlay of the QM predicted optimised structures (grey) of 0000972_dup4 with MM analogues (blue) extracted from a TD simulation at angles of -90° (a, c) and 180° (b, d). Where the top (a and b) and bottom (c and d) correspond to the fitting of the two separate main flexible bonds.

in the Parsley FF) and their effect has not been extensively studied in the class one functional where it is assumed that angle terms will mitigate most errors. We may well then need to also optimise such terms to help minimise the RMSD. While the proper torsion fitting method outlined here could be extended to such cases this is beyond the scope of this thesis and a subject for future investigation.

In some cases with multiple flexible bonds however, sequentially fitting the dihedrals was not able to improve similarly large RMSD issues. Figure 6.23 shows the alignment of 0000972_dup4 QM (grey) and MM (blue) structures from TD at angles of -90° (a, c) and 0° (b, d) during the fitting of two main rotatable bonds. Here we originally tried to optimise the torsion involving the phenyl ring and found that the second flexible bond of the molecule would allow a significant change in geometry that the FF could not recreate. This was also found to be true during the sequential optimisation of both bonds. However, this seems to be caused by the symmetry of the molecule as the QM optimisation and MM optimisation have

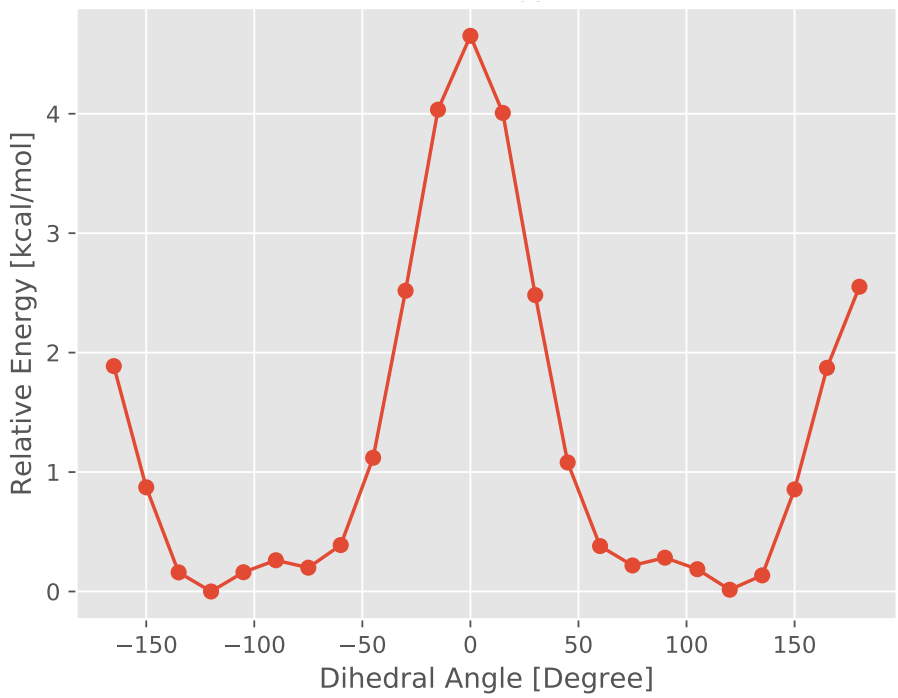

Figure 6.24: The symmetric PES from scanning the top flexible bond in 0000972_dup4 while holding the phenyl torsion at -90° using the sequentially optimised parameters and OpenMM.

favoured mirror image poses for the top of the molecule as can be seen in panels c and d of figure 6.23. To confirm this, another TD simulation was performed holding the phenyl torsion fixed at -90° while fully scanning the other flexible bond and the corresponding symmetric PES is shown in figure 6.24. The large RMSD error in this case is then erroneous and an artefact of the symmetry which should be further considered in future improvements to avoid such cases. The final RMSE for the molecule after sequential fitting is 0.315 kcal/mol which is well within the expected performance range.

In general the remaining errors not caused by symmetry issues and that can not be fixed by fitting multiple flexible bonds maybe an indication that the restrictive functional form of the class one FF is limiting the maximum achievable accuracy. This has been long recognised in spectroscopy, where coupling or cross-terms between angle-bending and torsion strain contributions were found to be essential to accurately represent torsional energetics [201]. Before moving to more advanced functional forms, allowing other terms to optimise such as the bond-stretching, angle-bending and even short-range non-bonded repulsive parameters during fitting may help improve accuracy, but could be detrimental to other properties and thus has been

avoided in this method. However with the careful weighting of multiple QM data points, such as the PES and normal modes, future optimisation strategies could simultaneously increase PES accuracy while maintaining performance in other critical aspects [20] and will need investigation.

6.4 Conclusions

To aid the widespread adoption of the QUBE FF we have implemented a series of new features into QUBEKit-V2 ahead of its release. These features expand on the capabilities of the original software and allow users to create template FF files which should allow for easier custom parametrisation and development. We have demonstrated how this method can be used to extract known parameters for molecules containing boron and silicon which would cause traditional transferable parametrisation routes to fail. As such, most MM simulations of systems containing boron and silicon, concerning porous materials and semiconductors, have required environment-specific parameters [202] and potentials [192] derived from QM calculations similar to the QUBE FF. By incorporating these elements into the QUBE FF we allow the automated derivation of MM parameters for such molecules within the class 1 functional form, reducing the parametrisation complexity of studying these systems. Furthermore, provided users have experimental data points to fit the required R_{free} parameters, they can potentiality extend or re-fit the QUBE FF themselves with minor effort and minimal modification of the source files. While the reduced empiricism of this model does reduce the complexity of fitting it can also limit the accuracy in some regards as we have seen with the silicon and boron containing molecules investigated here. Due to the sparse amount of experimental data available for the elements the environments of the atoms can differ significantly leading to very different L-J parameter requirements to align the predicted and experimental properties. Further work is needed to study the suitability of the FF functional form, and T-S scaling relations used for boron and silicon before widespread adoption of these parameters. Future work should also investigate the suitability of the suggested R_{free} values derived in this section for other thermodynamic properties such as the hydration free energy, which is a commonly used performance metric. However due to the limited amount of experimental data in this area, benchmarks may have to move straight to a drug discovery based setting similar to that presented in chapter 5 for which a small range of binding affinity data is

available [187, 188]. Hopefully the bonded parameters presented here are also helpful for the creation of general transferable FF parameters which should accelerate the rate at which we see silicon and boron containing molecules studied in CADD.

We have also presented our new open-source torsion parametrisation scheme which has many benefits over its predecessor such as: the ability to control the scanning range, the use of an open source QM engine to calculate the reference data and the ability to use different optimisation algorithms. The symmetry based parameter clustering technique we have introduced has also been shown to recover common FF dihedral types and importantly can indicate when to introduce new terms as initial FFs may lack specific coverage in some areas which, as we have shown, can potentially limit accuracy.

The choice of optimiser has also been shown to have a substantial effect on the resulting parameters and interestingly wider parameter space searching is often computationally expensive and rarely results in better parameters than the faster N-M and BFGS methods. Importantly we provide a wide range of optimisation routines and community best practices such as temperature weighting and regularisation that can be used on a case by case basis to determine the optimal parameters. Under normal automated operation we achieve the lowest RMSD between QM reference and MM predicted structures and RMSE between the QM and MM PES when using the BFGS optimisation method and have thus made this the default. The low mean error achieved with the BFGS method is very encouraging considering the complexity of the molecules presented in the test case, and the fact that we are only optimising the torsional terms within the standard functional form. Another limiting factor during the fitting may be due to the equilibrium bond lengths and angles along with the modified Seminario predicted force constants as they are all based on the fully optimised geometry of the molecule. This may over constrain the bonds and angles which may deviate from these ideal values during the rotation of the flexible bonds in the QM reference structures. The expansion of the FF functional form to include cross terms could account for these interdependencies in internal coordinates [47], but is beyond the scope of this thesis.

Chapter 7

Conclusions

The insight that can be gained from the atomistic study of complex biological systems has led to MM simulations becoming a vital component of CADD. General transferable FFs have a long-standing successful history in this domain due to their ease of use, years of refinement and wide range of potential application. However, due to the vastness of chemical space the transferability of the parameters, derived from experimental and QM data for a small representative set of molecules, has come into question. Furthermore, due to the size of these parameter libraries and underlying interdependencies (due to the iterative fitting methods traditionally used), accuracy improvements are incremental, labour-intensive, error-prone and beginning to stagnate.

A fundamentally different approach to this "parametrise once and transfer" philosophy is that of a transferable parametrisation strategy based on the most fundamental property of the molecules, the electronic structure. With increasing access to high-performance computing facilities and advances in DFT, accurate electronic structure calculations have become routine and can be applied to a wide range of systems prospectively. Thus without any prior chemical knowledge of a system, one can predict its fundamental properties using QM, and with the use of an appropriate parameter derivation method, one can infer accurate and system-specific MM parameters.

In this thesis, we have brought together a collection of such methods which have been newly developed to derive almost all of the required FF parameters to model a system using the class 1 additive FF functional form, known as the quantum mechanical bespoke (QUBE) FF. QUBE relies on the use of the modified Seminario method to derive all bond and angles terms from the QM optimised structure and

Hessian matrix. All non-bonded parameters (charges, L-J terms and virtual-sites) are derived from a single ground-state electron density via AIM partitioning. As these methods are often developed in isolation and involve various programs we aimed to streamline the process as much as possible by reducing user input and automating their derivation. This resulted in the creation of QUBEKit, an open-source software package written in python that automates the calculation of QM reference data and subsequent derivation of MM parameters reducing the complexity of parametrisation to that of a transferable FF.

In chapter 4 we set out to validate the performance of the QUBE FF when modelling small organic molecules using standard FF performance metrics such as the calculation of pure liquid densities, heats of vaporisation and free energies of hydration. Overall we found that the QUBE FF achieves competitive accuracy with established transferable FFs such as OPLS on a test set of over 100 molecules, which is promising as OPLS has been extensively fit to reproduce such experimental data (see section 4.3.1). We also see that due to our inclusion of an implicit solvent model during the calculation of the electron density we can explicitly capture solvent polarisation effects into our point charges without the need for common post-processing techniques such as charge scaling or bond charge corrections. Local environment polarisation effects are also thought to affect the strength of the van der Waals interactions and are traditionally represented by a range of atom types for each element corresponding to the different combinations of the ϵ and σ parameters. However, as we also derive the L-J terms directly from the same electron density using the T-S method the parameters naturally capture this response to their environment. We also found that the inclusion of virtual-sites was vital in cases where the electron density showed signs of anisotropy and therefore could not be faithfully represented by a single point charge. Vitally all of these non-bonded parameters are derived in an automated fashion from a single QM calculation which demonstrates the range of information that can be gained routinely from molecule specific calculations. Despite the careful consideration of each of these properties, general transferable FFs were still found to achieve a greater level of accuracy in regards to hydration free energies. The prediction of this property is largely regarded as an estimation of the performance of the FF in a CADD setting due to its links with the binding free energy and so any possible accuracy improvements are vital. To this end, we identified multiple avenues of follow up work that could potentially improve the accuracy of the QUBE FF within the current functional form, including a QUBE

specific water model, due to the improvements seen by using the TIP4P-D model, replacement of the AIM partitioning method and the effect of the combination rules used. Moving beyond the current functional form it is thought that a more physical description of the van der Waals interactions including higher-order terms may significantly improve accuracy as has been shown for a set of 11 alkanes [75]. Vitally we have provided a parametrisation platform in QUBEKit that can be easily extended and adapted to facilitate the investigation of such methods.

In chapter 5 we looked at the suitability of the QUBE FF in a typical drug design setting via the retrospective calculation of the relative binding free energies of 17 drug-like inhibitors of p38α MAP Kinase. This example system has been widely used in methodology and FF performance benchmarks and represents a typical lead-optimisation situation in which a general transferable FF would normally be used. Here we also employed the biological variation of the QUBE FF which can be used to model proteins, it consists of a library of custom bonded parameters whose accuracy has been extensively validated elsewhere [60]. The non-bonded parameters, however, are derived specifically for the system under study using the same AIM based scheme employed in QUBEKit to ensure compatibility of the receptor and ligand terms. Due to the drug-like nature of the molecules, in terms of size and flexibility, their parametrisation is a significant challenge for any QM based derivation method. The QUBE FF derived for each of the molecules, however, seemed to recreate the QM predicted PES around the two pose defining flexible bonds well, along with the relative conformer energies sampled within protein-ligand and ligand-solvent simulations. This resulted in satisfying correlations between MM and QM predicted relative energies that are consistent with the values achieved for small molecules (see section 4.3.4). This is also further validated by considering the predicted binding pose preferences of the molecules, in particular, we saw that the QUBE FF recovered an experimentally confirmed configuration despite starting from a different structure (see section 5). This is reassuring when considering the calculation of binding affinities as their accuracy depends on the physicality of the predicted binding poses. In regards to the free energies, we see that the QUBE FF can achieve the desired < 1 kcal/mol MUE in both relative and absolute [123] binding free energy calculations, that is thought to be effective to guide a drug design campaign [10]. These results indicate that this first generation of the QUBE FF has reached a sufficient level of accuracy and accessibility where it can be regularly used in support of a medicinal chemistry campaign.

In chapter 6 we revisited the design of QUBEKit and demonstrated how the QUBE FF can be routinely extended to new elements such as boron, silicon and phosphorous and improved by implementing a new torsion parametrisation routine that makes use of more open-source software. QUBE has the potential to pave the way to the regular simulation of systems containing boron and silicon which are gaining significant interest in drug design but cannot currently be explored using conventional transferable FFs. Future work in this area should then concentrate on the thorough assessment of the robustness of parameters derived for molecules containing such elements using the R_{free} terms presented here. Other methods to derive the ϵ and σ terms of the L-J potential based on the polarisabilities of the atoms should also be investigated as a replacement for the current scaling method. Such methods have also been identified as being more physical as they allow the non-bonded parameters to fully react to their environment unlike the T-S scheme used here which assigns each occurrence of an element the same well depth [203]. In regards to the new torsion optimisation scheme we have shown that other software can easily be integrated into QUBEKit resulting in a powerful and automated optimisation toolkit which improved the torsional parameters of a range of complex small molecules. Overall we were able to decrease the mean RMSE in the PES predicted using the OFF assigned starting parameters from 2.28 to 0.542 kcal/mol in the case of the BFGS optimiser. We have also seen how the choice of optimiser can play a crucial role in parameter fitting and future work regarding this is envisaged to determine the best general torsional parameter optimisation routine, including measuring the direct effect the parameters have on the structural RMSD during optimisation.

In summary this work has shown that QUBE molecule-specific FFs are a viable alternative to the general transferable ones commonly used in CADD and can now be routinely derived using QUBEKit with little user input, as has been done for the for the 400+ structures parametrised during this thesis. While the accuracy of the QUBE FF has been shown to be very competitive with transferable FF which have undergone consistent improvement for almost 40 years, it is important to note that this is just the first iteration of the QUBE FF. Due to its lack of empiricism (i.e. only 1 fitted parameter per element) we can routinely improve the accuracy of the FF by taking advantage of new exchange-correlation functionals, implicit solvent models or more complex functional forms with minimal effort. One limiting factor of the general adoption of the FF however is the compute time needed to derive the

parameters, with the drug-like molecules studied in chapter 5 in particular taking as long to parametrise as the subsequent FEP simulations. Advances in machine learning however could help relieve this computational burden as we have already started to see the successful use of this technique in the recreation of QM derived charges on a wide range of molecules [185] at a fraction of the cost. It is then envisaged that a complete set of QUBE molecule specific FF parameters could be accurately predicted using machine learning with a computational cost on par with transferable FF parameter assignment.

Chapter 8

Appendix A

8.1 Bonds and Angles

OPLS bond type	k_r (kcal/mol/Å^2)	equilibrium bond length (Å)
CY-C	211.6	1.491

Table 8.1: The missing OPLS bond type is shown with an estimate for the force constant and equilibrium bond length predicted by QUBE. This bond type was assigned to 1-cylcopropylethanone.

OPLS angle type	k_θ (kcal/mol/rad^2)	equilibrium angle (degrees)
CY-CY-C	75.6	117.5
C -CY-HC	34.0	116.3
CY-C -CT	72.6	116.1
CY-C -O	52.1	121.8
CA-C=-CT	65.4	117.3
Cl-CM-Cl	31.1	114.2

Table 8.2: The missing OPLS angle types are shown with estimates for the force constants and equilibrium angles predicted by QUBE. These missing angles were found in molecules 1-cylcopropylethanone and 1,1-dichloroethene.

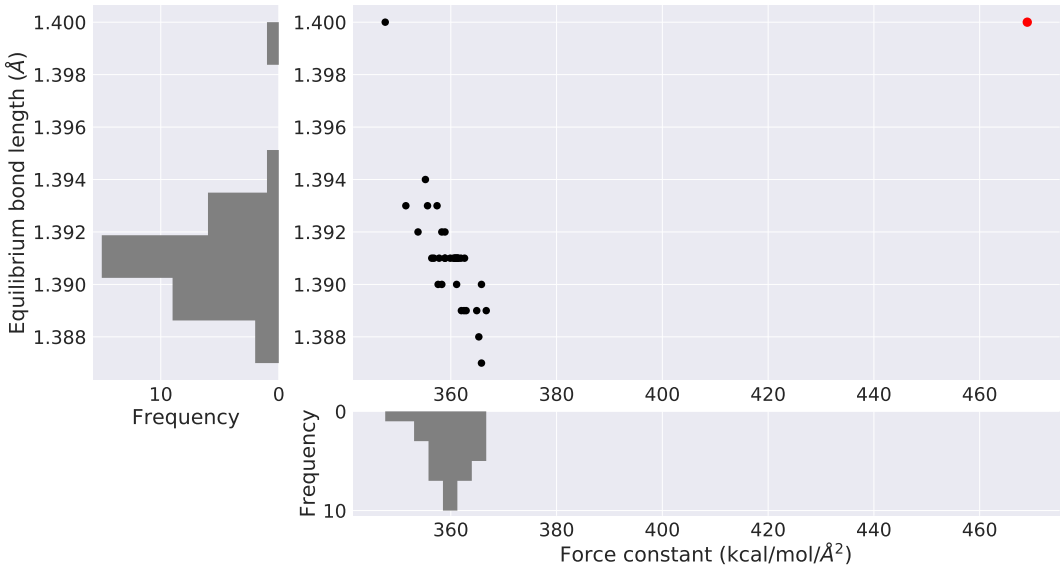

Figure 8.1: The QM predicted equilibrium bond length is compared to the associated derived force constant of each molecule they appear in for the OPLS CA-CA bond type with the OPLS values shown in red.

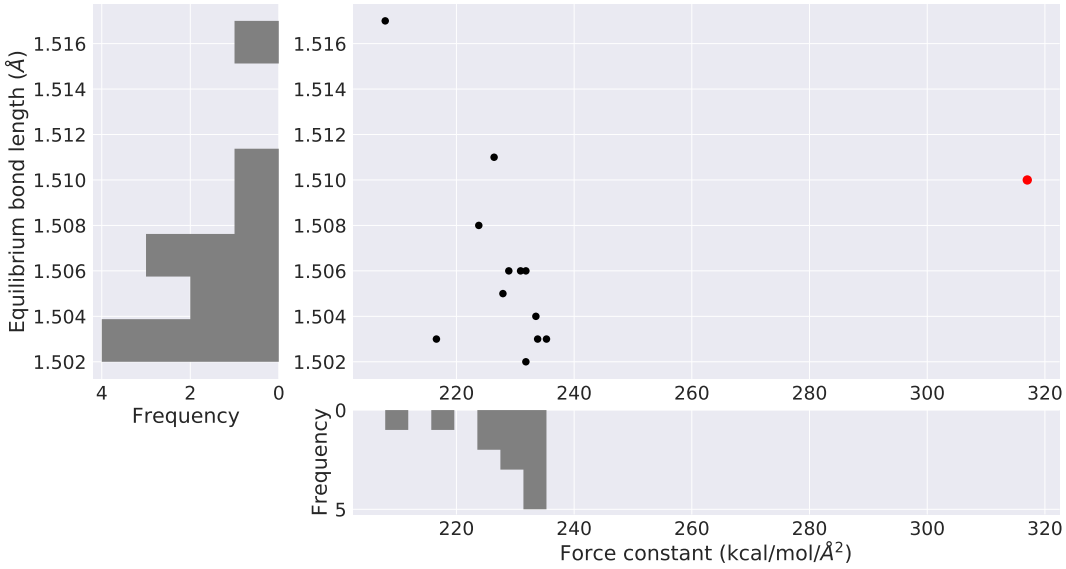

Figure 8.2: The QM predicted equilibrium bond length is compared to the associated derived force constant of each molecule they appear in for the OPLS CA-CT bond type with the OPLS values shown in red.

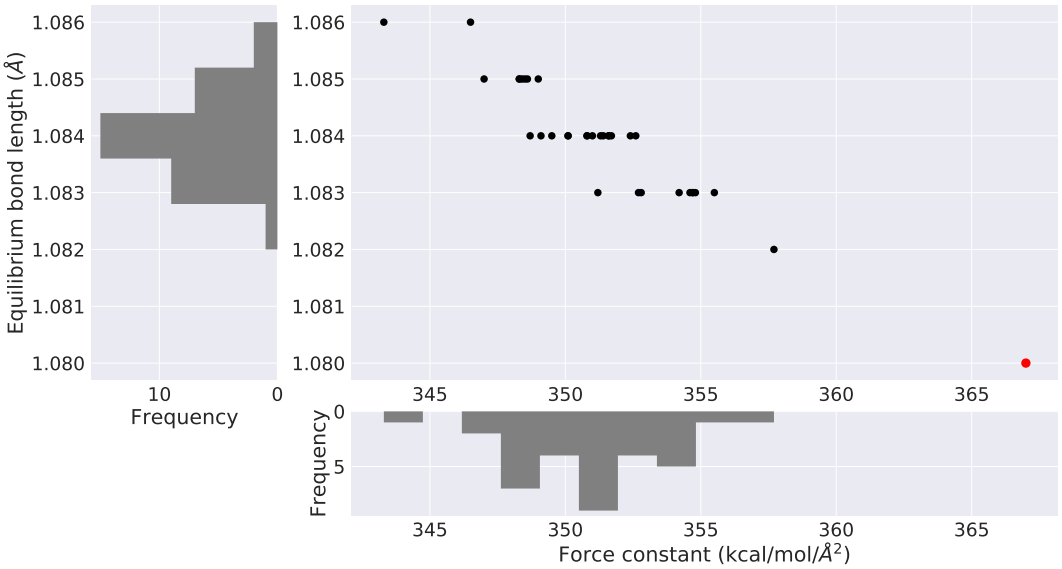

Figure 8.3: The QM predicted equilibrium bond length is compared to the associated derived force constant of each molecule they appear in for the OPLS CA-HA bond type with the OPLS values shown in red.

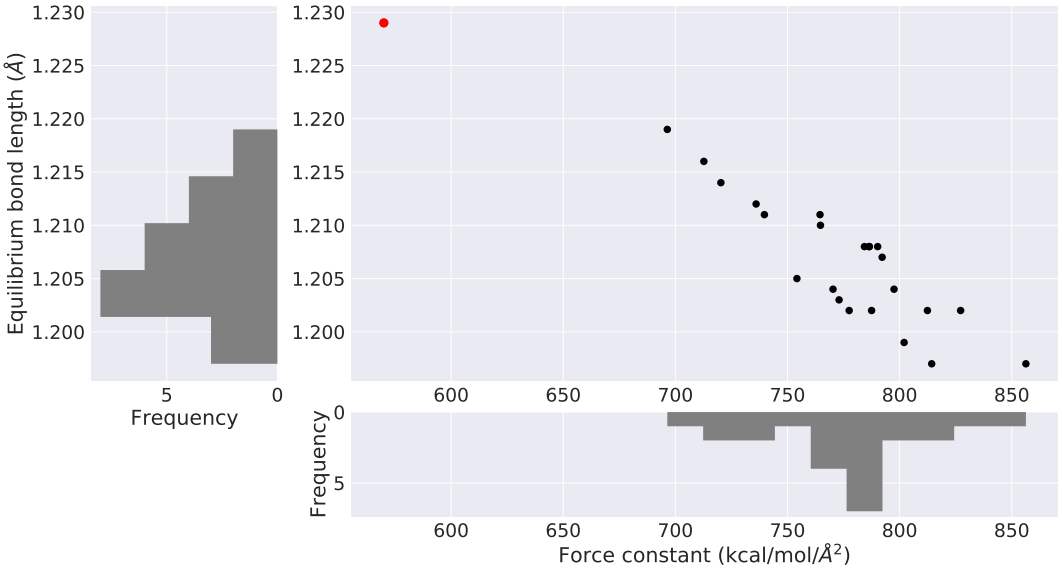

Figure 8.4: The QM predicted equilibrium bond length is compared to the associated derived force constant of each molecule they appear in for the OPLS C-O bond type with the OPLS values shown in red.

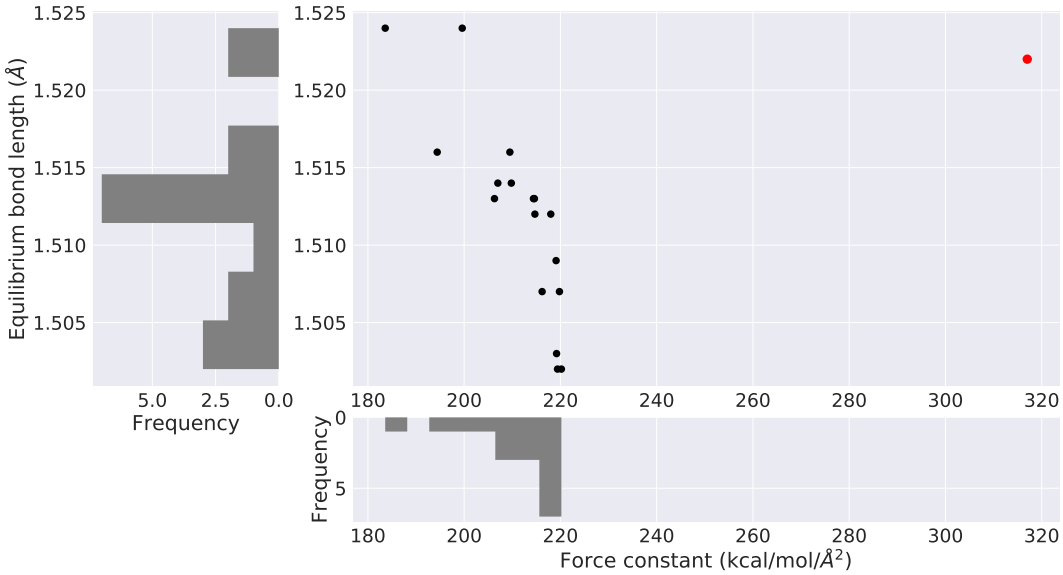

Figure 8.5: The QM predicted equilibrium bond length is compared to the associated derived force constant of each molecule they appear in for the OPLS CT-C bond type with the OPLS values shown in red.

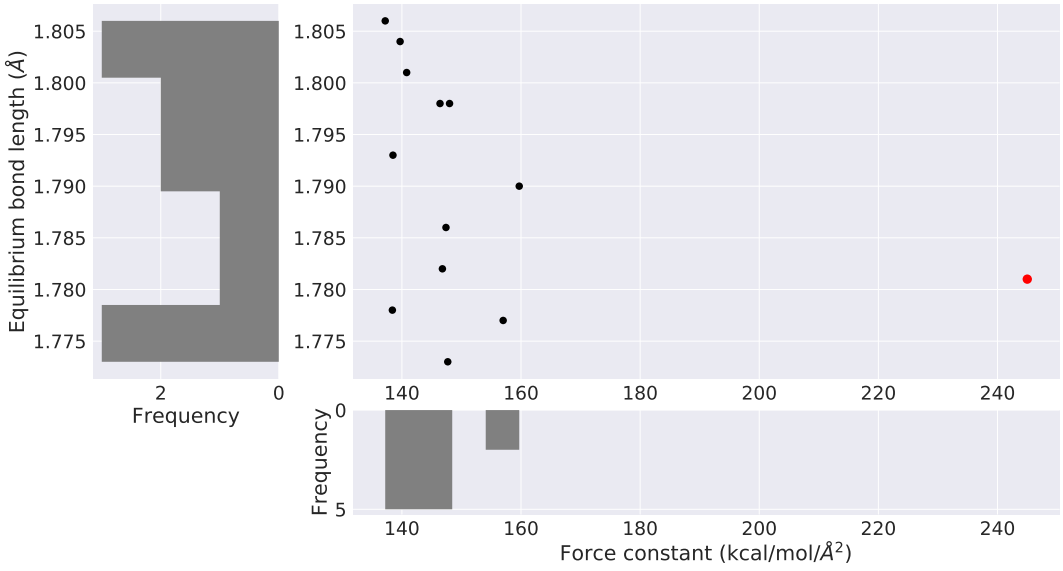

Figure 8.6: The QM predicted equilibrium bond length is compared to the associated derived force constant of each molecule they appear in for the OPLS CT-Cl bond type with the OPLS values shown in red.

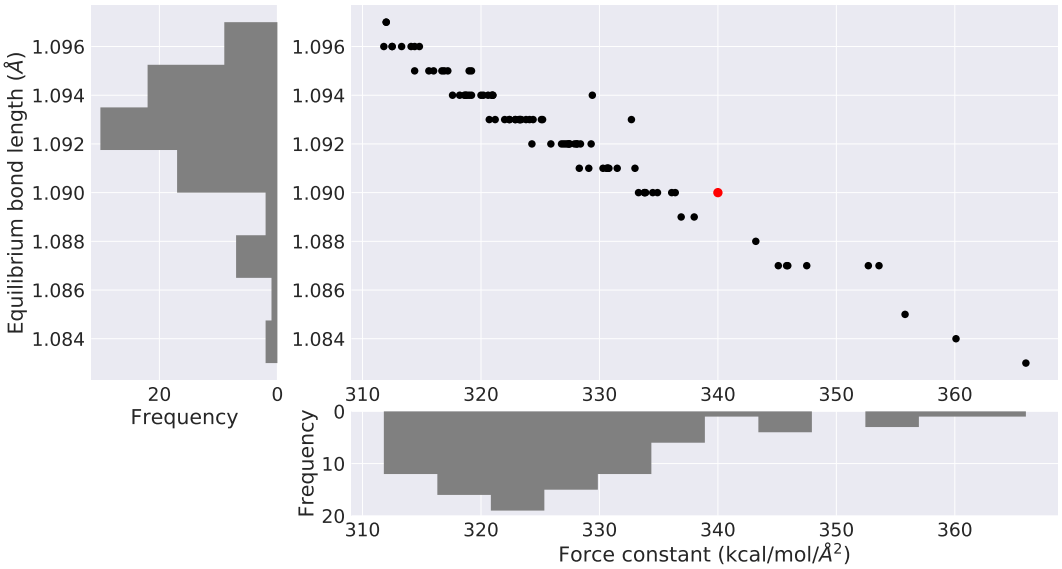

Figure 8.7: The QM predicted equilibrium bond length is compared to the associated derived force constant of each molecule they appear in for the OPLS CT-HC bond type with the OPLS values shown in red.

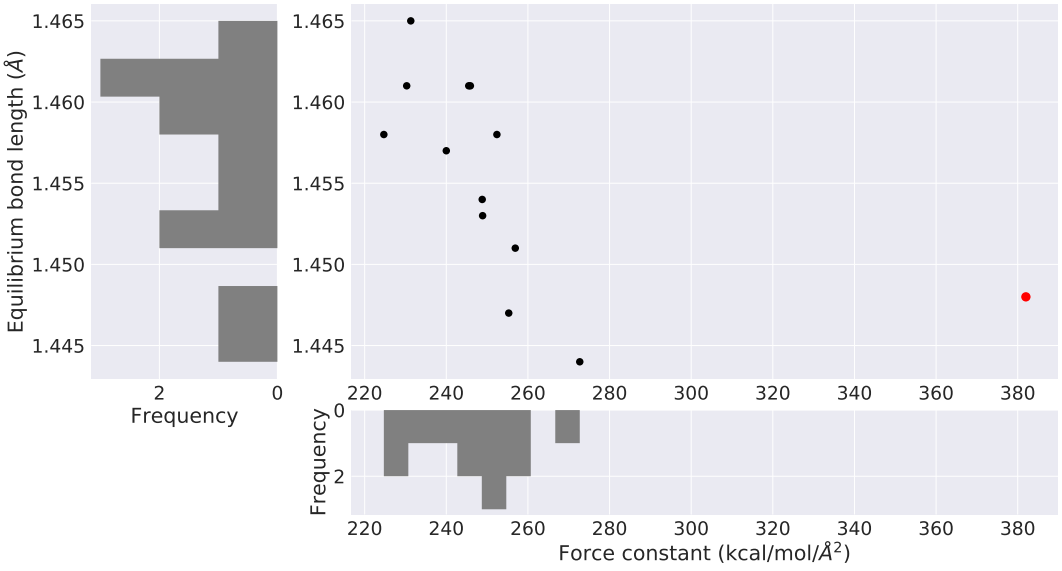

Figure 8.8: The QM predicted equilibrium bond length is compared to the associated derived force constant of each molecule they appear in for the OPLS CT-NT bond type with the OPLS values shown in red.

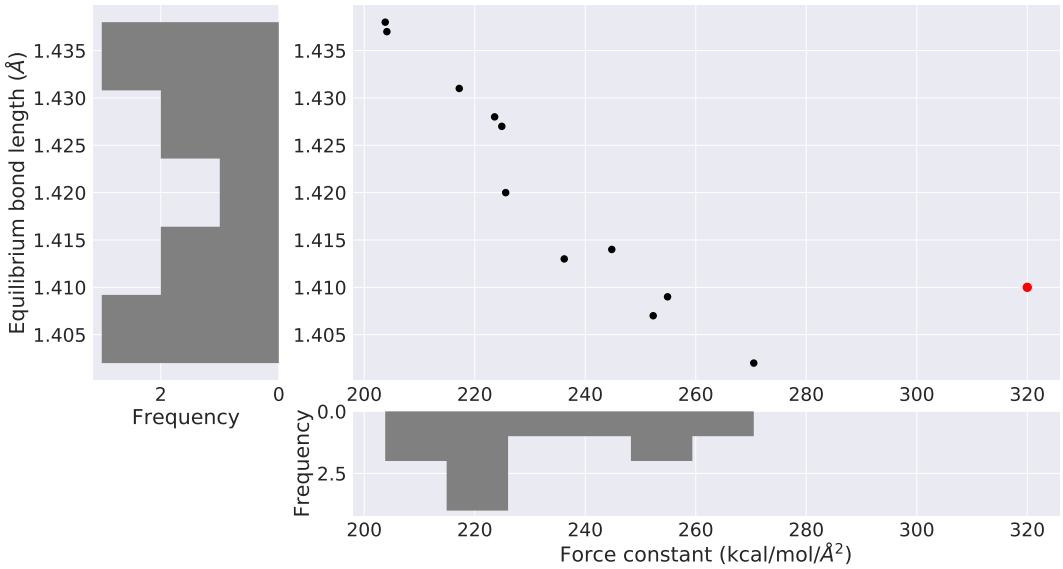

Figure 8.9: The QM predicted equilibrium bond length is compared to the associated derived force constant of each molecule they appear in for the OPLS CT-OS bond type with the OPLS values shown in red.

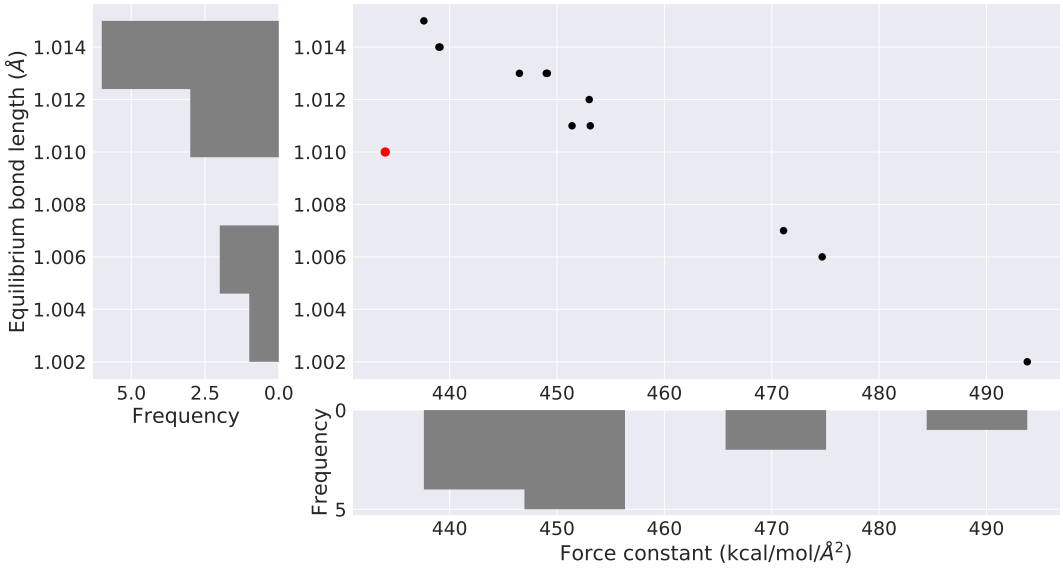

Figure 8.10: The QM predicted equilibrium bond length is compared to the associated derived force constant of each molecule they appear in for the OPLS NT-H bond type with the OPLS values shown in red.

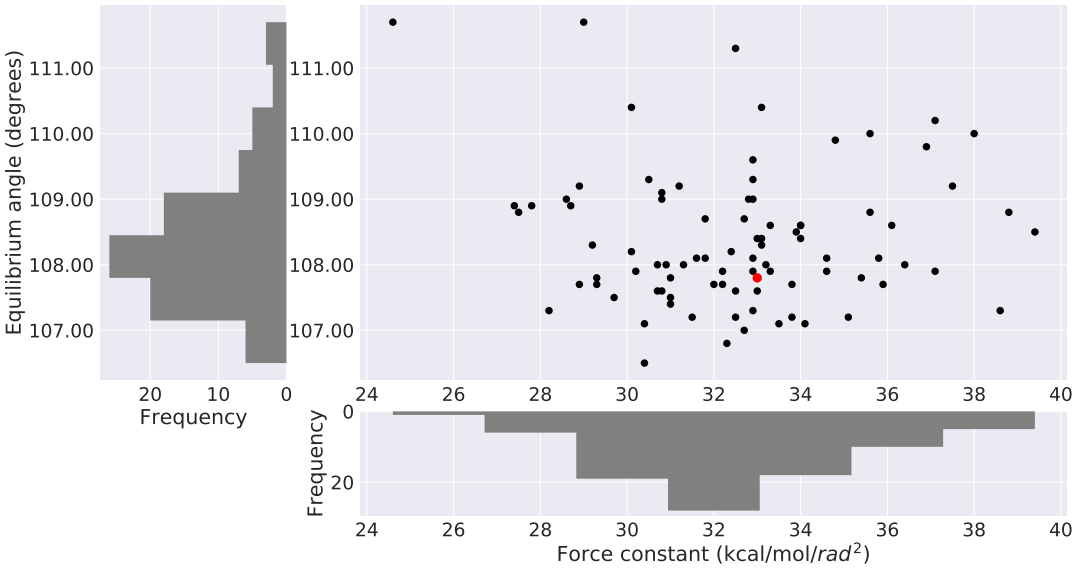

Figure 8.11: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS HC-CT-HC angle type with the OPLS values shown in red.

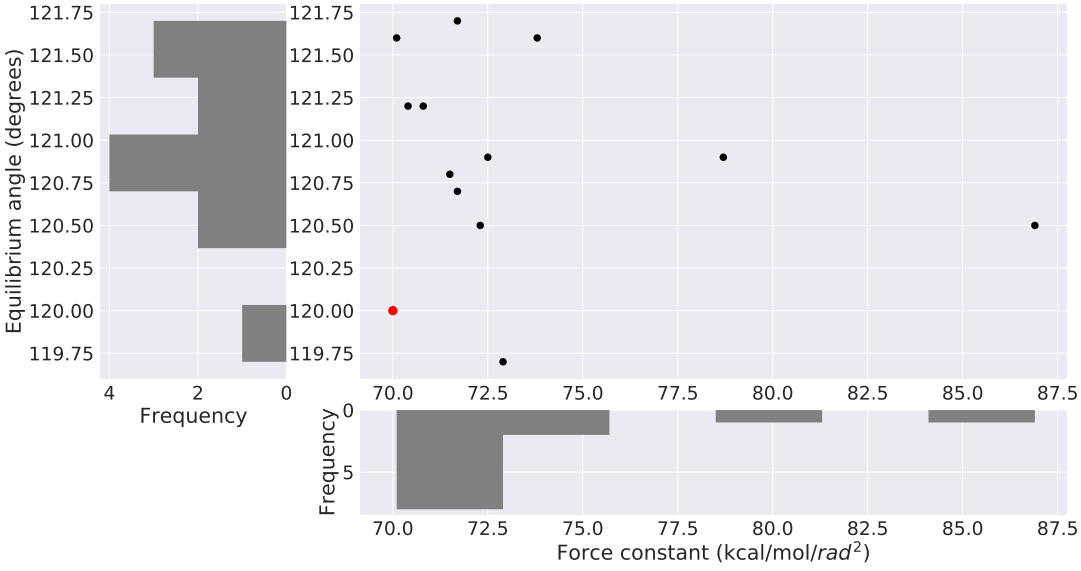

Figure 8.12: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS CA-CA-CT angle type with the OPLS values shown in red.

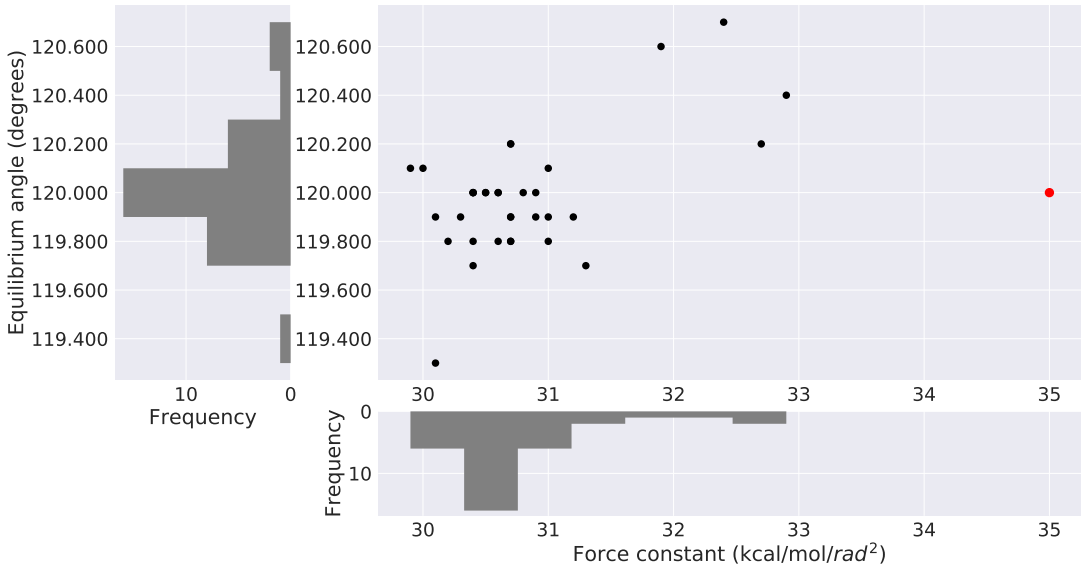

Figure 8.13: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS CA-CA-HA angle type with the OPLS values shown in red.

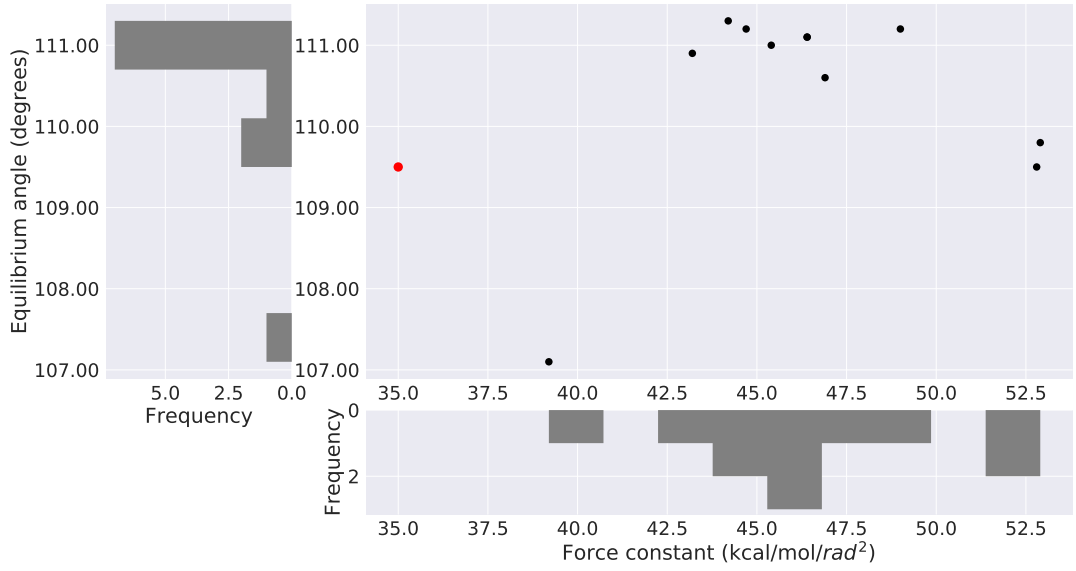

Figure 8.14: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS CA-CT-HC angle type with the OPLS values shown in red.

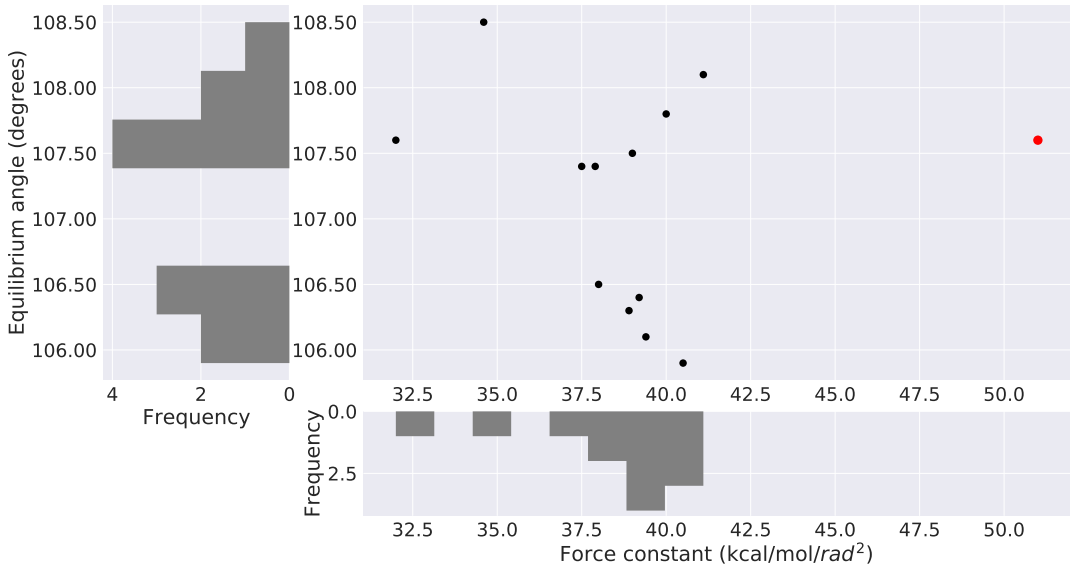

Figure 8.15: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS Cl-CT-HC angle type with the OPLS values shown in red.

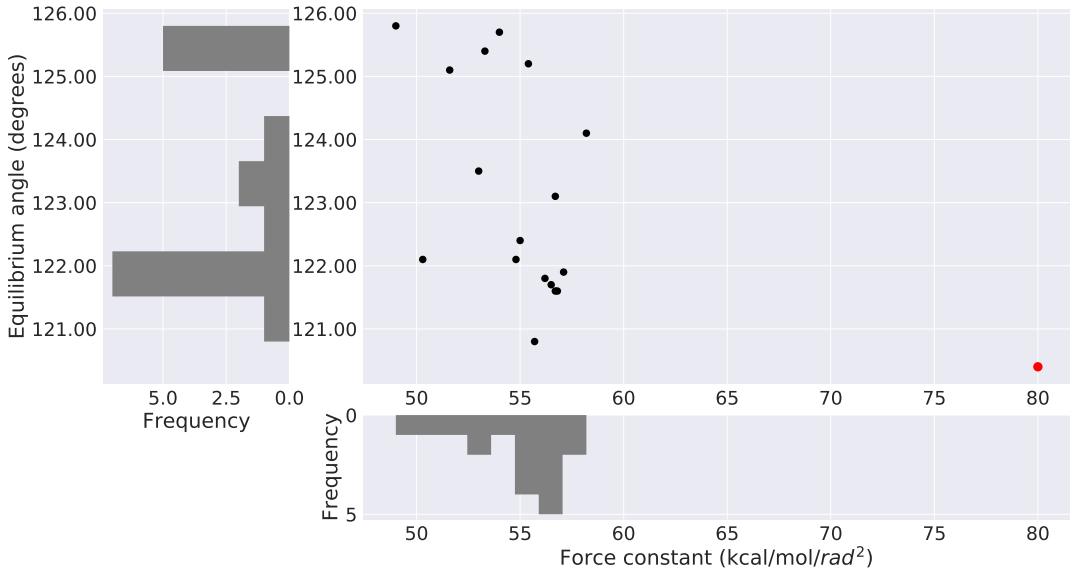

Figure 8.16: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS CT-C-O angle type with the OPLS values shown in red.

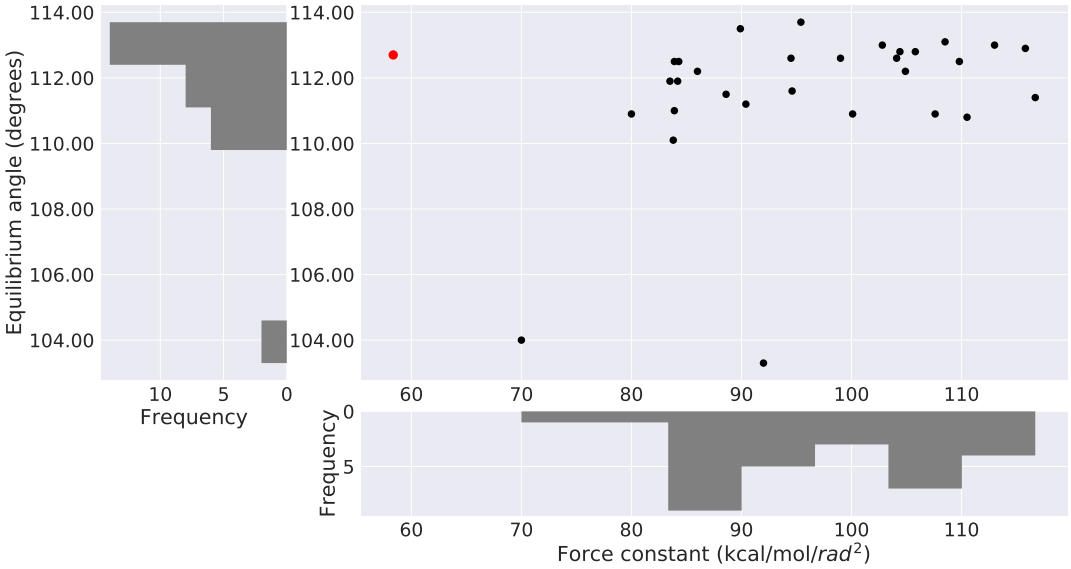

Figure 8.17: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS CT-CT-CT angle type with the OPLS values shown in red.

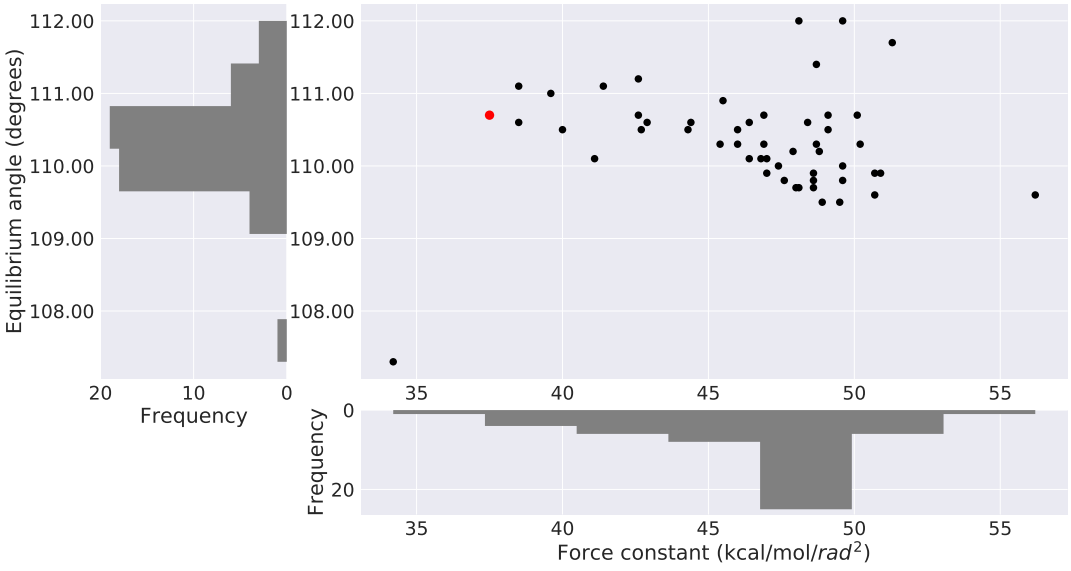

Figure 8.18: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS CT-CT-HC angle type with the OPLS values shown in red.

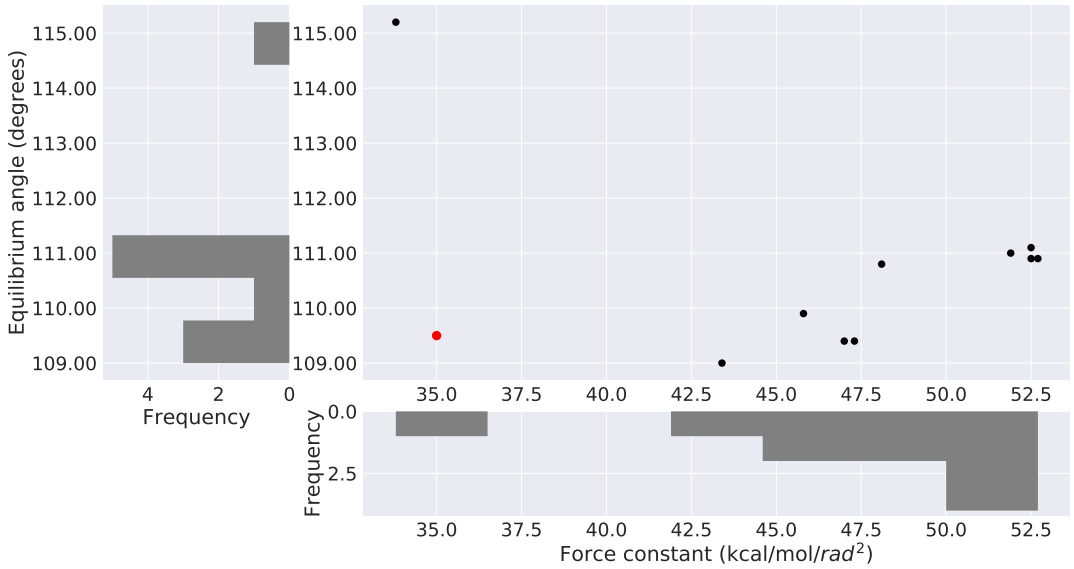

Figure 8.19: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS CT-NT-H angle type with the OPLS values shown in red.

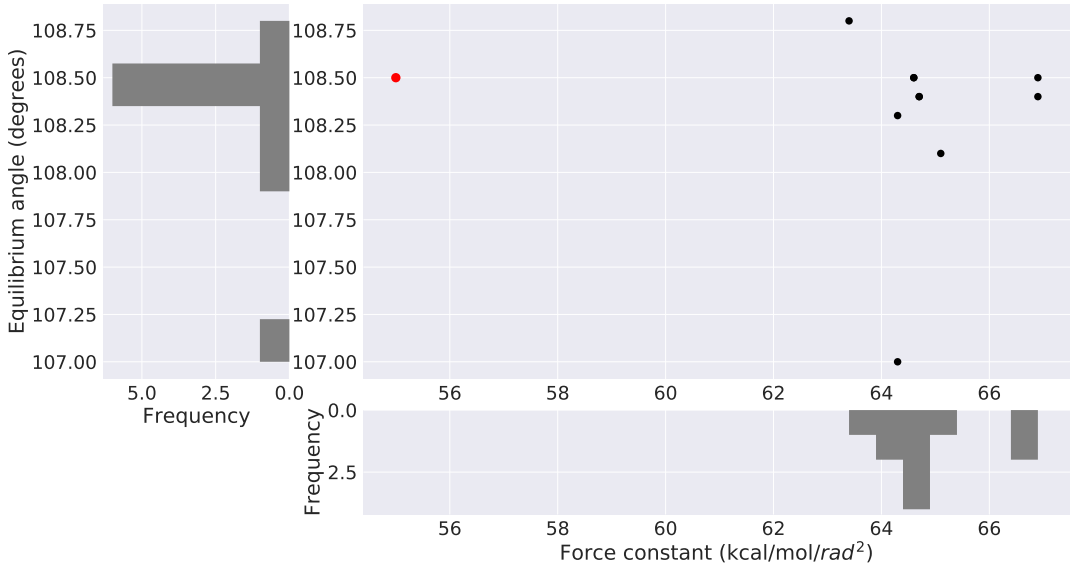

Figure 8.20: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS CT-OH-HO angle type with the OPLS values shown in red.

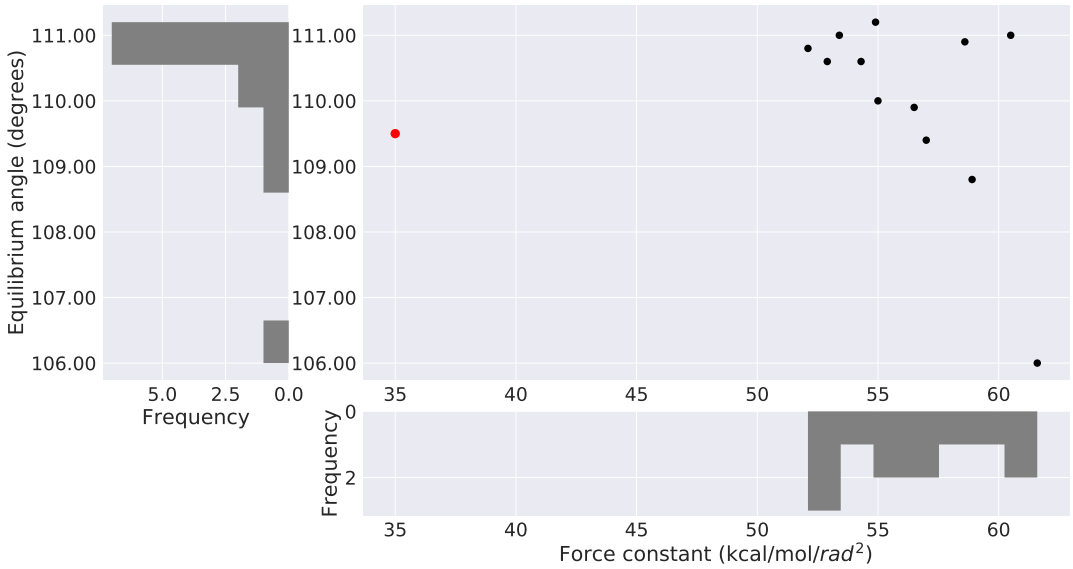

Figure 8.21: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS NT-CT-HC angle type with the OPLS values shown in red.

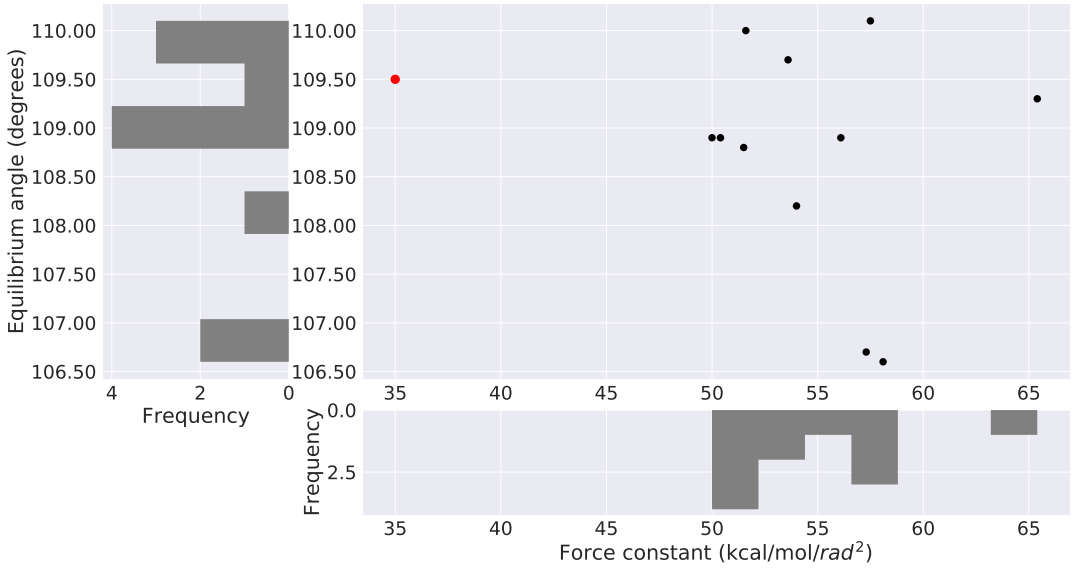

Figure 8.22: The QM predicted equilibrium angle is compared to the associated derived force constant of each molecule they appear in for the OPLS OS-CT-HC angle type with the OPLS values shown in red.

8.2 Dihedrals

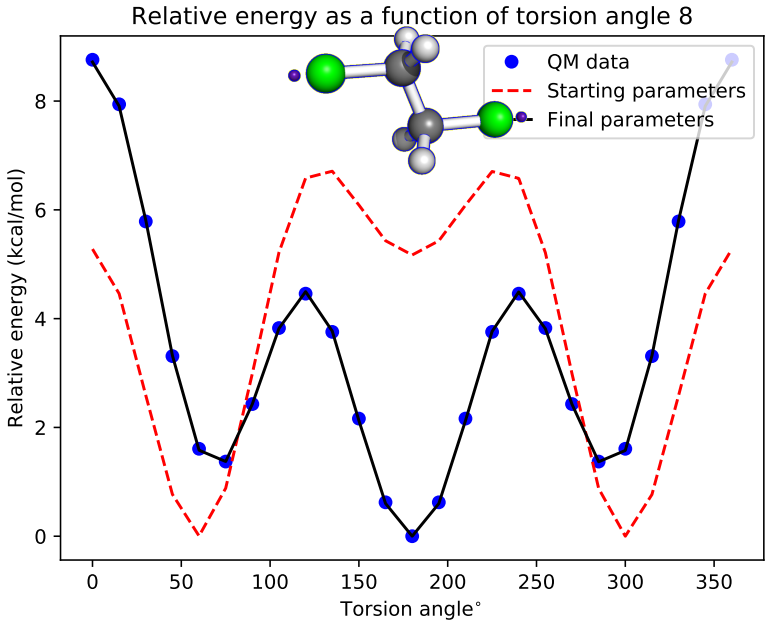

Figure 8.23: The QUBEKit generated torsional scan is shown for 1,2-dichloroethane. Where the QM data is calculated using the ωB97X-D[1] DFT functional and 6-311++G(d,p) basis set in Gaussian09 [2], the starting parameters are taken from OPLS and the final parameters are found using QUBEKit. Final error = 0.670 kcal/mol, bias = 0.655 kcal/mol

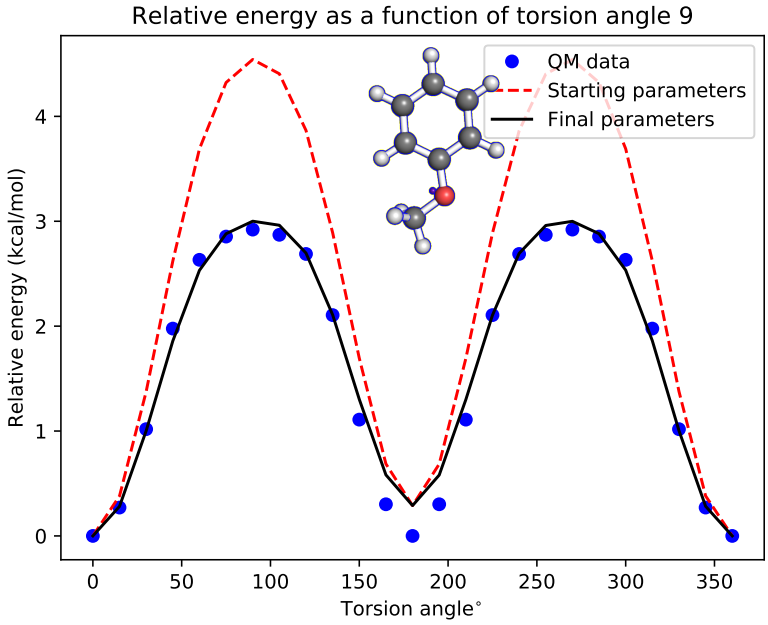

Figure 8.24: The QUBEKit generated torsional scan is shown for anisole. Where the QM data is calculated using the ωB97X-D[1] DFT functional and 6-311++G(d,p) basis set in Gaussian09 [2], the starting parameters are taken from OPLS and the final parameters are found using QUBEKit. Final error = 0.193 kcal/mol, bias = 0.079 kcal/mol.

8.3 Extra sites

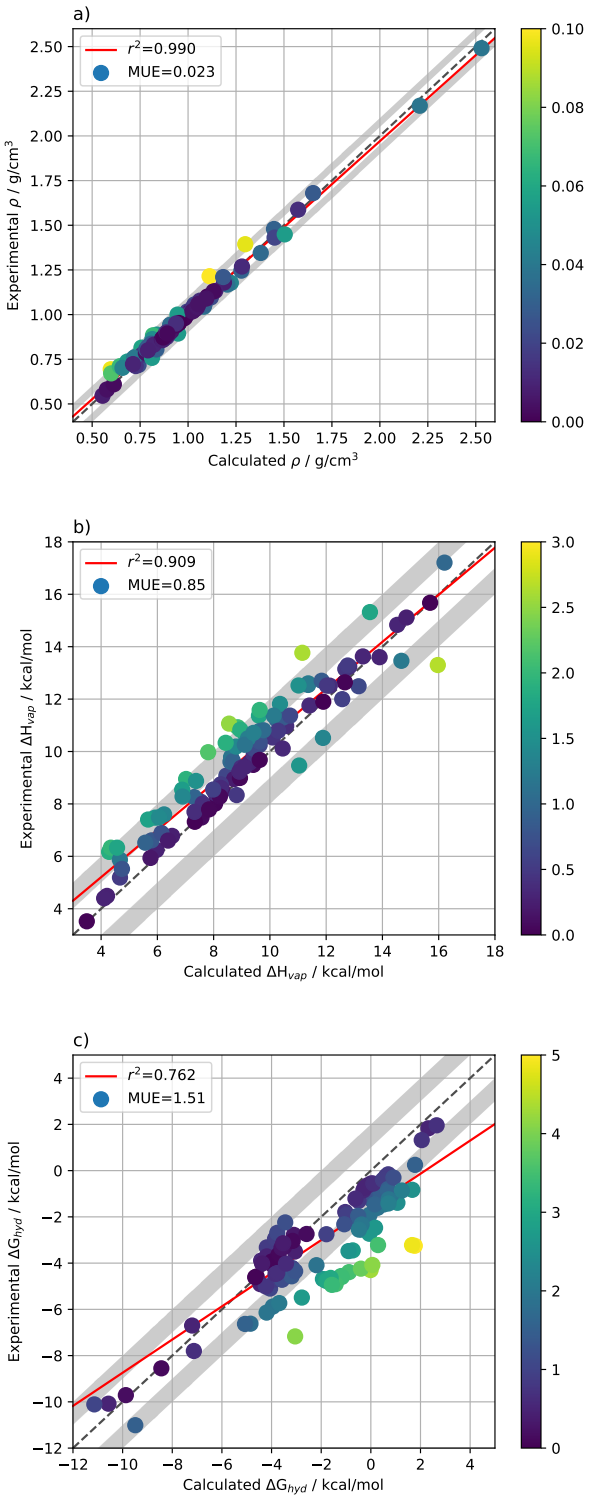

Figure 8.25: The standard force field liquid property metrics (a) liquid density, (b) heat of vaporization (c) free energy of hydration. Calculated for the organic molecule test using QUBE FF parameters with no virtual sites. Mean unsigned error (MUE) compared to experiment and r^2 correlation are included.

FF parameterization	ρ (g/cm^3)	ΔH_{vap} (kcal/mol)	ΔG_{hyd} (kcal/mol)
DDEC/OPLS	1.095	8.65	0.28
QUBE	1.101	8.62	0.55
Experiment	1.101	9.79	-1.12

Table 8.3: The predicted liquid density and thermodynamic properties for chlorobenzene are shown along with experimental values for two different parameterizations.

8.4 Transition pathways

All FEP transitions were performed relative to ligand **1** via the following pathways.

$2 \rightarrow 1$

$3 \rightarrow 1$

$4 \rightarrow 9 \rightarrow 10 \rightarrow 7 \rightarrow 1$

$5 \rightarrow 1$

$6 \rightarrow 3 \rightarrow 1$

$7 \rightarrow 1$

$8 \rightarrow 1$

$9 \rightarrow 10 \rightarrow 7 \rightarrow 1$

$10 \rightarrow 7 \rightarrow 1$

$11 \rightarrow 8 \rightarrow 1$

$12 \rightarrow 18 \rightarrow 1$

$13 \rightarrow 3 \rightarrow 1$

$14 \rightarrow 18 \rightarrow 1$

$15 \rightarrow 2 \rightarrow 1$

$16 \rightarrow 17 \rightarrow 18 \rightarrow 1$

$17 \rightarrow 18 \rightarrow 1$

$18 \rightarrow 1$

Bibliography

[1] J. D. Chai, M. Head-Gordon, *Phys. Chem. Chem. Phys.* **10**, 6615 (2008).

[2] M. J. Frisch, *et al.* Gaussian Inc. Wallingford CT 2009.

[3] J. Luccarelli, J. Michel, J. Tirado-Rives, W. L. Jorgensen, *J. Chem. Theory Comput.* **6**, 3850 (2010).

[4] D. L. Mobley, *et al.*, *J. Chem. Theory Comput.* **14**, 6076 (2018). PMID: 30351006.

[5] T. A. Manz, N. Gabaldon Limas, Chargemol program for performing ddec analysis. Available from http://ddec.sourceforge.net (accessed September 2018).

[6] X. Chu, A. Dalgarno, *J. Chem. Phys.* **121**, 4083 (2004).

[7] D. L. Mobley, J. P. Guthrie, *J. Comput.-Aided Mol. Des.* **28**, 711 (2014).

[8] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).

[9] P. Robustelli, S. Piana, D. E. Shaw, *P. Natl. Acad. Sci. USA* **115**, E4758 (2018).

[10] D. L. Mobley, P. V. Klimovich, *J. Chem. Phys.* **137**, 230901 (2012).

[11] W. L. Jorgensen, *Science (New York, N.Y.)* **303**, 1813 (2004).

[12] M. De Vivo, M. Masetti, G. Bottegoni, A. Cavalli, *J. Med. Chem.* **59**, 4035 (2016).

[13] Z. Cournia, B. Allen, W. Sherman, *J. Chem. Inf. and Model.* **57**, 2911 (2017). PMID: 29243483.

[14] J. Wang, W. Wang, P. A. Kollman, D. A. Case, *J. Comput. Chem* **25**, 1157 (2004).

[15] K. Vanommeslaeghe, *et al.*, *J. Comput. Chem.* **31**, 671 (2010).

[16] W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, *J. Am. Chem. Soc.* **118**, 11225 (1996).

[17] M. Bollini, *et al.*, *J. Med. Chem.* **54**, 8582 (2011).

[18] W.-G. Lee, *et al.*, *J. Am. Chem. Soc.* **135**, 16705 (2013).

[19] E. Harder, *et al.*, *J. Chem. Theory Comput.* **12**, 281 (2016).

[20] L. P. Wang, T. J. Martinez, V. S. Pande, *J. Phys. Chem. Lett.* **5**, 1885 (2014).

[21] W. L. Jorgensen, K. P. Jensen, A. N. Alexandrova, *J. Chem. Theory Comput.* **3**, 1987 (2007).

[22] A. Jakalian, B. L. Bush, D. B. Jack, C. I. Bayly, *J. Comput. Chem.* **21**, 132 (2000).

[23] A. Jakalian, D. B. Jack, C. I. Bayly, *J. Comput. Chem.* **23**, 1623 (2002).

[24] J. W. Storer, D. J. Giesen, C. J. Cramer, D. G. Truhlar, *J. Comput. Aided Mol. Des.* **9**, 87 (1995).

[25] A. Tkatchenko, M. Scheffler, *Phys. Rev. Lett.* **102**, 6 (2009).

[26] A. Tkatchenko, R. A. DiStasio Jr, R. Car, M. Scheffler, *Phy. Rev. Lett.* **108**, 236402 (2012).

[27] E. Clementi, G. Corongiu, G. Ranghino, *J. Chem. Phys.* **74**, 578 (1981).

[28] G. C. Lie, E. Clementi, *J. Chem. Phys.* **62**, 2195 (1975).

[29] H. Kistenmacher, H. Popkie, E. Clementi, R. O. Watts, *J. Chem. Phys.* **60**, 4455 (1974).

[30] P. Linse, S. Engström, B. Jönsson, *Chem. Phys. Lett.* **115**, 95 (1985).

[31] G. Karlstroem, P. Linse, A. Wallqvist, B. Joensson, *J. Am. Chem. Soc.* **105**, 3777 (1983).

[32] I. Cacelli, G. Cinacchi, G. Prampolini, A. Tani, *J. Am. Chem. Soc.* **126**, 14278 (2004).

[33] G. Prampolini, P. R. Livotto, I. Cacelli, *J. Chem. Theory Comput.* **11**, 5182 (2015).

[34] L. Greff da Silveira, M. Jacobs, G. Prampolini, P. R. Livotto, I. Cacelli, *J. Chem. Theory Comput.* **14**, 4884 (2018).

[35] B. Waldher, J. Kuta, S. Chen, N. Henson, A. E. Clark, *J. Comp. Chem.* **31**, 2307 (2010).

[36] I. Cacelli, C. F. Lami, G. Prampolini, *J. Comp. Chem.* **30**, 366 (2009).

[37] P. Xu, E. B. Guidez, C. Bertoni, M. S. Gordon, *J. Chem. Phys.* **148**, 090901 (2018).

[38] J. G. McDaniel, J. Schmidt, *J. Phys. Chem. A.* **117**, 2053 (2013).

[39] M. J. Van Vleet, A. J. Misquitta, A. J. Stone, J. R. Schmidt, *J. Chem. Theory Comput.* **12**, 3851 (2016).

[40] S. Vandenbrande, M. Waroquier, V. V. Speybroeck, T. Verstraelen, *J. Chem. Theory Comput.* **13**, 161 (2016).

[41] S. Grimme, *J. Chem. Theory Comput.* **10**, 4497 (2014).

[42] J. Yin, *et al.*, *J. Comput. Aided Mol. Des.* **31**, 1 (2017).

[43] D. L. Mobley, M. K. Gilson, *Annu. Rev. Biophys.* **46**, 531 (2017).

[44] T. B. Steinbrecher, *et al.*, *J. Chem. Inf. Model.* **55**, 2411 (2015).

[45] A. E. A. Allen, M. C. Payne, D. J. Cole, *J. Chem. Theory Comput.* **14**, 274 (2018).

[46] J. M. Seminario, *Int. J. Quantum Chem.* **60**, 1271 (1996).

[47] A. Hagler, *J. Chem. Theory Comput.* **11**, 5555 (2015).

[48] L.-P. Wang, *et al.*, *J. Phys. Chem. B.* **121**, 4023 (2017).

[49] T. Verstraelen, D. Van Neck, P. Ayers, V. Van Speybroeck, M. Waroquier, *J. Chem. Theory Comput.* **3**, 1420 (2007).

[50] E. Boulanger, L. Huang, C. Rupakheti, A. D. MacKerell, B. Roux, *J. Chem. Theory Comput.* **14**, 3121 (2018).

[51] S. K. Burger, *et al.*, *J. Chem. Theory Comput.* **8**, 554 (2012).

[52] S. Dasgupta, W. A. Goddard III, *J. Chem. Phys.* **90**, 7207 (1989).

[53] S. Dasgupta, K. A. Brameld, C.-F. Fan, W. A. Goddard III, *Spectrochimica Acta Part A: Mol. Biomol. Spectrosc.* **53**, 1347 (1997).

[54] I. Cacelli, G. Prampolini, *J. Chem. Theory Comput.* **3**, 1803 (2007).

[55] V. Barone, *et al.*, *Phys. Chem. Chem. Phys.* **15**, 3736 (2013).

[56] L. Hu, U. Ryde, *J. Chem. Theory Comput.* **7**, 2452 (2011).

[57] R. Wang, M. Ozhgibesov, H. Hirao, *J. Comput. Chem.* pp. 2349–2359 (2016).

[58] L. P. Lee, D. J. Cole, C. K. Skylaris, W. L. Jorgensen, M. C. Payne, *J. Chem. Theory Comput.* **9**, 2981 (2013).

[59] D. J. Cole, J. Z. Vilseck, J. Tirado-Rives, M. C. Payne, W. L. Jorgensen, *J. Chem. Theory Comput.* **12**, 2312 (2016).

[60] A. E. Allen, M. J. Robertson, M. C. Payne, D. J. Cole, *ACS Omega* **4**, 14537 (2019).

[61] J. T. Horton, A. E. A. Allen, L. S. Dodda, D. J. Cole, *J. Chem. Inf. Model.* **59**, 1366 (2019). PMID: 30742438.

[62] J. M. Turney, *et al.*, *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2**, 556 (2012).

[63] P. Eastman, *et al.*, *PLoS Comput. Biol.* **13**, e1005659 (2017).

[64] D. M. Ceperley, B. Alder, *Phys. Rev. Lett.* **45**, 566 (1980).

[65] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).

[66] J.-L. Fattebert, F. Gygi, *J. Comput. Chem.* **23**, 662 (2002).

[67] P. J. Stephens, F. Devlin, C. Chabalowski, M. J. Frisch, *J. Phys. Chem-US* **98**, 11623 (1994).

[68] A. D. Becke, *Phys. Rev, vol. A* **38**, 3098 (1988).

[69] C. Lee, W. Yang, R. Parr, *Phys. Rev, vol. B* **37**, 785 (1988).

[70] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* **132**, 154104 (2010).

[71] Q. Hill, C.-K. Skylaris, *P. Roy. Soc. A-Math. Phy.* **465**, 669 (2009).

[72] A. M. Reilly, A. Tkatchenko, *Chem. Sci.* **6**, 3289 (2015).

[73] J. Hermann, R. A. DiStasio Jr, A. Tkatchenko, *Chem. Rev.* **117**, 4714 (2017).

[74] S. Grimme, *Wires. Comput. Mol. Sci.* **1**, 211 (2011).

[75] K. M. Visscher, D. P. Geerke, *J. Chem. Theory Comput.* **15**, 1875 (2019).

[76] D. Bowler, T. Miyazaki, *Rep. Prog. Phys.* **75**, 036503 (2012).

[77] J. Dziedzic, H. H. Helal, C. K. Skylaris, A. A. Mostofi, M. C. Payne, *Europhys. Lett.* **95** (2011).

[78] J. Tomasi, M. Persico, *Chem. Rev.* **94**, 2027 (1994).

[79] A. Klamt, G. Schüürmann, *J. Chem. Soc. Perk. T. 2* pp. 799–805 (1993).

[80] J. Tomasi, B. Mennucci, R. Cammi, *Che. Rev.* **105**, 2999 (2005).

[81] V. Barone, M. Cossi, J. Tomasi, *J Chem. Phys.* **107**, 3210 (1997).

[82] D. J. Cole, N. D. M. Hine, *J. Phys. Condens. Matter* **28**, 393001 (2016).

[83] N. L. Allinger, K. Chen, J. Lii, *J. Comput. Chem.* **17**, 642 (1996).

[84] M. J. Robertson, J. Tirado-Rives, W. L. Jorgensen, *J. Chem. Theory Comput.* **11**, 3499 (2015).

[85] P. M. Todebush, G. Liang, J. P. Bowen, *Chirality* **14**, 220 (2002).

[86] J. W. Ponder, *et al.*, *J. Phys. Chem. B* **114**, 2549 (2010).

[87] J. A. Lemkul, J. Huang, B. Roux, A. D. MacKerell Jr, *Chem. Rev.* **116**, 4983 (2016).

[88] S. Patel, C. L. Brooks III, *J. Comp. Chem* **25**, 1 (2004).

[89] P. Dziedzic, *et al.*, *J. Am. Chem. Soc.* **137**, 2996 (2015).

[90] A. A. Hare, *et al.*, *Bioorg. Med. Chem. Lett.* **20**, 5811 (2010).

[91] Z. Cournia, *et al.*, *J. Med. Chem.* **52**, 416 (2008).

[92] O. Acevedo, *et al.*, *Curr. Pharm. Design.* **18**, 1199 (2012).

[93] S. Zheng, *et al.*, *J. Chem. Inf. Model.* **56**, 811 (2016).

[94] P. Li, K. M. Merz, *J. Chem. Inf. Model.* **56**, 599 (2016).

[95] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *J. Am. Chem. Soc.* **107**, 3902 (1985).

[96] L. S. Dodda, I. C. De Vaca, J. Tirado-Rives, W. L. Jorgensen, *Nucleic Acids Res.* **45**, 331 (2017).

[97] L. S. Dodda, J. Z. Vilseck, J. Tirado-Rives, W. L. Jorgensen, *J. Phys. Chem. B* **121**, 3864 (2017).

[98] W. L. Jorgensen, J. Tirado-Rives, *Proc. Natl. Acad. Sci. U. S. A.* **102**, 6665 (2005).

[99] W. Damm, A. Frontera, J. T. Rives, W. L. Jorgensen, *J. Comp. Chem.* **18**, 1955 (1997).

[100] T. A. Manz, D. S. Sholl, *J. Chem. Theory Comput.* **6**, 2455 (2010).

[101] T. A. Manz, D. S. Sholl, *J. Chem. Theory Comput.* **8**, 2844 (2012).

[102] L. S. Dodda, J. Z. Vilseck, K. J. Cutrona, W. L. Jorgensen, *J. Chem. Theory Comput.* **11**, 4273 (2015).

[103] C. Kramer, A. Spinn, K. R. Liedl, *J. Chem. Theory Comput.* **10**, 4488 (2014).

[104] O. T. Unke, M. Devereux, M. Meuwly, *J. Chem. Phys.* **147**, 161712 (2017).

[105] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* **79**, 926 (1983).

[106] X. C. Yan, M. J. Robertson, J. Tirado-Rives, W. L. Jorgensen, *J. Phys. Chem. B* **121**, 6626 (2017).

[107] M. Macchiagodena, G. Mancini, M. Pagliai, V. Barone, *Phys. Chem. Chem. Phys.* **18**, 25342 (2016).

[108] M. Macchiagodena, G. Mancini, M. Pagliai, G. Del Frate, V. Barone, *Chem. Phys. Lett.* **677**, 120 (2017).

[109] K. H. Dubay, *et al.*, *J. Chem. Theory Comput.* **8**, 4556 (2012).

[110] S. Chen, S. Yi, W. Gao, C. Zuo, Z. Hu, *J. Comput. Chem.* **36**, 376 (2015).

[111] J. Wildman, P. Repiščák, M. J. Paterson, I. Galbraith, *J. Chem. Theory Comput.* **12**, 3813 (2016).

[112] M. K. Dahlgren, P. Schyman, J. Tirado-Rives, W. L. Jorgensen, *J. Chem. Inf. Model.* **53**, 1191 (2013).

[113] M. Zgarbová, A. M. Rosnik, F. J. Luque, C. Curutchet, P. Jurečka, *J. Comput. Chem.* **36**, 1874 (2015).

[114] K. Vanommeslaeghe, M. Yang, A. D. Mackerell, *J. Comput. Chem.* **36**, 1083 (2015).

[115] H. C. Andersen, *J. Chem. Phys* **72**, 2384 (1980).

[116] M. P. Allen, D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, New York, NY, USA, 1989).

[117] J. Meller, *Encyclopedia of life sciences* (2001).

[118] G. Sliwoski, S. Kothiwale, J. Meiler, E. W. Lowe, *Pharmacol. Rev.* **66**, 334 (2014).

[119] W. L. Jorgensen, J. Tirado-Rives, *J. Comput. Chem.* **26**, 1689 (2005).

[120] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, *J. Chem. Phys* **21**, 1087 (1953).

[121] L. Wang, R. A. Friesner, B. Berne, *J. Phys. Chem. B* **115**, 9431 (2011).

[122] D. J. Cole, J. Tirado-Rives, W. L. Jorgensen, *J. Chem. Theory Comput.* **10**, 565 (2014).

[123] D. J. Cole, I. C. de Vaca, W. L. Jorgensen, *Chem. Commun.* (2019).

[124] R. W. Zwanzig, *J Chem. Phys.* **22**, 1420 (1954).

[125] M. Ciordia, L. Perez-Benito, F. Delgado, A. A. Trabanco, G. Tresadern, *J. Chem. Inf. Model.* **56**, 1856 (2016).

[126] W. L. Jorgensen, L. L. Thomas, *J. Chem. Theory Comput.* **4**, 869 (2008).

[127] W. L. Jorgensen, C. Ravimohan, *J Chem. Phys.* **83**, 3050 (1985).

[128] S. Piana, A. G. Donchev, P. Robustelli, D. E. Shaw, *J. Phys. Chem. B.* **119**, 5113 (2015).

[129] M. Riquelme, *et al.*, *J. Chem. Inf. Model.* **58**, 1779 (2018).

[130] G. Prampolini, M. Campetella, N. De Mitri, P. R. Livotto, I. Cacelli, *J. Chem. Theory Comput.* **12**, 5525 (2016).

[131] F. Zahariev, N. De Silva, M. S. Gordon, T. L. Windus, M. Dick-Perez, *J. Chem. Inf. Model.* **57**, 391 (2017).

[132] L. Huang, B. Roux, *J. Chem. Theory Comp.* **9**, 3543 (2013).

[133] M. S. S. Institute, The molssiquantum chemistry archive.

[134] A. S. Rose, P. W. Hildebrand, *Nucleic Acids Res.* **43**, W576 (2015).

[135] C. K. Skylaris, P. D. Haynes, A. A. Mostofi, M. C. Payne, *J. Chem. Phys.* **122** (2005).

[136] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).

[137] J. C. Womack, *et al.*, *J. Chem. Theory Comput.* **14**, 1412 (2018).

[138] L. P. Lee, *et al.*, *J. Chem. Theory Comput.* **10**, 5377 (2014).

[139] U. Essmann, *et al.*, *J. Chem Phys.* pp. 8577–8593 (1995).

[140] J. Wang, T. Hou, *J. Chem. Theory Comput.* **7**, 2151 (2011).

[141] M. J. Abraham, *et al.*, *SoftwareX* **1-2**, 19 (2015).

[142] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, W. F. van Gunsteren, *Chem. Phys. Lett.* **222**, 529 (1994).

[143] B. Hess, H. Bekker, H. J. C. Berendsen, G. E. M. F. Johannes, *J. Comput. Chem.* **18**, 1463 (1997).

[144] C. H. Bennett, *J. Comput. Phys.* **22**, 245 (1976).

[145] C. Caleman, *et al.*, *J. Chem. Theory Comput.* **8**, 61 (2012).

[146] J. Cerezo, G. Prampolini, I. Cacelli, *Theor. Chem. Acc.* **137**, 80 (2018).

[147] S. J. Weiner, *et al.*, *J. Am. Chem. Soc.* **106**, 765 (1984).

[148] B. Doherty, X. Zhong, O. Acevedo, *J. Phys. Chem. B* **122**, 2962 (2018).

[149] C. M. Breneman, K. B. Wiberg, *J. Comput. Chem.* **11**, 361 (1990).

[150] T. A. Manz, T. Chen, D. J. Cole, N. G. Limas, B. Fiszbein, *RSC Advances* **9**, 19297 (2019).

[151] T. A. Manz, N. G. Limas, *RSC Advances* **6**, 47771 (2016).

[152] N. G. Limas, T. A. Manz, *RSC Advances* **6**, 45727 (2016).

[153] T. A. Manz, *RSC Advances* **7**, 45552 (2017).

[154] N. G. Limas, T. A. Manz, *RSC Advances* **8**, 2678 (2018).

[155] T. Manz, T. Chen, D. J. Cole, N. G. Limas, B. Fiszbein (2018). 10.26434/chemrxiv.7205693.v1.

[156] A. J. Stone, *J. Science* **321**, 787 (2008).

[157] E. T. Walters, M. Mohebifar, E. R. Johnson, C. N. Rowley, *J. Phys. Chem. B* **122**, 6690 (2018).

[158] A. J. Misquitta, A. J. Stone (2018). Arxiv:1806.06737.

[159] M. Mohebifar, E. R. Johnson, C. N. Rowley, *J Chem. Theory Comp.* **13**, 6146 (2017).

[160] A. D. Becke, E. R. Johnson, *J. Chem. Phys.* **127**, 154108 (2007).

[161] K. M. Visscher, D. P. Geerke, *J. Phys. Chem. B* **124**, 1628 (2020).

[162] C. G. Mayne, J. Saam, K. Schulten, E. Tajkhorshid, J. C. Gumbart, *J. Comp. Chem.* **34**, 2757 (2013).

[163] S. Kraner, G. Prampolini, G. Cuniberti, *J. Phys. Chem. C.* **121**, 17088 (2017).

[164] O. Andreussi, I. G. Prandi, M. Campetella, G. Prampolini, B. Mennucci, *J. Chem. Theory Comput.* **13**, 4636 (2017).

[165] W. L. Jorgensen, *Acc. Chem. Res.* **42**, 724 (2009).

[166] L. Wang, *et al.*, *J. Am. Chem. Soc.* **137**, 2695 (2015).

[167] J. D. Chodera, *et al.*, *Curr. Opin. Struc. Biol.* **21**, 150 (2011).

[168] R. C. Bernardi, M. C. Melo, K. Schulten, *BBA-Gen. Subjects* **1850**, 872 (2015).

[169] V. Spiwok, Z. Sucur, P. Hosek, *Biotechnol. Adv.* **33**, 1130 (2015).

[170] S. Riniker, *J. Chem. Inf. Model.* **58**, 565 (2018).

[171] D. J. Cole, J. T. Horton, L. Nelson, V. Kurdekar, *Future Med. Chem.* **11**, 2359 (2019).

[172] P. S. Nerenberg, T. Head-Gordon, *Curr. Opin. Struc. Biol.* **49**, 129 (2018).

[173] R. M. Betz, R. C. Walker, *J. Comput. Chem.* **36**, 79 (2015).

[174] I. Cacelli, A. Cimoli, P. R. Livotto, G. Prampolini, *J. Comput. Chem.* **33**, 1055 (2012).

[175] M. J. Robertson, J. Tirado-Rives, W. L. Jorgensen, *J. Phys. Chem. Lett.* **7**, 3032 (2016).

[176] C. D. Christ, T. Fox, *J. Chem. Inf. Model.* **54**, 108 (2013).

[177] C. E. Fitzgerald, *et al.*, *Nat. Struct. Mol. Biol* **10**, 764 (2003).

[178] G. Barreiro, *et al.*, *J. Med. Chem.* **50**, 5324 (2007).

[179] J. Michel, J. Tirado-Rives, W. L. Jorgensen, *J. Phys. Chem. B* **113**, 13337 (2009).

[180] I. Cabeza de Vaca, Y. Qian, J. Z. Vilseck, J. Tirado-Rives, W. L. Jorgensen, *J. Chem. Theory Comput.* **14**, 3279 (2018).

[181] P. Liu, B. Kim, R. A. Friesner, B. Berne, *P. Natl. Acad. Sci. USA* **102**, 13749 (2005).

[182] L. Wang, B. Berne, R. A. Friesner, *P. Natl. Acad. Sci. USA* **109**, 1937 (2012).

[183] Y. Qian, *et al.*, *J. Phys. Chem. B* **123**, 8675 (2019).

[184] Z. Liu, S. J. Barigye, M. Shahamat, P. Labute, N. Moitessier, *J. Chem. Inf. Model.* **58**, 194 (2018).

[185] P. Bleiziffer, K. Schaller, S. Riniker, *J. Chem. Inf. Model.* **58**, 579 (2018).

[186] R. Tacke, S. Dörrich, *Workshop on Embracing Global Computing in Emerging Economies* (Springer, 2015), pp. 29–59.

[187] R. Tacke, *et al.*, *Organometallics* **23**, 4468 (2004).

[188] H. Lee, M. Fischer, B. K. Shoichet, S.-Y. Liu, *J. Am. Chem. Soc.* **138**, 12021 (2016).

[189] P. G. Campbell, A. J. Marwitz, S.-Y. Liu, *Angew. Chem. Int. Edit.* **51**, 6074 (2012).

[190] M. J. Bosdet, W. E. Piers, *Can. J. Chem.* **87**, 8 (2009).

[191] Z. Liu, T. B. Marder, *Angew. Chem. Int. Edit.* **47**, 242 (2008).

[192] S. Munetoh, T. Motooka, K. Moriguchi, A. Shintani, *Comp. Mater. Sci.* **39**, 334 (2007).

[193] E. R. Cruz-Chu, A. Aksimentiev, K. Schulten, *J. Phys. Chem. B* **110**, 21497 (2006).

[194] T. A. Hilder, *et al.*, *Micro & Nano Letters* **5**, 150 (2010).

[195] M. Schauperl, *et al.* (2019).

[196] L.-P. Wang, C. Song, *J Chem. Phys.* **144**, 214108 (2016).

[197] R. M. Parrish, *et al.*, *J. Chem. Theory Comput.* **13**, 3185 (2017).

[198] F. Gao, L. Han, *Comput. Optim. Appl.* **51**, 259 (2012).

[199] R. Storn, K. Price, *J. Global Optim.* **11**, 341 (1997).

[200] D. Mobley, The open force field 1.0 small molecule force field, our first optimized force field (codename "parsley") (2019).

[201] A. T. Hagler, *J. Comput Aid. Mol. Des.* **33**, 205 (2019).

[202] S. Amirjalayer, R. Q. Snurr, R. Schmid, *J. Phys. Chem. C* **116**, 4921 (2012).

[203] S. Kantonen, H. S. Muddana, N. M. Henriksen, L.-P. Wang, M. Gilson (2019).