UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Sciences

School of Chemistry

Methods for Accurate and Efficient Simulation of the Conformational Landscape of Ligands

by

João Pedro dos Santos Morado

MSc

A thesis for the degree of

Doctor of Philosophy

August 2022

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Chemistry

Doctor of Philosophy

Methods for Accurate and Efficient Simulation of the Conformational Landscape of Ligands

by João Pedro dos Santos Morado

Ligand modelling is an essential element of drug discovery. To accurately simulate chemical and physical phenomena, it is necessary to employ molecular models that provide reliable results in a timely fashion. The gold standard method in ligand modelling remains quantum mechanics (QM). Owing to the high computational cost of QM methods, their use in *ab initio* simulations is limited to all but the simplest systems. Molecular mechanics force fields (MM FFs) have also been around for decades. They stand as the cheapest alternative to QM methods, despite their widely-known accuracy limitations. A promising new alternative to FFs are the machine-learning (ML) potentials. ML potentials are molecular models based on artificial intelligence, seemingly more flexible and accurate than FFs, although more computationally costly.

For a given FF functional form, the quality of the parameterisation is crucial and determines how accurately observable properties can be computed from simulations. Whilst accurate FF parameterisations are available for biomolecules, the parameterisation of novel drug candidates is particularly challenging, as these may involve functional groups and interactions for which accurate parameters are not available. To address the problem of FF accuracy, we developed ParaMol, software that has the capability of reparameterising class I FFs with a special focus on druglike molecules. We demonstrate that, within the constraints of a FF

functional form, ParaMol can derive near-ideal FF parameters. Additionally, we illustrate the best practices to follow when employing specific parameterisation routes; the sensitivity of different fitting data sets, such as relaxed dihedral scans and configurational ensembles, to the parameterisation procedure; and the features of the various weighting methods available to weight configurations.

Monte Carlo (MC) and molecular dynamics (MD) simulations can be performed using FFs, ML potentials, or QM methods. The higher the level of theory used in MD or MC simulations, the more reliable the structural information extracted from them will be, despite the increase in computational cost. To combine the accuracy of *ab initio* simulations with the efficiency of classical ones, we present a multilevel MC method that allows quantum configurational ensembles to be generated while keeping the computational cost at a minimum. We show that FF reparameterisation is an efficient route to generate FFs that reproduce QM results more closely, which in turn can be used as low-cost models to approach the gold standard QM accuracy. We demonstrate that the MC acceptance rate is strongly correlated with various phase space overlap measurements, constituting a robust metric to evaluate the similarity between any two levels of theory. As more advanced applications, we apply the nMC-MC algorithm to generate the QM/MM distribution of a ligand in aqueous solution and present a self-parameterising version of the method.

Recently, ML potentials have emerged as an alternative to FFs. However, owing to their newness, there are many unanswered questions concerning their applicability that must be addressed. To this end, we present a comparative study that evaluates the performance of a ML potential, a traditional FF, and an optimally tuned FF in the modelling of a set of 10 γ-fluorohydrins that exhibit a complex interplay between intra- and intermolecular interactions in determining conformer stability. For this set of molecules, we benchmark the performance of each molecular model, evaluating their energetic, geometric, and sampling accuracy relative to quantum mechanical data, both in the gas phase and chloroform solution. We also assess the performance of the aforementioned molecular models in estimating J-coupling constants by comparing their predictions to experimental

data available in chloroform. We then discuss and highlight the strengths and weaknesses of each model, providing guidelines for future development of FFs and ML potentials.

The complexity and scope of the problems addressed in this thesis preclude complete or definitive solutions. Even so, we believe the outcomes of this work may have implications in different areas of chemistry and biology, especially for those interested in modelling the conformational landscape of small organic molecules. The overall conclusions of this thesis are: FFs can be reliably parameterised in an automated fashion using ParaMol; optimally tuned FFs can work as gateways to generate QM ensembles, at least for small molecules in the gas phase; despite the ability of ML potentials to reproduce their training data, the transferability of ML potentials to other domains is limited, and conventional FFs still play an important role in molecular simulations.

Contents

List of Figures

List of Tables

Acronyms

CG	Contracted Gaussian
CSD	Cambridge Structural Database
CV	Collective Variable
DOF	Degree of Freedom
DFT	Density Functional Theory
DFTB	Density Functional Based Tight Binding
ESP	Electrostatic Potential
FF	Force Field
FFTs	Fast Fourier Transforms
GAFF	General AMBER Force Field
GGA	Generalised Gradient Approximation
GTO	Gaussian-Type Orbital
HB	Hydrogen Bond
HF	Hartee-Fock
hMC	Hybrid Monte Carlo
LDA	Local Density Approximation
LJ	Lennard-Jones
LLS	Linear Least Squares
IMHB	Intramolecular Hydrogen Bond
MC	Monte Carlo
MD	Molecular Dynamics
MetaD	Metadynamics
mGGA	Meta-Generalised Gradient Approximation

ML	Machine Learning
MM	Molecular Mechanics
NGWF	Nonorthogonal Generalised Wannier Function
nMC-MC	Nested Markov Chain Monte Carlo
NNP	Neural Network Potential
ONETEP	Order-N Electronic Total Energy Package
PBCs	Periodic Boundary Conditions
PCM	Polarisable Continuum Model
PES	Potential Energy Surface
PME	Particle Mesh Ewald
QM	Quantum Mechanics
QUBEKit	Quantum Mechanical Bespoke Force Field Toolkit
RDF	Radial Distribution Function
RESP	Restrained Electrostatic Potential
RMSD	Root-Mean-Square Deviation
RMSE	Root-Mean-Square Error
SAEP	Sum of the Absolute Error of the Populations
SCC	Self-Consistent Charges
STO	Slater-Type Orbital

Declaration of Authorship

I, João Morado, declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

    - Morado, J.; Mortenson, P. N.; Verdonk, M. L.; Ward, R. A.; Essex, J. W.; Skylaris, C.-K. ParaMol: A Package for Automatic Parameterization of Molecular Mechanics Force Fields. J. Chem. Inf. Model. 2021, 61 (4), 2026–2047. https://doi.org/10.1021/acs.jcim.0c01444

- Morado, J.; Mortenson, P. N.; Nissink, J. W. M.; Verdonk, M. L.; Ward, R. A.; Essex, J. W.; Skylaris, C.-K. Generation of Quantum Configurational Ensembles Using Approximate Potentials. J. Chem. Theory Comput. 2021, 17 (11), 7021–7042. https://doi.org/10.1021/acs.jctc.1c00532

Signed:..... Date:.....

Acknowledgements

First of all, I am endlessly grateful to my supervisors Professor Chris-Kriton Skylaris and Professor Jonathan W. Essex for their continuous guidance, support and patience during the time I spent doing this Ph.D. project. Without their advice and scientific contributions, this work would not have been possible. I would also like to thank my industrial supervisors Dr. J. Willem M. Nissink and Dr. Richard A. Ward from AstraZeneca, and Dr. Paul N. Mortenson and Dr. Marcel L. Verdonk from Astex, for all the insightful discussions we shared, which were instrumental for the development of this work and its application to real-world problems.

I also acknowledge the funding provided by AstraZeneca, and the support I received from the EPSRC Centre for Doctoral Training, Theory and Modelling in Chemical Sciences (TMCS CDT) under Grant EP/L015722/1.

I am very grateful to all the members of the TMCS CDT, the Skylaris group, and the Essex group for the enjoyable times we spent. It was a privilege to meet such interesting people, and I truly appreciated every gesture of help and friendship.

Finally, I would like to deeply thank my partner, Margarida, as well as my parents, Vitor and Matilde, for their unconditional love and support during my studies.

Chapter 1

Project Motivation and Thesis Outline

Ligand modelling is a key aspect of many disciplines in chemical sciences.[7] Various areas of intensive research in the field, such as free energy calculations,[8,9] molecular docking,[10,11] and conformational analysis,[12,13] require frequent modelling of different types of ligands. These applications are of paramount importance to the pharmaceutical industry, since many stages of the drug discovery pipeline heavily rely on theoretical analysis and computer simulation. Lead optimisation, in particular, when performed following a rational approach, can greatly benefit from the structural and energetic information that can be extracted from *in silico* experiments. These computational techniques can produce either novel predictions or corroborate data obtained from other sources.[14–16] Computational experiments not only have the advantage of being faster[17] to perform than *in vitro* and *in vivo* experiments but also of having a substantially lower cost, making their optimisation the way forward towards cheaper and more efficient drug discovery.[18]

The study of the conformational dynamics of molecules free in solution is essential for predicting molecular properties and guiding the rational development

of new pharmaceutical compounds. The latter application is of utmost importance for the pharmaceutical industry, as knowledge of the unbound state is vital to understand the fundamentals of molecular recognition.[12,19–22] Besides the displacement of water from protein binding sites,[23–25] one of the main phenomena that impacts binding affinity is the reorganisation of the unbound state ligand upon binding to its target, a process that is influenced by the change in intramolecular energy of the ligand in adopting the bioactive conformer, as well as the associated loss of entropy.[22] Minimisation of the free energy penalty associated with this structural change is vital to optimising ligand potency, requiring knowledge of the physical interactions that control conformational preferences and methods for conformational analysis if a rational strategy is to be employed.[22] There is a wide range of experimental structural information on pharmaceutical compounds bound to their protein targets.[26,27] However, as it has been emphasised in various studies, the conformations of unbound compounds are still poorly characterised.[12,20,21,28] Therefore, the scientific community must put effort into developing tools that allow fast and reliable characterisation of unbound molecular conformers as these can potentially provide the so-called "missing link" in structure-based drug discovery.[20,28]

The work presented in this thesis concerns the development of methods for accurate and efficient simulation of the conformational landscape of ligands. Reliable molecular simulations require two fundamental components: a molecular model that describes the physics underlying a system of interest, and a method to extensively sample its conformational space. The development of sampling methods is an area of intensive research,[29–34] as nowadays it is still difficult to thoroughly sample the conformational space of molecules. The sampling problem is more significant the more costly a molecular model is, leading to a negative correlation between accuracy and sampling efficiency. The current inability to use quantum mechanical (QM) methods to simulate ligands is entirely due to their computational cost. Unless quantum computers bring significant speed-ups to electronic structure calculations, the routine use of QM methods for simulating the conformational dynamics of molecules remains a mirage. The

quest towards quantum accuracy in ligand modelling is thus an ongoing effort that has no simple solution. It is, however, the gold standard to achieve to obtain reliable predictions from simulations. The way forward to solve this conundrum must necessarily involve developing and improving cheap molecular models that approximate the quantum level of theory. Molecular mechanics force fields (MM FFs) have been around for decades. They stand as the cheapest alternative to QM methods, despite their widely-known accuracy problems. FFs demand a constant improvement effort, justified by their ability to simulate large systems at long time scales. A promising new alternative to FFs are the machine-learning (ML) potentials. ML potentials are molecular models derived using artificial intelligence, seemingly more flexible and accurate than FFs, though more computationally costly. Owing to their newness, there are many unanswered questions concerning their applicability. These must be addressed if the ML potentials are to become the *de facto* alternative to FFs in ligand modelling.

This project was divided into three main research studies. The first study addressed the problem of FF accuracy. It consisted in the development of methods to parameterise molecules by fitting to *ab initio* data. It led to the development of ParaMol, software that aims to ease the process of FF parameterisation. The second study tackled the sampling problem at the QM level. It involved the development of a multilevel Monte Carlo (MC) method capable of generating quantum configurational ensembles while keeping the computational cost at a minimum. This approach aimed to combine the computational efficiency of FFs with the accuracy of the QM level. The third and final study focused on benchmarking and comparing the performance of ML potentials and FFs. It aimed to determine the current strengths and pitfalls of each model and evaluate the levels of accuracy that can be attained. Given their relevance, we believe that the applications and results presented in this thesis may have implications in different areas of chemical sciences with biological relevance, especially for the drug design community.

1.1 Outline

This thesis is structured in such a way that each chapter aims to be self-contained. The general theoretical aspects on which this work is based are first presented, with a special focus on discussing the fundamental aspects of quantum mechanics, statistical mechanics, simulations methods, and molecular mechanics. We then proceed to present the theory, methods, results, discussions, and conclusions of the three research studies that form the novel core of this thesis. The background theory and motivation for these studies are presented at the beginning of the respective chapters. A brief outline of the contents of each chapter is stated in what follows.

Chapter 2 presents the basic theory underlying quantum mechanics. QM methods were recurrently used in this work, as they provide a reference to which the accuracy of FFs and ML potentials can be compared. Thus, this chapter is concerned with the fundamental principles of QM, especially those important for performing QM calculations within a chemical context. After introducing the Born-Oppenheimer approximation, two fundamental electronic structure methods, *viz.*, Hartree-Fock (HF) and density functional theory (DFT), are discussed at length. We then end the chapter with a brief review of the common basis sets used in electronic structure calculations.

Chapter 3 gives an overview of the theory of statistical mechanics, relating it to that of standard simulation methods. The basic properties of thermodynamic ensembles are discussed, and the two main simulation methods used throughout this thesis, *viz.*, molecular dynamics (MD) and MC, are reviewed. The most relevant thermostats and barostats used by these simulation methods are also presented. We finalise with a brief review of enhanced sampling methods.

Chapter 4 presents the fundamentals of molecular mechanics, as this classical framework forms the core of the molecular models used in this thesis. We review the most important classes of FFs and thoroughly discuss their functional forms. Particular emphasis is given to the complexities of class I FFs, such as the general

AMBER force field (GAFF), since this FF class was used as the starting point from which more accurate models were developed. Usual schemes employed to calculate long-range interactions are also discussed, and finally an overview of FF parameterisation methods is presented.

Chapter 5 comprises the first research study of this thesis. It presents the theory, development, and implementation of ParaMol, software that we developed, which aims to ease the process of FF parameterisation. ParaMol has a special focus on the parameterisation of bonded and nonbonded terms of druglike molecules by fitting to *ab initio* data. We demonstrate the capabilities of the software by deriving bonded parameters of three widely-known drug molecules: aspirin, caffeine, and a norfloxacin analogue. Additionally, we illustrate the best practices to follow when employing specific parameterisation routes; the sensitivity of the fitted parameters to the fitting procedure; and the features of the various weighting methods available to weight configurations used in the fitting.

Chapter 6 introduces a multilevel MC method that allows quantum configurational ensembles to be generated while keeping the computational cost at a minimum. We present the theory and algorithm of the methodology and apply it to a set of relevant druglike molecules. We show that FF reparameterisation is an efficient route to accelerate QM-level sampling and discuss the implications and features of the method. As more advanced applications, we present a self-parameterising version of the algorithm, which combines sampling and FF parameterisation in one scheme, and adapt the MC method to generate the QM/MM distribution of a ligand in aqueous solution.

Chapter 7 attempts to answer the question: ”does a machine-learnt potential perform better than an optimally tuned traditional force field?”. Having developed a method to parameterise druglike molecules in Chapter 5, and an algorithm to generate quantum configurational ensembles at a low computational cost in Chapter 6, we apply these techniques to derive optimally tuned FFs, which are tested against an ML potential. To this end, we evaluate the performance of a

standard FF, an optimally tuned FF, and an ML potential in the modelling of a set of γ-fluorohydrins. We assess the performance of each molecular model by comparing its predictions to those obtained from QM methods and experiments. The current strengths and shortcomings of each model are then analysed, from which guidelines for improvement are drawn.

Chapter 8 summarises the main conclusions of this thesis and provides suggestions to guide future research efforts.

Chapter 2

Quantum Mechanics

QM emerged in the 1920s as an alternative to the classical description of systems, and since then it has revolutionised the way nature is understood. Although there is no evidence that the central quantity of QM, the wavefunction, exists, the quantum theory nevertheless provides the most successful representation there is thus far to describe the behaviour of the microscopic world, being widely applied to perform calculations in chemical sciences. Despite its remarkable accuracy, real-world problems are still challenging to be solved quantum mechanically mostly owing to their computational cost, which limits the size of the systems and the time scales that can be simulated. A wide variety of methods have been proposed over the last decades to perform quantum mechanical calculations, with those related to the description of the electronic structure of chemical systems being the focus of the discussion presented next.

2.1 Fundamental principles of quantum mechanics

In the QM representation of the motions of particles, systems are described using a wavefunction, Ψ, an entity that provides a complete QM description of a system. For an N-electron system, the wavefunction depends on a set of spatial ($\boldsymbol{r}_i$) and spin (s_i) coordinates, $\boldsymbol{x} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N, s_1, s_2, \ldots, s_N)$, and

on time t. The wavefunction itself is not a physical observable, but it has a physical interpretation that was suggested by Born. The Born interpretation of the wavefunction states that the probability density of finding an electron i with spin s_i at position $\boldsymbol{r}_i$ and at time t is given by[35]

$$P(\boldsymbol{r}_i, s_i, t) = \Psi(\boldsymbol{r}_i, s_i, t)\Psi^*(\boldsymbol{r}_i, s_i, t) = |\Psi(\boldsymbol{r}_i, s_i, t)|^2 \tag{2.1}$$

This probability, when integrated over all space, must be 1, *i.e.*,

$$\langle \Psi | \Psi \rangle = \int_{-\infty}^{\infty} d^3\boldsymbol{r}_i \, |\Psi(\boldsymbol{r}_i, s_i, t)|^2 = 1 \tag{2.2}$$

meaning that the electron must exist within space. The time evolution of a quantum system governed by a wavefunction is determined by the time-dependent Schrödinger equation, which reads[36]

$$i\hbar \frac{\partial \Psi}{\partial t} = \hat{H}\Psi \tag{2.3}$$

where $\hat{H}$ is the Hamiltonian operator, $\hbar$ is the Planck constant divided by 2π, and i is the imaginary unit. If $\hat{H}$ is time independent, the wavefunction can be separated into a time-independent spatial/spin part, $\psi(\boldsymbol{x})$, and a time-dependent part, $\tau(t)$, as follows

$$\Psi(\boldsymbol{x}, t) = \psi(\boldsymbol{x})\tau(t) \tag{2.4}$$

Moreover, the time-independent version of equation (2.3) only depends on the spatial/spin wavefunction, and it can be written as

$$\hat{H}\psi(\boldsymbol{x}) = E\psi(\boldsymbol{x}) \tag{2.5}$$

where E is the energy of the system, which corresponds to the eigenvalue of the eigenvector $\psi(\boldsymbol{x})$. By introducing equation (2.4) into equation (2.3), the expression of time-dependent part of the wavefunction is obtained, which reads

$$\tau(t) = \exp\left(-iEt/\hbar\right) \tag{2.6}$$

Most problems in quantum chemistry do not depend on time, and therefore one normally seeks to solve equation (2.5) instead of equation (2.3), where determining the energy E corresponds to solving an eigenvalue problem. Finally, the expectation value of a dynamical observable A can be calculated using

$$\langle A \rangle = \frac{\langle \Psi | \hat{A} | \Psi \rangle}{\langle \Psi | \Psi \rangle} \tag{2.7}$$

where the denominator corresponds to the normalisation integral, often required to be unity. Note that in equation (2.7), $\Psi \equiv \psi$ holds if the observable of interest A is static. The probability density, $|\psi(\boldsymbol{r}_i, s_i)|^2$, also retains the same interpretation as that of equation (2.1), though it is now stationary.

2.2 The Born-Oppenheimer approximation

Consider a stationary system composed of n nuclei and N electrons, positively and negatively charged, respectively. If relativistic effects and spin-orbit interactions are neglected, the Hamiltonian such a system can be written as

$$\hat{H}(\boldsymbol{r}, \boldsymbol{R}) = \hat{T}_n(\boldsymbol{R}) + \hat{V}_{nn}(\boldsymbol{R}) + \hat{H}_{el}(\boldsymbol{r}, \boldsymbol{R}) \tag{2.8}$$

where $\boldsymbol{r} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)$ is the set of electronic coordinates, $\boldsymbol{R} = (\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_n)$ is the set of nuclear coordinates, $\hat{H}_{el}(\boldsymbol{r}, \boldsymbol{R})$ is the electronic Hamiltonian, $\hat{T}_n(\boldsymbol{R})$ is the nuclear kinetic energy operator, and $\hat{V}_{nn}(\boldsymbol{R})$ is the

nuclear-nuclear repulsion operator. In atomic units, the latter two operators are of the form

$$\hat{T}_n(\boldsymbol{R}) = -\sum_{\alpha}^{n} \frac{1}{2M_\alpha} \hat{\nabla}_\alpha^2 \tag{2.9}$$

$$\hat{V}_{nn}(\boldsymbol{R}) = \sum_{\alpha}^{n} \sum_{\alpha>\beta}^{n} \frac{Z_\alpha Z_\beta}{R_{\alpha\beta}} \tag{2.10}$$

where the M_α is the mass and Z_α the charge of the nucleus α, and $R_{\alpha\beta} = |\boldsymbol{R}_\alpha - \boldsymbol{R}_\beta|$ is the distance between nuclei α and β. Furthermore, the electronic Hamiltonian of equation (2.8) is given by

$$\hat{H}_{el}(\boldsymbol{r}, \boldsymbol{R}) = \hat{T}_e(\boldsymbol{r}) + \hat{V}_{ee}(\boldsymbol{r}) + \hat{V}_{en}(\boldsymbol{r}, \boldsymbol{R}) \tag{2.11}$$

where $\hat{T}_e(\boldsymbol{r})$ is the electronic kinetic energy operator, $\hat{V}_{ee}(\boldsymbol{r})$ is the electron-electron repulsion operator, and $\hat{V}_{en}(\boldsymbol{r}, \boldsymbol{R})$ is the electron-nucleus attraction operator, which can be written as follows

$$\hat{T}_e(\boldsymbol{r}) = -\sum_{i}^{N} \frac{1}{2} \hat{\nabla}_i^2 \tag{2.12}$$

$$\hat{V}_{ee}(\boldsymbol{r}) = \sum_{i}^{N} \sum_{j>i}^{n} \frac{1}{r_{ij}} \tag{2.13}$$

$$\hat{V}_{en}(\boldsymbol{r}, \boldsymbol{R}) = -\sum_{\alpha}^{n} \sum_{i}^{N} \frac{Z_\alpha}{R_{i\alpha}} \tag{2.14}$$

where $R_{i\alpha} = |\boldsymbol{r}_i - \boldsymbol{R}_\alpha|$ is the distance between nucleus α and electron i, and $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$ is the distance between electrons i and j. In equations (2.9) and (2.12), $\hat{\nabla}_\alpha^2$ and $\hat{\nabla}_i^2$ are the Laplacian operators, given by the sum of second partial derivatives with respect to the coordinates of nucleus α and electron i, *i.e.*,

$$\hat{\nabla}_{\alpha}^{2} = \frac{\partial^2}{\partial r_{\alpha,x}^2} + \frac{\partial^2}{\partial r_{\alpha,y}^2} + \frac{\partial^2}{\partial r_{\alpha,z}^2} \tag{2.15}$$

$$\hat{\nabla}_{i}^{2} = \frac{\partial^2}{\partial r_{i,x}^2} + \frac{\partial^2}{\partial r_{i,y}^2} + \frac{\partial^2}{\partial r_{i,z}^2} \tag{2.16}$$

In equation (2.11), the electron-nucleus attraction operator, $\hat{V}_{en}(\boldsymbol{r}, \boldsymbol{R})$, is the term responsible for coupling the motion of nuclei and electrons, thus preventing the wavefunction to be written as a product of electronic, $\phi(\boldsymbol{x}, \boldsymbol{\xi})$, and nuclear, $\eta(\boldsymbol{\xi})$, wavefunctions as follows

$$\psi(\boldsymbol{x}, \boldsymbol{\xi}) = \phi(\boldsymbol{x}, \boldsymbol{\xi})\eta(\boldsymbol{\xi}) \tag{2.17}$$

where the wavefunctions are time-independent since the system is stationary by construction, and $\boldsymbol{x} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N, s_1, s_2, \ldots, s_N)$ and $\boldsymbol{\xi} = (\boldsymbol{R}_1, \boldsymbol{R}_2, \ldots, \boldsymbol{R}_n, S_1, S_2, \ldots, S_n)$ denote the sets of electronic and nuclear spatial/spin coordinates, respectively. To solve the wavefunction inseparability problem, the most commonly employed scheme is the Born-Oppenheimer approximation,[37] which separates electronic and nuclear motion by exploiting the fact that the nuclear masses are much larger than those of the electrons, causing nuclei to move at much slower speeds than electrons. These observations lead to two assumptions that are at the core of the Born-Oppenheimer approximation, *viz.*: the electrons instantaneously adjust to changes of the nuclei positions; and the nuclei move in a potential field set up by the electrons since the electronic energy varies smoothly as a function of the nuclei coordinates. The Born-Oppenheimer approximation, also known as the adiabatic approximation, permits the wavefunction to be written as

$$\psi(\boldsymbol{x}; \boldsymbol{\xi}) = \phi(\boldsymbol{x}; \boldsymbol{\xi})\eta(\boldsymbol{\xi}) \tag{2.18}$$

where $\phi(\boldsymbol{x};\boldsymbol{\xi})$ is now the adiabatic electronic wavefunction, as the nuclei are assumed to be fixed, making it possible to treat their positions parametrically. In this framework, the total energy of a system is given by the sum of the nuclear and electronic energies, such that

$$E = E_{el}(\boldsymbol{r};\boldsymbol{R}) + E_{nuc}(\boldsymbol{R}) \tag{2.19}$$

$$E_{nuc}(\boldsymbol{R}) = \sum_{\alpha}^{n}\sum_{\alpha>\beta}^{n}\frac{Z_\alpha Z_\beta}{R_{\alpha\beta}} \tag{2.20}$$

where $E_{el}(\boldsymbol{r};\boldsymbol{R})$ and $E_{nuc}(\boldsymbol{R})$ are the electronic and nuclear energies, respectively. From equation (2.20), it can be seen that the nuclear-nuclear repulsion is treated at the classical level. Furthermore, the nuclear kinetic energy vanishes because nuclei are assumed to be stationary. Finally, the electronic energy, $E_{el}(\boldsymbol{r};\boldsymbol{R})$, can be calculated from the following time-independent Schrödinger equation

$$\left[\hat{T}_e(\boldsymbol{r}) + \hat{V}_{ee}(\boldsymbol{r}) + \hat{V}_{en}(\boldsymbol{r};\boldsymbol{R})\right]\phi(\boldsymbol{x};\boldsymbol{\xi}) = E_{el}\phi(\boldsymbol{x};\boldsymbol{\xi}) \tag{2.21}$$

where all operators are given as previously defined. There are two families of quantum chemistry methods commonly employed to approximate the electronic time-independent Schrödinger equation: wavefunction- and DFT-based methods. As the latter family of methods is used throughout this work, the foundations of DFT are explained in detail in Section 2.4. Furthermore, since the nuclear-nuclear repulsion, the nuclear-electron attraction, and the uncorrelated electron-electron repulsion energies used in DFT are the same as those used in HF theory, in the next section we first explain the general features of the HF method.

2.3 The Hartree-Fock method

The HF method is extensively used to approximate both the wavefunction and the ground-state energy of a system.[38–41] HF is considered a mean-field approach, as it reduces an N-electron problem to a one-electron one by assuming independent electrons. In this picture, the interaction between a given electron and all other electrons is calculated in an average fashion, so that each electron interacts with a mean-field potential that represents the average of all electron-electron interactions.

In the HF method, the exact electronic wavefunction is approximated through a single Slater determinant composed of one-electron wavefunctions. This allows the state of an N-electron system occupying N spin orbitals, $(\chi_a, \chi_b, \ldots, \chi_c)$, to be written without the need to specify which electron is in which orbital. The properties of the determinants also ensure that the wavefunction is antisymmetric with respect to the interchange of any two electrons, and that the wavefunction vanishes if any two electrons occupy the same spin orbital, thereby fulfilling the Pauli exclusion principle.[36] The single-determinant antisymmetric electronic wavefunction of an N-electron system is written as[40,42]

$$\phi^{HF}(\boldsymbol{x}) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_a(\boldsymbol{x}_1) & \chi_b(\boldsymbol{x}_1) & \cdots & \chi_c(\boldsymbol{x}_1) \\ \chi_a(\boldsymbol{x}_2) & \chi_b(\boldsymbol{x}_2) & \cdots & \chi_c(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_a(\boldsymbol{x}_N) & \chi_b(\boldsymbol{x}_N) & \cdots & \chi_c(\boldsymbol{x}_N) \end{vmatrix} \tag{2.22}$$

$$= |\chi_a \chi_b \cdots \chi_c\rangle \tag{2.23}$$

where the normalisation factor, $1/\sqrt{N!}$, is implicitly included in the short-hand notation used in equation (2.23). Each spin orbital, $\chi_a(\boldsymbol{x}_i)$, is formed by the product of a spatial orbital, $\sigma_a(\boldsymbol{r}_i)$, and a spin function, which can be either $\alpha_a(s_i)$ (spin up, $\uparrow$) or $\beta_a(s_i)$ (spin down, $\downarrow$). Hence, $\chi_a(\boldsymbol{x}_i)$ can be written as

$$\chi_a(\boldsymbol{x}_i) = \begin{cases} \sigma_a(\boldsymbol{r}_i)\alpha_a(s_i) \\ or \\ \sigma_a(\boldsymbol{r}_i)\beta_a(s_i) \end{cases} \tag{2.24}$$

The spin functions are orthonormal, and so are the spin orbitals. Having stated how Slater determinants are defined, the HF ground-state wavefunction and energy can now be determined using the variational theorem, which states that the energy of a system, E, is higher than that of its ground state, E_0, unless the wavefunction corresponds to the true ground-state wavefunction. Therefore, the following relation holds

$$E[\phi] \geq E_0 \tag{2.25}$$

where the energy is explicitly denoted as a functional of the wavefunction since the variational flexibility is in the choice of the spin orbitals that compose the wavefunction. Since the best single-determinant wavefunction is the one that gives the lowest possible HF energy, the HF equation can be obtained by minimising the HF energy with respect to the choice of spin orbitals*, resulting in the following eigenvalue equation

$$\hat{f}_a(\boldsymbol{x}_i)\chi_a(\boldsymbol{x}_i) = \epsilon_a\chi_a(\boldsymbol{x}_i) \tag{2.26}$$

where ϵ_a is the energy associated with the orbital $\chi_a(\boldsymbol{x}_i)$, and $\hat{f}_a(\boldsymbol{x}_i)$ denotes the Fock operator that acts on that orbital, which, in turn, is given by

$$\hat{f}_a(\boldsymbol{x}_i) = -\frac{1}{2}\hat{\nabla}_i^2 - \sum_{\alpha}^{n}\frac{Z_\alpha}{R_{i\alpha}} + v_a^{HF}(\boldsymbol{x}_i) \tag{2.27}$$

*See Szabo and Ostlund [42], pp. 31–38, or Schatz and Ratner [36], pp. 6–7, for a detailed explanation of the variational method.

where the first and second terms form the so-called core Hamiltonian and correspond to the kinetic and electron-nuclear attraction operators, respectively, and $v_a^{HF}(\boldsymbol{x}_i)$ is the mean-field potential experienced by the electron on the ath orbital due to the presence of the other electrons.

Starting from an initial guess of spin orbitals, the HF method consists in self-consistently solving the set of one-electron eigenvalue problems defined by equation (2.26). The requirement for an iterative solution arises due to the dependence of $v_a^{HF}(\boldsymbol{x}_i)$ on the spin orbitals of the other $N-1$ electrons. The occurrence of this dependence can be seen by writing $v_a^{HF}(\boldsymbol{x}_i)$ in terms of the Coulomb, $\hat{J}$, and exchange, $\hat{K}$, operators as follows

$$v_a^{HF}(\boldsymbol{x}_i) = \sum_{b \neq a} \left[\hat{J}_b(\boldsymbol{x}_i) - \hat{K}_b(\boldsymbol{x}_i) \right] \tag{2.28}$$

where the sum runs over all the $N-1$ orbitals that are not χ_a. In equation (2.28), the Coulomb operator, $\hat{J}_b$, gives the average local potential at position $\boldsymbol{r}_i$ that arises due to the charge distribution of any electron in spin orbital χ_b. This Coulomb operator reads

$$\hat{J}_b(\boldsymbol{x}_i) = \int ds_j \int d^3\boldsymbol{r}_j \, \frac{\chi_b^*(\boldsymbol{x}_j)\chi_b(\boldsymbol{x}_j)}{r_{ij}} = \int d^3\boldsymbol{r}_j \, \frac{\sigma_b^*(\boldsymbol{r}_j)\sigma_b(\boldsymbol{r}_j)}{r_{ij}} \tag{2.29}$$

which only depends on the spatial orbitals because the integration over the spin variable is always 1. Furthermore, the exchange operator $\hat{K}_b$, which arises from the antisymmetry requirement of the wavefunction, is defined by its action on the spin orbital $\chi_a(\boldsymbol{x}_i)$. This exchange operator can be written as

$$\hat{K}_b(\boldsymbol{x}_i)\chi_a(\boldsymbol{x}_i) = \left[\int ds_i ds_j \int d^3\boldsymbol{r}_j \, \frac{\chi_b^*(\boldsymbol{x}_j)\chi_a(\boldsymbol{x}_j)}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} \right] \chi_b(\boldsymbol{x}_i) \tag{2.30}$$

The integral of equation (2.30) only needs to be evaluated for orbitals with parallel spins, as it vanishes for orbitals with anti-parallel spins, causing the

motion of electrons with opposite spin to be uncorrelated. The total HF energy can now be written as

$$E^{HF} = \left\langle \chi_a \left| \sum_a^N \frac{1}{2}\hat{\nabla}_i^2 - \sum_a^N \sum_\alpha^n \frac{Z_\alpha}{R_{i\alpha}} \right| \chi_a \right\rangle + \sum_{a \geq b} \left[\langle \chi_a | \hat{J}_b | \chi_a \rangle - \langle \chi_a | \hat{K}_b | \chi_a \rangle \right] \quad (2.31)$$

Note that the $a \neq b$ restriction imposed in equation (2.28) was removed since $\langle \chi_a | \hat{J}_b | \chi_a \rangle \equiv \langle \chi_a | \hat{K}_b | \chi_a \rangle$. The integration variables are assumed to be $\boldsymbol{x}_i$ for the one-electron operators (kinetic and nucleus-electron attractions), and $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ for the two-electron operators (Coulomb and exchange).

In summary, the HF method describes N-electron systems by considering a mean-field potential that includes the Coulomb and non-local exchange contributions. Hence, the lack of electronic correlation in HF leads to poor accuracy in systems where a proper description of correlation effects is critical. Many post-HF correlated methods have been developed over the years to improve the description of chemical systems, such as, *e.g.*, Møller–Plesset perturbation theory,[43] configuration interaction,[44] coupled cluster,[45] and multi-configurational self-consistent field approaches.[46–48] Despite their differences, all these methods attempt to recover the missing static and dynamic correlation by adding more Slater determinants to the HF single-determinant wavefunction. Alternatively, DFT can also be used to improve the HF results. This method is explained at length in the next section, as it is widely used in this thesis.

2.4 Density functional theory

DFT is based on two seminal papers of Hohenberg, Kohn and Sham.[49,50] The central quantity of DFT is the electronic density, $\rho(\boldsymbol{r}_i)$, which is defined as the integral over all spin coordinates and the spatial coordinates of $N-1$ electrons, *i.e.*,

$$\rho(\boldsymbol{r}_i) = N \int \ldots \int ds_1 d^4\boldsymbol{x}_2 d^4\boldsymbol{x}_3 \ldots d^4\boldsymbol{x}_N \, |\phi(\boldsymbol{r}; \boldsymbol{R})|^2 \tag{2.32}$$

which has the advantage of depending only on 3 degrees of freedom (DOFs). This simplification greatly reduces the dimensionality of quantum chemistry calculations, as to compute the ground-state energy there is no need to solve the Schrödinger equation and determine the $4N$ dimensional wavefunction. This allows to computationally treat much larger systems, especially now that linear-scaling DFT has become widespread, owing much to the development of the Order-N Electronic Total Energy Package (ONETEP),[51] which enables simulation of realistic systems of thousands of atoms.

The early foundations of DFT date back to 1927, when Thomas and Fermi independently developed an approach to solve many-body problems in which the electronic density is the central variable rather than the wavefunction.[52,53] In the Thomas-Fermi model, the energy functional is composed of three terms: the kinetic energy term, corresponding to that of the uniform electron gas; the electron-electron interaction term, approximated by the classical Coulomb potential like in HF; and the nuclear-electron attraction term. This model was later extended by Dirac, in 1930, which augmented it by introducing exchange effects.[54] It was not until 1964, however, that the modern foundations of DFT were laid out.[49] This was done by Hohenberg and Kohn, authors of the two fundamental theorems that have paved the way towards the development of accurate methods for electronic structure calculations using DFT. These two Hohenberg-Kohn theorems are the subject of the discussion presented in the next section.

2.4.1 The Hohenberg-Kohn theorems

The first Hohenberg-Kohn theorem proves the existence of a one-to-one mapping between the electronic density and the external potential.[49] Physically, the external potential represents the nuclear attraction part of the electronic

Hamiltonian operator. To demonstrate this unequivocal one-to-one mapping, the proof presented by Hohenberg and Kohn proceeds by *reductio ad absurdum*, as described below.

Assume the existence of two different external potentials, $v_{ext}(\boldsymbol{r}_i)$ and $v'_{ext}(\boldsymbol{r}_i)$, which differ by more than an additive constant but give rise to the same electronic density, $\rho(\boldsymbol{r}_i)$. The ground-state wavefunctions associated with these external potentials are given by Ψ_0 and Ψ'_0 (which are necessarily different since $v_{ext}(\boldsymbol{r}_i) - v'_{ext}(\boldsymbol{r}_i) \neq c$), and the associated Hamiltonian and ground-state energies are denoted as H and H', and E_0 and E'_0, respectively. The variational theorem stated in equation (2.25) establishes that the energy of a system is higher than that of its ground state unless the wavefunction is the true ground-state wavefunction. Therefore, the following relations holds

$$E'_0 = \left\langle \Psi'_0 \middle| H' \middle| \Psi'_0 \right\rangle < \langle \Psi_0 | H' | \Psi_0 \rangle = \langle \Psi_0 | H - v_{ext} + v'_{ext} | \Psi_0 \rangle \tag{2.33}$$

which can be rewritten as

$$E'_0 < E_0 + \int d^3\boldsymbol{r}_i \ \left[v_{ext}(\boldsymbol{r}_i) - v'_{ext}(\boldsymbol{r}_i) \right] \rho(\boldsymbol{r}_i) \tag{2.34}$$

By interchanging the primed and unprimed quantities, the following inequation is obtained

$$E_0 < E'_0 + \int d^3\boldsymbol{r}_i \ \left[v'_{ext}(\boldsymbol{r}_i) - v_{ext}(\boldsymbol{r}_i) \right] \rho(\boldsymbol{r}_i) \tag{2.35}$$

Finally, adding equations (2.34) and (2.35) gives

$$E'_0 + E_0 < E'_0 + E_0 \tag{2.36}$$

which is clearly an inconsistency. Therefore, the conclusion is that two external potentials that differ by more than a constant must necessarily give rise to

different electronic densities. A corollary of this theorem is that the external potential is a functional of the electronic density. Furthermore, since the number of electrons and the external potential completely define the Hamiltonian and, consequently, the wavefunction, this means that the wavefunction must also be a functional of the electronic density.

The second Hohenberg-Kohn theorem uses the corollaries of the first theorem to demonstrate that the ground-state energy can be obtained variationally by minimising the electronic density.[49] To show that his holds true for any N-electron system governed by an external potential, first recall that both the wavefunction and the Hamiltonian are a functional of the electronic density, and thus the following can be written

$$F[\rho] = \langle \Psi | \hat{T}_e + \hat{V}_{ee} | \Psi \rangle \tag{2.37}$$

where $F[\rho]$ is the so-called universal functional. Likewise, since there is a one-to-one mapping between $v_{ext}(\boldsymbol{r}_i)$ and $\rho(\boldsymbol{r}_i)$, the total energy can also be written as a functional of the electronic density, *i.e.*,

$$E[\rho] = \langle \Psi | \hat{T}_e + \hat{V}_{ee} + v_{ext} | \Psi \rangle \tag{2.38}$$

$$= F[\rho] + \int d^3\boldsymbol{r}_i \, v_{ext}(\boldsymbol{r}_i)\rho(\boldsymbol{r}_i) \tag{2.39}$$

Hence, since the energy is a functional of the electronic density, the energy can be minimised by varying ρ under the constraint of preservation of the total number of electrons N, a quantity related to ρ by

$$N[\rho] = \int d^3\boldsymbol{r}_i \, \rho(\boldsymbol{r}_i) = N \tag{2.40}$$

Moreover, similarly to what was done in equation (2.33), the variational principle asserts that

$$\langle \Psi | \hat{T}_e + \hat{V}_{ee} + v_{ext} | \Psi \rangle > \langle \Psi_0 | \hat{T}_e + \hat{V}_{ee} + v_{ext} | \Psi_0 \rangle \tag{2.41}$$

where $\Psi_0 = \Psi[\rho_0]$ is the ground-state wavefunction, which is a functional of the exact ground-state electronic density. This leads to the conclusion that

$$E[\rho] > E[\rho_0] = E_0 \tag{2.42}$$

which demonstrates that the ground-state energy may be found variationally using the electronic density as a variable. Despite the importance of the Hohenberg-Kohn theorems, a practical way of performing DFT calculations was only introduced one year later, in 1965, by Kohn and Sham. The general features of this method are described in what follows.

2.4.2 The Kohn-Sham equations

In this section, the Kohn-Sham DFT formalism is presented.[50] Kohn-Sham DFT is extensively used in quantum chemistry, as it introduces a practical way of performing electronic structure calculations. Before discussing the Kohn-Sham DFT formalism, recall that the universal functional contains the contributions of the kinetic energy, $T_e[\rho]$, the classical Coulomb interaction, $J[\rho]$, and the non-classical energy, $E_{ncl}[\rho]$. Therefore, the universal functional can be written as

$$\begin{aligned} F[\rho] &= T_e[\rho] + V_{ee}[\rho] \\ &= T_e[\rho] + J[\rho] + E_{ncl}[\rho] \end{aligned}$$

In this equation, the only term that is known is $J[\rho]$, which is the Hartree energy, corresponding to the HF-like electron-electron interactions. To determine expressions for the other two terms, the Kohn-Sham ansatz must be used, in which the

interacting system is replaced by a non-interacting one, in such a way that the ground-state density of the latter is the same as that of the former. Like in the HF theory, this ansatz corresponds to assuming independence of electrons, making it possible to decompose the wavefunction into an antisymmetric product of one-electron spin orbitals. This gives rise to a Slater determinant made of the so-called Kohn-Sham orbitals, which reads

$$\phi^{DFT}[\rho(\boldsymbol{r}_i)] = |\chi_a(\boldsymbol{r}_1)\chi_b(\boldsymbol{r}_2)\dots\chi_c(\boldsymbol{r}_N)\rangle = |\chi_a\chi_b\dots\chi_c\rangle \tag{2.43}$$

where the same short-hand notation previously presented in equation (2.23) was used. The kinetic energy of such a Kohn-Sham non-interacting system is thus given by

$$T_s[\rho] = -\frac{1}{2}\left[\sum_a^{N_\alpha}\langle\sigma_a^\alpha|\nabla_i^2|\sigma_a^\alpha\rangle + \sum_b^{N_\beta}\left\langle\sigma_b^\beta\middle|\nabla_i^2\middle|\sigma_b^\beta\right\rangle\right] \tag{2.44}$$

where the sums run over all N_α and N_β spatial orbitals σ_a^α and σ_b^β with α and β spin, respectively, and the subscript i in ∇_i^2 refers to the dummy variable of integration $\boldsymbol{r}_i$. Note that, however, this is not equal to the kinetic energy of the interacting system. Kohn and Sham accounted for that difference by introducing the following separation of the universal functional

$$F[\rho] = T_s[\rho] + J[\rho] + E_{xc}[\rho] \tag{2.45}$$

where the exchange-correlation functional, $E_{xc}[\rho]$, accounts for everything that is unknown and reads

$$E_{xc}[\rho] = (T_e[\rho] - T_s[\rho]) + (V_{ee}[\rho] - J[\rho]) \tag{2.46}$$

with $V_{ee}[\rho]$ representing the exact electron-electron interactions of the interacting system. The problem now is to determine the orbitals of the non-interacting

system so that they reproduce the ground-state density of the interacting one. Before doing this, let us first write the total Kohn-Sham energy as

$$E^{KS}[\rho] = T_s[\rho] + \frac{1}{2}\int\int d\boldsymbol{r}_i d\boldsymbol{r}_j \, \frac{\rho(\boldsymbol{r}_i)\rho(\boldsymbol{r}_j)}{r_{ij}} + \int d\boldsymbol{r}_i \, v_{ext}(\boldsymbol{r}_i)\rho(\boldsymbol{r}_i) + E_{xc}[\rho] \quad (2.47)$$

where the electronic density can be calculated using the following expression

$$\rho(\boldsymbol{r}_i) = \sum_{a}^{N_\alpha} |\sigma_a^\alpha(\boldsymbol{r}_i)|^2 + \sum_{b}^{N_\beta} |\sigma_b^\beta(\boldsymbol{r}_i)|^2 \quad (2.48)$$

From the second Hohenberg-Kohn theorem, it is known that the ground-state energy may be found variationally using the electronic density as a variable. Therefore, by applying the variational principle to minimise $E^{KS}[\rho]$ with respect to ρ, the following set of one-electron Kohn-Sham equations is obtained

$$\hat{h}^{KS}(\boldsymbol{r}_i)\sigma_a^s(\boldsymbol{r}_i) = \epsilon_a^s(\boldsymbol{r}_i)\sigma_a^s(\boldsymbol{r}_i) \quad (2.49)$$

where the s superscript denotes the spin of the spatial orbital, $\sigma_a^s(\boldsymbol{r}_i)$. The Kohn-Sham Hamiltonian, $\hat{h}^{KS}(\boldsymbol{r}_i)$, and the effective potential, $v_{eff}(\boldsymbol{r}_i)$, read

$$\hat{h}^{KS}(\boldsymbol{r}_i) = -\frac{1}{2}\nabla_i^2 + v_{eff}(\boldsymbol{r}_i) \quad (2.50)$$

$$v_{eff}(\boldsymbol{r}_i) = v_{ext}(\boldsymbol{r}_i) + \int d\boldsymbol{r}_j \, \frac{\rho(\boldsymbol{r}_j)}{r_{ij}} + v_{xc}(\boldsymbol{r}_i) \quad (2.51)$$

with the exchange-correlation potential of equation (2.51) having the form

$$v_{xc}(\boldsymbol{r}_i) = \frac{\delta E_{xc}[\rho]}{\delta\rho(\boldsymbol{r}_i)} \quad (2.52)$$

In practice, the Kohn-Sham DFT approach proceeds by first calculating the effective potential given by equation (2.51) from a provided initial guess of the electronic density. The set of Kohn-Sham equations (2.49) is then solved, resulting in new spin orbitals that are employed to calculate the electronic density using equation (2.48), and the Kohn-Sham energy resorting to equation (2.47). This procedure is repeated iteratively until convergence, when it is assumed that the ground-state density and energy have been found.

It is worth noting that the Kohn-Sham method turns the problem of interacting electrons in an external potential into that of Kohn-Sham non-interacting electrons in an effective potential. Therefore, from the first Hohenberg-Kohn theorem, it follows that there is a one-to-one mapping between the effective potential and the electronic density, implying that the exact ground-state electronic density may only be found if the exact form of $E_{xc}[\rho]$ is known. Even though Kohn-Sham DFT is a formally exact framework, the exact form of $E_{xc}[\rho]$ is not known except for the free electron gas. Therefore, it is necessary to approximate $E_{xc}[\rho]$ so that the missing exchange-correlation interactions are taken into account. In the next section, we delve into the world of exchange-correlation functionals and present their main categories.

2.4.3 The exchange-correlation functional

The world of exchange-correlation functionals is populated by various methods that attempt to somewhat capture the exchange and correlation effects that are missing from the universal functional. This is a rapidly- and ever-evolving field, in which several categories of exchange-correlation functionals have already been proposed, some more accurate than others, depending on the assumptions and approximations implied in their development.

The functionals based on the local density approximation (LDA) are amongst those with the simplest description of the electronic density. LDA functionals assume that the local electronic density is the same as that of the uniform

electron gas, thereby neglecting any information regarding the derivatives of the electronic density. These functionals are typically separated into exchange, $E_X^{LDA}[\rho]$, and correlation, $E_C^{LDA}[\rho]$, parts as follows

$$E_{xc}^{LDA}[\rho] = E_x^{LDA}[\rho] + E_c^{LDA}[\rho] \tag{2.53}$$

with the analytical form of $E_x^{LDA}[\rho]$ being given by

$$E_x^{LDA}[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \int d^3\boldsymbol{r}_i \, \rho^{4/3}(\boldsymbol{r}_i) \tag{2.54}$$

Various expressions have been proposed thus far to approximate $E_c^{LDA}[\rho]$, with a few examples of LDA functionals being VWN,[55] PW92,[56] and CAPZ.[57,58]

The next obvious step to increase the performance of LDA functionals is to include the description of some local features of the electronic density. This is the approach taken by the functionals based on the generalised gradient approximation (GGA), for which the exchange-correlation energy reads

$$E_{xc}^{GGA}[\rho] = \int d^3\boldsymbol{r}_i \, \epsilon_{xc}\left[\rho(\boldsymbol{r}_i), \nabla\rho(\boldsymbol{r}_i)\right] \tag{2.55}$$

where ϵ_{xc} is the exchange-correlation energy density, representing the energy per electron as a function of the spatial coordinates. As can be seen from equation (2.55), GGA functionals depend not only on the electronic density but also on its gradient, allowing more accurate representation of systems for which the electronic density is not constant. The most popular functionals belonging to this category are PBE,[59] PW91,[60] and BLYP.[61,62]

In addition to the gradient of the electronic density, meta-GGA (mGGA) functionals, such as TPSS[63] and M06-L,[64] further include its Laplacian, $\nabla^2\rho(\boldsymbol{r}_i)$. The general form of the mGGA functionals reads

$$E_{xc}^{mGGA}[\rho] = \int d^3\boldsymbol{r}_i \, \epsilon_{xc} \left[\rho(\boldsymbol{r}_i), \nabla\rho(\boldsymbol{r}_i), \nabla^2\rho(\boldsymbol{r}_i)\right] \tag{2.56}$$

Moreover, hybrid functionals, also known as adiabatic connection method functionals, besides the standard GGA or mGGA exchange-correlation, also include a contribution from the exact HF non-local exchange energy, E_x^{HF}, calculated as a functional of the Kohn-Sham orbitals. Hybrid functionals attempt to mitigate the so-called self-interaction error, and they have the following general functional form

$$E_H^{hyb}[\rho] = (1-a)E_{xc}^{(m)GGA}[\rho] + aE_x^{HF} \tag{2.57}$$

where a is a parameter that controls the amount of HF exchange energy that is introduced. Undoubtedly, the most popular hybrid functional is B3LYP,[65] but other options are available, such as, *e.g.*, B1PW91,[66] PBE0,[67] VV10,[68,69] and ωB97X.[70]

Finally, double-hybrid functionals are the most recent development in the field. These schemes add a second-order Møller-Plesset correlation energy term, E_c^{MP2}, obtained from the Kohn-Sham GGA or mGGA orbitals and eigenvalues, to the functional form of the hybrid functionals.[71] The general expression of double-hybrid functionals is given by

$$E_{xc}^{DH}[\rho] = a_x E_x^{HF} + (1-a_x)E_x^{(m)GGA}[\rho] + (1-a_c)E_c^{(m)GGA}[\rho] + a_c E_c^{MP2} \tag{2.58}$$

where a_x and a_c control the relative contributions of each term to the exchange and correlation parts, respectively. The first three terms are computed in a self-consistent fashion following the standard Kohn-Sham approach, while E_c^{MP2} is added *a posteriori*.[72] Typical double-hybrid functionals are B2-PLYP[73] and PBE0-DH.[74]

From the discussion presented above, it is clear that there is a large variety of exchange-correlation functionals. These are diagrammatically organised in terms of their accuracy in the so-called Jacob's ladder of density functional approximations for the exchange-correlation energy,[75] of which the lowest rung corresponds to the HF theory, and the highest one to the double-hybrid approaches. As a general rule, the higher one is on Jacob's ladder, the higher the accuracy and the computational cost of the DFT calculations. It is worth noting, however, that this is not always the case, as the performance of functionals is often system-specific, and it is good practice to benchmark their accuracy before applying them to systems of interest.

Alternatively to plain DFT methods, density functional-based tight-binding (DFTB) methods can also be used to perform electronic structure calculations in chemical sciences, avoiding the requirement to approximate the exchange-correlation functional. These are very efficient approximations of DFT, as they are based on expansions up to the third order of the total Kohn-Sham energy with respect to charge density fluctuations,[1,76–78] which may or may not include self-consistent charges (SCC). Owing to its success in describing the energetics and geometries of small organic molecule,[72,79–82] SCC-DFTB was used frequently in the research studies presented in this thesis.

In what follows, we end this chapter on quantum mechanics by giving a brief overview of the basis sets that are commonly employed to describe the spin orbitals of which the wavefunction is composed.

2.5 Basis sets

As previously discussed, in most electronic structure calculations the wavefunction is decomposed into a product of independent one-electron wavefunctions that represent spin orbitals. Since the analytical expression of these orbitals is usually unknown, it is necessary to express them as a linear combination of some known auxiliary functions - referred to as a basis set -, which span the Hilbert

space where the problem is solved. In this approach, each spin orbital, $\chi_a(\boldsymbol{r}_i)$, is expanded as a linear combination of N_b basis functions, $v_\gamma(\boldsymbol{r}_i)$, such that

$$\chi_a(\boldsymbol{r}_i) = \sum_{\gamma=1}^{N_b} c_{\gamma a} v_\gamma(\boldsymbol{r}_i) \tag{2.59}$$

where $c_{\gamma a}$ are the expansion coefficients of the spin orbital $\chi_a(\boldsymbol{r}_i)$. The choice of a basis set is critical when employing *ab initio* methods, as it has an impact on both the computational cost and accuracy of the calculations.[83]

In the early days of electronic structure calculations, Slater-type orbitals (STOs) were the basis set of choice due to the correctness of their short- and long-range behaviour.[84,85] The expression of an STO in Cartesian coordinates is given by

$$v_{abc}^{STO}(\boldsymbol{r}_i) = N x^a y^b z^c \exp\left[-\zeta \sqrt{x_i^2 + y_i^2 + z_i^2}\right] \tag{2.60}$$

where N is a normalisation constant, a, b, and c control the angular momentum $L = a + b + c$, and ζ defines the width of the orbitals. STOs are, however, unpractical due to their computational inefficiency. A much more efficient alternative are the Gaussian-type orbitals (GTOs).[86] Although GTOs do not correctly describe the form of the cusp at the nucleus and decay too fast, the ratio of the number of GTOs to the number of STOs required to obtain comparable accuracy pends to the side of the GTOs in terms of computational cost,[85] making them the preferred basis set of quantum chemists. The expression of a GTO in Cartesian coordinates is given by

$$v_{abc}^{GTO}(\boldsymbol{r}_i) = N x^a y^b z^c \exp\left[-\zeta \left(x_i^2 + y_i^2 + z_i^2\right)\right] \tag{2.61}$$

where all terms are as previously defined for STOs. In practice, in quantum chemistry, contracted Gaussians (CGs), which are linear combinations of primitive GTOs, are used to represent atomic orbitals. The simplest type of CGs are the STO-nG basis sets, which attempt to approximate STOs using n primitive

GTOs. STO-nG basis sets are, however, poor representations of atomic orbitals, since only one CG is included per atomic orbital. An extension to this picture is obtained when using more than one CG to represent the atomic orbitals of the valence electrons. This approximation results in the so-called split-valence basis sets, which can be double-, triple-, quadruple-zeta, and so on, depending on how many basis functions describe the valence electrons. For example, the 6-31G double-zeta basis set describes the core electrons using a contraction of six GTOs, while each valence electron is described using two basis functions: the first is composed of a contraction of three GTOs, and the second consists of a single uncontracted GTO. More flexibility can be introduced into basis sets by making use of polarisation and/or diffuse functions. Polarisation functions add CGs with an angular momentum higher than that naturally present in the valence shell of a given atom. A typical example of a polarised basis set is the 6-31G* basis set, which augments the 6-31G basis set through addition of polarisation functions to the heavy atoms. Diffuse functions, on the other hand, lead to, *e.g.*, the 6-31G+ basis set, and they are used to better represent the atomic orbitals at regions far from the nucleus.

Finally, plane-wave basis sets are also commonly employed in electronic structure calculations, especially in those that involve periodic systems, as they are solutions to the Schrödinger equation for a particle in a periodic box. The expression of a plane wave reads

$$v_{\boldsymbol{G}}(\boldsymbol{r}_i) = N \exp\left[i\boldsymbol{G} \cdot \boldsymbol{r}_i\right] \tag{2.62}$$

where i is the imaginary unit, and $\boldsymbol{G}$ is the wave vector.

2.6 Summary

In this chapter, we have presented the basics of quantum mechanics. We started from the fundamental principles of the theory and then proceeded to discuss

the specifics of electronic structure calculations, *viz.*, the Born-Oppenheimer approximation, the Hartree-Fock method, density functional theory, and basis sets.

In the next chapter, we present the fundamental principles of statistical mechanics and contextualize their use in standard simulation methods. Specifically, we discuss the main concepts of statistical mechanics, the features of the principal thermodynamic ensembles, and the fundamentals of the Monte Carlo and molecular dynamics simulation methods. Lastly, we conclude with a discussion on thermostats, barostats, and enhanced sampling methods.

Chapter 3

Statistical Mechanics and Simulation Methods

Statistical mechanics was one of the major advances in the physics of the 19^{th} century. Since then, it has revolutionised how problems involving a large number of particles, typically on the order of 10^{23}, are solved. Systems of such dimensionality, and particularly those involving interacting particles, pose several challenges to those interested in studying them. First, it is difficult to exactly define their state; and second, there are neither the mathematical tools required to analytically solve 10^{23} coupled differential equations nor the computational power to numerically handle them. Additionally, researchers are generally more interested in the collective behaviour rather than in the individual behaviour of particles, meaning that even if it were possible to computationally solve these N-particle problems, it would still be a waste of computational resources since much of the information would be either meaningless or unintelligible, not providing much relevant physical data. Hence, statistical mechanics is a key discipline that establishes bridges between the realms of the macroscopic and microscopic worlds. In doing so, it allows the derivation of macroscopic thermodynamic properties from microscopic descriptions of systems.

3.1 Fundamental principles of statistical mechanics

In statistical mechanics, a microstate refers to a specific microscopic configuration of a thermodynamic system with an associated probability of occurrence, and an ensemble is a collection of microstates compatible with a specific macrostate of a thermodynamic system. Two fundamental principles are required to derive all the framework of statistical mechanics. The first is the principle of *a priori* equal probability, which states that microstates in the same microcanonical ensemble, i.e., microstates that have the same number of particles, N, volume, V, and energy, E, must occur with the same frequency. This principle assumes that nature cannot distinguish between two identical microstates that have the same macroscopic values for N, V, and E.

The second principle, postulated by Gibbs, is based on Boltzmann's idea of ergodicity, used to define a system that moves on a constant-energy surface in phase space and that eventually will visit all its domain. This principle postulates that whatever ensemble is set up, it must be ergodic. An immediate corollary of this postulate is the ergodic hypothesis, which states that in equilibrium the ensemble average of a property A is the same as the time average of that property for a single system. Because of this, the following relation holds

$$\langle A \rangle_{time} = \langle A \rangle_{ensemble} \tag{3.1}$$

which can alternatively be expressed as[87]

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t dt' \; A(\boldsymbol{p}(t'), \boldsymbol{q}(t')) = \int_{-\infty}^{\infty} d^{3N}\boldsymbol{p} \int_{\Omega} d^{3N}\boldsymbol{q} \, \rho(\boldsymbol{p}, \boldsymbol{q}) A(\boldsymbol{p}, \boldsymbol{q}) \tag{3.2}$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_{3N})$ and $\boldsymbol{q} = (q_1, q_2, \ldots, q_{3N})$ are the momentum and position coordinates of a three-dimensional N-particle system, $\rho(\boldsymbol{p}, \boldsymbol{q})$ is the probability density of the microstates, and Ω is the region of space defined by the containing volume.

In practice, theoretical evaluation of either side of equation (3.2) must be done discretely. For example, time averages are often calculated with methods such as MD, which propagate a system by numerically integrating its equations of motion through finite and discrete time steps. Owing to the non-continuous nature of these techniques, time averages must be calculated using the following equation

$$\langle A \rangle_{time} \approx \frac{1}{N_t} \sum_{i=1}^{N_t} A(\boldsymbol{p}(t_i), \boldsymbol{q}(t_i)) \tag{3.3}$$

where i is an index that runs over all N_t time steps. Likewise, analytical calculations of ensemble averages are often impossible to perform owing to the form of $\rho(\boldsymbol{p}, \boldsymbol{q})$. Hence, stochastic methods, such as MC, must be employed to numerically approximate the right-hand side of equation (3.2), making estimations of ensemble averages using

$$\langle A \rangle_{ensemble} \approx \sum_{i=1}^{N_p} \rho(\boldsymbol{p}_i, \boldsymbol{q}_i) A(\boldsymbol{p}_i, \boldsymbol{q}_i) \tag{3.4}$$

where i runs over a collection of N_p distinct points in phase space. Note that equations (3.3) and (3.4) become exact as $N_t \to +\infty$ and $N_p \to +\infty$, respectively.

It is worth mentioning that many systems studied through theoretical simulations do not behave ergodically on the time scales that can be achieved computationally, the reason being that they often get trapped in energy basins bounded by high-energy barriers[88] or in loops within phase space.[89] This phenomenon is known as kinetic trapping, and it is the main factor preventing a thorough exploration of phase space. Since it has a huge impact on the accuracy of the results obtained in simulations, the development of methods to solve this issue is an area of intensive research that has led to the development of many enhanced sampling schemes. An overview of the state of the art of this field is presented in Section 3.6.

3.2 Thermodynamic ensembles

In the previous introduction to statistical mechanics, we briefly referred to ensembles without actually defining what they are. The concept of the ensemble is central in statistical mechanics, and it corresponds to a hypothetical representation of a set of identical copies of a system, in which each copy is in one possible microstate.[87,90,91] An ensemble must be in a well defined thermodynamic state, characterised by a small set of macroscopic variables that are fixed, such as, *e.g.*, the number of particles, N, the temperature, T, the chemical potential, μ, the pressure, P, or the total energy, E. There are various thermodynamic ensembles, with the most common ones being the microcanonical ensemble (NVE), the canonical ensemble (NVT), the grand canonical ensemble (μVT), and the isothermal-isobaric ensemble (NPT).

Furthermore, in relation to the previous discussion on the ergodic hypothesis, it remains to address how the values of the thermodynamic properties themselves are calculated. Since these are specific for each thermodynamic ensemble, we now turn our discussion to the features of the three ensembles used in this thesis, *viz.*, the microcanonical (NVE), the canonical (NVT), and the isothermal-isobaric (NPT). In what follows, we always consider three-dimensional classical systems with continuous energy levels composed of N distinguishable particles.

3.2.1 The microcanonical ensemble

The thermodynamic variables fixed in the microcanonical ensemble are the number of particles, N, the volume, V, and the total energy, E. It is rare to find experiments carried out in this ensemble, and there are only a few theoretical applications of it.[90] Despite this, the NVE ensemble is of great conceptual importance since it considers an isolated system, *i.e.*, a situation in which no work or heat is exchanged between the system and its environment.

In an isolated system in equilibrium, the probability density of finding a system with total energy E is proportional to[87]

$$\mathcal{N}(\boldsymbol{p},\boldsymbol{q}) \propto \delta\left[H(\boldsymbol{p},\boldsymbol{q}) - E\right] \tag{3.5}$$

where $H(\boldsymbol{p},\boldsymbol{q})$ is the Hamiltonian of the system, and the δ function is used to select those microstates of a N-particle system in a container of volume V that have the desired energy E. The microcanonical partition function is directly constructed by integrating the above probability density over all phase space volume, *i.e.*,

$$Q_{NVE} = \frac{1}{h^{3N}} \int_{-\infty}^{\infty} d^{3N}\boldsymbol{p} \int_{\Omega} d^{3N}\boldsymbol{q}\, \delta\left[H(\boldsymbol{p},\boldsymbol{q}) - E\right] \tag{3.6}$$

where h is the Planck constant, introduced to make Q_{NVE} dimensionless. Hence, considering the principle of *a priori* equal probability, the probability density stated in equation (3.5) can be rewritten as[92]

$$\mathcal{N}(\boldsymbol{p},\boldsymbol{q}) = \frac{\delta\left[H(\boldsymbol{p},\boldsymbol{q}) - E\right]}{Q_{NVE}} = \frac{1}{\mathcal{W}} \tag{3.7}$$

where $\mathcal{W}$ is the number of microstates with energy E. This statistical probability is related to the entropy through the Boltzmann's entropy formula, which reads

$$S = k_B \ln\left(\mathcal{W}\right) \tag{3.8}$$

where k_B is the Boltzmann constant.

3.2.2 The canonical ensemble

The macrostate of a system in the canonical ensemble is defined by the number of particles, N, the volume, V, and the temperature, T. The canonical ensemble

is often used in both experimental and theoretical settings, and it considers a system in contact with a heat bath at constant temperature. As different copies of a system in the NVT ensemble may have different energies, its probability density function is proportional to the Boltzmann distribution* and reads

$$\mathcal{N}(\boldsymbol{p},\boldsymbol{q}) = \frac{\exp\left[-\beta H(\boldsymbol{p},\boldsymbol{q})\right]}{Q_{NVT}} \tag{3.9}$$

where $\beta = (k_B T)^{-1}$ is the thermodynamic beta, and Q_{NVT} is the canonical partition function, which reads

$$Q_{NVT} = \frac{1}{h^{3N}} \int_{-\infty}^{\infty} d^{3N}\boldsymbol{p} \int_{\Omega} d^{3N}\boldsymbol{q}\ \exp\left[-\beta H(\boldsymbol{p},\boldsymbol{q})\right] \tag{3.10}$$

There are cases in which the Hamiltonian is separable into kinetic, $K(\boldsymbol{p})$, and potential energy, $U(\boldsymbol{q})$, terms, as shown in the following equation

$$H(\boldsymbol{p},\boldsymbol{q}) = U(\boldsymbol{q}) + K(\boldsymbol{p}) \tag{3.11}$$

$$= U(\boldsymbol{q}) + \sum_{i}^{N} \frac{|\boldsymbol{p}_i|^2}{2m_i} \tag{3.12}$$

where m_i is the mass of particle i. Whenever this separation is possible, the integration over the momentum coordinates can be carried out analytically, yielding a factor of $\left(h^2/2\pi m k_B T\right)^{1/2}$ for each of the $3N$ DOFs.[94] Therefore, the canonical partition function can be rewritten simply as a configurational partition function (often referred to as the configurational integral) that reads

$$\mathcal{Z}_{NVT} = \int_{\Omega} d^{3N}\boldsymbol{q}\ \exp\left[-\beta U(\boldsymbol{q})\right] \tag{3.13}$$

*See Pathria and Beale[93], pp. 41–44, for a detailed derivation of the probability density function for the canonical ensemble.

The canonical probability density function can also be rewritten considering only the potential energy part, *i.e.*,

$$\mathcal{N}(\boldsymbol{q}) = \frac{\exp\left[-\beta U\left(\boldsymbol{q}\right)\right]}{\mathcal{Z}_{NVT}} \tag{3.14}$$

Finally, as the entropy connects the macroscopic thermodynamics with the statistical interpretation in the microcanonical ensemble, in the canonical ensemble this connection is made by the Helmholtz free energy. The Helmholtz free energy is a thermodynamic state function that measures the useful work obtainable from a closed isothermal system and its expression is given by

$$A = -k_b T \ln\left(Q_{NVT}\right) \tag{3.15}$$

3.2.3 The isothermal-isobaric ensemble

Another fundamental ensemble is the isothermal-isobaric, in which the macrostate of a system is defined by the number of particles, N, the pressure, P, and the temperature, T. The NPT ensemble is suited to simulate biological processes or chemical reactions because in solution, *in vitro*, or *in vivo* these phenomena usually occur under conditions of constant pressure. The probability density of the NPT ensemble is given by[87]

$$\mathcal{N}(\boldsymbol{p}, \boldsymbol{q}, V) = \frac{\exp\left\{-\beta\left[H(\boldsymbol{p}, \boldsymbol{q}) + PV\right]\right\}}{Q_{NPT}} \tag{3.16}$$

where Q_{NPT} is the isothermal-isobaric partition function, which reads

$$Q_{NPT} = \frac{1}{V_0 h^{3N}} \int_0^\infty dV\, V^N \int_{-\infty}^{\infty} d^{3N}\boldsymbol{p} \int_\Omega d^{3N}\boldsymbol{q}\, \exp\left\{-\beta\left[H(\boldsymbol{p}, \boldsymbol{q}) + PV\right]\right\} \tag{3.17}$$

where V_0 is some basic unit of volume chosen to render Q_{NPT} dimensionless. Analogously to what was done for the canonical ensemble, it is possible separate the configurational properties from the kinetic ones to obtain the following isothermal-isobaric configurational partition and probability density functions

$$\mathcal{Z}_{NPT} = \frac{1}{V_0} \int dV \int_{\Omega} d^{3N}\boldsymbol{q} \exp\left\{-\beta\left[U(\boldsymbol{q}) + PV\right]\right\} \tag{3.18}$$

$$\mathcal{N}(\boldsymbol{q}, V) = \frac{V^N \exp\left\{-\beta\left[U(\boldsymbol{q}) + PV\right]\right\}}{Q_{NPT}} \tag{3.19}$$

Finally, the thermodynamic property related to the NPT partition function is the Gibbs free energy, which is given by

$$G = -k_b T \ln\left(Q_{NPT}\right) \tag{3.20}$$

3.3 Monte Carlo

MC refers to a class of methods that resorts to random sampling and statistical modelling to simulate the behaviour of complex systems and numerically estimate mathematical functions.[95] Its modern origins date back to the 18$^{\text{th}}$ century, when George Louis LeClerc, Comte de Buffon (1707-1788) estimated the value of π in his famous needle experiment. It was not until the 20$^{\text{th}}$ century, however, that MC started to be widely employed as a tool to simulate physical and chemical phenomena.[96,97] In chemical sciences, of great historical importance is the introduction,[98,99] in 1953, of the MC algorithm proposed by Metropolis *et al.*,[100] used to perform the first simulation of a hard-sphere "liquid". Since then, several MC algorithms have been suggested,[101–106] and these have been applied to solve different problems.[107–111]

MC methods normally attempt to perform two operations: generate S samples, $\boldsymbol{x}^S = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_S \in \mathbb{R}^n)$, which are here assumed to be continuous real vectors,

from a probability density function; and estimate expectation values of functions under this density function.[87,90,112] Each of these operations has its own caveats, which require careful consideration. Regarding the calculation of the expected value of a thermodynamic property of interest $A = A(\boldsymbol{x})$, suppose one wants to evaluate the following integral

$$\langle A \rangle = \int_{\Omega} d^n \boldsymbol{x}\, \rho(\boldsymbol{x}) A(\boldsymbol{x}) \tag{3.21}$$

where $\rho(\boldsymbol{x})$ is a probability density function that weights the values of $A(\boldsymbol{x})$, and Ω defines the integration domain. As previously discussed, equations of this type are often impossible to solve analytically owing to the mathematical form of the $\rho(\boldsymbol{x})$ term. It is possible, nonetheless, to employ methods like MC to numerically estimate such integrals. Hence, if samples can be drawn directly from $\rho(\boldsymbol{x})$, the MC estimator of $\langle A \rangle$ is given by the sample mean, *i.e.*,

$$\langle A \rangle \approx \hat{A} = \frac{1}{S} \sum_{i=1}^{S} A(\boldsymbol{x}_i) \tag{3.22}$$

However, in most cases, sampling directly from $\rho(\boldsymbol{x})$ is unfeasible because either its normalisation constant is unknown, or it is challenging to draw n-dimensional samples from $\rho(\boldsymbol{x})$. The solution for this problem is to introduce an auxiliary probability density function, $w(\boldsymbol{x})$, from which sampling is performed. To ensure that $\hat{A}$ still converges to $\langle A \rangle$ as $S \to \infty$, $w(\boldsymbol{x})$ should contain $\rho(\boldsymbol{x})$ and be non-zero everywhere where $\rho(\boldsymbol{x})$ is non-zero. When using this auxiliary function, instead of attempting to evaluate equation (3.21), the following integral is considered

$$\langle A \rangle = \int_{\Omega} d^n \boldsymbol{x}\, w(\boldsymbol{x}) \left[\frac{\rho(\boldsymbol{x}) A(\boldsymbol{x})}{w(\boldsymbol{x})} \right] \tag{3.23}$$

Therefore, similarly to what was done to estimate equation (3.21), the integral of equation (3.23) can be calculated using the following MC estimator

$$\langle A \rangle \approx \hat{A} = \frac{1}{S} \sum_{i=1}^{S} \frac{\rho(\boldsymbol{x}_i) A(\boldsymbol{x}_i)}{w(\boldsymbol{x}_i)} \tag{3.24}$$

Now, say the $w(\boldsymbol{x})$ function is chosen to be a hyperrectangle from which uniform sampling is performed. Since in this case $w(\boldsymbol{x})$ can be determined analytically, the MC estimator reads

$$\langle A \rangle \approx \hat{A} = \frac{V}{S} \sum_{i=1}^{S} A(\boldsymbol{x}_i) \rho(\boldsymbol{x}_i) \tag{3.25}$$

where $V = \int_\Omega d^n \boldsymbol{x}$ is the volume of the sample space defined by the n-orthotope or n-cube. In practice, uniform sampling is inefficient because it is prone to exploring sample space regions that have low importance for the integral of equation (3.23). For example, thinking in terms of configurational space, the regions that are the most relevant to explore are those where $\exp[-\beta H(\boldsymbol{q})]$ is the largest, and if uniform sampling is performed this subtlety is not taken into account. A possible solution to increase the efficiency of MC integration is to perform importance sampling, a technique in which the chosen $w(\boldsymbol{x}_i)$ is biased towards regions of sample space that have high importance. To understand the value of this method, consider a situation in which one is concerned with the estimation of $\langle A \rangle$ in the canonical ensemble, and that the $w(\boldsymbol{x})$ function is chosen to be the normalised canonical Boltzmann distribution. On these assumptions, the following relation holds

$$w(\boldsymbol{x}) = \rho(\boldsymbol{x}) = \frac{\exp[-\beta H(\boldsymbol{x})]}{Z_{NVT}} \tag{3.26}$$

and, therefore, equation (3.24) reduces to

$$\langle A \rangle \approx \hat{A} = \frac{1}{S} \sum_{i=1}^{S} A(\boldsymbol{x}_i) \tag{3.27}$$

This is one of the main ideas behind the Metropolis algorithm,[100] in which expectation values are calculated by simply averaging the $A(\boldsymbol{x}_i)$ values. The brilliance of the Metropolis method is that it allows to sample from $w(\boldsymbol{x}) = \rho(\boldsymbol{x})$ without requiring the value of the partition function to be known (see discussion in Section 3.3.1). Furthermore, besides importance sampling, other MC sampling techniques are available, and we refer the reader to the comprehensive literature of Lemieux[113], Landau and Binder[114], and Binder and Heermann[115] for further details.

To conclude the discussion on the general aspects of the MC method, it is worthwhile mentioning that another important feature of MC integration is that the variance $Var\left[\hat{A}\right]$ scales as $Var\left[\hat{A}\right] \sim 1/S$, thus decreasing as the number of samples S increases. This can be verified by using the definition of equation (3.21) to express $Var\left[\hat{A}\right]$ as follows

$$Var\left[\hat{A}\right] = Var\left[\frac{1}{S}\sum_{i=1}^{S} A(\boldsymbol{x}_i)\rho(\boldsymbol{x}_i)\right] = \frac{1}{S^2} Var\left[\sum_{i=1}^{S} A(\boldsymbol{x}_i)\rho(\boldsymbol{x}_i)\right] \tag{3.28}$$

Since MC draws are uncorrelated samples (random and independent), the variance of the sum is the sum of the variances, and therefore equation (3.28) can be rewritten as

$$Var\left[\hat{A}\right] = \frac{1}{S^2}\sum_{i=1}^{S} Var\left[A(\boldsymbol{x}_i)\rho(\boldsymbol{x}_i)\right] = \frac{1}{S} Var\left[A(\boldsymbol{x}_i)\rho(\boldsymbol{x}_i)\right] \tag{3.29}$$

leading to the conclusion that the accuracy of the MC estimator is independent of the dimensionality of the sample space, only depending on the number of samples, S. A similar derivation can be done to show that importance sampling, when properly performed, reduces the estimation variance in MC simulations.[90]

3.3.1 Acceptance criteria in Monte Carlo

The derivation of the acceptance criteria in standard MC simulations requires satisfying the detailed balanced principle, also known as the condition of microscopic reversibility, so that reversible Markov chains are constructed. In what follows, for the sake of simplicity, it is assumed that the state of the system is fully defined by the position of its particles, although note that this may not always be the case, since, for example, in the isothermal-isobaric ensemble the state of the system also depends on the volume. On this assumption, detailed balanced implies that in equilibrium the average number of accepted moves from a configuration $\boldsymbol{q}_i = (q_{i,1}, q_{i,2}, ..., q_{i,3N})$ to any other configuration $\boldsymbol{q}_f = (q_{f,1}, q_{f,2}, ..., q_{f,3N})$, where N is the number of particles, is exactly cancelled by the number of reverse moves, *i.e.*,

$$\mathcal{N}(\boldsymbol{q}_i)\pi(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \mathcal{N}(\boldsymbol{q}_f)\pi(\boldsymbol{q}_f \to \boldsymbol{q}_i) \tag{3.30}$$

where $\mathcal{N}(\boldsymbol{q}_i)$ is the probability of finding the system in configuration $\boldsymbol{q}_i$, and $\pi(\boldsymbol{q}_i \to \boldsymbol{q}_f)$ is the transition probability of going from configuration $\boldsymbol{q}_i$ to configuration $\boldsymbol{q}_f$. The latter probability is constructed by noting that a MC move consists of two stages. The first stage corresponds to performing a trial move from $\boldsymbol{q}_i$ to $\boldsymbol{q}_f$, with an associated probability of occurrence given by $\alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f)$. Moreover, the second stage corresponds to accepting or rejecting this trial move, with an acceptance probability denoted by $\theta(\boldsymbol{q}_i \to \boldsymbol{q}_f)$. Since the two stages are independent of each other, $\pi(\boldsymbol{q}_i \to \boldsymbol{q}_f)$ can be written as

$$\begin{cases} \pi(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f) \times \theta(\boldsymbol{q}_i \to \boldsymbol{q}_f) & \text{if } i \neq j \\ \pi(\boldsymbol{q}_i \to \boldsymbol{q}_i) = 1 - \sum_{j \neq i} \pi(\boldsymbol{q}_i \to \boldsymbol{q}_f) & \text{if } i = j \end{cases} \tag{3.31}$$

The detailed balance principle can be satisfied by carefully defining the form of the acceptance probability $\theta(\boldsymbol{q}_i \to \boldsymbol{q}_f)$. Noticing that in equilibrium $\theta(\boldsymbol{q}_i \to \boldsymbol{q}_f)$ is constant if microscopic reversibility is ensured, then the following relation holds

$$\theta(\boldsymbol{q}_i \to \boldsymbol{q}_f) + \theta(\boldsymbol{q}_f \to \boldsymbol{q}_i) = s(\boldsymbol{q}_i \to \boldsymbol{q}_f) \tag{3.32}$$

where $s(\boldsymbol{q}_i \to \boldsymbol{q}_f)$ is a symmetric function, i.e., $s(\boldsymbol{q}_i \to \boldsymbol{q}_f) = s(\boldsymbol{q}_f \to \boldsymbol{q}_i)$, chosen such that $0 \leq \alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f) \leq 1$. Hence, combining the relation of equation (3.32) with the results of equations (3.30) and (3.31) leads to the following expression

$$\mathcal{N}(\boldsymbol{q}_i)\alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f)\theta(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \mathcal{N}(\boldsymbol{q}_f)\alpha(\boldsymbol{q}_f \to \boldsymbol{q}_i)\theta(\boldsymbol{q}_f \to \boldsymbol{q}_i) \tag{3.33}$$

which holds for any i and j and can be rewritten as

$$\theta(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \frac{s(\boldsymbol{q}_i \to \boldsymbol{q}_f)}{1 + \frac{\mathcal{N}(\boldsymbol{q}_i)\alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f)}{\mathcal{N}(\boldsymbol{q}_f)\alpha(\boldsymbol{q}_f \to \boldsymbol{q}_i)}} \tag{3.34}$$

This is the general form of the acceptance probability $\theta(\boldsymbol{q}_i \to \boldsymbol{q}_f)$, as proposed by Hastings[116]. Various forms have been suggested for the function $s(\boldsymbol{q}_i \to \boldsymbol{q}_f)$. For example, Barker[117] suggested to take $s(\boldsymbol{q}_i \to \boldsymbol{q}_f) = 1$, which is simply a normalisation of the sum of equation (3.32). Furthermore, Metropolis *et al.*[100] suggested the following form for $s(\boldsymbol{q}_i \to \boldsymbol{q}_f)$

$$s(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \begin{cases} 1 + \frac{\mathcal{N}(\boldsymbol{q}_i)}{\mathcal{N}(\boldsymbol{q}_f)} \frac{\alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f)}{\alpha(\boldsymbol{q}_f \to \boldsymbol{q}_i)} & \text{if } \frac{\mathcal{N}(\boldsymbol{q}_i)}{\mathcal{N}(\boldsymbol{q}_f)} \frac{\alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f)}{\alpha(\boldsymbol{q}_f \to \boldsymbol{q}_i)} \geq 1 \\ 1 + \frac{\mathcal{N}(\boldsymbol{q}_f)}{\mathcal{N}(\boldsymbol{q}_i)} \frac{\alpha(\boldsymbol{q}_f \to \boldsymbol{q}_i)}{\alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f)} & \text{if } \frac{\mathcal{N}(\boldsymbol{q}_f)}{\mathcal{N}(\boldsymbol{q}_i)} \frac{\alpha(\boldsymbol{q}_f \to \boldsymbol{q}_i)}{\alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f)} < 1 \end{cases} \tag{3.35}$$

which reduces equation (3.34) to

$$\theta(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \min\left[1, \frac{\mathcal{N}(\boldsymbol{q}_f)}{\mathcal{N}(\boldsymbol{q}_i)} \frac{\alpha(\boldsymbol{q}_f \to \boldsymbol{q}_i)}{\alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f)}\right] \tag{3.36}$$

which can be recognised as the widely used Metropolis criterion. Nothing has yet been said about the form of $\mathcal{N}(\boldsymbol{q})$. Since this quantity has to be proportional to the probability density function of the considered ensemble, it is given by one

of the expressions that was previously stated, *viz.*, equation (3.7) for the micro-canonical ensemble, equation (3.14) for the canonical ensemble, and equation (3.19) for the isothermal-isobaric ensemble.

3.3.2 Monte Carlo algorithm structure

Although there are several types of MC algorithms, in general they all share common features that are now described. The general workflow of a typical MC algorithm is shown diagrammatically in Figure 3.1, and it can be summarised in the following steps:

1. Initialize the starting configuration $\boldsymbol{q}_i$.
2. Perform N_{MC} MC sweeps[†]:
   (a) Generate a new configuration $\boldsymbol{q}_f$ by making a perturbation to configuration $\boldsymbol{q}_i$.
   (b) For a given acceptance criterion, accept $\boldsymbol{q}_f$ if $\theta(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f)$ is greater or equal than a random number X sampled from a uniform distribution defined within the interval $[0, 1]$, and reject $\boldsymbol{q}_f$ otherwise.
   (c) If the configuration $\boldsymbol{q}_f$ were rejected, start the next iteration from configuration $\boldsymbol{q}_i$; otherwise start the next iteration from configuration $\boldsymbol{q}_f$, *i.e.*, set $\boldsymbol{q}_f = \boldsymbol{q}_i$.
3. Stop the simulation.

A large variety of choices are available for the perturbation done in step 2a.[106] Some common MC moves include simulation box scaling, thermodynamic perturbation, particle displacement, particle insertion or deletion, molecule rotation, and other types of stochastic or deterministic perturbations of the DOFs of the system.

[†]An MC sweep is the natural unit of simulation "time" of MC simulations and consists in a sequence of random moves that are accepted or rejected according to an acceptance criterion.

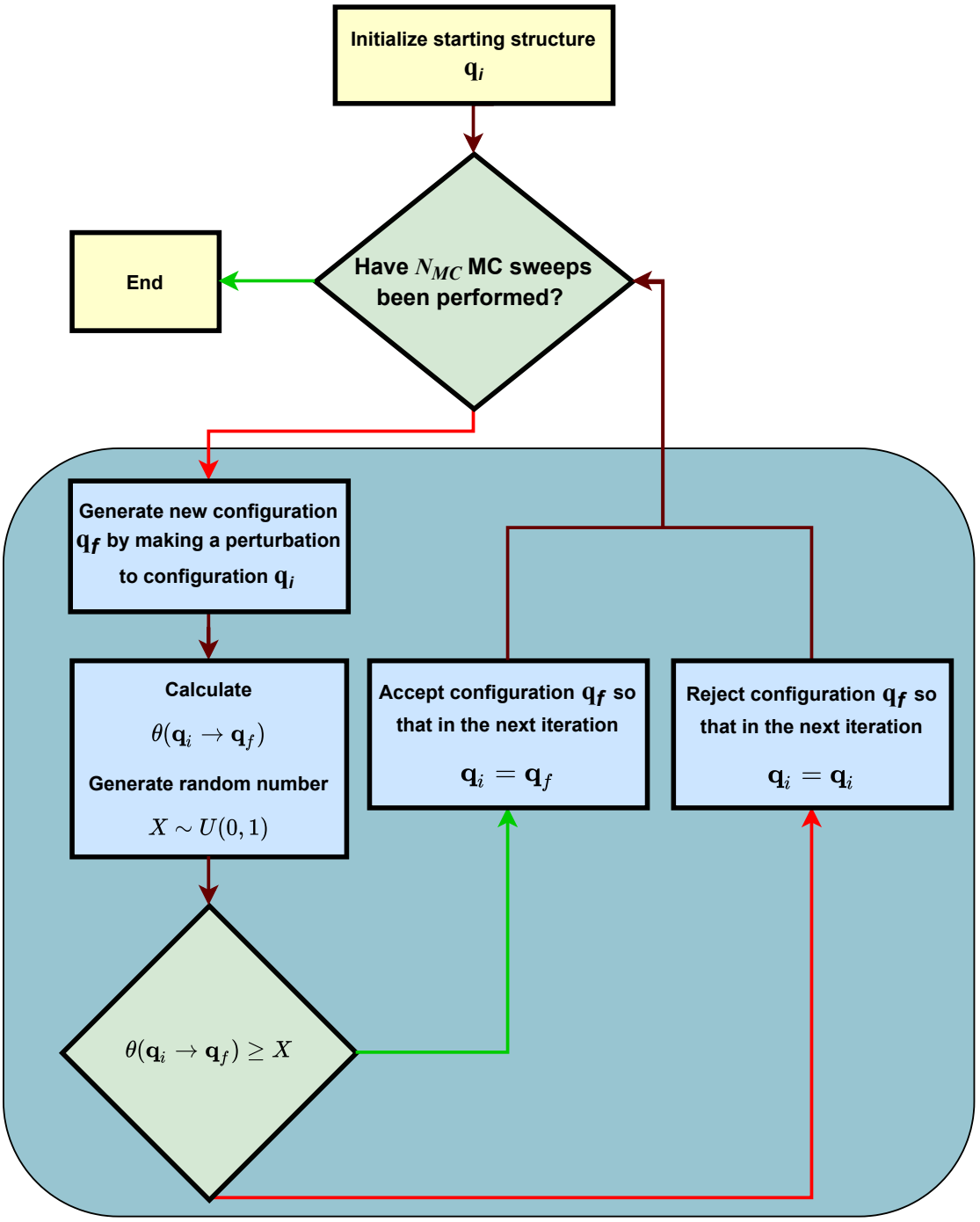

Figure 3.1: Diagram describing the general workflow of the MC algorithm. $U(0,1)$ denotes a random number between 0 and 1 sampled from a uniform distribution. The green arrows denote conditionals for which the evaluated condition is true, whereas the red arrows denote conditionals for which the evaluated condition is false.

3.4 Molecular dynamics

Molecular dynamics is the standard method for simulating the dynamical behaviour of systems. Contrary to MC algorithms, which are stochastic in nature, plain MD methods are deterministic and allow the time-dependent behaviour of systems to be studied by numerical integration of their equations of motion. The first MD simulations were performed in 1957 by Alder and Wainwright[118], using systems composed of hard spheres. Advances in MD algorithms and computational power permitted the first MD simulations of a realistic system, *viz.*, liquid water, to be performed in 1974 by Stillinger and Rahman[119], and a few years later of a protein, *viz.*, the bovine pancreatic trypsin inhibitor, which was run by McCammon *et al.*[120] in 1977. Since then, the use of MD simulations has become widespread much owing to the continuous development of molecular mechanics and, more recently, to the advent of GPUs that can perform GPU-accelerated MD. These advances have allowed simulations of larger systems for longer time scales.[121] MD simulations are nowadays routinely employed to study a wide range of problems in chemical sciences, with druglike molecules, proteins, and nucleic acids being some of the common systems of interest.

The underlying principles of MD were firstly derived by Sir Isaac Newton in 1687. In Newton's formalism of classical mechanics, a set of Cartesian coordinates, $\boldsymbol{r} = (r_1, r_2, r_3, ..., r_{3N})$, is usually employed, and the time evolution of one of these degrees of freedom is calculated using the following equation

$$F_i = m\frac{d\dot{r}_i(t)}{dt} = m\frac{d^2 r_i}{dt^2} \tag{3.37}$$

where dot notation is used to denote time derivatives, F_i represents the force acting upon the degree of freedom i at time t, and m the mass of the particle associated with that degree of freedom.

An alternative formalism of classical mechanism was proposed by Joseph-Louis Lagrange. In the Lagrangian picture, a set of generalised coordinates $\boldsymbol{q} = \{q_i\}$

that depends on the system of interest is used, and the equations of motions are rewritten for the appropriate configuration manifold such that any constraints are considered from the outset. The central quantity in this formalism is the Lagrangian, $L(\boldsymbol{r}, \dot{\boldsymbol{r}}, t)$, which is defined as the difference between the kinetic, $K(\dot{\boldsymbol{r}})$, and potential energies, $U(\boldsymbol{r})$. The expression of the Lagrangian reads

$$L(\boldsymbol{r}, \dot{\boldsymbol{r}}, t) = K(\dot{\boldsymbol{r}}) - U(\boldsymbol{r}) = \sum_{i}^{3N} \frac{1}{2} m \dot{r}_i^2 - U(\boldsymbol{r}) \tag{3.38}$$

where, for the sake of simplicity, it was assumed that $\boldsymbol{q} \equiv \boldsymbol{r}$, *i.e.*, that the generalised coordinates correspond to Cartesian coordinates. The equation of motion of the Lagrangian formalism reads

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{r}_i}\right) = \frac{\partial L}{\partial r_i} \tag{3.39}$$

By inserting the result of equation (3.38) into equation (3.39), the following relation is obtained

$$m \frac{d\dot{r}_i}{dt} = -\frac{\partial U(\boldsymbol{r})}{\partial r_i} \tag{3.40}$$

from which, by comparison with equation (3.37), the link between the Newtonian and Lagrangian formalisms is unraveled

$$F_i = -\frac{\partial U(\boldsymbol{r})}{\partial r_i} \tag{3.41}$$

demonstrating that the force can be expressed as the gradient of the potential energy. Equation (3.41) provides a fundamental result for MD, as it can be used to calculate forces from an arbitrary potential energy function. Moreover, this result also implies an interesting feature of MD: since each F_i depends on the potential energy, $U(\boldsymbol{r})$, which in turns depends on the position coordinates of all particles, the $3N$ differential equations that must be solved to propagate the

equations of motion are coupled whenever an interacting system is considered, and, therefore, generally, it is necessary to resort to numerical integration to obtain the dynamical behaviour of such an interacting system.

Hamiltonian mechanics can also be used to perform MD, consisting in a reformulation of the Lagrangian formalism that can be extended to QM. It describes classical systems using a set of canonical coordinates and conjugate momenta, $\{q_i, p_i\}$, defined in phase space. The fundamental quantity of this formalism is the Hamiltonian, $H(\boldsymbol{q}, \boldsymbol{p})$, which represents the total energy of the system calculated as the sum of the kinetic, $K(\dot{\boldsymbol{p}})$, and potential energies, $U(\boldsymbol{q})$. The expression of the Hamiltonian reads

$$H(\boldsymbol{q}, \boldsymbol{p}) = K(\dot{\boldsymbol{p}}) + U(\boldsymbol{q}) \tag{3.42}$$

The Hamiltonian equations of motion are given by

$$\frac{\partial H}{\partial p_i} = \dot{q}_i \tag{3.43}$$

$$\frac{\partial H}{\partial q_i} = -\dot{p}_i \tag{3.44}$$

Finally, it is worthwhile mentioning that initial velocities for MD simulations can be drawn from the Maxwell-Boltzmann distribution, which reads

$$p(v_i) = \left(\frac{m_i}{2\pi k_b T}\right)^{1/2} \exp\left[-\frac{m_i v_i^2}{2k_B T}\right] \tag{3.45}$$

where v_i is the velocity of the degree of freedom i, and the other variables are as previously defined.

Having discussed the most important formalisms that allow the dynamical behaviour of systems to be determined, it remains to address how the equations of motion can be numerically solved. Several numerical algorithms have

been developed for this purpose, the most popular ones being the leap-frog algorithm,[122] the Verlet algorithm,[123] the Beeman's algorithm,[124,125] and the velocity-Verlet algorithm.[126] The basic assumption in all these methods is that the positions, velocities and accelerations can be approximated by a Taylor series expansion. Since the velocity-Verlet algorithm is the one used in this work, it is discussed in detail in the next section.

3.4.1 The velocity-Verlet algorithm

The velocity-Verlet algorithm[126] is an improved version of the original Verlet algorithm.[123] The main advantage of the velocity-Verlet algorithm with respect to the leap-frog algorithm is that the former calculates the positions and velocities at the same instant of time. The basic equations of the velocity-Verlet algorithm are the following

$$q_i(t+dt) = q_i(t) + \dot{q}_i(t)dt + \frac{F_i(t)dt^2}{2m_i} + \mathcal{O}(dt^3) \tag{3.46}$$

$$\dot{q}_i(t+dt) = \dot{q}_i(t) + \frac{dt}{2m_i}\left[F_i(t+dt) + F_i(t)\right] + \mathcal{O}(dt^3) \tag{3.47}$$

where dt is the integration time step. From these equations, it can be seen that the local error in positions and velocities is $\mathcal{O}(dt^3)$. However, since a simulation of length t needs to perform t/dt time steps, the global error of this algorithm is $\mathcal{O}(dt^2)$, making velocity-Verlet a second-order method. Importantly, this integration algorithm preserves two fundamental properties of the classical equations of motion, *viz.*, time reversibility, and symplecticity[‡]. Sympleticity is essential to ensure the integrator conserves the energy of dynamical systems during propagation. Furthermore, in most implementations, equations (3.46) and (3.47) are further splitted into

[‡]Sympleticity implies phase space "volume" preservation over time, *i.e.*, that the density of microstates is held constant within the entire phase space over the trajectory time.

$$\dot{q}_i(t + \frac{dt}{2}) = \dot{q}_i(t) - \frac{F_i(t)}{2m_i}dt \tag{3.48}$$

$$q_i(t + dt) = q_i(t) + \dot{q}_i(t + \frac{dt}{2})dt \tag{3.49}$$

$$\dot{q}_i(t + dt) = \dot{q}_i(t + \frac{dt}{2}) - \frac{F_i(t + dt)}{2m_i}dt \tag{3.50}$$

The workflow of this algorithm can be summarised as follows:

1. For every degree of freedom i:
   (a) Given q_i at time t, calculate F_i using equation (3.41).
   (b) Using equation (3.48), calculate the half-step velocity using $\dot{q}_i$ and F_i at time t.
   (c) Update q_i for time $t + \delta t$ using equation (3.49).
   (d) Given q_i at time $t + \delta t$, calculate the new force F_i using equation (3.41).
   (e) Using F_i calculated in the previous step (1d), update the half-step $\dot{q}_i$ to to the full step $\dot{q}_i$ using equation (3.50);
   (f) Go back to step 1.

3.5 Thermostats and barostats

Thermostats and barostats are required to control the temperature and pressure of a system during an MD simulation. The use of thermostats and barostats allows the simulation of ensembles that match experimental conditions, making them essential components of MD simulations. The simplest thermostats available are the velocity-rescaling[127] and Berendsen.[128] Despite their computational convenience, the velocity-rescaling and Berendsen thermostats do not

sample any defined ensemble and are prone to producing the flying ice cube effect.[129] These drawbacks prevent the use of the velocity-rescaling and Berendsen thermostats in MD production runs unless more advanced velocity-rescaling schemes are employed.[130] The Nosé-Hoover,[131,132] Nosé-Hoover chain,[89] and Anderson[133] thermostats, on the other hand, correctly sample the NVT ensemble. The Nosé-Hoover thermostat, however, is not ergodic for small or stiff systems,[89] making the Nosé-Hoover chain and Anderson thermostats the most viable methods to effectively sample the canonical distribution. Lastly, the temperature of a system can also be controlled by performing Langevin dynamics.[134] Since Langevin dynamics was the temperature-controlling method mostly used throughout this thesis, it is described in detail in subsection 3.5.1.

MD simulations require barostats to simulate the NPT ensemble. The principles underlying pressure coupling methods are very similar to those of temperature coupling methods. For example, the algorithm underlying the Berendsen barostat is the same as that of the thermostat with the same name,[128] though, instead of scaling the velocities, the barostat works by scaling the cell vectors and system coordinates at each MD step. Similarly, the Monte Carlo barostat also controls the pressure through scaling of the cell vectors and system's coordinates, although this scaling is done stochastically and scaling attempts are accepted or rejected according to an acceptance criterion. As the Monte Carlo barostat[135–137] was the pressure-controlling scheme used throughout this thesis, it is described in detail in subsection 3.5.2. Other popular pressure-controlling methods are the Parrinello-Rahman barostat,[138–140] which, besides volume scaling, allows the simulation box to change its shape, and the Anderson[133,141,142] barostat, which works by mimicking the action of a piston that can compress or decompress a system to which it is coupled.

3.5.1 Langevin dynamics

The Langevin equation is a stochastic differential equation that can be used to sample an ensemble at a fixed temperature T.[87,134] The Langevin equation is

the fundamental equation behind Langevin dynamics and can be constructed by adding a dissipative force and a noise term to the Hamiltonian equations of motion. These added terms mimic the function of a thermostat by allowing energy to flow into or out of the system.[143] The equations that govern Langevin dynamics are given by

$$\dot{p}_i = -F_i\left[q_i(t)\right] - \gamma_i p_i(t) + \sqrt{2m_i\gamma_i k_B T}\frac{dW(t)}{dt} \tag{3.51}$$

$$\dot{q}_i = \frac{p_i(t)}{m_i} \tag{3.52}$$

where F_i is the deterministic force acting on the degree of freedom i, γ_i is the friction coefficient associated with that degree of freedom, and $dW(t)$ is a Wiener noise that satisfies the conditions $\langle dW(t)dW(t')\rangle = \delta(t-t')$ and $\langle dW(t)\rangle = 0$. The noise term can be interpreted as a random force, or fluctuation, which brings energy into the system. This energy is dissipated through the dissipative force, $-\gamma_i p(t)$, which can be interpreted as a friction force arising due to the solvent.

Care has to be taken when choosing the friction coefficient, γ_i, as too small values may lead to ineffective dissipation of energy, causing the system to heat. This is of particular concern for non-equilibrium situations since small values of γ_i may lead to the breakdown of the system.[144] On the other hand, in the high friction limit, which occurs when $\gamma_i \to \infty$, Langevin dynamics becomes Brownian dynamics because the deterministic force is negligible in comparison to the other terms.[145] Several algorithms have been proposed for the numerical integration of the Langevin equations, and we refer the reader to the comprehensive review made by Leimkuhler and Matthews[146], Chapter 7, for further details.

3.5.2 The Monte Carlo barostat

The MC barostat can be used to simulate the effects of constant pressure by stochastically adjusting the size of a periodic simulation box. Although there are

advanced versions of this algorithm,[137] its basic idea is to perform a simulation box rescaling at a chosen frequency and after a regular MD time step. This perturbation consists in proposing a trial volume change, $\Delta V = \xi \Delta V_{max}$, generated using a random number, ξ, sampled from a uniform distribution defined within the interval $[-1, 1]$. The limit ΔV_{max} is chosen so that the MC acceptance ratio is typically about 40-50%. The box lengths and centre of mass coordinates of each molecule are then scaled according to[135,136]

$$l_i' = l_i \left(\frac{V'}{V} \right)^{1/3} \tag{3.53}$$

$$r_i' = (r_i - c_i) \left(\frac{V'}{V} \right)^{1/3} + c_i \tag{3.54}$$

where l_i is the size of the box along dimension i, r_i is coordinate of the degree of freedom i, $V' = V + \Delta V$, and c_i denotes the coordinate of the centre of the periodic box along dimension i. To derive the acceptance criterion for this MC move, consider the form of the Metropolis criterion stated in equation (3.36). If $\alpha(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) = \alpha(\boldsymbol{q}_f \rightarrow \boldsymbol{q}_i)$ is chosen by imposing symmetry on the probability of occurrence of the move, equation (3.36) reduces to

$$\theta(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) = \min \left[1, \frac{\mathcal{N}(\boldsymbol{q}_f, V')}{\mathcal{N}(\boldsymbol{q}_i, V)} \right] \tag{3.55}$$

Furthermore, since constant pressure simulations must be performed in the NPT ensemble, $\mathcal{N}$ is given by the corresponding probability density function already stated in equation (3.19). Therefore, $\mathcal{N}(\boldsymbol{q}_i, V)$ and $\mathcal{N}(\boldsymbol{q}_f, V')$ are given by

$$\mathcal{N}(\boldsymbol{q}_f, V') = \frac{V'^N \exp\left\{-\beta \left[U(\boldsymbol{q}_f) + PV'\right]\right\}}{Q_{NPT}} \tag{3.56}$$

$$\mathcal{N}(\boldsymbol{q}_i, V) = \frac{V^N \exp\left\{-\beta \left[U(\boldsymbol{q}_i) + PV\right]\right\}}{Q_{NPT}} \tag{3.57}$$

Hence, by inserting equations (3.56) and (3.57) into equation (3.55), the following acceptance criterion is obtained

$$\theta(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) = \min\left[1, \exp\left\{-\beta\left[\Delta U + P\Delta V\right] - N\ln\left(\frac{V'}{V}\right)\right\}\right] \tag{3.58}$$

where $\Delta U = U(\boldsymbol{q}_f) - U(\boldsymbol{q}_i)$. The volume move is then accepted if $\theta(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) \geq X$, where X is a random number sampled from a uniform distribution defined within the interval $[0, 1]$, and rejected otherwise. Note that an isotropic MC barostat was considered here, but this does not have to necessarily be the case since the above procedure can easily be adapted to take anisotropy into account.[147]

3.6 Enhanced sampling methods

Enhanced sampling methods are nowadays routinely employed to solve the sampling problem in MD simulations. With the currently available MD algorithms and computational power, it is possible to simulate mesoscale systems containing millions of atoms on the nanosecond time scale. In this regard, in 2019 Jung *et al.*[148] reported the first atomistic MD simulation of an entire gene, a system composed of 1 billion atoms, which was simulated during approximately 1 ns. Another remarkable example is the atomistic 121 ns MD simulation of the H1N1 viral envelope (*ca.* 160 million atoms) performed by Durrant *et al.*[149] While these noteworthy cases are representative of the limits imposed on the simulation of mesoscale systems by present-day software and hardware architectures, smaller systems composed of hundreds of thousands of atoms can now be simulated on time scales that can go up to the microsecond.[150] Despite this, MD simulations of this size and length still demand computational resources that are not available to all research groups, making enhanced sampling methods the only feasible way to simulate certain phenomena of interest in chemical sciences.

The fastest molecular motions are vibrations and rotations, which occur on time scales ranging from femtosecond to picosecond, and picosecond to nanosecond, respectively.[151] These motions are easily captured by atomistic MD simulations, in which the standard time step used is 1 fs. Complex systems, however, are characterised by a multiple time scale nature, in which the dynamics of fast movements takes place in parallel to slow motions that occur on a microsecond to millisecond time scale. Events of such kind are, for example, enzyme catalysis, protein-ligand binding, protein folding, signal transduction, and allosteric regulation.[88,152] These biomolecular phenomena are poorly captured by MD simulations, being considered as rare events on the currently accessible simulation times. Furthermore, the occurrence of some conformational changes even in small-to-medium-sized organic molecules often depends upon the emergence of unlikely fluctuations, since the free energy landscapes of such systems are characterised by high barriers separating long-lived metastable states.[153] Most MD simulations are, consequently, not truly ergodic because they cannot explore every available point in phase space. Enhanced sampling methods are, therefore, the way forward towards efficient and thorough sampling in MD simulations.

Enhanced sampling methods can be broadly separated into two families: those that depend on collective variables (CVs), and those that are independent of CVs.[88] CVs, also referred to as order parameters or reaction coordinates, are defined as functions, generally non-linear, of the atomic coordinates, $\boldsymbol{q}$, such that a set of CVs $\boldsymbol{s}(\boldsymbol{q})$ is defined as $\boldsymbol{s}(\boldsymbol{q}) = (\boldsymbol{s}_1(\boldsymbol{q}), \boldsymbol{s}_2(\boldsymbol{q}), \ldots, \boldsymbol{s}_d(\boldsymbol{q}))$. This CV set should be able to describe the key features of the physical behaviour of interest, distinguish between all relevant metastable states, and include all the slow DOFs.[154] The equilibrium distribution, $P(\boldsymbol{s})$, of the CVs, $\boldsymbol{s}(\boldsymbol{q})$, is thus given by

$$P(\boldsymbol{s}) = \int d^{3N}\boldsymbol{q}\, \delta\left[\boldsymbol{s} - \boldsymbol{s}(\boldsymbol{q})\right] \mathcal{N}(\boldsymbol{q}) = \langle \delta\left[\boldsymbol{s} - \boldsymbol{s}(\boldsymbol{q})\right] \rangle \tag{3.59}$$

where $\mathcal{N}(\boldsymbol{q})$ is, for example, the canonical probability density function defined in equation (3.14). The free energy surface is defined as the logarithm of this

distribution, and therefore it reads

$$F(\boldsymbol{s}) = -\frac{1}{\beta} \ln \left[P(\boldsymbol{s}) \right] \tag{3.60}$$

Of the family of methods that depend on CVs, the most popular ones are umbrella sampling,[34] and metadynamics (MetaD),[29] though there are many more available such as J-walking,[155] adaptive biasing force method,[156] or conformational space annealing,[157,158] to name just a few. Umbrella sampling was the first method of this kind to be proposed. It works by introducing a bias potential $U_{bias}\left[\boldsymbol{s}(\boldsymbol{q})\right]$ designed to enhance the sampling of CV space, such that the equilibrium distribution of the CVs, taking into account the biasing potential, reads

$$P_{bias}(\boldsymbol{s}) = \int_{\Omega} d^{3N}\boldsymbol{q}\, \delta\left[\boldsymbol{s} - \boldsymbol{s}(\boldsymbol{q})\right] \mathcal{N}(\boldsymbol{q}) \mathcal{N}_{bias}(\boldsymbol{q}) \tag{3.61}$$

where $\mathcal{N}_{bias}(\boldsymbol{q})$ is the probability density function of the biasing potential, which is given by

$$\mathcal{N}_{bias}(\boldsymbol{q}) = \frac{\exp\left\{-\beta U_{bias}\left[\boldsymbol{s}(\boldsymbol{q})\right]\right\}}{Z_{bias}} = \frac{\exp\left\{-\beta U_{bias}\left[\boldsymbol{s}(\boldsymbol{q})\right]\right\}}{\int_{\Omega} d^{3N}\boldsymbol{q}\, \exp\left\{-\beta U_{bias}\left[\boldsymbol{s}(\boldsymbol{q})\right]\right\}} \tag{3.62}$$

Therefore, the unbiased free energy surface can be obtained using

$$F(\boldsymbol{s}) = -\frac{1}{\beta} \ln \left[P_{bias}(\boldsymbol{s}) \right] = -\frac{1}{\beta} \ln \left[\frac{P(\boldsymbol{s})}{Z_{bias}} \right] - U_{bias}\left(\boldsymbol{s}\right) \tag{3.63}$$

where $P(\boldsymbol{s})/Z_{bias}$ is the distribution sampled in the biased simulation. Hence, from equation (3.63) it follows that the unbiased free energy surface in umbrella sampling can be recovered by reweighting the distribution from the biased simulation.

MetaD works by constructing an history-dependent bias potential $U_{bias}(\boldsymbol{s}, t)$ that gradually accumulates multivariate Gaussian repulsive potentials $\mathcal{G}\left[\boldsymbol{s}, \boldsymbol{s}(t')\right]$ at the visited CV, *i.e.*,[88]

$$U_{bias}(\boldsymbol{s}, t) = \int_0^t dt' \, \mathcal{G}\left[\boldsymbol{s}, \boldsymbol{s}(t')\right] \tag{3.64}$$

$$\mathcal{G}\left[\boldsymbol{s}, \boldsymbol{s}(t')\right] = \omega \exp\left\{-\frac{1}{2}\left[\boldsymbol{s} - \boldsymbol{s}(t')\right]^T \boldsymbol{\Sigma}^{-1}\left[\boldsymbol{s} - \boldsymbol{s}(t')\right]\right\} \tag{3.65}$$

where ω is the height of the Gaussian, and $\boldsymbol{\Sigma}^{-1}$ is a symmetric covariance matrix. By doing this, metaD disfavours states that were already visited during the simulation and guides the system towards new regions in CV space. The goal of metaD simulations is to obtain a sampled distribution that becomes uniform as $t \to \infty$, therefore allowing the free energy to be recovered, as it is proportional to the negative of the bias potential. However, since repulsive potentials never stop being added to this bias potential, metaD simulations never converge, presenting a systematic error. The solution for this problem is called well-tempered metaD, a method in which the height of the Gaussians becomes time-dependent, decreasing as the biasing potential increases, so that

$$\omega(t) = \omega \exp\left[-\frac{\beta U_{bias}(\boldsymbol{s}, t)}{\gamma - 1}\right] \tag{3.66}$$

where $\gamma > 1$ is a constant called bias factor. Hence, since as $t \to \infty$, $\omega(t) \to 0$, the bias potential converges as

$$U_{bias}(\boldsymbol{s}) \propto -\left(1 - \frac{1}{\gamma}\right) F(\boldsymbol{s}) \tag{3.67}$$

allowing convergence of the free energy surface.

On the other hand, of the family of methods that are independent of CVs, the most popular ones are parallel tempering (or replica exchange)[32,159,160] and its

derivatives.[161,162] In the most basic version of this method, R parallel replicas of a system are considered, and these replicas are concomitantly simulated at different temperatures using MD. The lowest temperature replica represents the target ensemble for which sampling is to be enhanced, while the high-temperature replicas are used to accelerate the exploration of conformational space. Enhanced sampling is then achieved by performing MC exchange attempts between the configurations of adjacent replicas every N^t MD steps. A diagram describing the general workflow of these replica-exchange-based methods is shown in Figure 3.2.

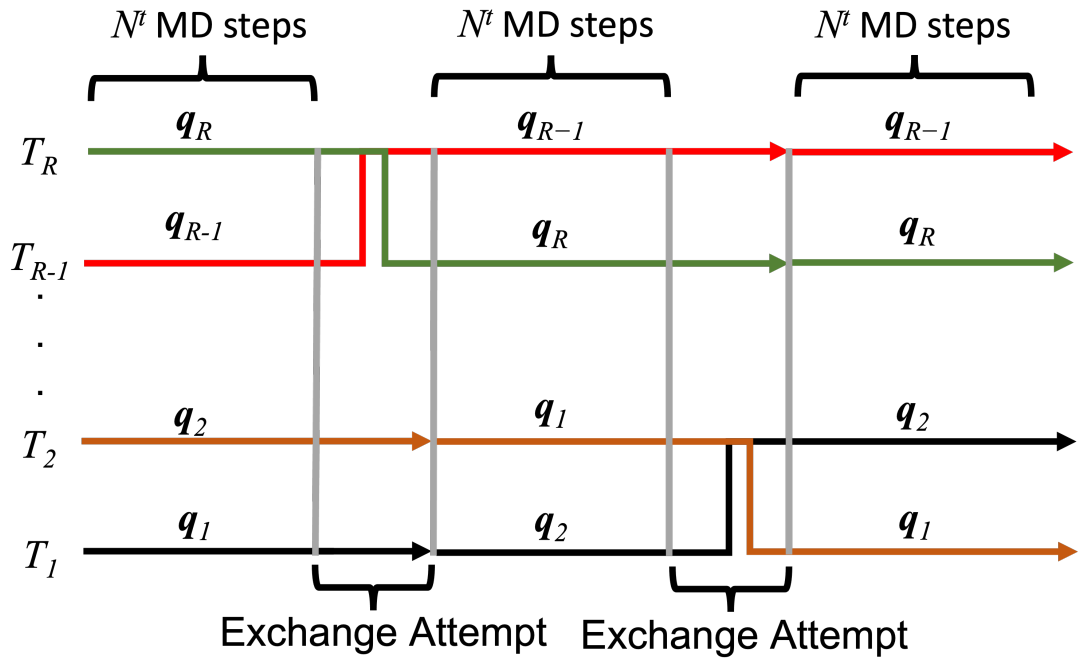

Figure 3.2: Diagram describing the general workflow of replica-exchange-based methods. A supersystem composed of R replicas is represented, for which exchange attempts between adjacent replicas are attempted every N_t MD steps. These exchange attempts consist in configuration exchanges, which are accepted or rejected according to the acceptance criterion given by equation (3.71). The temperature of the replicas increases monotonically from T_1 to T_R, allowing for enhanced sampling to be achieved.

To derive the MC acceptance criterion for the exchange of configurations between two replicas, consider the procedure presented in Section 3.3.1. The first step involves imposing detailed balance, which forces the number of exchanges from replica i to f to be exactly cancelled by the number of reverse moves, *i.e.*,

$$\mathcal{N}_i(\boldsymbol{q}_i)\mathcal{N}_f(\boldsymbol{q}_f)\pi[i(\boldsymbol{q}_i) \leftrightarrow f(\boldsymbol{q}_f)] = \mathcal{N}_i(\boldsymbol{q}_f)\mathcal{N}_f(\boldsymbol{q}_i)\pi[i(\boldsymbol{q}_f) \leftrightarrow f(\boldsymbol{q}_i)] \tag{3.68}$$

where $\mathcal{N}_i(\boldsymbol{q}_i)$ is the probability of finding the system with configuration $\boldsymbol{q}_i$ on the ith replica, and $\pi[i(\boldsymbol{q}_i) \leftrightarrow f(\boldsymbol{q}_f)]$ is the transition probability for the exchange of configurations between replicas i and f. As in standard MC moves, $\pi[i(\boldsymbol{q}_i) \leftrightarrow f(\boldsymbol{q}_f)]$ consists of two stages: the first stage corresponds to performing a trial configuration exchange $i(\boldsymbol{q}_i) \leftrightarrow f(\boldsymbol{q}_f)$, which has as an associated probability of occurrence given by $\alpha[i(\boldsymbol{q}_i) \leftrightarrow f(\boldsymbol{q}_f)]$. Furthermore, the second stage corresponds to the acceptance (or rejection) of this exchange attempt, being its acceptance probability denoted by $\theta[i(\boldsymbol{q}_i) \leftrightarrow f(\boldsymbol{q}_f)]$. If α is symmetric, i.e., $\alpha[i(\boldsymbol{q}_i) \leftrightarrow f(\boldsymbol{q}_f)] = \alpha[i(\boldsymbol{q}_f) \leftrightarrow f(\boldsymbol{q}_i)]$, equation (3.68) can be written as

$$\mathcal{N}_i(\boldsymbol{q}_i)\mathcal{N}_f(\boldsymbol{q}_f)\theta[i(\boldsymbol{q}_i) \leftrightarrow f(\boldsymbol{q}_f)] = \mathcal{N}_i(\boldsymbol{q}_f)\mathcal{N}_f(\boldsymbol{q}_i)\theta[i(\boldsymbol{q}_f) \leftrightarrow f(\boldsymbol{q}_i)] \tag{3.69}$$

which can be rearranged to give

$$\frac{\theta[i(\boldsymbol{q}_i) \leftrightarrow f(\boldsymbol{q}_f)]}{\theta[i(\boldsymbol{q}_f) \leftrightarrow f(\boldsymbol{q}_i)]} = \frac{\mathcal{N}_i(\boldsymbol{q}_f)\mathcal{N}_f(\boldsymbol{q}_i)}{\mathcal{N}_i(\boldsymbol{q}_i)\mathcal{N}_f(\boldsymbol{q}_f)} \tag{3.70}$$

Using the Metropolis version of the acceptance criterion, equation (3.70) reduces to

$$\theta[i(\boldsymbol{q}_i) \leftrightarrow f(\boldsymbol{q}_f)] = \min\left[1, \frac{\mathcal{N}_i(\boldsymbol{q}_f)\mathcal{N}_f(\boldsymbol{q}_i)}{\mathcal{N}_i(\boldsymbol{q}_i)\mathcal{N}_f(\boldsymbol{q}_f)}\right] \tag{3.71}$$

where $\mathcal{N}(\boldsymbol{q})$ is given by equation (3.7) for the microcanonical ensemble, equation (3.14) for the canonical ensemble, and equation (3.19) for the isothermal-isobaric ensemble. The exchange move is then accepted if $\theta[i(\boldsymbol{q}_i) \leftrightarrow f(\boldsymbol{q}_f)] > X$, where X is a random number sampled from a uniform distribution defined within the interval [0, 1], and rejected otherwise.

We refer readers to the reviews of Yang *et al.*[88] and Bernardi *et al.*[163] for more information about enhanced sampling methods.

3.7 Summary

In this chapter, we have presented the basic theory behind statistical mechanics and simulation methods. We began with the fundamental principles of statistical mechanics and proceeded to discuss the features of the main thermodynamic ensembles, the fundamentals of the Monte Carlo and molecular dynamics simulation methods, and the theory underlying thermostats, barostats, and enhanced sampling methods.

The next chapter is devoted to molecular mechanics, the main classical method used to model molecular systems in this thesis. We introduce the concept of potential energy surface, discuss the main classes and functional forms of force fields, review the methods for calculating long-range interactions, and present the details of the general AMBER force field. Lastly, we conclude with a review of the current state of the art of force field parameterisation strategies.

Chapter 4

Molecular Mechanics

MM refers to the use of classical descriptions to model molecular systems. Unlike QM methods, which explicitly consider electronic motions, in MM methods the energy of a system is a function of nuclear positions only, since electrons are either ignored or only partially incorporated at a mean-field level in polarisable models. This approximation greatly simplifies the calculations that must be performed to determine the behaviour of systems of interest, therefore reducing the computational cost required for their simulation. MM models are often referred to as FFs. FFs have been continuously developed since the 1960s, providing nowadays the principal model with which complex chemical systems are described. Although MM methods cannot provide any information about electronic properties, they still have, in some cases, advantages in terms of accuracy. A clear example in which MM models have been historically superior to QM methods is in the description of dispersion interactions, which current *ab initio* methods can only accurately reproduce at the MP2 or higher levels of theory if no empirical MM-like corrections are employed.[164] Despite their indubitable success, MM models are based on empirical energy functions and require a set of FF parameters to work, making their development an ongoing process since the diversity of systems of interest never stops posing new challenges.

4.1 Potential energy surfaces

The potential energy surface (PES) describes the energy of a system as a function of some (or all) of its geometric parameters. The concept of the PES is not exclusive to MM, as it is used in any method (*e.g.*, QM methods or machine-learning potentials) that can calculate the energy of a system in terms of structural parameters. PESs can be determined owing to the Born-Oppenheimer approximation, which, by decoupling electronic and nuclear motion, makes the concept of geometry meaningful.[165]

The stationary points are the key features of a PES. Stationary points are defined as points where the gradient of the energy with respect to all structural parameters is zero. Three types of stationary points are relevant in chemistry: minima, transition states, and higher-order saddle points. Minima correspond to stable (global minima) or metastable (local minima) molecular structures. Transition states are first-order saddle points in the energy map connecting two minima. Higher-order saddle points are less important than transition states, although some reactions or transitions may occur through them, especially through second order saddle points. A representation of a PES is shown in Figure 4.1.

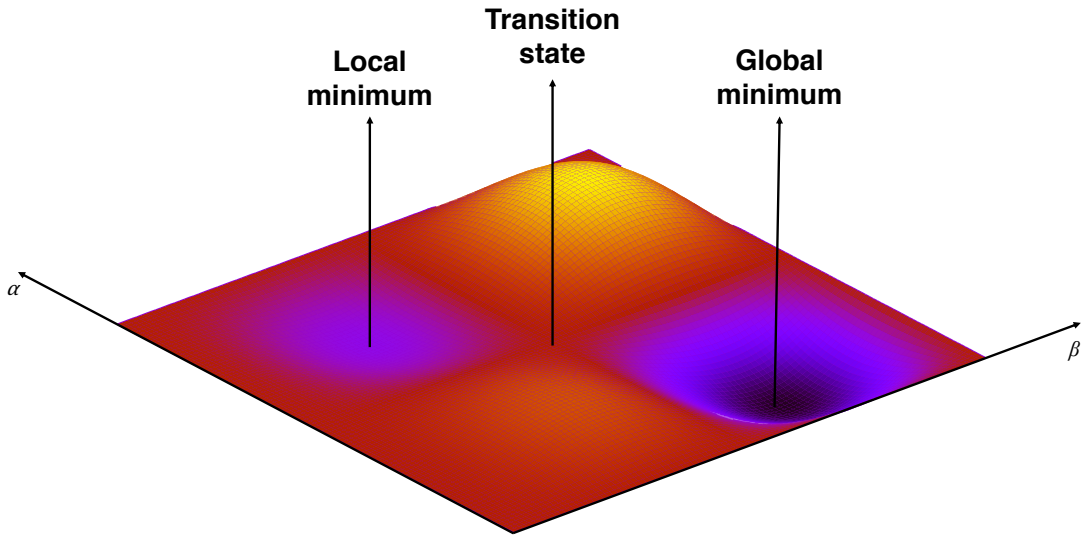

Figure 4.1: Simplified PES as a function of structural parameters α and β.

To characterise the stationary points of a PES, recall that for a linear molecule with N atoms there are $p = 3N - 5$ vibrational DOFs. Furthermore, for a non-linear molecule, $p = 3N - 6$ independent variables are necessary to define its internal coordinates, $\boldsymbol{q} = (q_1, q_2, \ldots, q_p)$. Taking this into account, the Hessian matrix of such a system reads

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 U}{\partial q_1^2} & \cdots & \frac{\partial^2 U}{\partial q_1 \partial q_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 U}{\partial q_p \partial q_1} & \cdots & \frac{\partial^2 U}{\partial q_p^2} \end{bmatrix} \tag{4.1}$$

The nature of a stationary point can be determined by calculating the eigenvalues of the Hessian at that point. If the eigenvalues are all positive, the stationary point is a minimum. If the eigenvalues are all negative, the stationary point is a maximum. Otherwise, the stationary point is a saddle point, of which first-order saddle points, which are a minimum in all DOFs except one, are particularly important in chemistry since they correspond to transition states.[166]

4.2 Force fields

FFs are the state-of-the-art models employed in the simulation of systems in chemical sciences. The core of any FF is the potential energy function, or functional form, used to map the molecular representation, $\boldsymbol{q}$, of a system of interest to its potential energy, U.[164] Besides the FF functional form, a set of FF parameters is usually required, which can be obtained either from the various databases available or bespoke-derived. The detail of the molecular representation defines the granularity of the FF. In this regard, molecular modelling can be done using either all-atom,[95,167] united-atom,[168,169] or coarse-grained FFs.[170–172] The atomistic representation explicitly includes every atom, being, therefore, the most accurate at the MM level. The united-atom model does not explicitly represent nonpolar hydrogens, capturing their steric effect by modifying the Lennard-Jones (LJ) parameters of the parent atom.[173] Coarse-grained FFs further reduce

the DOFs of a system by representing groups of atoms as beads, allowing cheaper simulations at the cost of lower accuracy. There are also hybrid MM models that combine, for example, the all-atom and coarse-grained representations,[174,175] in which the atomistic details are applied to the most relevant parts of the system, and coarse-graining is used for the less important ones, permitting a balance to be achieved between accuracy and computational cost. Since this work focus on all-atom FFs, in what follows we present a discussion about the main variations of this type of model, which can be broadly divided into four classes:[164,173,176,177]

- Rule-based: FFs in this category resort to a minimal set of parameters that covers all elements in the periodic table. The functional forms of rule-based FFs typically contain terms found in class I and class II FFs (*e.g.*, bond stretching, dihedral, and bond-angle terms), although rule-based FFs do not require sets of interaction-specific parameters as required for class I and class II FFs. Instead, the parameters governing all interactions in a system are generated using generic rules. Rule-based FFs are usually universal in applicability and can virtually describe any system at the cost of lower accuracy. Examples of such FFs are UFF,[178] DREIDING,[179] and ESFF.[180]

- Class I: FFs belonging to this class are usually called "harmonic" since they use harmonic potentials to model both bond stretching and angle bending. The LJ 12-6 potential is used to describe the dispersion interactions, and the Coulomb potential is used for electrostatics. Class I FFs are also often called "diagonal" FFs, as no cross-terms coupling different functional form terms (also known as off-diagonal terms) are included. Note, however, that class I FFs may contain harmonic couplings, such as the Urey-Bradley potential. Class I FFs are commonly employed in biomolecular simulations, as these FFs have been specifically derived to model organic molecules, peptides, proteins, and nucleic acids. The most popular class I FFs are AMBER,[181,182] CHARMM,[183] GROMOS,[184] and OPLS.[185–188]

- Class II: This class of FFs employs functional forms that include cross-terms that capture couplings between DOFs. They also include cubic or quartic terms to model bond stretching and angle bending, and the dispersion interactions are treated using exponential-type potentials.[177] Owing to this, class II FFs are more computationally expensive than class I FFs, but generally more accurate at predicting energies, geometries, and vibrational frequencies.[176,177] Class II FFs are commonly applied in material science and biomolecular simulations, with MM3,[189–191] CFF,[192,193] COMPASS,[194,195] and MMFF94[196–200] being examples of popular FFs that belong to this class.

- Class III: FFs in this class represent the most sophisticated models in MM. Besides anharmonic and cross-terms, class III FFs also include description of non-additive electrostatic effects, such as hyperconjugation, and polarisation. Polarisation is included through either induced dipoles (AMOEBA[201,202]), the drude model (Polarisable CHARMM[203,204]), or fluctuating charges (fluc-q[205–207]).

As can be seen from this summary, each FF class has its advantages and drawbacks, and their choice depends on the type of system to be used and the degree of accuracy to be achieved. In what follows, we explicitly define the potential functions typically used in FFs, with emphasis on those employed by class I FFs, which were frequently used throughout this thesis.

4.2.1 Bonded potentials

The simplest FF functional forms comprise three types of bonded interactions: bond stretching term, angle bending terms, and torsional terms. These are pictorially represented in Figure 4.2.

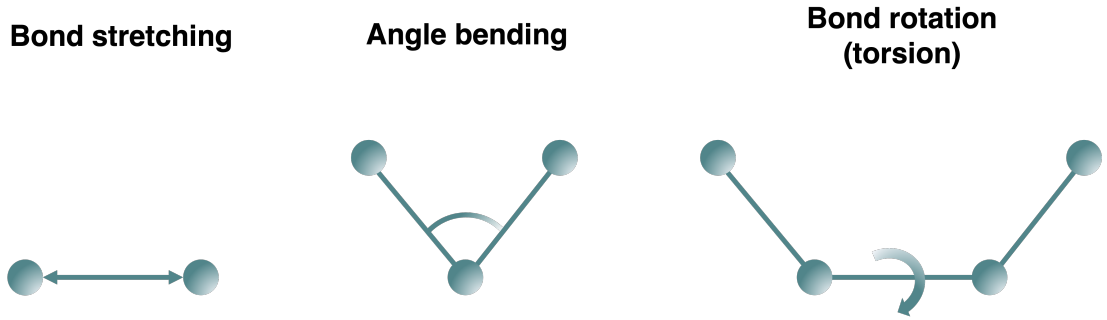

Figure 4.2: Pictorial representation of the three main contributions of the bonded terms of a MM FF, *viz.*, bond stretching, angle bending, and bond rotation (torsion).

Bond stretching terms define how the energy of a bond changes with its length. The energy of a bond can be written as a Taylor expansion about the point $r = r_{eq}$, with r_{eq} being the reference bond length, such that

$$U_{bond}(r) = U_{bond}\left(r_{eq}\right) + \left.\frac{dU_{bond}(r)}{dr}\right|_{r=r_{eq}} \left(r - r_{eq}\right) + \frac{1}{2}\left.\frac{d^2U_{bond}(r)}{dr^2}\right|_{r=r_{eq}} \left(r - r_{eq}\right)^2 + \mathcal{O}\left(r^3\right) \tag{4.2}$$

where we have neglected all terms of order higher than two. In equation (4.2), since the energy reference is arbitrary, $U_{bond}(r_{eq})$ can be set to zero. Furthermore, since the force at equilibrium is zero, the second term of the expansion is also zero. Hence, equation (4.2) can be rewritten as

$$U_{bond}(r) = \frac{1}{2}\left.\frac{d^2U_{bond}(r)}{dr^2}\right|_{r=r_{eq}} \left(r - r_{eq}\right)^2 + \mathcal{O}\left(r^3\right) \tag{4.3}$$

where the second derivative gives the curvature of the potential about the equilibrium bond length and is usually called the harmonic bond force constant K_b. Therefore, in the harmonic approximation, the potential function that models bond stretching is given by

$$U_{bond}(r) = K_b \left(r - r_{eq}\right)^2 \tag{4.4}$$

An analogous derivation can be done to show that the harmonic angle bending term reads

$$U_{angle}(\theta) = K_\theta \left(\theta - \theta_{eq}\right)^2 \tag{4.5}$$

where θ_{eq} is the reference valence angle, and K_θ is the harmonic angle force constant. Note, however, that the harmonic approximation in insufficient for situations in which the bonds and angles deviate far from their equilibrium values. Most vibrational motions are also anharmonic, requiring terms beyond the harmonic approximation for an accurate description. An accurate representation of bond stretching is given by the Morse potential,[208] which reads

$$U_{bond}(r) = D_e \left\{1 - \exp\left[-\alpha \left(r - r_{eq}\right)\right]\right\}^2 \tag{4.6}$$

where D_e is the dissociation energy or well depth, and $\alpha = \sqrt{K_b/2D_e}$. Owing to its computational cost and requirement of three parameters (D_e, K_B, and r_{eq}), the Morse potential is rarely used in molecular simulations. Its shape is instead approximated by class II and class III FFs, which truncate equation (4.2) at orders higher than two. For example, a typical quartic energy function for bond stretching is given by[173]

$$U_{bond}(r) = K_b \left(r - r_{eq}\right)^2 + K_b^{'} \left(r - r_{eq}\right)^3 + K_b^{''} \left(r - r_{eq}\right)^4 \tag{4.7}$$

where $K_b^{'}$ and $K_b^{''}$ correspond to the second and third derivatives of the energy with respect to the bond length, evaluated at the reference bond length. Despite their higher accuracy in reproducing PESs and vibrational spectra, these anharmonic functions also introduce more parameters in the FF, making parameterisation more difficult.

Bond rotation terms, also known as torsional potentials, are commonly expressed as a sum of cosine functions with multiplicities $n = 1, 2, 3, \ldots$, and amplitudes V_n, such that

$$U_{torsion}(\phi) = \sum_n V_n[1 + cos(n\phi - \gamma_n)] \tag{4.8}$$

where γ_n are the phase factors that determine where the torsional potential passes through its minimum. There are also improper dihedrals terms, which either prevent molecules from flipping to their mirror images or keep planar groups planar. These are applied to arrangements of four atoms as represented in Figure 4.3.

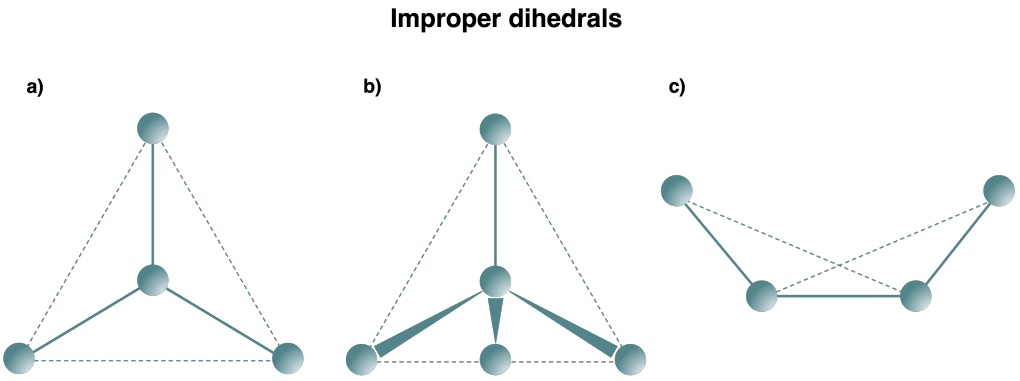

Figure 4.3: Pictorial representation of improper dihedrals used to impose planar (a) or tetrahedral (b) geometries, or to prevent out of plane bending for rings (c).

Improper dihedrals are routinely modelled using the periodic potential stated in equation (4.8). Alternatively, they can be modelled using a simple harmonic potential that reads

$$U_{improper}(\phi) = K_\phi \left(\phi - \phi_{eq}\right)^2 \tag{4.9}$$

where ϕ is the angle between the two planes formed by the triads of atoms 1-2-3 and 2-3-4 (see Figure 4.3), ϕ_{eq} is the improper dihedral equilibrium value, and K_ϕ is the improper dihedral amplitude. Finally, cross-terms that couple bonds,

angles, and torsions are also included in class II and class II FFs. For example, a typical bond-bond cross-term is given by

$$U_{bond}(r_1, r_2) = K_{b_1} K_{b_2} \left(r_1 - r_{eq_1}\right) \left(r_2 - r_{eq_2}\right) \tag{4.10}$$

Other types of cross-terms, such as those describing bond-angle, angle-angle, bond-torsion, and angle-torsion couplings can be constructed in a similar way.[173] In class I FFs, a common bond-angle cross-term is the Urey-Bradley potential, which is mainly used by the CHARMM FF.[183] The Urey-Bradley potential accounts for corrections to angle bending that arise due to bond stretching. This term is important for the proper description of in-plane deformations and for the separation of symmetric and asymmetric bond stretching vibrations.[209] The Urey-Bradley potential has a simple harmonic form that reads

$$U_{UB}(S) = K_{UB} \left(S - S_{eq}\right)^2 \tag{4.11}$$

where K_{UB} and S_{eq} are the Urey-Bradley force constant and equilibrium distance, respectively, of a virtual bond formed between atoms 1 and 3. Another important cross-term with particular relevance for the modelling of proteins is the CMAP,[164,210] which is a torsion-torsion term that couples the Φ and Ψ protein backbone torsion angles. It is used in the CHARMM FF[183] and corresponds to a Ramachandran-like plot that represents the differences between the MM and QM energies at every point in the grid.[173]

4.2.2 Nonbonded potentials

The electrostatic interactions in class I FFs are handled by the Coulomb potential. This treatment implies a point charge model in which each atom is assigned a fixed partial charge. The Coulomb potential reads

$$U_{Coulomb}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{4.12}$$

where q_i and q_j are the partial charges assigned to atoms i and j, respectively, r_{ij} denotes the distance between these atoms, and ϵ_0 is the vacuum permittivity. This "additive" treatment of electrostatics assumes that the charges do not affect each other, and therefore, that all electrostatic interactions can be summed to yield the total electrostatic energy.[173] It is well known, however, that especially in the condensed phase, molecules have higher dipoles than in the gas phase.[164] This phenomenon, which is not captured unless polarisation is taken into account, is the main driving force behind the development of class III FFs. In class I and class II FFs, in which partial charges are usually derived for the gas phase, a possible workaround for this issue consists in calculating partial charges in the dielectric medium of interest. By doing so, it is possible to mitigate some of the undesired consequences that arise due to the use of gas-phase partial charges in condensed-phase situations. Another common workaround to mimic condensed-phase settings is to overpolarise the molecules using low QM levels of theory.[211,212] A "real" treatment of electrostatic interactions, however, must necessarily go beyond the point charge model, even when polarisation is taken into account. This is the approach used by, for example, the AMOEBA FF,[201,202] in which multipole-multipole interactions are considered instead of the usual charge-charge interactions. Using this framework, AMOEBA represents the charge distribution of each atom by the permanent atomic monopole (charge), dipole, and quadrupole moments, of which only the dipole moment is polarisable. This combination of permanent atomic multipoles and atomic induced dipoles leads to AMOEBA's great performance when considering complex systems for which polarisation and accurate electrostatics are critical.[201]

Van der Waals interactions in class I FFs are handled using the LJ 12-6 potential, which reads

$$U_{LJ}(r_{ij}) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{4.13}$$

where ϵ_{ij} is the well depth of the LJ interaction between atoms i and j, and σ_{ij} the distance at which said interaction vanishes. The ϵ and σ values can be calculated using the Lorentz-Berthelot combination rules,[213,214] which are given by

$$\sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2} \tag{4.14}$$

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \tag{4.15}$$

In practice, a cutoff is used to truncate the LJ 12-6 interactions at a given distance. However, as a sharp truncation at the cutoff distance leads to discontinuities in the potential, which may cause the energy not to be conserved, a switching function is usually employed to make the interaction go smoothly to zero over a finite distance range. For example, OpenMM[215] uses the following switching function

$$S = 1 - 6x^5 + 15x^4 - 10x^3 \tag{4.16}$$

where $x = (r - r_s)/(r_c - r_s)$, and r_s refers to the distance from which equation (4.16) is used to multiply the LJ 12-6 energy. For a situation in which both a cutoff and a switching function are employed, the LJ 12-6 equation can be written as

$$U_{LJ}(r_{ij}) = \begin{cases} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], & r_{ij} \leq r_s \\ 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] S, & r_c > r_{ij} > r_s \\ 0, & r_{ij} \geq r_c \end{cases} \tag{4.17}$$

As an alternative to switching functions, shift functions, which shift the LJ 12-6 potential in the entire range by an amount equal to $U_{LJ}(r_c)$, can also be applied. Shifting the LJ 12-6 potential by a constant is a simple way to ensure that there is no discontinuity in the energy at $r_{ij} = r_c$. This simple shifting scheme does not ensure that the force goes to zero at the cutoff distance, however. To solve this problem, several shifted-force LJ 12-6 potentials have been proposed.[216–219] Shifted-force LJ 12-6 potentials add a small linear term that makes the LJ 12-6 derivative equal to zero at r_c, consequently removing the previously mentioned problems in both energies and forces.[87]

The use of cutoffs requires energy corrections that take into account missing dispersion interactions. These terms approximate the energy contribution of all interactions with $r_{ij} \geq r_c$. For an isotropic fluid with homogeneous LJ sites, the LJ 12-6 energy correction is given by[220]

$$U_{LJ}^{corr} = \frac{8\pi N^2}{V}\left(\frac{\langle \epsilon_{ij}\sigma_{ij}^{12}\rangle}{9r_c^9} - \frac{\langle \epsilon_{ij}\sigma_{ij}^{6}\rangle}{3r_c^3}\right) \tag{4.18}$$

where V is the volume of the simulation box, N is the number of particles of the system, and $\langle \ldots \rangle$ denotes an average over all pairs of particles in the system. Similar terms can also be applied to correct the pressure.[220] For inhomogeneous systems outside the cutoff distance, Ewald-family methods can be used to capture the long-range dispersion interactions. Some of the Ewald-family methods used for the treatment of electrostatic interactions are presented in the next section.

Lastly, most FFs do not include Coulomb and LJ 12-6 interactions between atoms separated by one or two bonds and use modified parameters for atoms separated by three bonds, which give rise to the so-called 1-4 interactions. The 1-4 interactions can be interpreted as scaled nonbonded interactions, since the LJ and electrostatic interactions are damped by a scale constant that prevents the molecule from breaking or deforming due to the occurrence of geometries that lead to high repulsions.

4.3 Long-range interactions

Realistic simulations of biological systems necessarily require the inclusion of long-range interactions. These are normally calculated while employing periodic boundary conditions (PBCs), which are used to suppress surface and finite-size effects that arise due to reducing systems that are infinite or very large to smaller, computationally-tractable simulation cells. When using PBCs, the original unit cell is infinitely replicated in all directions to form a periodic lattice, such that if a particle leaves the original unit cell during the simulation, then a copy of that particle enters the cell from the opposite side. To calculate nonbonded interactions for systems with PBCs, the simplest scheme available is the minimum-image convention. In the minimum image convention, each particle only interacts with the closest image formed by the remaining particles. For computational efficiency, modifications of the minimum-image convention scheme have been developed, in which it is combined with truncation at spherical cutoffs. Despite being conceptually simple and computationally convenient, schemes revolving around the (truncated) minimum-image convention pose several problems for simulations, as they introduce non-negligible errors and artificial behaviour.[221–223] To realistically simulate periodic systems, it is, therefore, necessary to resort to methods that properly capture long-range interactions. In what follows, we focus our discussion on the most important methods used to treat long-range electrostatic interactions, *viz.*, the Ewald and particle mesh Ewald (PME) methods.

In an infinitely periodic system composed of N particles in a cubic box of size L, the total Coulomb energy is given by[222]

$$U_{Coulomb} = \frac{1}{2} \sum_{\boldsymbol{n}}{}' \sum_{i,j} \frac{q_i q_j}{r_{ij,\boldsymbol{n}}} \tag{4.19}$$

where $\boldsymbol{n} = (n_1, n_2, n_3)$ runs over all copies of the periodic cell, the prime indicates that $i = j$ is omitted for $\boldsymbol{n} = 0$, the indices i and j run over all particles, q_i and q_j

are the partial charges, and $r_{ij,\boldsymbol{n}}$ is the distance between a particle in the original cell and a particle at an image cell $\boldsymbol{n}$. For the sake of simplicity, all $4\pi\epsilon_0$ factors are omitted in equation (4.19), which corresponds to reducing charges by a factor of $(4\pi\epsilon_0)^{1/2}$.[87] Calculation of long-range Coulomb interactions using this equation is computationally demanding, as well as slowly and conditionally convergent (depends on the order in which the sum is made). Methods that efficiently calculate the long-range interactions of systems with PBCs must therefore be employed for proper treatment of the electrostatics. The Ewald summation is one of the techniques of choice to converge long-range contributions. It works by recasting equation (4.19) as a sum of two rapidly converging series plus a constant term, *i.e.*,[222]

$$U_{Ewald} = E_{direct} + E_{reciprocal} + E_{self} \tag{4.20}$$

where E_{direct} is the direct (real) space sum, $E_{reciprocal}$ is the reciprocal (imaginary) space sum, and E_{self} is the self-energy term. Each of these terms can be defined as follows

$$E_{direct} = \frac{1}{2}{\sum_{\boldsymbol{n}}}' \sum_{i,j} q_i q_j \frac{\operatorname{erfc}\left(\alpha r_{ij,\boldsymbol{n}}\right)}{r_{ij,\boldsymbol{n}}} \tag{4.21}$$

$$E_{reciprocal} = \frac{1}{2\pi V} \sum_{i,j} q_i q_j \sum_{\boldsymbol{m}\neq 0} \frac{\exp\left[-(\pi \boldsymbol{m}/\alpha)^2 + i2\pi \boldsymbol{m}\cdot(\boldsymbol{r}_i - \boldsymbol{r}_j)\right]}{\boldsymbol{m}^2} \tag{4.22}$$

$$E_{self} = -\frac{\alpha}{\sqrt{\pi}} \sum_i q_i^2 \tag{4.23}$$

where V is the volume of the unit cell, $\boldsymbol{m}$ is a reciprocal-space vector, and α is an internal parameter that is defined next. In the Ewald method, each point charge of a neutral charge system is surrounded by a charge distribution, typically Gaussian-shaped, of equal magnitude and opposite sign, which reads

$$\rho_i(\boldsymbol{r}) = q_i\alpha^3 \exp\left(-\alpha^2|\boldsymbol{r}|^2\right) / \sqrt{\pi^3} \tag{4.24}$$

where α determines the width of the Gaussian distribution, and $\boldsymbol{r}$ is the position relative to the centre of the distribution. This charge distribution screens the interactions between neighbour point charges, making them short-range in nature, allowing, therefore, for rapid convergence of the direct space sum. Since the erfc function tends to zero as $r_{ij,\boldsymbol{n}}$ increases, if α in equation (4.21) is chosen to be large enough, the direct space sum reduces to the minimum image convention.[87] To counteract the screening charge distribution and recover the original charge distribution, a second Gaussian charge distribution is added to each point charge to cancel the screening charge distribution in the real space. This cancelling distribution is summed in reciprocal space and then transformed back to real space through a Fourier transform, which decays rapidly since equation (4.22) is a smooth function.[222]

The PME method is an alternative way of converging the long-range interactions. Similarly to the Ewald method, PME also recasts the potential energy into direct and reciprocal sums and uses Gaussian charge distributions.[221,224] The main distinctive feature of PME is that the reciprocal sum is approximated using fast Fourier transforms (FFTs) with convolutions on a grid in which charges are interpolated to the grid points.[222] For example, OpenMM distributes the charges onto nodes of a rectangular mesh using 5^{th} order B-splines.[215] The use of FFT permits the PME algorithm to scale as $\mathcal{O}\left(N\log(N)\right)$.

4.4 The general AMBER force field

The FF used throughout this work is the GAFF, which has parameters for almost all organic molecules.[182] It uses the following class I functional form

$$U = \sum_{bonds} \frac{K_b}{2}(r - r_{eq})^2 + \sum_{angles} \frac{K_\theta}{2}(\theta - \theta_{eq})^2 + \sum_{dihedrals} V_n \left[1 + \cos\left(n\phi - \gamma_n\right)\right] + \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{R_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{4.25}$$

where all terms are as previously defined. The charge method used in GAFF is the HF/6-31G* RESP charge model, which is presented in detail in Section 5.2.5. Alternatively, the AM1-BCC method,[225,226] which, like RESP, also emulates the HF/6-31G* electrostatic potential of a molecule, can also be applied to compute the atomic charges. The GAFF van der Waals parameters are the same as those used by the traditional AMBER FF. To derive the reference bond lengths, GAFF resorted to data obtained through X-ray and neutron diffraction experiments, as well as theoretical MP2/6-31G* calculations. The reference angle values were determined from previous FFs parameters, MP2/6-31G* calculations, Cambridge Structural Database (CSD) data, and empirical rules. Furthermore, bond and angle force constants were derived using empirical functions. Finally, torsional parameters were determined by fitting the series of cosine functions to torsional profiles obtained by performing geometry optimisations//single-point calculations at the MP4/6-311G(d,p)//MP2/6-31G* level of theory.

4.5 Force field parameterisation

As previously discussed, a FF consists of a functional form and a set of unknown FF parameters that enters the functional form. While the potential energy function is universal for a given FF class (see Section 4.2), and the FF parameters may be transferable within specific families of molecules (*e.g.*, proteins, nucleic acids, ligands), it is frequently necessary to derive new parameters for novel chemical moieties or challenging molecular interactions. FF parameterisation is, therefore,

the process with which optimal FF parameters are derived so that the FFs reproduce target theoretical and/or experimental properties. In the context of the research presented in this thesis, FF parameterisation is of particular importance for the parameterisation of novel ligands, for which transferable parameters are often inaccurate.

The idea of transferability of FF parameters is based upon spectroscopic observations that indicate that there is sufficient similarity between certain natural parameters within a given molecular class. For example, C-C bonds in alkanes are similar enough so that the same bond length can be used to describe, *e.g.*, pentane or hexane.[227] The concept of transferability is also related to that of atom types, which are assigned to the different elements of a molecule considering their hybridisation states and chemical environment.[173] Since the larger the number of atom types, the lower the generality of a FF, a balance must be achieved between transferability and accuracy. Although handy and successful in many situations, atom types make it difficult to systematically extend FFs, as their diversity is virtually unlimited. Moreover, an exhaustive definition of every possible atom type would result in redundant parameters, and, therefore, in practice, FFs limit their atom type definitions to those commonly encountered in most chemical and biological systems. This approach, however, does not come without issues to FF development, having led to the proposal of new approaches of assigning FF parameters that escape atom types.[228,229]

FF parameterisation typically involves two stages: calculation or selection of a target theoretical and/or experimental data set to be used in the parameterisation; and optimisation of the FF parameters. There is a wide range of experimental data that can be used for FF parameterisation, such as dielectric constants and thermodynamics properties, as well as data obtained from X-ray crystallography, IR spectroscopy, and NMR spectroscopy. Additionally, any property that can be computed by *ab initio* methods can also be included in the target data set, such as, *e.g.*, energies, forces, electrostatic potentials, vibrations, among others.[173,230] In what follows, we focus our discussion on the optimisation of FF parameters using *ab initio* data, as it was the approach taken in this work.

Of great historical importance in the fitting of FF parameters to QM data is the the force-matching method proposed by Ercolessi and Adams[231] in 1994. In this method, the set of unknown FF parameters, $\boldsymbol{p}$, is determined by minimising an objective function that measures the difference between the FF and *ab initio* forces for a large set of different configurations. The objective function of the force-matching method reads

$$\chi_F(\boldsymbol{p}) = \sum_{i}^{N_s} w_i \sum_{j}^{3N} \left(F_{i,j}^{MM}(\boldsymbol{p}) - F_{i,j}^{QM} \right)^2 \tag{4.26}$$

where i runs over all N_s configurations, j runs over all $3N$ atomic force components, w_i is the weight of the ith configuration, $F_{i,j}^{MM}$ is the force on the jth atomic component of configuration i as predicted by the FF, and $F_{i,j}^{QM}$ is the reference force obtained from an *ab initio* calculation. The weighting methods that can be used to modify the relative importance of the configurations of the target data set are discussed at length in Section 5.2.4. The force-matching method objective function of equation (4.26) can be readily recognised as a least squares problem of which the solution is the optimal set of FF parameters. Incidentally, over the past 30 years, solving least squares problems has been the main FF parameterisation strategy, as the force-matching method can be generalised to include other experimental or theoretical properties of interest, as well as to include regularisation terms that prevent overfitting. This generalisation is thoroughly discussed in Section 5.2.1, as it forms the basis of ParaMol, software that we developed and aims to ease the process of FF parameterisation by fitting to *ab initio* data (see Chapter 5). Parameterisation using least squares strategies has the advantage of being analytically solvable and not requiring an initial guess of FF parameters if the least squares problem is linear. On the other hand, for non-linear problems, local minimisation algorithms must be employed, though they do not ensure that a global minimum is found and require an initial guess of parameters.

Before the advent of least squares strategies for FF parameterisation, there was no

mathematically defined way to perform systematic and automatic optimisation of FFs. FF parameterisation, thus, involved informed trial and error procedures that were used to construct the first FF parameter sets.[227] As an advancement to least squares fitting strategies, alternative objective functions that did not require minimisation of the sum of squared residuals began to be employed. For these problems, genetic algorithms have stood out as one of the most viable global optimisation approaches.[232–234]

Parameterisation strategies not resorting to least squares fitting have also been developed. These strategies have a more physics-driven rather than data-driven philosophy. A striking example of such a physics-based parameterisation strategy is the one presented by the Quantum Mechanical Bespoke Kit (QUBEKit),[235] which forms the basis of the QUBE FF. In the QUBE FF, equilibrium bond lengths and angles are determined from QM optimised geometries; harmonic bond and angle force constants are derived from the QM Hessian matrix using a modified version of the Seminario method;[236,237] charge and LJ nonbonded parameters are derived from atoms-in-molecules partitioning of the QM electron density;[238] and the derivation of the torsional terms follows a more traditional least squares fitting approach. Besides enabling bespoke FFs to be derived and demonstrating high accuracy, one of the advantages of QUBEKit is its scaling performance, which allows systems composed of thousands of atoms, such as proteins, to be treated.[239]

Recently, Bayesian inference began to be used as a statistical formalism to perform FF parameterisation.[240–243] This statistical technique is based on the Bayes' theorem, which relates the conditional and marginal probabilities of two stochastic events. To understand how Bayesian inference works, consider the conditional probability of observing a set of parameters, $\boldsymbol{p}$, given a target data set, $\mathcal{D}$, which is proportional to

$$P(\boldsymbol{p}|\mathcal{D}) \propto P(\mathcal{D}|\boldsymbol{p})P(\boldsymbol{p}) \tag{4.27}$$

where $P(\boldsymbol{p}|\mathcal{D})$ is the posterior distribution, $P(\mathcal{D}|\boldsymbol{p})$ is the likelihood, and $P(\boldsymbol{p})$ is the prior distribution. FF parameterisation can then be seen as the problem of maximising the posterior distribution with respect to the FF parameters.[244] Hence, to find the optimal set of FF parameters, the likelihood and prior distributions must be determined. Evaluating the likelihood requires calculating the FF physical properties as a function of the FF parameters and comparing those properties to reference data through an error model, typically chosen so that the errors are Gaussian-distributed. The FF outputs can be calculated using, *e.g.*, standard MM simulations, though in many situations this is a costly procedure, and thus a surrogate model, such as a Gaussian process, is employed to estimate the likelihood. This involves cheaply approximating the response surface of the physical properties with respect to the FF parameters and comparing those values to the reference data through the error model. Furthermore, the prior distribution must be chosen so that it restraints the parameters to physically sensible regions using previous FF parameters, physical constraints, or physical intuition. Finally, since posteriors are generally non-analytical, MC sampling schemes are employed to generate trial moves in parameter space. The advantages of Bayesian inference are that it can be used to obtain FF parameters with uncertainty estimates for each parameter,[241,244] as well as to compare the fitness of various FF functional forms,[240] thus providing a complete framework that will be at the core of future FFs.

4.6 Summary

In this chapter, we have presented the fundamentals of molecular mechanics models. We introduced the concept of potential energy surface, discussed the main classes and functional forms of force fields, reviewed methods for calculating long-range interactions, and presented the details of the general AMBER force field. Lastly, we concluded with a review of the current state of the art of force field parameterisation strategies.

The next chapter comprises the first research study of this thesis. It presents the theory, development, implementation, and validation of ParaMol, software that aims to ease the process of FF parameterisation. ParaMol has a special focus on the parameterisation of bonded and nonbonded terms of druglike molecules by fitting to *ab initio* data. We demonstrate the capabilities of the software by deriving bonded parameters of three widely-known drug molecules: aspirin, caffeine, and a norfloxacin analogue. Additionally, we illustrate the best practices to follow when employing specific parameterisation routes; the sensitivity of the fitted parameters to the fitting procedure; and the features of the various weighting methods available to weight configurations used in the fitting.

Chapter 5

ParaMol: A Package for Automatic Parameterisation of Molecular Mechanics Force Fields

The ensemble of structures generated by MM simulations is determined by the functional form of the FF employed and its parameterisation. For a given functional form, the quality of the parameterisation is crucial and determines how accurately observable properties can be computed from simulations. Whilst accurate FF parameterisations are available for biomolecules, such as proteins or DNA, the parameterisation of new molecules, such as drug candidates, is particularly challenging as these may involve functional groups and interactions for which accurate parameters are not available. In this chapter, in an effort to address this problem, we present ParaMol, a Python package that has a special focus on the parameterisation of bonded and nonbonded terms of druglike molecules by fitting to *ab initio* data. We demonstrate the software by deriving bonded parameters of three widely-known drug molecules: aspirin, caffeine, and a norfloxacin analogue. For these molecules, we show that, within the constraints of the functional form, the methodologies implemented in ParaMol are able to derive near-ideal parameters. Additionally, we illustrate the best practices to follow when employing specific parameterisation routes; the sensitivity of

different fitting data sets, such as relaxed dihedral scans and configurational ensembles, to the parameterisation procedure; and the features of the various weighting methods available to weight configurations. Owing to ParaMol's capabilities, we propose that this software can be introduced as a routine step in the protocol normally employed to parameterise druglike molecules for MM simulations.

This chapter has been published as an article in the Journal of Chemical Information and Modeling:

- Morado, J.; Mortenson, P. N.; Verdonk, M. L.; Ward, R. A.; Essex, J. W.; Skylaris, C.-K. ParaMol: A Package for Automatic Parameterization of Molecular Mechanics Force Fields. J. Chem. Inf. Model. 2021, 61 (4), 2026–2047. https://doi.org/10.1021/acs.jcim.0c01444

5.1 Introduction

MM-based simulation methods such as MD and MC are commonly employed to solve many problems in chemistry, physics, biochemistry, and condensed matter.[7] The ability of these MM-based methodologies to correctly model systems of interest relies mainly on two aspects: their capacity to extensively sample the configurational space and the accuracy of the underlying FF.

The sampling problem is still an area of intensive research, with many enhanced sampling methods being proposed in the past decades, *e.g.*, MetaD,[29,30] Hamiltonian replica-exchange,[31–33] and umbrella sampling.[34] On the other hand, the accuracy of MM simulations relies on the underlying FF, which comprises a functional form and a set of parameters. The functional form consists of a function that defines the potential energy of the system and allows the calculation of forces, which enables equations of motion to be numerically solved. Amongst the most commonly used fixed-charge FF functional forms are AMBER,[182,245]

GROMOS,[184] CHARMM,[183] and OPLS.[185–188] These FFs already contain extensive databases of parameters for different types of molecules. Even so, many applications require the parameterisation of novel molecules or levels of accuracy in the conformations and energetics that are unattainable using default FF parameters. Of great importance is the parameterisation of molecules for drug-design applications and for the calculation of quantum corrections to classical free energies, which were shown to converge faster if MM descriptions more similar to the quantum level are employed.[246–248] Here, in an effort to address the problem of FF accuracy, we present ParaMol, a software package that is capable of deriving bespoke FF parameters in an automated fashion by fitting to *ab initio* data.

Different software packages have already been released for the purpose of automatic FF parameterisation. Each has its features, specific methods, and design choices. For example, Paramfit[249] is capable of parameterising the bonded parameters in the AMBER equation by fitting to *ab initio* forces and energies; QUBEKit uses a physics-driven parameterisation methodology that enables bespoke FFs to be derived for systems composed of thousands of atoms;[235] ffTK[250] (VMD plugin) and GAAMP[251] were designed specifically to derive CHARMM-compatible parameters for small molecules and permit the parameterisation of charges and bonded parameters; the CPMD software package[252] also contains a QM/MM force-matching implementation and can derive charges and bonded terms parameters for the AMBER and GROMOS96 equations; Schrödinger's proprietary software is capable of parameterising the OPLS FF and systematically generating missing torsional parameters;[185,187,188,253] finally, ForceBalance[230,254] stands out due to its generality, as it is capable of parameterising different FF functional forms to experimental data and has many optimisation algorithms available.

ParaMol can be used as a stand-alone package and as a Python package to create user-customised parameterisation protocols. It differs from other parameterisation software packages in some of its implementation choices and in its special

focus on the parameterisation of druglike molecules from first-principles quantum mechanics. ParaMol aims to ease all steps in a standard parameterisation workflow: it automates configurational sampling, the calculation of reference data, and the procedure of obtaining the optimal FF parameters. Therefore, ParaMol can be easily introduced as a routine step in the standard workflow used to prepare druglike molecules for MM simulations. The software can also be extended to accommodate new objective functions, fitting properties, and FF functional forms. ParaMol also has parallel capabilities that allow distributing the calculation of the objective function and *ab initio* training data amongst the available computational resources. Currently, the package is able to derive parameters for class I additive potential energy functions, such as those used by AMBER, CHARMM, and OPLS FFs.[176]

As application examples, we assessed the limits of accuracy that can be attained by fitting bonded parameters of the GAFF functional form to QM calculations. For this purpose, we chose three widely-known drug molecules: aspirin, caffeine, and a norfloxacin analogue. To illustrate the dihedral scan functionality that is available in ParaMol, we optimised the dihedral parameters associated with the main rotatable bond of a norfloxacin analogue. Furthermore, for aspirin, we explored the advantages and limitations of the use of dihedral scans against configurational ensembles generated MD simulations, as ways of exploring the PES, and optimised aspirin's bond, angle, and dihedral parameters; finally, we employed adaptive parameterisation to derive new bonded parameters for caffeine.

This chapter is structured as follows: we first present the basic theory underlying the implementation of the ParaMol package, *viz.*, the generalisation of the force-matching method,[231] the restrained electrostatic potential (RESP) model,[211,255,256] the optimisation algorithms available, as well as some remarks about regularisation and parameter preconditioning; then we describe the organisation of the software package and its functionalities. Finally, we conclude by presenting the application of different parameterisation protocols to the previously mentioned test cases.

5.2 Theory and methods

5.2.1 Generalisation of the force-matching method

A generalisation of the original force-matching method[231] can be formulated in which, instead of only fitting forces, the aim is to fit the FF to reproduce, within the constraints of the functional form, any target experimental or theoretical property of interest. In this context, the optimisation procedure can be seen as a mathematical problem in the space of FF parameters, which are denoted as $\boldsymbol{p}$, being $\boldsymbol{p}$ a vector containing all the optimisable parameters. The optimisation aims to determine the optimal set of parameters that minimise an objective function, here denoted as X. The objective function contains the squares of the residuals, and in its general form reads

$$X(\boldsymbol{p}) = X_F(\boldsymbol{p}) + \sum_{\{A\}} X_A(\boldsymbol{p}) + \Theta(\boldsymbol{p}) \tag{5.1}$$

where X_F corresponds to the term of the objective function by which MM forces are fitted to reference values, X_A accounts for the fitting of any other property of interest A to reference data (*e.g.*, potential energy, electrostatic potential), and $\Theta(\boldsymbol{p})$ is a regularisation term that can be optionally included in order to prevent overfitting (discussed in detail in Section 5.2.6). Specifically, two different types of force-matching terms, X_F^I and X_F^{II}, are implemented in ParaMol. The type I force-matching term fits the norm of the atomic forces to reference data and it has the following form[231]

$$X_F^I(\boldsymbol{p}) = \frac{1}{3N_a} \sum_i^{N_s} \omega_i \sum_j^{N_a} \frac{\left|\Delta \boldsymbol{F}_{i,j}\right|^2}{\mathrm{Var}(\boldsymbol{F}^{ref})} \tag{5.2}$$

where $\Delta \boldsymbol{F}_{i,j} = \boldsymbol{F}_{i,j}^{MM}(\boldsymbol{p}) - \boldsymbol{F}_{i,j}^{ref}$, $\boldsymbol{F}_{ij}^{ref}$ and $\boldsymbol{F}_{ij}^{MM}$ are the QM (reference) and MM force vectors, respectively, of atom j in conformation i, ω_i is the weight of the ith conformation, N_s is the number of structures provided, and N_a the number of

atoms of the system. The type II force-matching term fits every component of the atomic forces to reference data and it is given by[254]

$$X_F^{II}(\boldsymbol{p}) = \frac{1}{3N_a}\sum_i^{N_s}\omega_i\sum_j^{N_a}\left[\Delta\boldsymbol{F}_{i,j}(\boldsymbol{p})^T\langle\boldsymbol{F}_{i,j}^{ref}\otimes\boldsymbol{F}_{i,j}^{ref}\rangle^{-1}\Delta\boldsymbol{F}_{i,j}(\boldsymbol{p})\right] \tag{5.3}$$

It is worth noting that the variance, $\mathrm{Var}(\boldsymbol{F}^{ref})$, and covariance, $\langle\boldsymbol{F}_{i,j}^{ref}\otimes\boldsymbol{F}_{i,j}^{ref}\rangle$, are used in X_F^I and X_F^{II}, respectively, so that the residuals in the objective function are dimensionless and maximally of unit magnitude. Furthermore, in equation (5.1), X_A is a general expression for the fitting of any property of interest A to reference data, which for the case of a global property is given by

$$X_A(\boldsymbol{p}) = \sum_i^{N_s}\omega_i\frac{\left(A_i^{MM}(\boldsymbol{p}) - A_i^{ref}\right)^2}{\mathrm{Var}(A^{ref})} \tag{5.4}$$

Furthermore, similarly to what was done in equations (5.2) and (5.3), if A is an atom-based property, the appropriate sum over all atoms and normalisation constant must be introduced. It is worth mentioning that a special case of equation (5.4) is considered when the property to be fitted is the energy, *i.e.*, when $A = E$. In this case, since different levels of theory have different energy references (*e.g.*, the QM and MM energies usually differ by several orders of magnitude), the expression used for X_E reads

$$X_E(\boldsymbol{p}) = \sum_i^{N_s}\omega_i\frac{\left(E_i^{MM}(\boldsymbol{p}) - E_i^{ref} - \langle\Delta E\rangle\right)^2}{\mathrm{Var}(E^{ref})} \tag{5.5}$$

where E_i^{ref} and E_i^{MM} are the QM (reference) and MM potential energies, and $\langle\Delta E\rangle = \frac{1}{N_s}\sum_i^{N_s}\left(E_i^{ref} - E_i^{MM}\right)$ is a term that brings the two distributions together by subtracting the average difference between the reference and MM energies from the energy residuals.

5.2.2 Data set generation: dihedral scans and configurational ensembles

Regarding the schemes through which reference data sets of configurations and their respective properties of interest can be generated for parameterisation purposes, two methods are routinely employed to explore the PES of small organic molecules: dihedrals scans, and configurational ensembles.

The most common method to explore the PES is by performing one-dimensional relaxed scans of the DOFs of interest (*e.g.*, dihedrals), in which only the DOFs not explicitly being constrained are allowed to relax (by default all the DOFs not being scanned). There are mainly two disadvantages associated with this methodology. First, the energy can change dramatically if a substituent group falls into a different molecular configuration due to concerted motions, which causes discontinuities in the energy profiles. Second, if there are non-negligible couplings between DOFs, *i.e.*, if the DOFs are non-orthogonal, then the energy landscape is not correctly described by a one-dimensional surface, demanding higher-dimensional scans that quickly become prohibitive.[257,258]

Alternatively to the use of dihedral scans, it is also possible to use either MD or MC simulations to generate configurational ensembles. Whilst the disadvantages of the relaxed scans are not present in this case, generating configurational ensembles requires sufficiently long simulations that guarantee exploration of the relevant parts of the PES, or specific techniques that guarantee sufficient coverage of sampling (*e.g.*, replica-exchange algorithms[31–33]).

5.2.3 Dihedral fitting approaches

Although the derivation of dihedral parameters can be performed using configurational ensembles, computationally it is often less costly and more convenient to use dihedral scans. We implemented in ParaMol two different dihedral fitting approaches that use dihedral scans as the fitting data sets. In what follows,

we use the notation $E^{A,B}$, where A and B refer to the levels of theory used to calculate the single-point energies and perform the geometry optimisations, respectively.

A commonly employed dihedral fitting approach,[249,251,259] hereinafter referred to as QM-relaxed, is to derive the dihedral parameters by determining differences between the MM single-point energies ($E^{MM,ref}$) and the QM single-point energies ($E^{ref,ref}$), obtained in vacuum and using the same QM geometry for both the MM and QM calculations. In this case, the objective function reads

$$X_{dih}(\boldsymbol{p}) = \sum_{i}^{N_s} \omega_i \frac{\left(E_i^{MM,ref}(\boldsymbol{p}) - E_i^{ref,ref} - \langle \Delta E \rangle\right)^2}{\mathrm{Var}(E^{ref,ref})} \tag{5.6}$$

However, as pointed out by other authors,[250,260,261] an approach that is often underappreciated and that yields more adequate parameters, hereinafter referred to as MM-relaxed, is obtained when a further MM optimisation (with the proper constraints) is also carried out for every conformation of the dihedral scan. Therefore, since in this case the MM single point-energy ($E^{MM,MM}$) is calculated based on the MM-relaxed geometry rather than the QM-relaxed geometry, the objective function reads

$$X_{dih,relaxed}(\boldsymbol{p}) = \sum_{i}^{N_s} \omega_i \frac{\left(E_i^{MM,MM}(\boldsymbol{p}) - E_i^{ref,ref} - \langle \Delta E \rangle\right)^2}{\mathrm{Var}(E^{ref,ref})} \tag{5.7}$$

The rationale underlying the MM-relaxed approach is that the MM energy is highly influenced by the intramolecular energy of terms associated with parameters that are not being optimised. QM optimisations can result in geometries that are deformed from the point of view of the MM level because the FF parameters may have been obtained by fitting to experimental data or QM levels that are different from those used to perform the dihedral scans. Consequently, the MM and QM dihedral profiles may present significant differences in regions of the PES that are of primary interest for proper modelling, such as minima

and transition states. Hence, by using QM geometries that acquire MM energies (QM-relaxed approach) rather than MM geometries that acquire MM energies (MM-relaxed approach), the parameterisation procedure is likely to generate biased parameters, as it attempts to correct for differences in the dihedral profiles that are unrelated to the FF parameters associated with the dihedral(s) being scanned. Interestingly, the MM-relaxed approach can also be used to take into account more complex relaxation situations, such as, *e.g.*, the environment-related effects that occur in solution or a protein environment.[261]

Overall, it is recommended to use the MM-relaxed approach as long as the resultant MM-optimised geometries do not significantly differ from the QM-optimised ones. As a rule of thumb, if the global conformational preferences of the molecule do not change after the MM optimisation, then the MM-relaxed approach is preferred. Finally, it is also worth mentioning that the QM-relaxed approach is a good approximation whenever the DOFs not being scanned match in the QM and MM optimised geometries, or whenever the remaining FF terms do not contribute significantly to the dihedral profile. This concern is particularly important for hard DOFs (bonds and angles), for which small differences in value lead to large energy changes due to large force constants. Therefore, it is recommended to relax those DOFs before deriving dihedral parameters, as otherwise biased parameters are likely to be obtained.

5.2.4 Weighting methods

We implemented a variety of weighting methods in ParaMol that give more importance to some conformations than others. Currently, the weighting methods available in ParaMol are the following:

- **Uniform weighting**: this is the simplest weighting method that is possible to apply. It assigns equal weight to all conformations, such that the weight of any two conformations i and k is given by

$$\omega_i = \omega_k = \frac{1}{N_s} \tag{5.8}$$

This weighting method may be problematic if very high-energy conformations are present because, in order to minimise the errors in their description, the fitting procedure may adversely affect the description of highly-populated low-energy conformations. This usually happens due to constraints of the functional form. A practical solution for this problem is to use ParaMol to prune out conformations of which the reference energy is larger than a given value relative to the minimum energy conformation (*e.g.*, 10.0 kcal/mol).

- **Boltzmann weighting**: Boltzmann weighting based on the reference (QM) energies gives more importance to low-energy conformations than to high-energy conformations. This non-uniform weighting is achieved by weighting each conformation by the factor

$$\omega_i = \frac{\exp\left[-\beta(E_i^{QM} - < E^{QM} >)\right]}{\sum_j^{N_s} \exp\left[-\beta(E_j^{QM} - < E^{QM} >)\right]} \tag{5.9}$$

The disadvantage of Boltzmann weighting is that it usually leads to inaccurate energies for conformations located at or near high-energy barriers. This often compromises the dynamics of the model, preventing its use in standard simulation methods.[258]

- **Non-Boltzmann weighting**: The non-Boltzmann weighting method implemented in ParaMol is the one proposed by Wang *et al.*,[262] in which the expression used for the weight of a conformation i is given by

$$\omega_i = \frac{\exp\left[-\beta(E_i^{MM} - E_i^{QM} - < \Delta E >)\right]}{\sum_j^{N_s} \exp\left[-\beta(E_j^{MM} - E_j^{QM} - < \Delta E >)\right]} \tag{5.10}$$

where $\Delta E = E^{MM} - E^{QM}$. This weighting method gives larger weights to conformations in which the MM energy is underestimated ($E^{MM} -$

$E^{QM} < 0$) than to conformations in which the MM energy is overestimated ($E^{MM} - E^{QM} > 0$), with respect to the reference (QM) energy. Hence, as pointed out by its original authors,[263] configurations with negative $E^{MM} - E^{QM}$ have a spuriously large thermodynamic weight in the MM representation and are more likely to appear during MM sampling, which could lead to incorrect equilibrium averages due to incorrect equilibrium structures. On the other hand, configurations with positive $E^{MM} - E^{QM}$ have a spuriously small weight in the MM representation, which could result in overestimation of transition-state energies and underestimation of fluctuations. Therefore, by heavily penalizing configurations with MM energies that are lower than QM energies, this weighting procedure avoids the creation of spurious MM minima and forces the fitting errors into the high-energy regions, which are, in a sense, higher-order errors than the incorrect equilibrium averages.

- **Manual weighting**: This weighting method allows the user to choose the weights of each conformation, which will be constant throughout the whole optimisation. This may be of special importance if the user knows which conformations should be given more or less importance. Other publications have suggested that weights of less than or equal to five are typically appropriate for the underrepresented conformations, assuming weights of unity for the rest of the target data.[260]

5.2.5 Charge fitting to electrostatic potential: the RESP model

ParaMol can derive atom-centred point charges by fitting to a reference electrostatic potential (ESP).[255] Specifically, ParaMol contains an implementation of the RESP model.[211,256] The objective function used in the multiconformational RESP fit reads

$$X_{RESP}(\boldsymbol{q}) = \sum_{i}^{N_s} \omega_i \sum_{k}^{N_{grid}} \left(\sum_{j}^{N_{charges}} \frac{q_j}{r_{jk,i}} - V_{k,i}^{QM} \right)^2 + \lambda_1 \left(\sum_{j}^{N_{charges}} q_j - q_{tot} \right) + \sum_{m=2}^{N_{constraints}} \lambda_m f_m(\boldsymbol{q}) + \Theta(\boldsymbol{q}) \tag{5.11}$$

where ω_i is the weight of the ith conformation, $\boldsymbol{q} = (q_1, ..., q_j)$ is the vector of charges allowed to vary during the fitting, V_k^{QM} is the value of the calculated ESP at the grid point k, and r_{jk} is the distance between the atomic centre j and the grid point k. Furthermore, λ_1 corresponds to the Lagrange multiplier used to constraint the sum of the charges to the total molecular charge, and λ_m (with $m > 1$) corresponds to the Lagrange multipliers used to impose other types of constraints such as, for instance, symmetry constraints.[256] Finally, as in equation (5.1), $\Theta(\boldsymbol{q})$ defines the penalty function optionally applied so that the fit becomes restrained.

ParaMol is able to perform charge fitting by using SciPy's[264] implementation of the COBYLA,[265] SLSQP[266] or Trust Region[267] algorithms. Moreover, we also implemented an analytical solution of the system of equations that arises from taking the derivatives of equation (5.11) with respect to the charges and Lagrange multipliers. More information about the implementation of this analytical solution can be found in Refs. 255 and 211.

5.2.6 Preconditioning of optimisable parameters and regularisation

In order to avoid overfitted parameterisations, which may occur whenever the amount of reference data used in the optimisation is not extensive enough, regularisation has to be applied so that, during the optimisation, the parameters remain within a range of values that makes physical sense. This is done through the inclusion of the penalty functions Θ in equations (5.1) and (5.11). In Bayesian

statistics, penalty functions correspond to the negative logarithm of a prior distribution, and the regularised objective function corresponds to the posterior distribution.[230] Hence, it is possible to design penalty functions by making assumptions regarding the prior distribution of the parameters. We implemented in ParaMol different regularisation methods. For instance, if the user wants to apply L1 regularisation, *i.e.*, if the prior distribution of a parameter p is assumed to be given by $P(p) = \exp\left(-\frac{|p-p^0|}{\gamma}\right)$, where γ controls the width of the distribution and p^0 is the parameter initial guess, then the penalty function reads

$$\Theta_{L1}(\boldsymbol{p}) = \alpha \sum_{m}^{N_p} \frac{|p_m - p_m^0|}{\gamma_m} \tag{5.12}$$

where α is an adjustable parameter that controls the strength of the regularisation. Similarly, if the user wants to apply L2 regularisation, *i.e.*, if the prior distribution of the parameters is assumed to be Gaussian, a harmonic penalty function is then employed, which reads

$$\Theta_{L2}(\boldsymbol{p}) = \alpha \sum_{m}^{N_p} \frac{(p_m - p_m^0)^2}{\gamma_m^2} \tag{5.13}$$

The widths of the prior distributions can be automatically generated or manually chosen by the user using physical knowledge. Regarding the automatic generation of these hyperparameters, ParaMol uses a procedure in which either the arithmetic or geometric mean is calculated for classes of FF parameters (*e.g.*, bond force constants, dihedral phases, etc.). All parameters within the same class will then use this mean value as the width of their prior distributions. This is similar to the approach followed by ForceBalance.[254] Moreover, the procedure used to automatically generate the prior widths can also be used to construct the Jacobi preconditioner, which scales the parameters so that they are all treated on the same footing by the optimisation algorithm. Specifically, the Jacobi (diagonal) preconditioner used in ParaMol is given by $\boldsymbol{P} = \gamma_m \delta_{mm}$.

Finally, if charges are being fitted, it is also possible to apply a hyperbolic regularisation term that prevents the charges from deviating too much from a target charge of zero.[211] This hyperbolic penalty function is given by

$$\Theta_{HB}(\boldsymbol{q}) = \alpha \sum_{m}^{N_{charges}} \left[(q_m^2 + \beta^2)^{1/2} - \beta \right] \tag{5.14}$$

where α and β are adjustable hyperparameters that define the asymptotic limits of the strength of the restraint and the tightness of the hyperbola around its minimum, respectively.

5.2.7 Optimisation algorithms

We implemented in ParaMol global and local optimisation algorithms that perform non-linear minimisation of the objective function, *viz.*, non-reversible Monte Carlo,[268] gradient descent,[269] stochastic gradient descent,[270] and simulated annealing.[271] Furthermore, ParaMol also interfaces with the Python SciPy package, from which several minimisation algorithms can be used (*e.g.*, Nelder-Mead, Powell, BFGS, L-BFGS-B, SLSQP, COBYLA, and Trust Region). Since ParaMol has no implementation of analytical derivatives of the objective function with respect to the set of parameters being optimised, whenever necessary the Jacobian matrix is calculated using numerical derivatives, and the Hessian matrix is approximated using BFGS or SR1 updates.[264,272]

In addition to the non-linear iterative optimisers previously described, ParaMol also offers analytical linear least squares (LLS) solutions to the parameterisation of the bonded parameters (bond, angle, and dihedral parameters) of class I FFs.[258,259] This fitting approach can be employed alongside any of the available regularisation schemes, though currently only to find the minimum of the squared deviations of the energies, as shown in equation (5.5). The LLS solver does not support the use of the non-Boltzmann weighting given by equation (5.10), as the dependence of this weighting method on the MM energies makes it

suited to be solved only through non-linear optimisation. The main "disadvantage" of the LLS fitting approach is that it provides a single deterministic answer, whereas a scatter of possible solutions with nearly the same quality concerning the objective function usually exists. On the other hand, iterative methods, such as the stochastic Monte Carlo or gradient-based optimisations, can find other nearby solutions, which may have value if they produce different simulation outcomes that may be preferred in specific cases (*e.g.*, produce the right helical propensity, or orientation of a drug molecule in a protein binding site). Nevertheless, it should be stressed that these solutions should only be fielded once the absolute optimum obtained by the LLS fitting has been attempted.

5.2.8 ParaMol package structure

ParaMol is designed to be used as a Python package that can be easily extended by the user to include extra functionalities or other parameterisation protocols. The Python (sub)subpackages and modules that comprise ParaMol's top-level package, as well as the main interactions between them, are depicted in Figure 5.1. ParaMol uses OpenMM[215] as its MM engine and has implemented wrappers of AMBER, DFTB+[273,274] and ASE.[275] The ASE package allows single-point or geometry-optimisation calculations to be performed using any of the calculators or optimisers available in it. Moreover, ParaMol also offers symmetrisers that enable subjecting reparameterisations to the symmetries defined in the topology files used by traditional MM packages, such as AMBER, CHARMM, or GROMACS. Interfaces to read and write input files for these packages are also available.

In order to set up a custom parameterisation protocol using ParaMol, firstly a ParaMol's representation of the system of interest must be created by resorting to the *ParaMolSystem* object defined in the *system* module of the *System* subpackage. An instance of this object stores the reference data, contains the MM (modules in *MM_engines* subpackage) and QM engines (wrappers defined in the modules of

the *QM_engines* subpackage) used by the system, and stores ParaMol's representation of the FF (modules in *ForceField* subpackage). Furthermore, an instance of the *ObjectiveFunction* object must also be created. This object is defined in the *objective_function* module of the *ObjectiveFunction* subpackage. The *ObjectiveFunction* object requires some properties (objects defined in the modules of the *ObjectiveFunction.Properties* subsubpackage) to be fitted to the reference data, and an instance of the *ParameterSpace* object defined in the *parameter_space* module of the *ParameterSpace* subpackage, which stores the vector space of optimisable parameters. Lastly, one of the optimisers available in the modules of the *optimisers* subpackage must be used to perform the minimisation of the objective function.

Alternatively, it is possible to resort to one of the tasks already implemented in the modules of the *Tasks* subpackage to perform specific parameterisation protocols (these tasks are described in detail in the next subsection). ParaMol tasks greatly simplify the use of the software because they usually only require instancing of *ParaMolSystem* objects and of the desired task. More information about the ParaMol package and examples of how to use the code can be found at ParaMol's website, `https://paramol.readthedocs.io`.

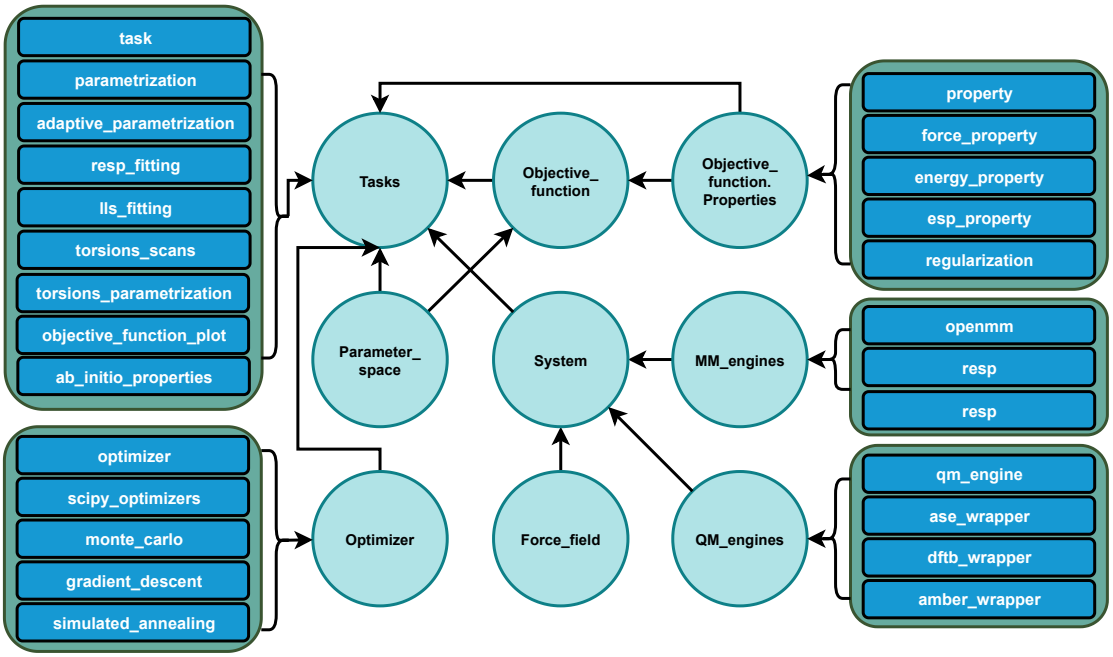

Figure 5.1: Overview of the structure of the ParaMol Python top-level package. Paramol's (sub)subpackages are represented as cyan circles and the respective modules as blue rectangles. The most relevant interactions between (sub)subpackages are represented with arrows. The direction of the arrows indicates that modules from the destination (sub)subpackage require modules from the source (sub)subpackages (*e.g.*, the modules of the *System* subpackage require modules from the *Force_field*, *MM_packages* and *QM_packages* subpackages). Some (sub)subpackages and modules are not shown for the sake of conciseness.

5.2.9 ParaMol tasks

ParaMol includes built-in tasks that perform specific parameterisation protocols and also routines that aid the parameterisation protocols themselves as, *e.g.*, utilities that assess convergence of the optimisation procedure and calculate *ab initio* reference data. Currently, the parameterisation tasks available in ParaMol are the following:

- **Parameterisation** (*ParaMol.Tasks.parametrization*): This tasks performs ParaMol's standard parameterisation protocol. Specifically, the parameterisation task creates the parameter space, the objective function, and the optimiser that are used in the optimisation of the FF parameters. It also preconditions the optimisable parameters and defines the constraints to which

the optimisation is subjected, *e.g.*, total charge or symmetry constraints. Regarding the symmetry constraints, these have to be defined manually by the user so that physical-based symmetries are retained. Alternatively, it is also possible to apply ParaMol's AMBER, CHARMM, or GROMACS symmetrisers, which subject the optimisation to the symmetries defined in the respective topology files.

During the optimisation procedure itself, the objective function typically has to be evaluated hundreds to thousands of times. This evaluation can be done either in serial or in parallel. For the latter case, it is possible to use OpenMM's support of different platforms and distribute the computation of the MM properties amongst the available CPUs or GPUs by using Python's multiprocessing package.

- **LLS fitting** (*ParaMol.Tasks.lls_fitting*): This task is very similar to the parameterisation task, except that it does not support the non-Boltzmann weighting scheme and can only parameterise the bond, angle, and dihedral parameters of class I FFs by minimising the squared deviations of the energies. The LLS solution is calculated by resorting to Numpy's[276] numpy.linalg.lstsq function.

- **Adaptive parameterisation** (*ParaMol.Tasks.adaptive_parametrization*): This task performs adaptive parameterisation, which consists of a self-consistent loop in which, at each iteration, configurational sampling and parameter optimisation are carried out. First, given an initial guess of FF parameters, a set of configurations is generated using any integrator available in OpenMM, and the reference *ab initio* data for this data set is calculated. Then, a new set of optimal parameters is determined by resorting to the Parameterisation task. Finally, the convergence of the self-consistent procedure is assessed, which is assumed to occur when the root-mean-square deviation (RMSD) of the current parameters with respect to parameters of the previous iteration is less than a user-defined threshold. The correction to the weights of the conformations described in Ref. 254 can be optionally applied in every iteration. This correction removes the bias introduced

by the fact that conformations at different iterations are sampled using different FFs.

- **Dihedral scans** (*ParaMol.Tasks.torsions_scans*): This task performs 1D or 2D relaxed dihedral scans. Specifically, for the 1D case (2D scans follow the same approach), this task requires specification of the quartet of atoms a-b-c-d for which the potential energy scan will be performed by rotating the b-c bond. By default, at a given step of the scan, only the dihedral angle being scanned is fixed, allowing the remaining DOFs to relax during the geometry optimisation. However, it is also possible to further constrain other DOFs, such as, *e.g.*, bonds, angles, or other dihedrals, during a scan, a feature that resorts to the capabilities of the ASE package.[275]

- **Automatic soft dihedral parameterisation** (*ParaMol.Tasks.torsions_parametrization*): This task allows automatic parameterisation of soft (rotatable) dihedrals in a way inspired by the protocol used by GAAMP.[251] This approach is of particular importance because soft dihedrals, which have small energy barriers, are the ones that control the conformational preferences of a molecule. Therefore, accurate dihedral parameters are important since they crucially determine the topology of the PES. The first step of this task concerns the identification of soft bonds, here defined as bonds that contain soft dihedrals, which is done resorting to the RDKit package.[277] ParaMol then iterates over all soft bonds and generates relaxed scans of their soft dihedrals. If a soft bond has more than two soft dihedrals of the same type, *i.e.*, soft dihedrals that share exactly the same atom types, a relaxed dihedral scan is only performed at the first encounter with this soft dihedral type. In addition, if two or more soft bonds have exactly the same soft dihedrals types, they are considered to belong to the same soft bond type, and thus scans are only performed at the first encounter with a soft bond of this type. Furthermore, every time a new soft dihedral type is scanned, optimisation of the parameters of that soft dihedral type is performed. This step is important to generate smoother energy profiles because, by default, ParaMol performs a MM geometry

optimisation before the QM geometry optimisation, a "preconditioning" that substantially decreases the computational cost of the high-level calculation. Hence, by having a gradually better MM representation, the MM optimisations are more likely to find QM-like energy minima that lower the cost of the QM optimisations. Finally, once ParaMol finishes iterating over the soft bonds, concomitant parameterisation of all soft dihedral parameters is performed using the calculated relaxed dihedral scans. In this final optimisation, the optimised parameters generated in the intermediate reparameterisations are forgotten, as ParaMol performs the final reparameterisation starting from the originally provided MM parameters. A diagram describing the workflow of this task is shown in Appendix A, Figure A.18.

- **RESP charge fitting** (*ParaMol.Tasks.resp_fitting*): This task performs charge fitting to a reference ESP that can be obtained from quantum chemistry packages, as previously described in subsection 5.2.5. ParaMol currently can extract the ESP directly from a Gaussian output. The output of other software has to be converted by the user to the format read by ParaMol.

- **Calculation of *ab initio* reference data** (*ParaMol.Tasks.ab_initio_properties*): This task calculates *ab initio* reference data by using any QM calculator available in the ASE package, or one of the wrappers of QM packages implemented in ParaMol. These calculations can be performed in serial or in parallel, the latter by distributing the workload amongst the available CPUs by using Python's multiprocessing package.

5.3 Application examples

In what follows, we present examples of reparameterisation of drug molecules. For this purpose, we used the GAFF, which was already presented in Section 4.4. The reparameterisations were performed using SciPy's SLSQP optimiser, and they were deemed to be converged whenever the objective function between two

successive iterations did not change by more than 10^{-6}, *i.e.*, $X_{n+1} - X_n < 10^{-6}$ (10^{-8} for the norfloxacin analogue example). Furthermore, GAFF parameters were used as the initial guess for the optimisations, except when stated otherwise. L2 (harmonic) regularisation was applied with prior widths inspired by the values reported in Ref. 263 (see Table 5.1). The objective function included as targets either forces - equation (5.3) - or energies - equation (5.5) -, or both at the same time. The initial parameterisation of the drug molecules was performed using Antechamber packages, which are part of AmberTools. AM1-BCC charges were calculated after the geometry was optimised at the target level of theory, which was either the DFTB+[273,274] implementation of SCC-DFTB including the D3 dispersion correction[278] with Becke-Johnson damping;[279] the non-local van der Waals DFT functional VV10,[68,69] as implemented in the linear-scaling DFT package ONETEP;[51,280,281] or the Psi4[282] implementation of the long-range corrected hybrid DFT functional ωB97X-D[70] with the 6-31G* basis set. The choice of these QM levels relies on the evidence that they perform quite well in determining conformations and respective energies.[72,79–82] The topology and coordinates files used as inputs to ParaMol were created using LEaP. Atom type symmetries were preserved during reparameterisation, unless otherwise indicated.

Table 5.1: ParaMol default prior width values for each parameter type.

Parameter type	Prior width
bond length	0.05 nm
bond force constant	10^5 kJ mol^{-1} nm^{-2}
bond angle	$\pi/16$ rad
angle force constant	10^2 kJ mol^{-1} rad^{-2}
dihedral phase	π rad
dihedral amplitude	16.736 kJ mol^{-1}
Lennard-Jones 12-6 ϵ	0.30 kJ mol^{-1}
Lennard-Jones 12-6 σ	0.20 nm
charge	0.5 e
1–4 electrostatic scaling factor	1.0
1–4 Lennard-Jones scaling factor	1.0

5.3.1 Details of the QM calculations

The SCC-DFTB calculations were performed using the 19.1 version of the DFTB+[273,274] package. These calculations included the D3 dispersion correction[278] with Becke-Johnson damping.[279] The dispersion values used for the Becke-Johnson damping were $a_1 = 0.5719$, $a_2 = 3.6017$, $s_6 = 1.0$, and $s_8 = 0.5883$ (as stated in the DFTB+ manual). The DFTB parameters employed were stored in Slater-Koster files that belong to the DFTB parameter set mio-1-1.[1]

The calculations that employed the non-local van der Waals DFT functional VV10[68,69] were performed using the 5.3.1.18 version of the linear-scaling DFT package ONETEP.[280,281] These calculations used a nonorthogonal generalised Wannier function (NGWF) radius of 8 Bohr and a kinetic energy cutoff of 700 eV. The convergence threshold for the root-mean-square gradient of the NGWFs was set to 1×10^{-4}.

The calculations that employed the ωB97X-D/6-31G* level of theory were performed using the 1.3.2 version of the Psi4 package[282] with default settings.

5.3.2 Dihedral scans: norfloxacin analogue

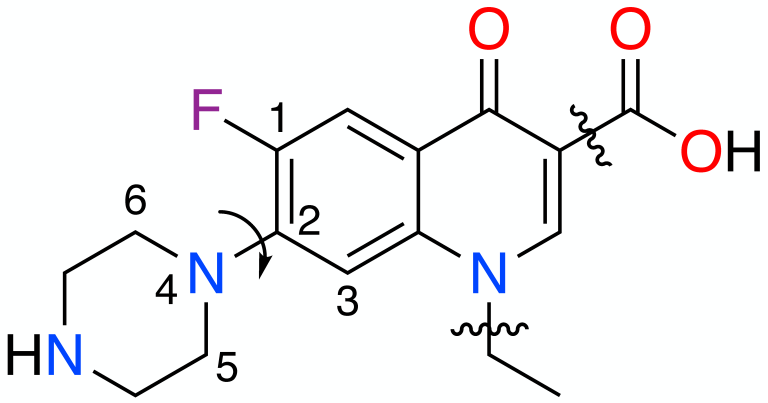

Figure 5.2: Molecular structure of norfloxacin. Owing to the unavailability of fluorine parameters in the mio-1-1 set,[1] the molecule used in this example is a norfloxacin analogue, in which we substituted the fluorine attached to the C1 carbon by a hydrogen atom. The differences in the torsional preferences introduced by this change are negligible, as it is shown in Figures A.5 and A.6 of Appendix A using data from the CSD.[2] Additionally, as a typical example of a fragment-based approach, we cut the molecule at the positions indicated by the wavy lines. All dangling bonds were capped with hydrogen atoms.

The first example of this chapter concerns the dihedral scan functionality implemented in ParaMol. In order to illustrate the procedure and the issues that may arise when reparameterising the dihedrals of a drug molecule, we optimised the parameters (force constants and phase constants) of the dihedrals associated with the main rotatable bond of a norfloxacin analogue (C2-N4, see Figure 5.2). As this molecule is achiral, to increase the transferability of the parameters we constrained the phase constants to be fixed to 180° (0° would be equivalently valid).[183,259] In the FF topology, two dihedrals contain C2 and N4 as inner atoms: C5-N4-C2-C1 and C5-N4-C2-C3. Both have the same atom types and, therefore, share the same set of parameters. GAFF models this dihedral type (c3-nh-ca-ca) by including only one term with periodicity $n = 2$. Nevertheless, to increase the

flexibility of the FF, we included all terms in the Fourier expansion with periodicities from $n = 1$ to $n = 6$. The numerical experiments performed here aimed to assess the performance of the weighting methods implemented in ParaMol, *viz.*, uniform, Boltzmann, and non-Boltzmann weighting when attempting to reproduce a target dihedral energy profile, as well as to illustrate the differences between the MM-relaxed and QM-relaxed approach.

Before analysing the results obtained, it is worth discussing some considerations about the QM dihedral energy profiles. Owing to the substantial discontinuities obtained in one-dimensional dihedral scans, we opted to perform fittings using a two-dimensional PES. The observed discontinuities were related to conformational changes that occurred in the piperazinyl ring as the N4-C2 bond was rotated, and they were caused by a flip of the pyramidal geometry of the N4 centre, which led to sudden energy variations. This phenomenon is seen by following the profile defined by the red stars in Figure 5.3, which indicate the minimum energy structure for a given ϕ angle. To avoid the energy discontinuities obtained in the one-dimensional dihedral scan, we opted to use a two-dimensional PES, generated by varying the C5-N4-C2-C1 (ϕ) dihedral angle from to -180° to 170° in steps of 10°, whilst concomitantly varying the C2-C6-N4-C5 (ψ) improper dihedral angle from 120° to 180°, and from -178° to -120° in steps of 2°. A total of 2196 geometry optimisations were performed, resulting in the 2D PES represented in Figure 5.3.

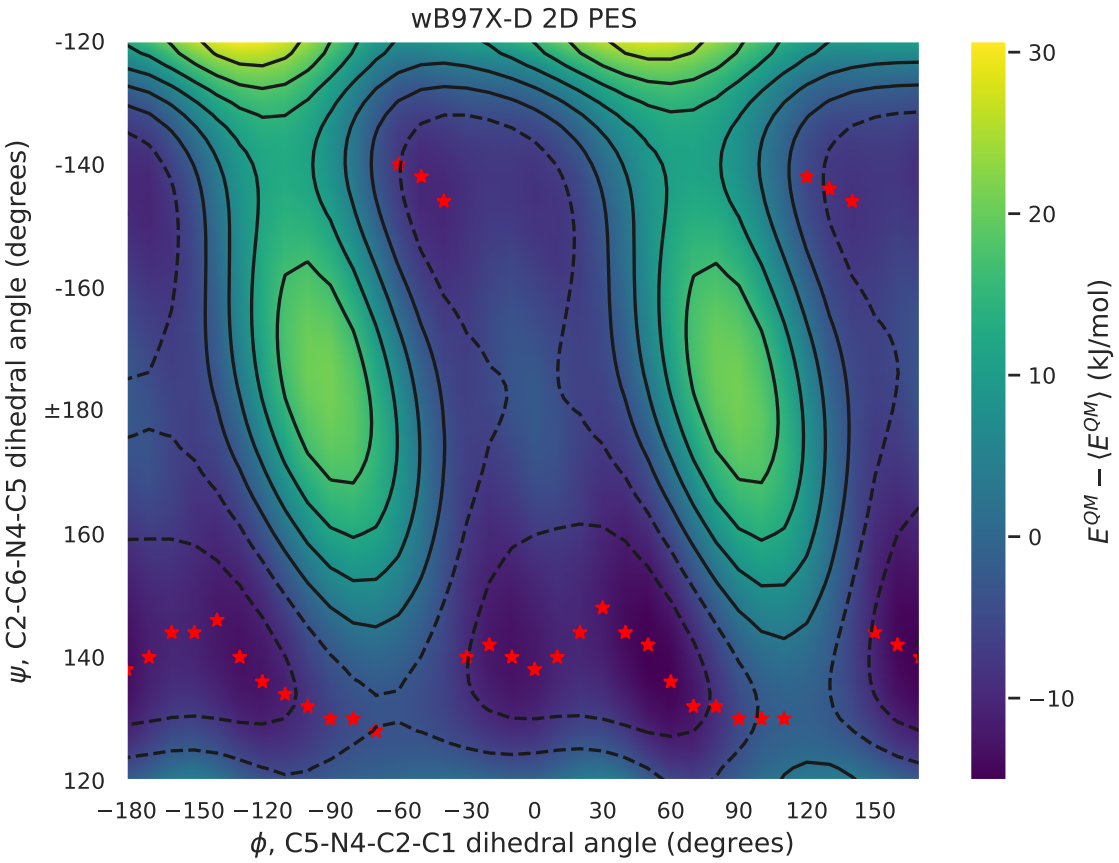

Figure 5.3: ωB97X-D/6-31G* PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) 2D dihedral scan for the norfloxacin analogue fragment. The red stars correspond to the minimum energy structure for a given ϕ dihedral angle value.

Besides the spontaneous conformational changes that may occur when rotating a chosen dihedral, another issue that commonly produces discontinuities in relaxed dihedral scans is hysteresis in the energy associated with the dihedral being relaxed.[258] This is prone to occur when optimising one data point after another, since non-orthogonal DOFs may be put under strain, accumulating potential energy that is released once the molecule crosses a given threshold, causing the strained DOFs to relax. A commonly employed solution to identify and correct this issue is to perform scans in both directions and pick the one that yields the more physically sensible profile. This practice is important because sudden physically-based changes in energy, like the ones seen in Figure 5.3, can be easily mistaken by discontinuities resultant of path-related hysteresis, and if the latter are artificial, the former are desirable to be captured (ideally exhaustively scanned by performing high-dimensional scans). Additionally, in

some applications, it is possible to impose symmetries by constraining specific DOFs, so that conformational changes and energy jumps are avoided. However, it is important to keep in mind that doing so yields a constrained adiabatic PES that misrepresents the (local) energy minimum for a given dihedral angle value. Consequently, due to this reason, we decided to proceed with two-dimensional relaxed scans, as they are a better representation of the PES of the molecule.

The results obtained for the ωB97X-D/6-31G* MM-relaxed dihedral fittings are shown in Figure 5.4, and the final optimised parameters are shown in Table 5.2 (the results of the SCC-DFTB-D3 reparameterisations are shown in Appendix A, Figures A.1, A.2, A.3, and A.4, and Table A.1). The final parameters obtained using the QM-relaxed approach of equation (5.6) are shown in Table 5.2, and the ωB97X-D/6-31G* dihedral energy profiles are shown in Figure 5.5. All fittings were performed using the objective function of equation (5.7) with an additional L2 regularisation term ($\alpha = 0.1$). The weighting temperature used was 500 K, and both the SLSQP SciPy optimiser and the LLS fitting approach were employed to optimise the dihedral parameters.

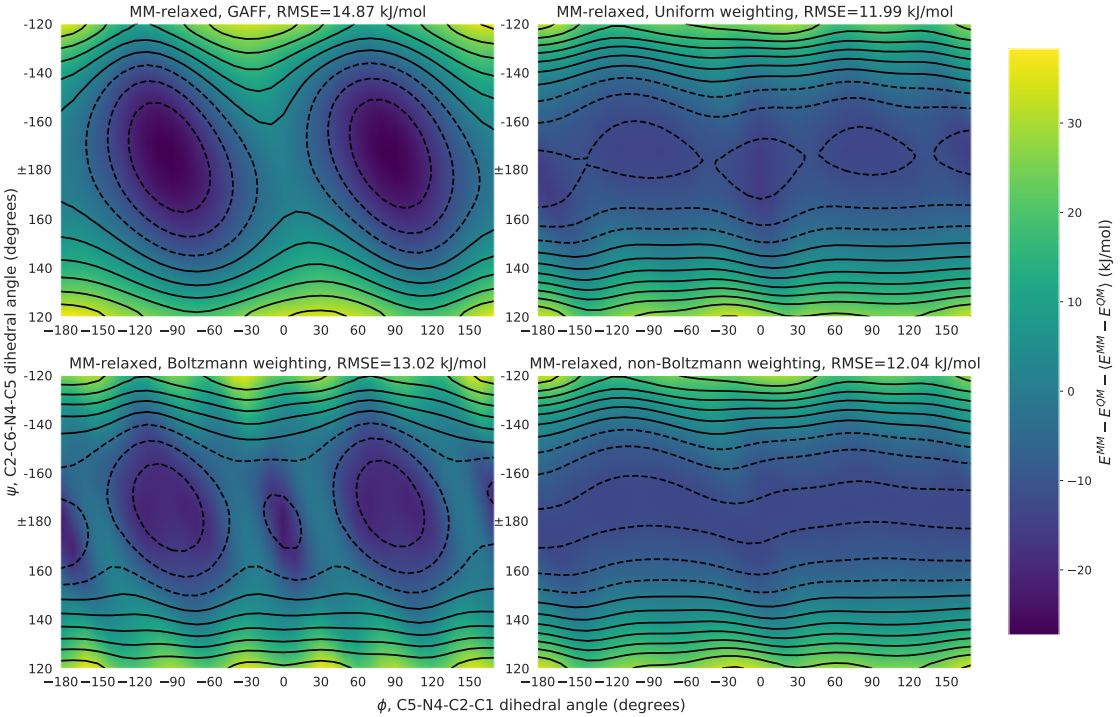

Figure 5.4: Relative errors of the MM FFs (GAFF, uniform, Boltzmann, and non-Boltzmann weightings) with respect to the target (ωB97X-D/6-31G*) PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) 2D dihedral scan. The MM-relaxed approach was employed to optimise the FFs.

Through the analysis of the fitting curves shown in Figure 5.4 and of the SCC-DFTB-D3 MM-relaxed fittings shown in Appendix A, we conclude that a significant improvement with respect to GAFF is observed for the three sets of parameters. Specifically, uniform and non-Boltzmann weighting performed the best in reproducing both the ωB97X-D/6-31G* and SCC-DFTB-D3 levels of theory, leading to fittings with root-mean-square errors (RMSEs) of 11.99/11.47 kJ mol^{-1} and 12.04/11.49 kJ mol^{-1}, respectively, for ωB97X-D/SCC-DFTB-D3, while Boltzmann weighting performed slightly worse in terms of RMSEs (13.02/12.52 kJ mol^{-1}). Furthermore, non-Boltzmann weighting led to an overall robust description of the QM minima and, more importantly, showed a tendency to skew the distribution of the errors towards positive values, being overall the weighting scheme with less negative relative errors (Figure 5.4). On the other hand, since Boltzmann weighting emphasises having a good description of the QM minima, it was the scheme that performed the worst for conformations located near high-energy barriers (see, *e.g.*, regions located at $\phi = [-120°, -90°]$

and $\phi = [90°, 120°]$). Boltzmann weighting underestimated the energies of the transition-state conformations by as much as *ca.* 20 kJ mol^{-1} for both ωB97X-D/6-31G* and SCC-DFTB-D3. Hence, although uniform weighting led to the best RMSEs, the residuals of this scheme tend to be symmetrically distributed around zero (Figure 5.4 and SCC-DFTB-D3 fittings in Appendix A), leading to the creation of artifacts in the PES, such as spurious minima. On the other hand, as non-Boltzmann weighting emphasises correcting regions of the PES for which the MM energy is lower than the QM energy, the creation of spurious MM minima is substantially mitigated by this weighting method, a feature that leads us to advocate for its use.

	GAFF	**Uniform**	**Boltzmann**	**non-Boltzmann**
			SciPy SLSQP solution	
V_1	0.00	-2.19 / -3.88	0.16 / -1.73	-1.67 / -4.21
V_2	17.57	7.98 / 8.26	6.86 / 7.22	7.87 / 8.62
V_3	0.00	-7.64 / -4.18	-3.52 / -2.12	-6.97 / 0.60
V_4	0.00	0.15 / 0.92	1.67 / 2.03	-0.29 / 0.61
V_5	0.00	0.50 / 0.66	-1.94 / 1.34	0.56 / 1.37
V_6	0.00	0.27 / 0.13	1.60 / 1.64	0.10 / 0.32
			LLS solution	
V_1	0.00	-2.18 / -3.88	0.21 / -1.74	-
V_2	17.57	7.98 / 8.26	6.86 / 7.22	-
V_3	0.00	-7.64 / -4.18	-3.50 / -2.07	-
V_4	0.00	0.15 / 0.92	1.67 / 2.03	-
V_5	0.00	0.50 / 0.66	-1.94 / 1.34	-
V_6	0.00	0.27 / 0.13	1.60 / 1.64	-

Table 5.2: Dihedral force constants (kJ mol^{-1}) derived using the MM-relaxed/QM-relaxed approach. The fittings were performed using the ωB97X-D/6-31G* PES.

Through the analysis of the values of the final optimised parameters shown in Table 5.2, we conclude that the SLSQP SciPy optimiser and the LLS fitting

gave identical results in terms of accuracy. Since the LLS optimisation problem is always convex, it follows that LLS fitting always finds a global, although not necessarily unique, solution. Thus, this demonstrates that within the constraints of the functional form, the methodologies implemented in ParaMol are able to derive near-ideal parameters for small organic molecules. Moreover, the parameters did not stray far away from physically sensible values, which robustly indicates that the regularisation applied was strong enough to avoid non-physical force constants. Therefore, even though not applying regularisation may allow the optimisation procedure to reproduce small details in the dihedral profile correctly, we advocate for the use of regularisation, as it helps generate parameters that are more suited to be used in standard MM simulations. It is also worth noting that by imposing L2 regularisation, the optimised parameters depend on their initial guesses and, therefore, the results here presented might be potentially improved by starting the optimisations from better initial guesses. Nevertheless, since in many cases it is not straightforward to postulate good initial parameters, we decided to test the limiting case where $V_i = 0.0$ kJ mol^{-1} for all i, for which we demonstrated that even a blind guess led to improved FFs. Our experience suggests $V = 0.0$ kJ mol^{-1} is usually a good initial guess and, consequently, it is the one we recommend using by default in the absence of better ones.

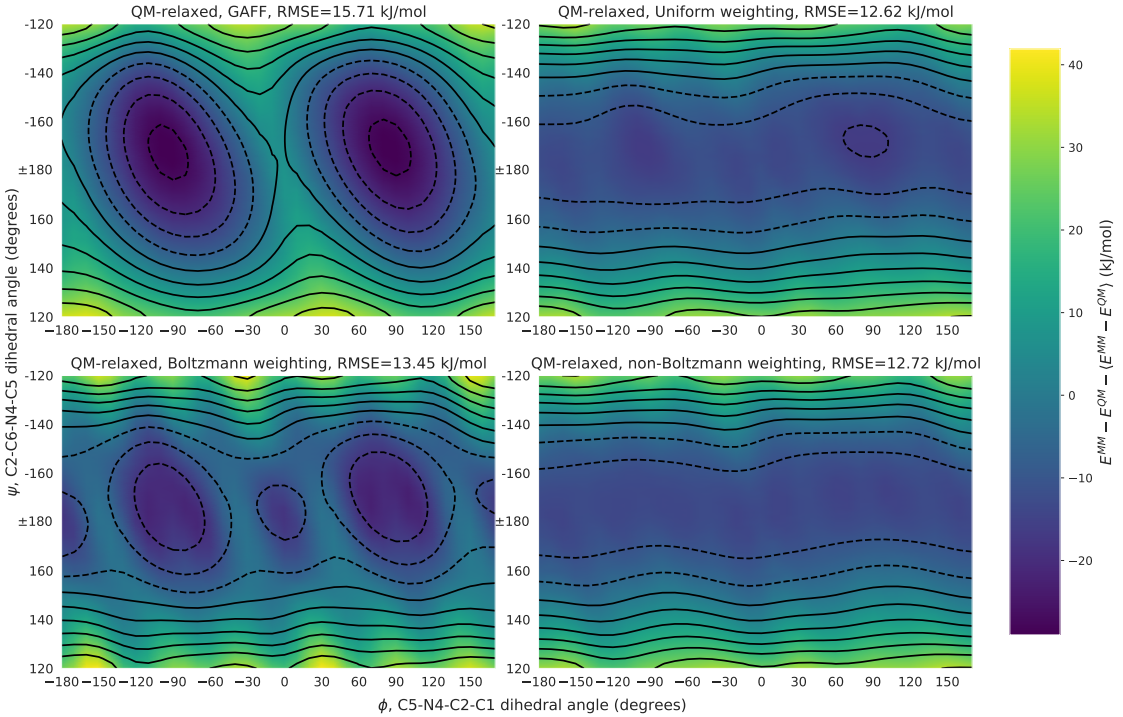

Figure 5.5: Relative errors of the MM FFs (GAFF, uniform, Boltzmann, and non-Boltzmann weightings) with respect to the target (ωB97X-D/6-31G*) PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) 2D dihedral scan. The QM-relaxed approach was employed to optimise the FFs.

Finally, with regards to the QM-relaxed approach, it is worth discussing the bias that this methodology introduces in the final derived parameters. A naive analysis of the obtained QM-relaxed energy profiles can lead us to consider them as correct: the fittings shown in Figure 5.5 exhibit a similar agreement with the target ωB97X-D/6-31G* PES as the ones obtained for the MM-relaxed approach, and the derived FF parameters are within physically sensible ranges (Table 5.2). Despite this, the artifacts introduced by this approach manifest themselves when MM-relaxed energy profiles are calculated using the QM-relaxed-derived FFs. We proceeded to perform this extra MM-relaxation of the QM-relaxed energy profiles, for which the results obtained are shown in Figure 5.6. Through the analysis of these plots, it can be seen that the QM-relaxed approach led to the creation of non-negligible artifacts in the PES as, *e.g.*, the spurious minima observed at *ca.* 0° and ±180°. Hence, since the QM-relaxed approach is critically dependent on the other intramolecular FF parameters,[261] it substantially biased the derived FF parameters and, therefore, we advocate against its use. The

artifacts arising from using the QM-relaxed approach are normally more serious the lower the dimensionality of the PES used in the fitting, as the chances of not covering substantial mismatches between the MM and QM levels increase.

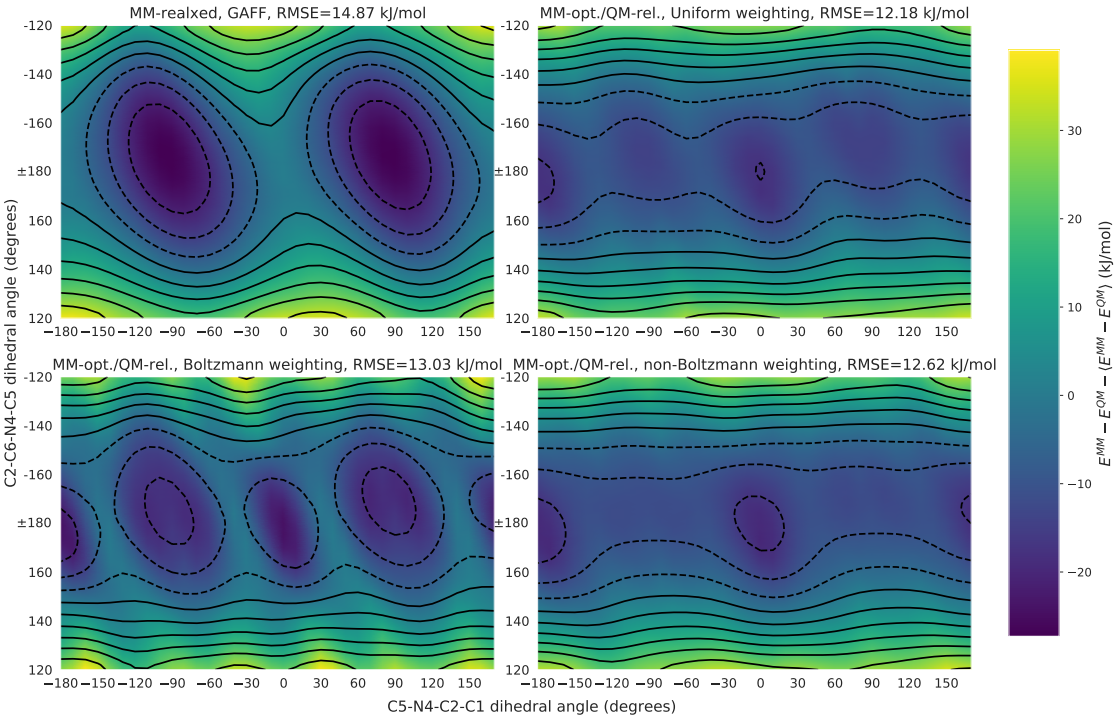

Figure 5.6: Relative errors of the MM FFs (GAFF, uniform, Boltzmann, and non-Boltzmann weightings) with respect to the target (ωB97X-D/6-31G*) PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) 2D dihedral scan. The MM PESs used to calculate the relative errors were obtained by MM optimisation of the QM-relaxed PESs used in Figure 5.5.

5.3.3 Parameterisation of aspirin

As a second example of the parameterisation methodologies implemented in ParaMol, we present and discuss the results obtained in the parameterisation of aspirin. We parameterised aspirin using both relaxed dihedral scans and a configurational ensemble generated by an MD simulation. The main aim of these parameterisation experiments was to reproduce the conformational preferences of aspirin at the SCC-DFTB-D3 level of theory.

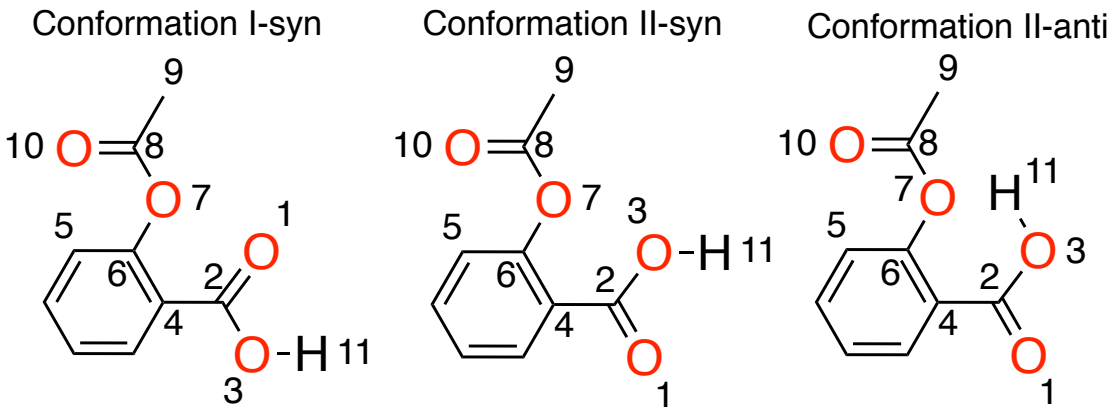

Figure 5.7: Molecular structures of aspirin. SCC-DFTB-D3 MD simulations sample mostly configurations that occur through rotation of the C6-O7 bond in conformations I-syn and II-syn. Moreover, even though conformation II-anti was not sampled in the SCC-DFTB-D3 MD simulation, it is shown here because it was sampled in some reparameterised FFs. In solution, the carboxylic acid of aspirin assumes predominantly its deprotonated state (pKa=3.49-3.6 at 25 °C[3]). Therefore, the results and discussions of this section concerning the II-syn and II-anti conformations are simply illustrations of the features of the parameterisation protocols employed, as these are mainly relevant in the gas phase.

As shown in Figure 5.8, the most striking difference between the populations predicted by the GAFF and SCC-DFTB-D3 is seen for the dihedrals involved in the rotation of the C6-O7 soft bond (C4-C6-O7-C8 and C5-C6-O7-C8). These dihedrals are originally modelled by the GAFF using only one dihedral term with periodicity $n = 2$, leading to the two minima observed in the dihedral distributions, in contrast with the four minima predicted by SCC-DFTB-D3. Hence, as the number of minima of the SCC-DFTB-D3 distributions do not match the number of minima predicted by the FF, we increased the flexibility of GAFF in our reparameterisation by including all terms in the Fourier expansion with periodicities from $n = 1$ up to $n = 4$. Furthermore, although dihedrals C4-C6-O7-C8 and C5-C6-O7-C8 share the same set of FF parameters, their SCC-DFTB-D3 dihedral populations are substantially different, implying that if we were to use a single potential to model both dihedrals, we would have to rely on the nonbonded terms to implicitly break their symmetry. This is, however, a clear example of the inability of the nonbonded terms of equation (4.25) to correctly model the intramolecular interactions that occur in aspirin, and especially the

weak hydrogen bond that can be formed between atoms O10 and H11, which is predominantly of electrostatic character since it occurs at distances of ca. 4.0 Å (see SCC-DFTB-D3 distribution in Figure 5.9).[283] This is concluded by examining the populations of the C5-C6-O7-C8 and C4-C6-O7-C8 dihedrals at *ca.* $\pm 130^{\circ}$ and $\pm 60^{\circ}$, respectively, for which the GAFF populations decay smoothly, not skewing towards the weakly hydrogen-bonded conformations that are present in the SCC-DFTB-D3 distribution.

The configurational distributions generated by simple reparameterisation of the original GAFF predicted wrong global minima and generally gave poor agreement with respect to the SCC-DFTB-D3 distribution (see Appendix A, Figure A.7). As a workaround for this issue, we decided to break the symmetry of the C5-C6-O7-C8 and C4-C6-O7-C8 dihedrals to artificially compensate for the limitations of the nonbonded (especially electrostatic) terms of the FF. In practice, this is equivalent to introducing a new atom type at, *e.g.*, position C4, as this carbon is linked to the carboxylic group and, consequently, its nature is very different from the C5 carbon atom. This breaking of symmetry makes it possible to independently optimise the parameters of each of these dihedrals, a step which proved to be essential to reproduce the target SCC-DFTB-D3 configurational distribution, as the simple augmentation of the number of dihedral terms was insufficient to do so. Lastly, for the two dihedral types involved in the rotation around the C8-C9 bond, GAFF assigns three terms with periodicities $n = 1, 2, 3$ to the O7-C8-C9-H dihedrals (o-c-c3-hc type), and one term with periodicity $n = 2$ to the O10-C8-C9-H dihedrals (os-c-c3-hc type). Despite this, we only assigned terms with periodicity $n = 3$ since this is a multiplicity not forbidden by symmetry. These dihedrals have a sp^3 carbon as one of the inner atoms (C1), which has three identical hydrogen substituents, and, therefore, all terms with multiplicity that is not a multiple of 3 vanish.[258]

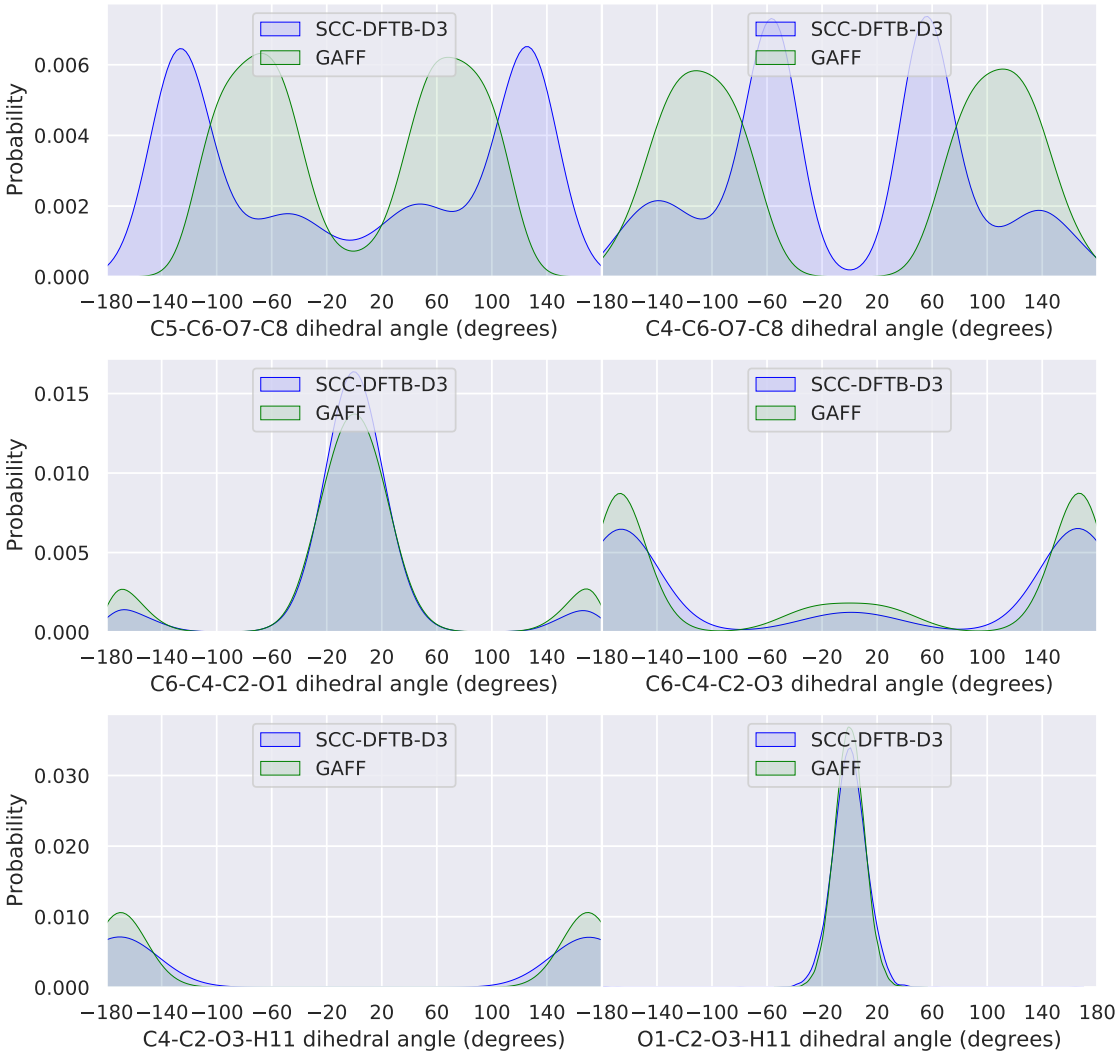

Figure 5.8: Kernel density estimations of the populations of the soft dihedrals of aspirin obtained from MD simulations using SCC-DFTB-D3 and the original GAFF. The soft dihedrals here presented (C5-C6-O7-C8, C4-C6-O7-C8, C6-C4-C2-O1, C6-C4-C2-O3, C4-C2-O3-H11, and O1-C2-O3-H11) are the ones for which parameters were optimised using relaxed dihedral scans.

5.3.3.1 Reparameterisation using a configurational ensemble

Regarding the reparameterisations performed using a SCC-DFTB-D3 configurational ensemble, these were designed to assess the performance of the weighting methods and the impact of the regularisation strength. To generate the SCC-DFTB-D3 configurational ensemble, we performed a gas-phase MD simulation using the DFTB+ package during 10 ns, in which snapshots were collected every 1 ps, resulting in a total of 10000 configurations. In this MD simulation, the

Nosé-Hoover thermostat was applied to maintain the temperature at 350 K with a coupling strength of 3400 cm^{-1}, a value close to the calculated highest vibrational frequency of the molecule (3670.75 cm^{-1}). We reparameterised all intramolecular parameters of aspirin, such that the vector of optimsable parameters was $\boldsymbol{p} = (\boldsymbol{K_b}, \boldsymbol{r_{eq}}, \boldsymbol{K_\theta}, \boldsymbol{\theta_{eq}}, \boldsymbol{V_n}, \boldsymbol{\gamma_n})$. In order to do this, we employed an objective function that fits both energies and forces through the use of the equations (5.5) and (5.3), respectively, plus the additional L2 regularisation term of equation (5.13). The total number of FF parameters concomitantly optimised was 108. The original GAFF parameters were used as initial guesses. After new FF parameters were derived, the optimised FFs were used to perform 100 ns of MD simulations, which were carried out using a Langevin integrator with a friction coefficient of 2 ps^{-1}, a time-step of 1 fs, and at a temperature of 350 K. Snapshots were collected every 10 ps, amounting for a total of 10000 snapshots for each simulation.

The configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle obtained using the non-Boltzmann weighting with a weighting temperature of 500 K is shown in Figure 5.9, and the distributions obtained for weighting temperatures of 300 K, 1000 K, 2000 K are shown in Appendix A, Figures A.8, A.9, and A.10, respectively. By examining the cases for which the weighting temperature was either 300 K or 500 K, we conclude that for strong regularisation strengths ($\alpha = 1$ and $\alpha = 0.1$), the optimised FFs reproduced quite well the general features of the SCC-DFTB-D3 distribution, as these FFs were able to sample the 4 minima that occur through rotation of the C6-O7 bond of conformation I-syn (outer edge of the distribution), as well as the states for which aspirin assumes the conformation II-syn (inner edge of the distribution). On the other hand, when intermediate regularisation strength ($\alpha = 0.01$) was employed, configurations in which aspirin assumes the II-anti conformation were sampled for the FF derived using a weighting temperature of 500 K, even though these configurations are not observed in the SCC-DFTB-D3 distribution. This spurious sampling was further aggravated when the weakest regularisation strength ($\alpha = 0.001$) was employed, for which the simulations became kinetically

trapped in these conformations, despite being started from the global minimum geometry (conformation I-syn with a C5-C6-O7-C8 dihedral angle value of *ca.* 130°).

The sampling of spurious conformations is a frequent issue when using reparameterised FFs in MM simulations, and it may occur whenever the spurious geometries are absent from the data set that was used to perform the fitting. Hence, in this case, since the fitting procedure had no information about the SCC-DFTB-D3 forces and energies of the II-anti conformations and transition states that lead to them, the barrier heights for the conversion of the carboxylic group from syn to anti were underestimated. A possible solution for this issue is to further reoptimise the FF, including the sampled spurious conformations so that the optimisation procedure also takes them into account. By doing this, it could be possible to prevent the oversampling of the spurious geometries, as the features of non-Boltzmann weighting lead to a tendency to overestimate barrier heights and/or equilibrium energies. Note, however, that this problem was not present for the strongly-regularised FFs, clearly indicating the importance of regularisation, which, by not allowing the FF parameters to stray away too much from physically sensible values, helps in preventing the creation of spurious minima.

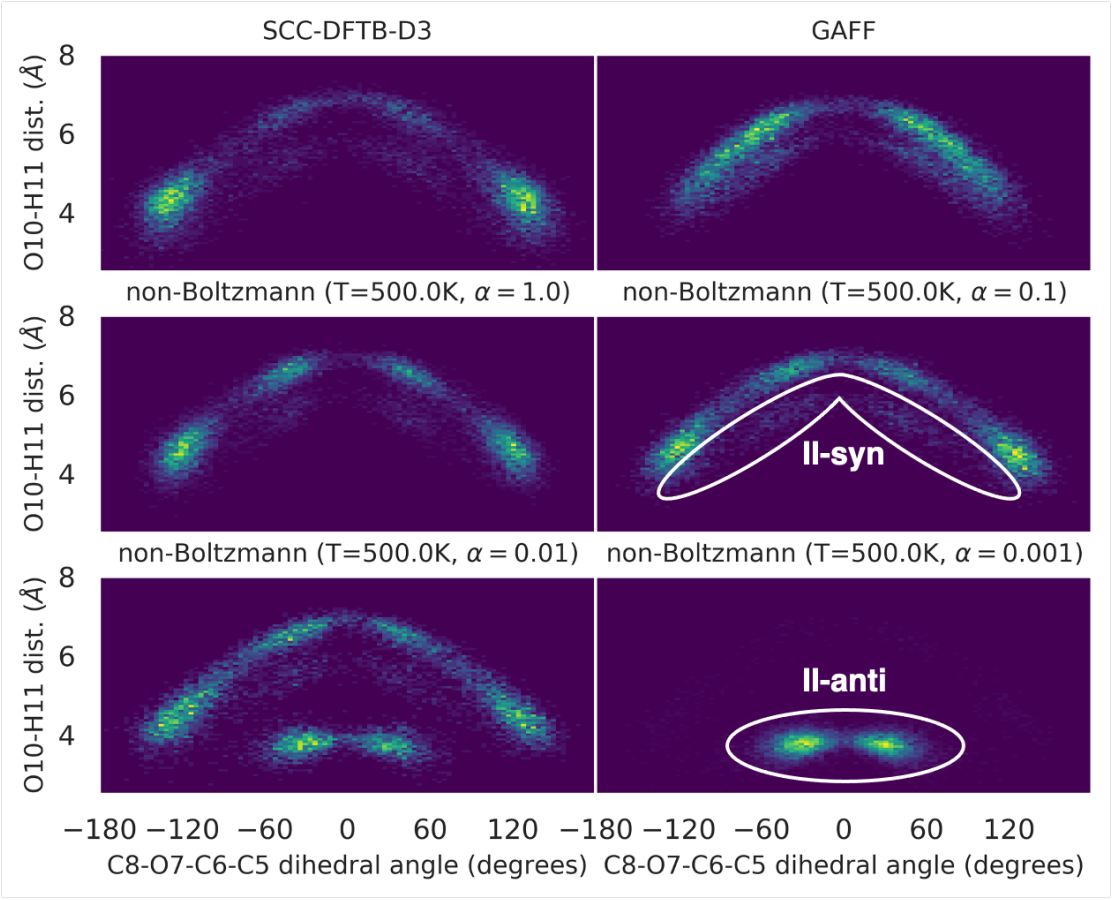

Figure 5.9: Configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle of aspirin obtained from MD simulations using SCC-DFTB-D3, the GAFF, and the GAFF.MOD (reparameterised) FFs. The latter were derived employing non-Boltzmann weighting, with a weighting temperature of 500 K and using different regularisation strengths ($\alpha = (1.0, 0.1, 0.01, 0.001)$). The data set used in the reparameterisation was the SCC-DFTB-D3 configurational ensemble. All represented distributions contain 10000 configurations.

The configurational distributions obtained using Boltzmann weighting with strong regularisation ($\alpha = 1.0$) and at different weighting temperatures are shown in Appendix A, Figure A.11. Through their analysis, we conclude that the overall agreement to the SCC-DFTB-D3 distribution was poor. The I-syn and II-syn minima were sampled (except for the FF derived with a weighting temperature of 500 K), but with incorrect frequency. Furthermore, it can also be seen that the distributions are highly asymmetric and show sampling of the spurious II-anti conformation, suggesting an overestimation of the barrier heights between the different minima and an underestimation of the syn-to-anti energy barrier. We do not show the configurational distributions of the FFs

derived with other regularisation strengths because they sampled unphysical configurations. This observation strongly indicates that Boltzmann weighting requires strong regularisation to produce FFs that can be potentially used in MM modelling. The reason for this may be attributed to the fact that Boltzmann weighting emphasises the description of the QM minima. Hence, if the regions of the PES that correspond to these QM minima are overfitted at the cost of poorly describing the remaining of the energy landscape, as soon as the molecule moves away from these minima regions, the PES becomes unphysical, ultimately leading to distorted geometries and wrong dynamics.

Finally, the configurational distributions obtained using uniform weighting with different regularisation strengths, $\alpha = (1.0, 0.1, 0.01, 0.001)$, are shown in Appendix A, Figure A.12. As for the Boltzmann weighting, the FFs derived using uniform weighting led to configurational distributions that poorly agree with the SCC-DFTB-D3 distribution. Furthermore, all FFs except the one derived using $\alpha = 1.0$ were kinetically trapped at the global minimum conformation from which the MD simulations were started. This indicates either overstabilisation of this minimum or overestimation of the transition states to which it is connected. The overstabilisation of the minimum is justified by the fact that uniform weighting equally allows for positive and negative $E^{MM} - E^{QM}$ values, which may lead to regions with negative $E^{MM} - E^{QM}$ values that have spuriously large thermodynamics weights. The asymmetries that might be imposed on the PES by equally allowing for positive and negative errors are an issue that was already reported and discussed by other authors.[263] On the other hand, overestimation of transition-state energies is likely to occur if the global minimum geometries are the most populated in the data set used in the fitting. This situation can lead the optimisation procedure to overfitting overrepresented configurations at the expense of misdescribing other configurations. Whenever this occurs, underrepresented configurations, such as transition states, are generally poorly described.

	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.001$
GAFF	53.24 / 140.95	53.24 / 140.95	53.24 / 140.95	53.24 / 140.95
non-Boltzmann (T=300 K)	44.03 / 94.68	41.81 / 87.29	40.61 / 83.25	39.75 / 84.81
non-Boltzmann (T=500 K)	43.85 / 93.87	41.73 / 87.15	40.38 / 82.62	39.60 / 83.12
non-Boltzmann (T=1000 K)	43.63 / 93.97	41.05 / 87.87	39.59 / 82.55	38.81 / 81.57
non-Boltzmann (T=2000 K)	43.61 / 95.13	40.39 / 88.91	38.61 / 83.38	37.75 / 81.20
Boltzmann (T=300 K)	49.78 / 124.33	47.27 / 110.25	46.88 / 109.13	48.66 / 123.05
Boltzmann (T=500 K)	44.73 / 124.62	38.97 / 110.61	37.15 / 113.44	36.81 / 123.93
Boltzmann (T=1000 K)	42.35 / 111.84	37.02 / 103.67	34.12 / 102.17	33.53 / 103.05
Boltzmann (T=2000 K)	42.29 / 107.81	37.12 / 100.01	33.95 / 95.15	32.92 / 95.45
Uniform	42.28 / 104.50	37.11 / 96.75	33.92 / 91.54	32.84 / 91.44

Table 5.3: RMSE of the energies (kJ mol^{-1}) / Average RMSE of the atomic force (kJ mol^{-1} Å^{-1} atom^{-1}). The RMSEs were calculated for the SCC-DFTB-D3 configurational ensemble data set, and they represent the energies and forces errors between the SCC-DFTB-D3 level of theory and the reparameterised FFs. The formula used to compute them is given by $RMSE(E) = \sqrt{\frac{\sum_i^{N_s}\left(E_i^{QM} - E_i^{MM} - \langle \Delta E \rangle\right)^2}{N_s}}$ with $\langle \Delta E \rangle = \frac{1}{N_s}\sum_i^{N_s}\left(E_i^{QM} - E_i^{MM}\right)$.

Overall, as a general guideline to follow when fitting to configurational ensembles, we recommend the use of non-Boltzmann weighting as this weighting scheme seems to be generally less sensitive to the regularisation strength and yielded the best performance in reproducing the SCC-DFTB-D3 distribution. Furthermore, strong regularisation ($\alpha = 1.0$ and $\alpha = 0.1$) seems to result in more reliable parameters than intermediate ($\alpha = 0.01$) and weak ($\alpha = 0.001$) regularisation, as a closer agreement to the SCC-DFTB-D3 distribution was obtained for the FFs derived using strong regularisation. Additionally, strong regularisation also prevents the FF parameters from deviating much from their original values, enabling the FF parameters to be kept within a range of physically sensible values.

Intermediate and weak regularisation strengths led to FFs that have generally lower energy RMSEs and lower average atomic force RMSEs (see Table 5.3) than their strongly-regularised counterparts (except when Boltzmann weighting is

employed with low weighting temperatures, as, in these situations, there is a tendency to overfit the QM minima). However, these putative better fittings came at the cost of creating artifacts in the PES, such as, *e.g.*, spurious minima. Finally, progressively employing higher weighting temperatures (1000 K and 2000 K) in the non-Boltzmann and Boltzmann schemes led to results that became gradually similar to the ones that were obtained upon using the uniform weighting scheme, as expected. Hence, since uniform weighting did not perform particularly well when using a configurational ensemble as the fitting data set, we do not recommend the use of high weighting temperatures. Therefore, unless one has a specific reason to do so, temperatures in a range between 300 K to 500 K are preferable.

5.3.3.2 Reparameterisation using dihedral scans

Let us now turn our discussion to the results obtained by reparameterising the soft dihedrals of aspirin using relaxed dihedral scans. The dihedral energy profiles of aspirin were created from 36-point one-dimensional relaxed scans in which each point was spaced by 10°. These dihedrals scans were performed for all soft dihedrals except those associated with the rotation of the methyl group (the QM energy profile of such dihedrals generally can be reproduced reasonably by the GAFF[251]) and those involved in the rotation of the O7-C8 bond, as it is fairly rigid (see Figure 5.7). Specifically, the soft dihedrals scanned were C5-C6-O7-C8, C6-C4-C2-O1, C6-C4-C2-O3, C4-C2-O3-H11, and O1-C2-O3-H11. Furthermore, for optimisation purposes, the C4-C6-O7-C8 dihedral was also included, as we broke the symmetry of the dihedrals involved in the rotation of the C6-O7 bond. All geometry optimisations were performed while fixing only the dihedral being scanned, and they were deemed to be converged when the force on all atoms was less 1×10^{-2} eV Å^{-1}. The reparameterisations were performed using the MM-relaxed approach of equation (5.7) with strong regularisation ($\alpha = 1.0$) and with a weighting temperature of 500 K. The vector of optimisable parameters that entered in the optimisation was $\boldsymbol{p} = (\boldsymbol{V_n}, \boldsymbol{\gamma_n})$, in

which any pair of V_n and γ_n belongs to a term of the previously mentioned soft dihedrals. The 26 dihedral parameters were all optimised concomitantly. The fittings obtained when employing Boltzmann weighting are shown in Figure 5.10, and the fittings for the non-Boltzmann and uniform weighting methods are shown in Appendix A, Figures A.13 and A.14.

Through the analysis of these results, we conclude that most of the improvement in the fittings occurred for the C5-C6-O7-C8 dihedral, an observation that supports the argument that the main source of the mismatch seen between the GAFF and the SCC-DFTB-D3 distributions comes from the soft dihedrals that model the rotation about the C6-O7 bond. For the remaining dihedrals, modest improvements or even slight worsening were obtained. The latter situation occurs because the optimisation procedure may sacrifice some accuracy in specific dihedrals to obtain a better global agreement. Furthermore, through the analysis of the configurational distribution represented in Figure 5.11, we conclude that, independently of the weighting scheme applied, the agreement obtained to the target distribution was quite good. It is also interesting to notice the sampling, even though very rarely, of the II-anti conformation and, surprisingly, of the I-anti conformation, which was visited even less often, when using non-Boltzmann and uniform weighting.

Overall, all weighting schemes performed similarly when fitting was performed using dihedral scans. Nevertheless, as a general guideline, we recommend the use of non-Boltzmann weighting due to its features. This recommendation, however, is less strict than that previously made for the fitting performed using configurational ensembles. Furthermore, with regards to the regularisation strength, our experience indicates that, in most cases, this reparameterisation approach requires strong-to-intermediate regularisation strengths ($\alpha = 1.0$ or $\alpha = 0.1$), as attempts to use weaker regularisation strengths resulted, in general, in unstable FFs that tend to be unsuitable for MD simulations.

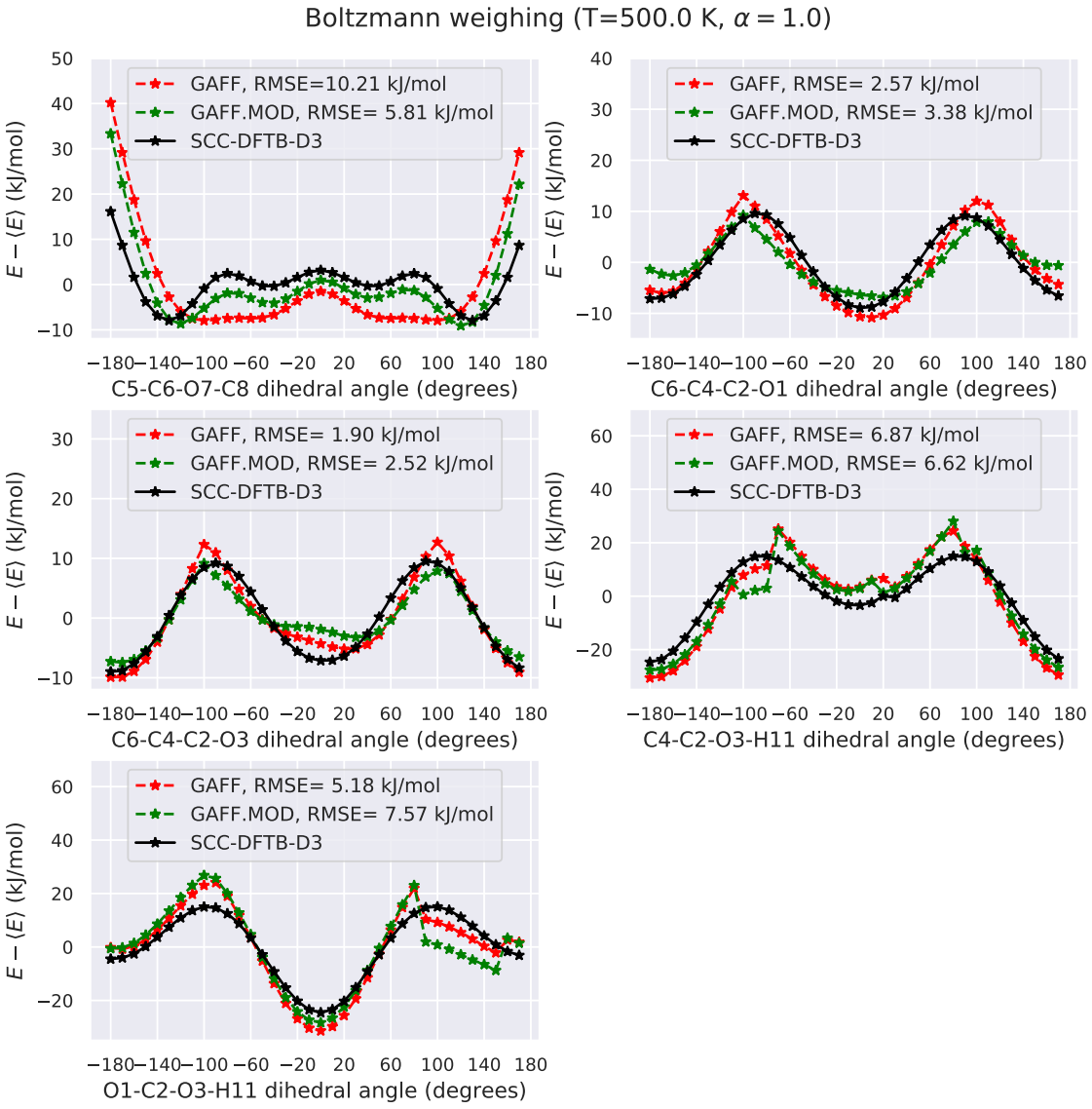

Figure 5.10: Comparison of the SCC-DFTB-D3, GAFF, and GAFF.MOD (reparameterised FF) dihedral energy profiles for the C5-C6-O7-C8, C6-C4-C2-O1, C6-C4-C2-O3, C4-C2-O3-H11, and O1-C2-O3-H11 dihedral angles. The GAFF curves correspond to MM-relaxed energy profiles. The GAFF.MOD FF was obtained by employing the MM-relaxed approach with Boltzmann weighting (T=500.0 K, $\alpha = 1.0$). The parameters of the dihedrals represented in this Figure were concomitantly optimised along those of the C4-C6-O7-C8 dihedral using the ParaMol's automatic soft dihedral parameterisation task.

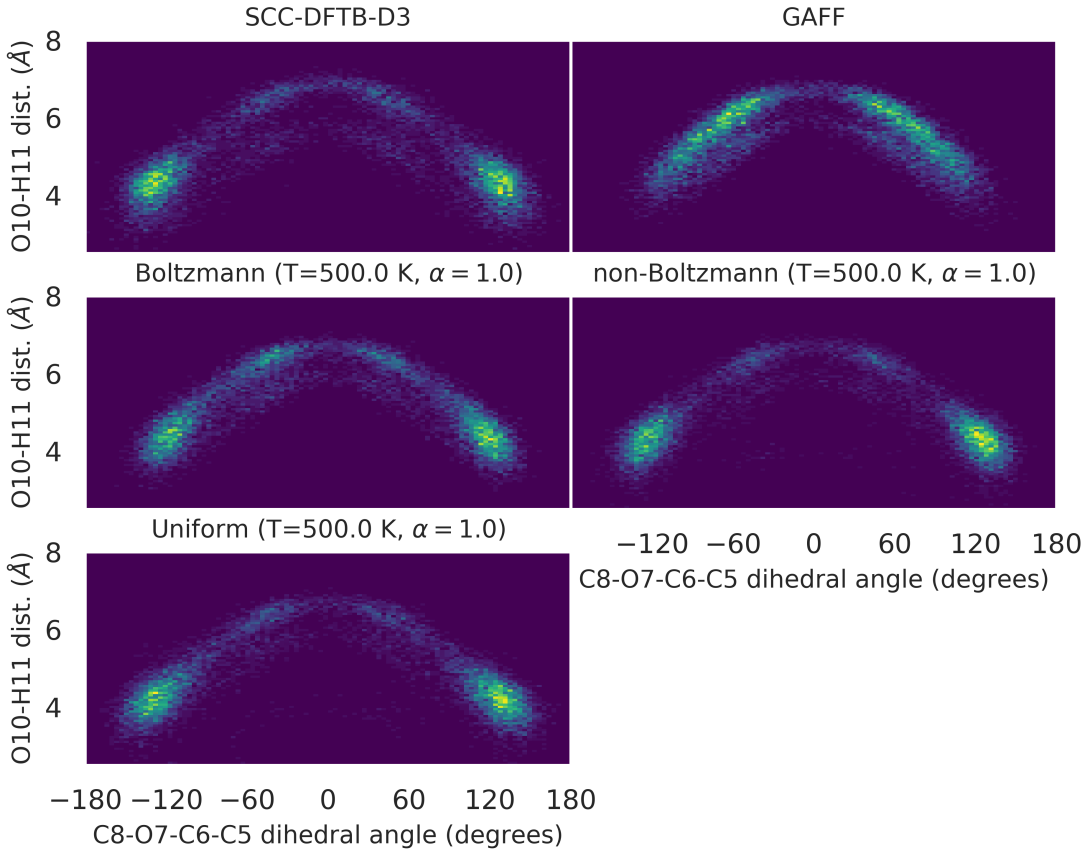

Figure 5.11: Configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle of aspirin obtained from MD simulations using SCC-DFTB-D3, the GAFF, and the GAFF.MOD (reparameterised) FFs. The latter were derived through reparameterisation of the soft dihedrals employing Boltzmann (Figure 5.10), non-Boltzmann, and uniform weighting methods (Appendix A, Figures A.13 and A.14, respectively), with a weighting temperature of 500 K and a regularisation strength of $\alpha = 1.0$. All represented distributions contain 10000 configurations.

Lastly, we also attempted to reproduce the SCC-DFTB-D3 distribution by simply reparameterising the C5-C6-O7-C8 and C4-C6-O7-C8 dihedrals, *i.e.*, the soft dihedrals that model the rotation around the C6-O7 bond, as these dihedrals are the main source of the mismatch seen between the GAFF and the SCC-DFTB-D3 distributions. These reparameterisations were performed using the MM-relaxed approach of equation (5.7) with strong regularisation ($\alpha = 1.0$) and with a weighting temperature of 500 K. As seen in the dihedral energy profiles of Figure 5.12, the dihedral energy profiles of the reparameterised FFs are in excellent agreement with SCC-DFTB-D3. Reparameterisation led to a decrease in the

energy RMSE from 10.21 kJ mol^{-1} (GAFF) to 3.19 kJ mol^{-1} (Boltzmann weighting), 3.24 kJ mol^{-1} (non-Boltzmann weighting), and 1.29 kJ mol^{-1} (uniform weighting). Furthermore, through the analysis of the configurational distribution represented in Figure 5.13, we conclude that regardless of the weighting scheme applied, the agreement to the SCC-DFTB-D3 distribution is quite good. Despite this, a small underrepresentation of the configurations of the II-syn conformation is seen, as well as very rare sampling of the II-anti conformation for the Boltzmann-weighted FF. All in all, parameterisation of the dihedrals associated with the main faulty soft bond proved to be an efficient route to correctly model the conformational dynamics of aspirin, especially given that it is computationally cheap and all weighting schemes performed similarly.

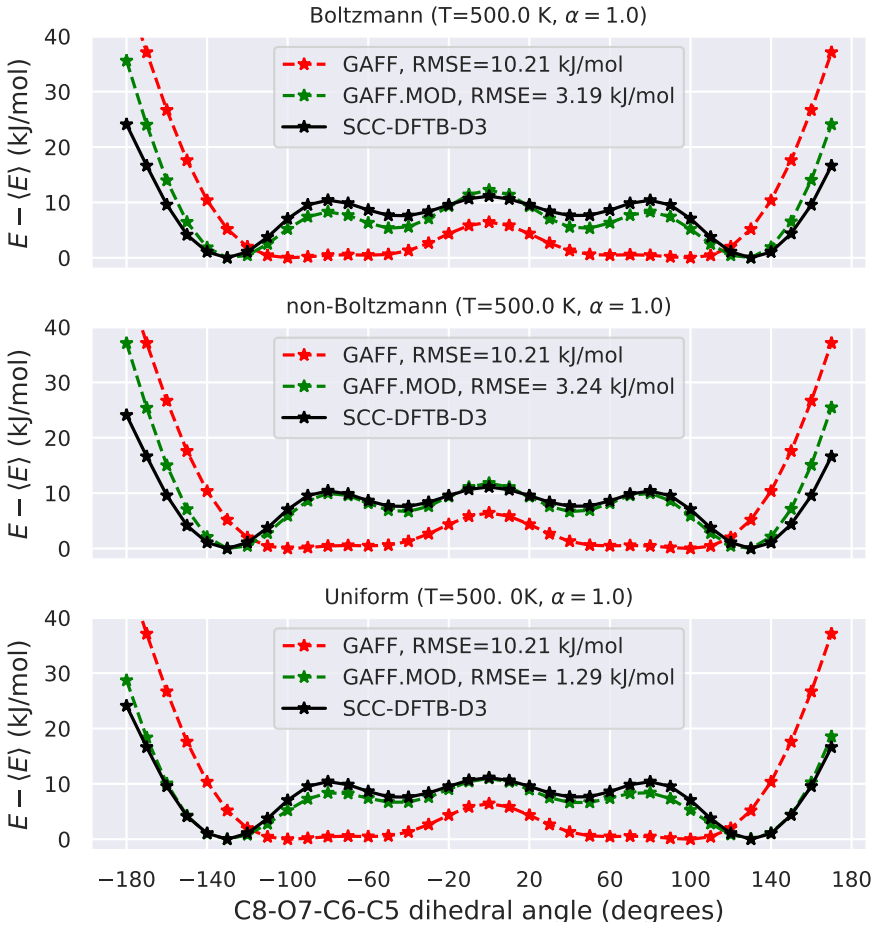

Figure 5.12: Comparison of the SCC-DFTB-D3, GAFF, and GAFF.MOD (reparameterised FF) energy profiles of the C5-C6-O7-C8 dihedral. The GAFF curves correspond to MM-relaxed energy profiles. The GAFF.MOD FF was obtained by employing the MM-relaxed approach to optimise the parameters of the C5-C6-O7-C8 dihedral. The weighting methods and regularisation strength used are indicated on top of each plot.

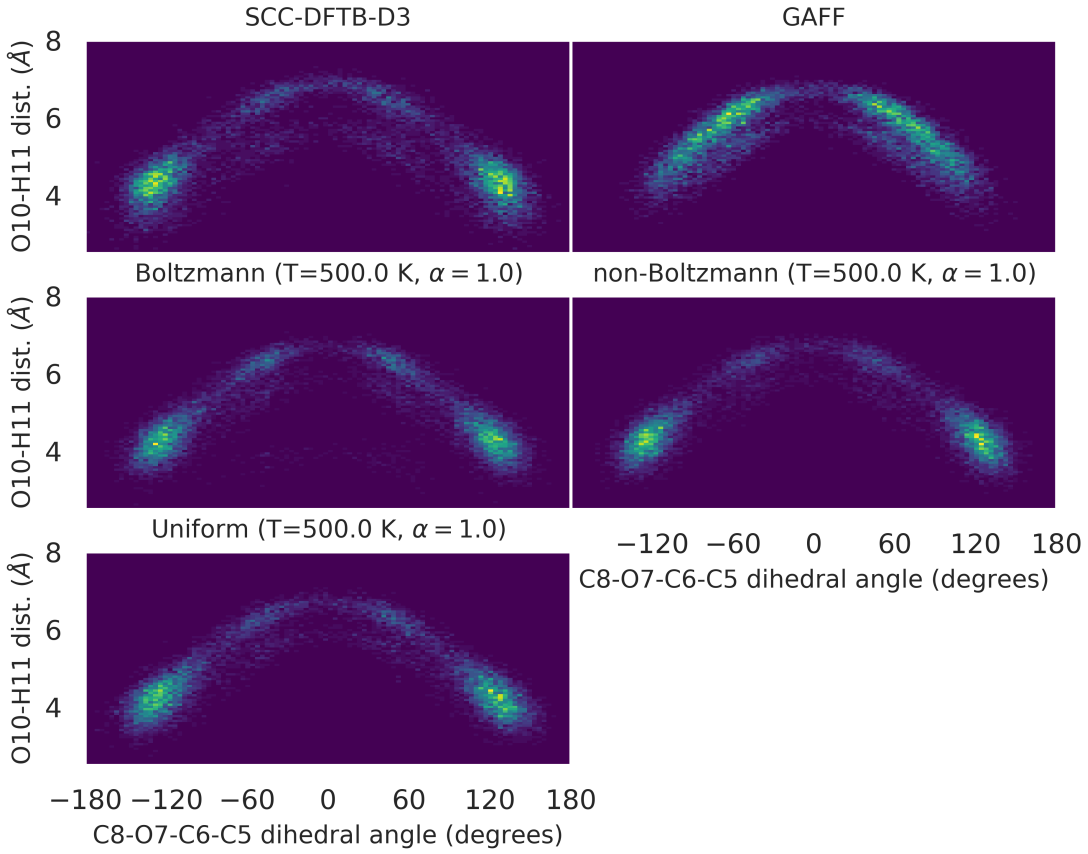

Figure 5.13: Configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle of aspirin obtained from MD simulations using SCC-DFTB-D3, the GAFF, and the GAFF.MOD FFs of Figure 5.12. The latter were obtained by reparameterisation of the dihedrals associated with the main faulty soft bond of aspirin (see Figure 5.12). All represented distributions contain 10000 configurations.

5.3.4 Adaptive parameterisation of caffeine

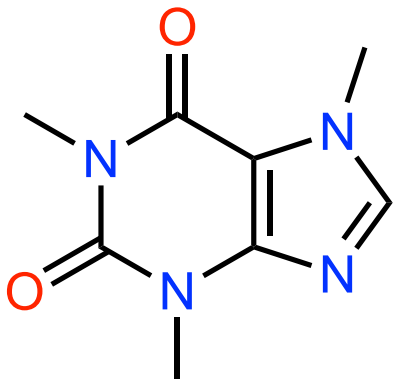

Figure 5.14: Molecular structure of caffeine.

As a last illustrative example of ParaMol's parameterisation capabilities, we reparameterised all intramolecular parameters of caffeine (Figure 5.14) to the VV10 level of theory using adaptive parameterisation (the SCC-DFTB-D3 results are shown in Appendix A, Figures A.15, A.16, and A.17). Specifically, the vector of parameters that entered in the optimisation was $\boldsymbol{p} = (\boldsymbol{K_b}, \boldsymbol{r_{eq}}, \boldsymbol{K_\theta}, \boldsymbol{\theta_{eq}}, \boldsymbol{V_n}, \boldsymbol{\gamma_n})$. The total number of optimisable parameters was 156. The minimised objective function included an energy, force, and regularisation terms, as given by equations (5.3), (5.5), and (5.13).

In every iteration of the adaptive parameterisation procedure, 100 new configurations separated 0.5 ps from each other were generated and added to the previous ones. These configurations were obtained using Langevin dynamics with a friction coefficient of 2 ps^{-1}, a time-step of 1 fs, and a temperature of 300 K. No special sampling technique was employed to explore the PES of caffeine since it is mostly planar and does not have much conformational flexibility. The adaptive parameterisation procedure was deemed to be converged when the RMSD of the parameters between two successive iterations was less than 10^{-4}. The adaptive parameterisation performed 23 iterations in total until convergence, corresponding to a total of 2300 structures in the last iteration.

The plots of the RMSD of the parameters and the components of the objective function as a function of the iteration number are shown in Figure 5.15. Through their analysis, we can see that most of the improvement in the objective function occurred in the first 2-3 iterations, as shown by the θ_{L2} regularisation term, which remained practically steady afterwards, indicating that only small adjustments to the FF parameters occurred. After this substantial initial refinement, the steady increase of the X_E energy term may be attributed to the convergence of the relative populations of the configurational ensemble used in the parameterisation, as the initial and final X_E values were practically the same. Interestingly, this steady increase is not seen in the X_F term, suggesting that X_F was less sensitive to the completeness of the configurational ensemble. Furthermore, not much variation in any term occurred after the 15th iteration, as can be seen through the stabilisation of the objective function terms, which is a robust indication that,

at this point, both the sampling and the parameter optimisation were practically converged.

Finally, to evaluate the improvement of the energies and forces, we generated two testings data sets composed of 1000 configurations using either the FF before (GAFF) and after reparameterisation (GAFF.MOD). This was done by performing short MD simulations with the same settings as before (snapshots were collected every picosecond). Since each data set was generated by sampling from its respective FF, this analysis evaluates how close to the target level of theory each FF samples. The plot that shows the correlation between the QM energies and the MM energies is represented in Figure 5.16, and the atomic forces errors are shown in the molecular structures of Figure 5.17. The RMSE of the energies before and after reparameterisation was 12.82 kJ mol^{-1} and 6.73 kJ mol^{-1}, respectively, which reveals that GAFF.MOD samples conformations that are closer in energy to the VV10 level of theory. On other hand, the average RMSE of the atomic forces improved from 83.61 kJ mol^{-1} $Å^{-1}$ $atom^{-1}$ to 50.95 kJ mol^{-1} $Å^{-1}$ $atom^{-1}$ after reparameterisation, a clear indication that GAFF.MOD is an improved FF in relation to GAFF since it also predicts better forces.

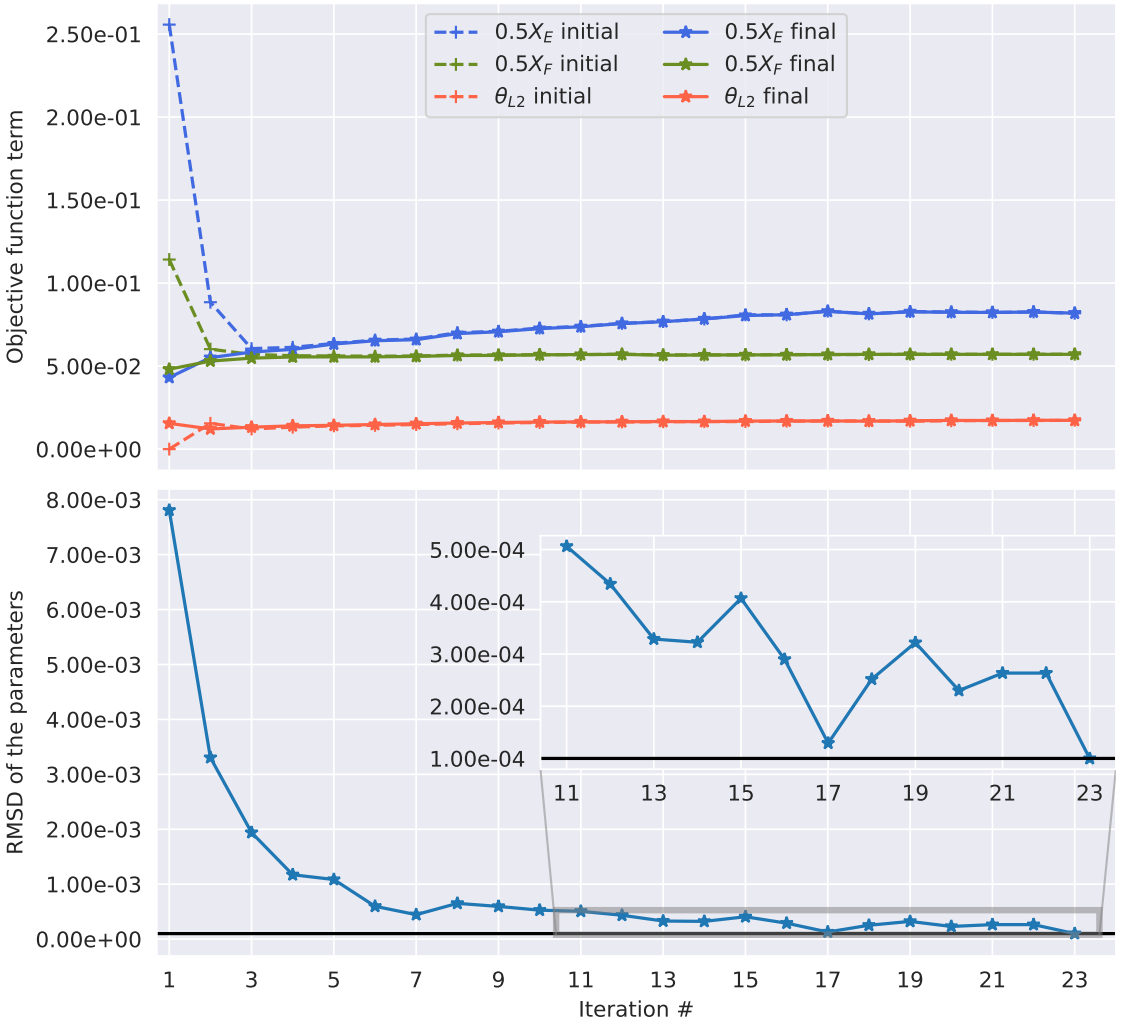

Figure 5.15: Top panel: Plot of the values of each term included in the objective function at the beginning (dashed lines) and end (solid lines) of each iteration. X_E corresponds to the energy term, X_F to the forces term, and θ_{L2} to the regularisation term. Bottom panel: Plot of the RMSD of the parameters as a function of the iteration number.

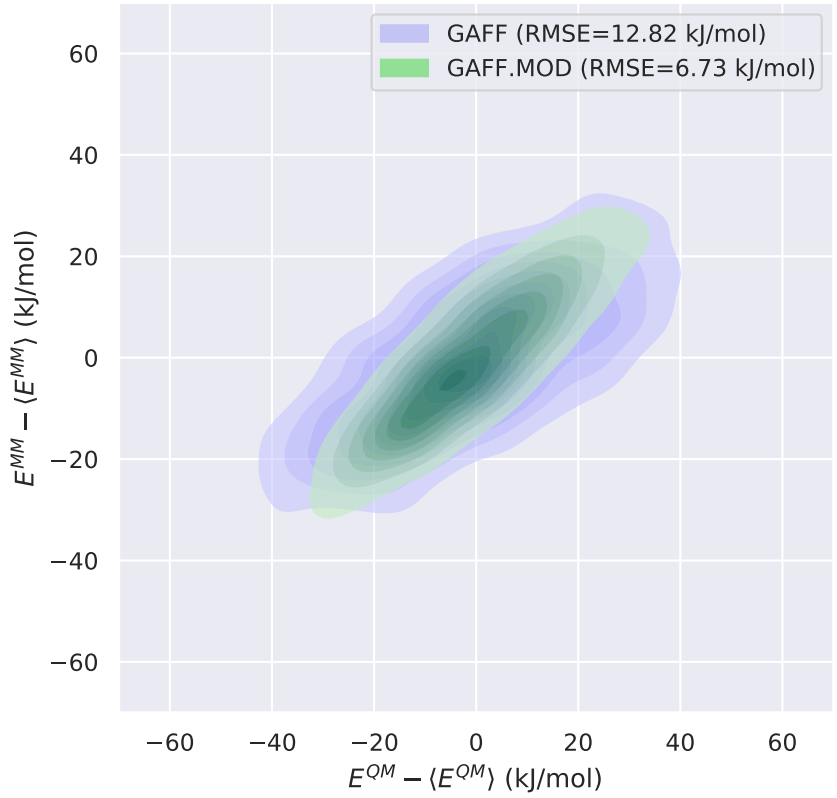

Figure 5.16: Correlation between the QM energies and the MM energies of caffeine before (GAFF) and after (GAFF.MOD) the adaptive reparameterisation procedure. Each data sets consists of 1000 configurations generated through a short MD simulation that used the respective FF.

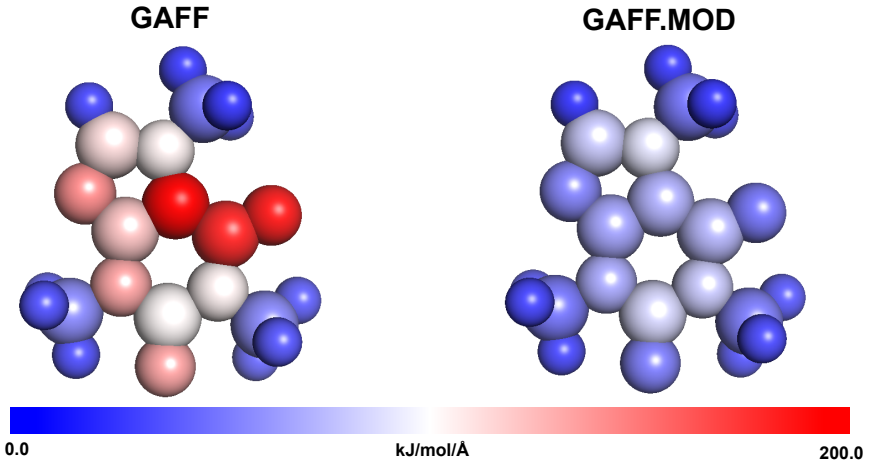

Figure 5.17: Atomic force errors before (GAFF, left) and after (GAFF.MOD, right) reparameterisation, calculated using $RMSE(F_j) = \sqrt{\frac{\sum_{i=1}^{3} \left| F_{i,j}^{QM} - F_{i,j}^{MM} \right|^2}{3}}$. The average RMSE of the atomic forces improved from 83.61 kJ mol^{-1} Å^{-1} atom^{-1} (GAFF) to 50.95 kJ mol^{-1} Å^{-1} atom^{-1} after reparameterisation (GAFF.MOD).

5.4 Conclusions

In this chapter, we have presented ParaMol, software that has the capability of reparameterising class I force field with a special focus on druglike molecules. As explained and demonstrated in the examples provided, ParaMol has many automated capabilities that allow the reparameterisation of molecules through the use of different protocols. Its application may have implications in different areas of chemistry with biological relevance that require FFs with high levels of accuracy. The results obtained demonstrate that, within the constraints of the functional form, the methodologies implemented in ParaMol are able to derive near-ideal parameters for small organic molecules.

We demonstrated that the use of MM-relaxed dihedral scans is a robust way to reparameterise the parameters of dihedrals and that this methodology is not very sensitive to the weighting method, although it requires strong-to-intermediate regularisation strengths. On the other hand, since fittings to QM-relaxed dihedrals scans are critically dependent on the intramolecular FF parameters, these fittings substantially bias the derived FF parameters and, therefore, the use of QM-relaxed dihedral scans should be avoided. Furthermore, configurational ensembles generated through standard MM simulation methods may also be used as parameterisation data sets, even though they make the optimisations more sensitive to the weighting method. In this context, the best results were obtained when using non-Boltzmann weighting, which proved to be the most reliable weighting scheme, despite its tendency to overestimate transition-state energies and underestimate fluctuations. Moreover, Boltzmann weighting, which emphasises the description of QM minima, tends to overfit low energy regions of the PES at the cost of poorly describing the remainder of the energy landscape. Hence, it requires strong regularisation to produce FFs that can be potentially used in MM modelling. Finally, since uniform weighting allows for positive and negative $E^{MM} - E^{QM}$ values, it is prone to the creation of asymmetries in the PES, which often lead to spurious minima due to artificially large thermodynamics weights and poor description of underrepresented configurations (*e.g.*,

transition states). As in Boltzmann weighting, uniform weighting also requires strong regularisation to mitigate some of these undesirable features.

Sampling of spurious conformations is a common issue that arises when reparameterising FFs, and it may occur whenever weak regularisation is employed or the spurious geometries are absent in the data set used to perform the fitting. A possible solution for this issue is to further reoptimise the FFs including the spurious conformations so that the optimisation procedure has information about them. Owing to the features of non-Boltzmann weighting, it is the indicated method to apply in these situations, as it tends to overestimate barrier heights and/or equilibrium energies, which are features that, ultimately, prevent the oversampling of "artificial" geometries.

When using configurational ensembles as parameterisation data sets, temperatures in a range between 300 and 500 K should be applied if using Boltzmann or non-Boltzmann weighting, as progressively employing higher temperatures leads to results that become gradually similar to the ones that are obtained when using uniform weighting (which does not perform particularly well in this setting). Alternatively, it is also possible to resort to ParaMol's soft dihedral parameterisation task, which identifies and concomitantly parameterises all dihedrals associated with the rotatable bonds of a molecule. This method has a computational cost significantly lower than the configurational ensemble approach, whilst inheriting all features implicit to dihedral scans. Finally, adaptive parameterisation is also an attractive and useful way to optimise parameters, as it combines self-consistent sampling and parameter optimisation in a single protocol.

In general, most of the parameterisation routines implemented in ParaMol can be performed automatically. However, care has to be taken when performing parameterisations using a non-linear iterative optimiser at the expense of the LLS fitting approach, as the former may become trapped in local minima, whereas the latter is deterministic and ensures obtaining the global minimum. Consequently, whenever possible and suitable, the LLS solution is preferred. Moreover, manual

quality checks may be required to identify poor data and outliers in the data set used in the parameterisation, which is of particular importance since most of the FF optimisation problems arise as a result of low-quality fitting data.

Owing to its potential, we suggest that ParaMol can be introduced as a routine step in the protocol normally employed to parameterise drug molecules for MM simulations. The software is licensed under the MIT open source license. The code is available at GitHub at `https://github.com/JMorado/ParaMol`, and the documentation can be found at `https://paramol.readthedocs.io`.

In the next chapter, we introduce a multilevel MC method that allows quantum configurational ensembles to be generated while keeping the computational cost at a minimum. We present the theory and algorithm of the methodology and apply it to a set of relevant druglike molecules. We show that FF reparameterisation is an efficient way to accelerate the QM-level sampling and discuss the implications and features of the method. As more advanced applications, we apply the nMC-MC algorithm to generate the QM/MM distribution of a ligand in aqueous solution and present a self-parameterising version of the method, which combines sampling and FF parameterisation in one scheme.

Chapter 6

On the Generation of Quantum Configurational Ensembles Using Approximate Potentials

Conformational analysis is of paramount importance in drug design: it is crucial to determine pharmacological properties, understand molecular recognition processes, and characterise the conformations of ligands when unbound. MM simulation methods, such as MC and MD, are usually employed to generate ensembles of structures due to their ability to extensively sample the conformational space of molecules. The accuracy of these MM-based schemes strongly depends on the functional form of the FF and its parameterisation, components that often hinder their performance. High-level methods, such as *ab initio* MD, provide reliable structural information but are still too computationally expensive to allow for extensive sampling. Therefore, to overcome these limitations, in this chapter we present a multilevel MC method that is capable of generating quantum configurational ensembles while keeping the computational cost at a minimum. We show that FF reparameterisation is an efficient route to generate FFs that reproduce QM results more closely, which in turn can be used as low-cost models to achieve the gold standard QM accuracy. We demonstrate that the MC acceptance rate is strongly correlated with various phase space overlap

measurements and that it constitutes a robust metric to evaluate the similarity between the MM and QM levels of theory. As a more advanced applications, we present a self-parameterising version of the algorithm, which combines sampling and FF parameterisation in one scheme, and adapt the MC algorithm to generate the QM/MM distribution of a ligand in aqueous solution.

This chapter has been published as an article in the Journal of Chemical Theory and Computation:

- Morado, J.; Mortenson, P. N.; Nissink, J. W. M.; Verdonk, M. L.; Ward, R. A.; Essex, J. W.; Skylaris, C.-K. Generation of Quantum Configurational Ensembles Using Approximate Potentials. J. Chem. Theory Comput. 2021, 17 (11), 7021–7042. https://doi.org/10.1021/acs.jctc.1c00532

6.1 Introduction

The study of the conformational dynamics of molecules free in solution is essential for predicting molecular properties and to guide the rational development of new pharmaceutical compounds. The latter application is of utmost importance for the pharmaceutical industry since knowledge of the unbound state is vital to understand the fundamentals of molecular recognition.[12,19–22] Besides the displacement of water from protein binding sites,[23–25] one of the main phenomena that impacts binding affinity is the reorganisation of the unbound state ligand upon binding to its target, a process that is influenced by the change in intramolecular energy of the ligand in adopting the bioactive conformer, as well as the associated loss of entropy.[22] Minimisation of the free energy penalty associated with this structural change is vital to optimising ligand potency, requiring knowledge of the physical interactions that control conformational preferences and methods for conformational analysis if a rational strategy is to be employed.[22] There is a wide range of experimental structural information on pharmaceutical compounds bound to their protein targets.[26,27] However, as it has

been emphasised in various studies, the conformations of unbound compounds are still poorly characterised.[12,20,21,28] Therefore, the scientific community must put effort into developing tools that allow fast and reliable characterisation of unbound molecular conformers as these can potentially provide the so-called "missing link" in structure-based drug discovery.[20,28]

The most widely used experimental method to elucidate unbound conformational ensembles is NMR spectroscopy, which is often utilised in drug design to complement X-ray protein–ligand structural information.[22,28,284] Additionally, MM simulation methods, such as MD and MC, are also commonly employed to predict thermophysical molecular properties and generate structures for conformational analysis.[7,21] In particular, static properties, such as, *e.g.*, optical spectra, NMR spectra, and solvation free energies, can be determined from the relative populations of the free state conformers,[285] which are usually possible to estimate in MM-based simulations since, in many instances, these permit ergodic sampling.[29–34] Although these methods allow extensive sampling of the configurational space of molecules, the functional form used by the FF affects the sampling quality, and parameterisation must be adequate to ensure accurate results. High-level simulation schemes, such as, *e.g.*, *ab initio* MD, have become the gold standard for simulation purposes as they provide reliable structural information at the quantum level, but are still too computationally expensive to allow achieving the time scales typically required for convergence of the simulations.[286–290] Hence, to attain extensive and reliable sampling, it is necessary to find a compromise between the efficiency of the MM-based methods and the accuracy of the QM level of theory.

Several approaches have already been proposed to sample molecular conformations with QM accuracy at a nearly MM cost. In this context, Rosa *et al.*[285] proposed a postprocessing method in which, through the use of conformational clustering and thermodynamic perturbation theory, it is possible to estimate the QM populations by correcting MM populations. Others have attempted to explore the conformational landscape of bioactive small molecules by using a combination of classical Hamiltonian replica exchange with high-level QM

calculations.[291] In this chapter, we attempt to bridge the gap between the MM efficiency and the QM accuracy by presenting a methodology that is based on an *ab initio* MC algorithm. This approach enables recovery of the correct QM ensembles while keeping the computational cost at a minimum. It is also capable of self-parameterising FFs to a target level of theory in an iterative and on-the-fly fashion, a feature that can be applied whenever generation of high-quality FFs is required.

The method we propose in this chapter consists in a nested Markov chain Monte Carlo (nMC-MC) algorithm that combines sampling at the MM level with periodic switching attempts to the QM level. This nMC-MC algorithm works by first resorting to the hybrid Monte Carlo (hMC) scheme to rigorously generate configurations that belong to a target MM ensemble.[292–294] These are subsequently used as trial states for a second Markov chain, in which they are accepted or rejected according to a correction step based on the difference between the MM and QM potentials.[295,296] In this way, it is possible to generate quantum configurational ensembles using approximate potentials (*e.g.*, MM FFs). This multilevel *ab initio* MC algorithm has already been applied in various contexts, such as in the fitting of FF dihedral angles,[297] in a "stepping stone" approach for obtaining quantum free energies of hydration,[298] and in a MC resampling approach for the calculation of hybrid classical/quantum free energies;[299] to improve the efficiency of Born models in MC simulations,[300] to model reactivity in small molecules,[301] and to enhance the conformational sampling of disordered regions of proteins.[302,303]

It is widely known that the key factor for convergence of multilevel approaches is ensuring a favourable overlap between the energy distributions of different levels of theory.[248,304,305] Otherwise, as FFs often predict conformations and energies that substantially deviate from the QM level, low acceptance rates are obtained when attempting to sample from the MM to the QM chain. Ultimately, the mismatch between the MM and QM descriptions becomes a bottleneck because it prevents a thorough exploration of relevant regions of the QM PES, slowing convergence of the sampling of the target quantum configurational

distributions. Different strategies have already been proposed to improve the overlap between the probability distributions of the energies associated with different levels of theory.[304] One possibility is to artificially broaden the MM distribution by manipulating the thermodynamic variables (*e.g.*, pressure and temperature) characterising the reference system[306–309] or using Tsallis statistics.[310] It is also possible to introduce an intermediate level of theory (*e.g.*, a semiempirical QM method) to bridge the gap between the MM and QM chains or employ an arbitrary number of intermediate potential energy layers with sufficient overlap between their probability distributions.[311] Another option is to increase the overlap between the distributions by improving the MM description so that it becomes more QM-like, which is the approach followed in this study. This increased overlap can be achieved through either FF reparameterisation (typically by force-matching[231,248,305]), by using ML potentials,[312,313] or through fitting of *ad hoc* potentials.[306,314–316]

As application examples, we tested the proposed methodology on a set of small organic molecules of increasing complexity, which are representative fragments of molecules found in drug discovery programs. Specifically, we attempted to generate quantum configurational ensembles of aniline, acetanilide, biphenyl, diphenyl ether, and sulfanilamide. As a relevant druglike example, we investigated a fragment of cpd 26, which is the core of an efficacious low nM antagonist of the inhibitor of apoptosis proteins cIAP1 and XIAP.[15] Furthermore, as proof of principle, we used octahydrotetracene to demonstrate that the nMC-MC algorithm can be coupled with a reparameterisation step, allowing for iterative optimisation of the molecule's FF parameters using the on-the-fly QM-generated ensemble. This self-parameterisation nMC-MC algorithm is similar in philosophy to the methods presented in some past applications,[312,314–317] though these studies did not use MM FFs or druglike molecules. Finally, as a more advanced application, we applied the nMC-MC algorithm to generate the QM/MM[318] distribution of aniline in aqueous solution.

This chapter is structured as follows: we first present the basic theory underlying the proposed approach, *viz.*, the hMC method, the nMC-MC algorithm, the FF

reparameterisation approach, the phase space overlap metrics, and the numerical experiments protocol. We then present applications of the algorithms to the previously mentioned test cases and conclude with final remarks.

6.2 Theory and methods

6.2.1 Hybrid Monte Carlo

hMC is an exact sampling approach that combines the features of the MC and MD simulation methods in such a way that the trial steps of the MC algorithm are short MD runs. Therefore, hMC inherits the advantages of both algorithms, such as the tendency of MD to move the system towards regions of configuration space that are energetically favourable, and the possibility to relax the restriction on the size of the MD time step, dt, through the application of a MC step.[92] hMC also prevents the numerical instabilities that arise due to the numerical integration algorithms used by MD simulations.[319]

At the start of every iteration, hMC draws new velocities from the Maxwell-Boltzmann distribution at a chosen temperature T_K, a step which is performed using the Marsaglia polar method.[298,320] Then, a short MD simulation in the microcanonical ensemble (NVE) is run using a sympletic integrator that preserves detailed balance[321] (*e.g.*, the velocity-Verlet algorithm,[322,323] the integrator used in this study) during M steps. Finally, the final configuration of the system is accepted or rejected according to a given acceptance criterion, which for a canonical ensemble (NVT) at temperature T_U reads[292]

$$\phi(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) = \min\left\{1, \exp\left[-\beta_K \Delta K - \beta_U \Delta U^{MM}(\boldsymbol{q}_i, \boldsymbol{q}_f)\right]\right\} \tag{6.1}$$

where $\Delta K = K_f - K_i$ is the difference of the kinetic energy between the final and initial states of the short MD run, $\Delta U^{MM} = U^{MM}(\boldsymbol{q}_f) - U^{MM}(\boldsymbol{q}_i)$ is the difference between the potential energy of the system at configurations $\boldsymbol{q}_f$ and $\boldsymbol{q}_i$,

and β_K and β_U are the thermodynamics betas corresponding to the temperatures T_K and T_U, respectively. The latter do not need have the same value, a feature that can be used as a means of increasing the conformational sampling efficiency by, for example, using high T_K values for the kinetic energy component. It is worth mentioning that M and dt are hyperparameters that have to be chosen properly in order to ensure sampling of uncorrelated snapshots while keeping the wall time required for each MD run manageable.[298] These hyperparameters may have an impact on the acceptance rates.[294,298,324]

The disadvantage of hMC is that its acceptance probability decays exponentially with the system size because the RMSE of the energy increases with $N_a^{1/2}$,[90,294] where N_a is the number of atoms of the system. There have been many attempts to circumvent this bottleneck, the most widely studied being sampling from shadow Hamiltonians.[294,325] Nevertheless, owing to the relatively small size of the molecular systems covered in this study, this issue does not pose a problem for the current application. As it is discussed in Section 6.2.2, hMC can also be embedded in an nMC-MC algorithm, in which the hMC moves are used as the trial steps of an *ab initio* MC algorithm.

6.2.1.1 Acceptance criterion derivation

To derive the acceptance criterion for the hMC algorithm, assume that a simulation starts from configuration $\boldsymbol{q}_i$. The probability that configuration $\boldsymbol{q}_f$ is reached after M MD steps is proportional to the initial momenta, $\boldsymbol{p}_i$, and is given by

$$\alpha\left(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f\right) \propto \prod_{\gamma=1}^{N_a}\left(\frac{\beta_K}{2m_\gamma\pi}\right)^{-1/2} \exp\left[-\beta_K \sum_{\gamma=1}^{N_a}\frac{\boldsymbol{p}_{i,\gamma}^2}{2m_\gamma}\right] \tag{6.2}$$

where γ runs over all N_a atoms, β_K is the thermodynamic beta, and m_γ the mass of γth atom. Noting that the argument of the exponential function corresponds to the kinetic energy, K, equation (6.2) can be rewritten as

$$\alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f) \propto \exp\left[-\beta_K K(\boldsymbol{p}_i)\right] \tag{6.3}$$

And similarly to what we did in equations (6.2) and (6.3), the probability of reaching $\boldsymbol{q}_i$ from $\boldsymbol{q}_f$ reads

$$\alpha(\boldsymbol{q}_f \to \boldsymbol{q}_i) \propto \exp\left[-\beta_K K(\boldsymbol{p}_f)\right] \tag{6.4}$$

where $\boldsymbol{p}_j$ is the negative of the momentum of the system after M MD steps, chosen as so in order to impose the detailed balance condition. It is now possible to use the result of equation (3.33) to derive the hMC acceptance criterion for the canonical ensemble, for which the probability density is stated in equation (3.14). Knowing that the transition probabilities are given by equations (6.3) and (6.4), the hMC acceptance criterion reads

$$\theta(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \frac{\alpha(\boldsymbol{q}_f \to \boldsymbol{q}_i)}{\alpha(\boldsymbol{q}_i \to \boldsymbol{q}_f)} \frac{\mathcal{N}(\boldsymbol{q}_f)}{\mathcal{N}(\boldsymbol{q}_i)} \tag{6.5}$$

$$= \frac{\exp\left[-\beta_K K(\boldsymbol{p}_f)\right]}{\exp\left[-\beta_K K(\boldsymbol{p}_i)\right]} \frac{\exp\left[-\beta_U U(\boldsymbol{q}_f)\right]}{\exp\left[-\beta_U U(\boldsymbol{q}_i)\right]} \tag{6.6}$$

$$= \exp\left[-\beta_K \Delta K - \beta_U \Delta U^{MM}(\boldsymbol{q}_i, \boldsymbol{q}_f)\right] \tag{6.7}$$

Finally, as shown in equation (3.36), by making the Metropolis choice for the acceptance criterion, the following hMC acceptance criterion is obtained

$$\phi(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \min\left\{1, \exp\left[-\beta_K \Delta K - \beta_U \Delta U^{MM}(\boldsymbol{q}_i, \boldsymbol{q}_f)\right]\right\} \tag{6.8}$$

which corresponds to the result previously stated in equation (6.1).

6.2.2 Sampling from approximate potentials

As it was formalised by Gelb in his seminal work about sampling from approximate potentials, it is possible to create an nMC-MC simulation by coupling hMC with a correction step based on the difference between the MM and QM potentials.[295,296] The expression of this correction step reads

$$\theta(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) = \min\left\{1, \exp\left[-\beta'_U \Delta U^{QM}(\boldsymbol{q}_i, \boldsymbol{q}_f) + \beta_U \Delta U^{MM}(\boldsymbol{q}_i, \boldsymbol{q}_f)\right]\right\} \quad (6.9)$$

where $\Delta U^{MM} = U^{MM}(\boldsymbol{q}_f) - U^{MM}(\boldsymbol{q}_i)$, $\Delta U^{QM} = U^{QM}(\boldsymbol{q}_f) - U^{QM}(\boldsymbol{q}_i)$, and β_U and β'_U are the thermodynamic betas of the MM and QM ensembles, respectively. Note that, as in the hMC algorithm, the beta thermodynamic parameters do not need to be the same in both Markov chains, a feature that can be exploited as a way of increasing the overlap between the MM and QM levels.[306,307]

The nMC-MC algorithm works by first generating a trial structure through the hMC algorithm, which is then attempted to be sampled into the QM level by applying the acceptance criterion of equation (6.9). If the structure is accepted, the next hMC run starts from this configuration; otherwise, the hMC run starts from the last accepted configuration. A detailed diagram of the workflow of this algorithm is shown in Figure 6.1.

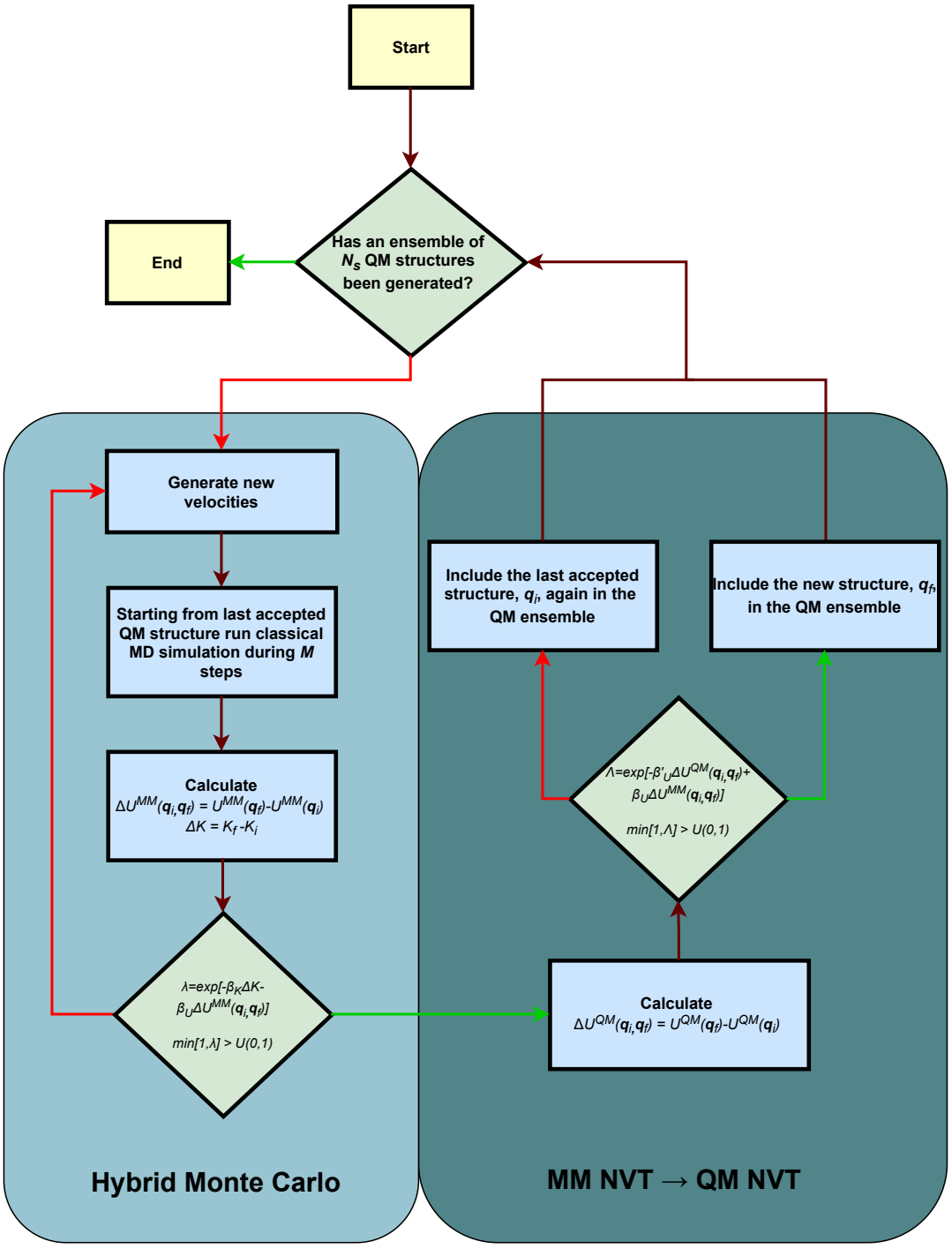

Figure 6.1: Diagram describing the workflow of nMC-MC algorithm as implemented in ParaMol.[4] The hMC part of the algorithm is used to generate an exact NVT ensemble (left), while the sampling from approximate potentials part is used as a switching step between the MM and QM levels of theory (right). $U(0,1)$ denotes a random number between 0 and 1 sampled from a uniform distribution, and the i and f subscripts refer to the initial and final states of a given iteration. The green arrows denote conditionals for which the evaluated condition is true, whereas the red arrows denote conditionals for which the evaluated condition is false.

6.2.2.1 Acceptance criterion derivation

To derive the acceptance criterion for the nMC-MC algorithm, consider a first Markov chain, identified by the QM superscript, and a second Markov chain, identified by the MM superscript, which uses the approximate potential that generates new trial states for the first Markov chain. Now suppose that a molecular simulation at the MM chain generates the state f given that it has started at state i. In such a situation, the probability of occurrence of a trial move from $\boldsymbol{q}_i$ to $\boldsymbol{q}_f$ in the QM Markov chain is given by

$$\alpha^{QM}(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \pi^{MM}(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \alpha^{MM}(\boldsymbol{q}_i \to \boldsymbol{q}_f)\theta^{MM}(\boldsymbol{q}_i \to \boldsymbol{q}_f) \tag{6.10}$$

Hence, by substituting this result into equation (3.36), *i.e.*, by making the Metropolis choice for the acceptance criterion, we obtain

$$\theta^{QM}(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \min\left[1, \frac{\mathcal{N}^{QM}(\boldsymbol{q}_f)}{\mathcal{N}^{QM}(\boldsymbol{q}_i)} \frac{\alpha^{QM}(\boldsymbol{q}_f \to \boldsymbol{q}_i)}{\alpha^{QM}(\boldsymbol{q}_i \to \boldsymbol{q}_f)}\right] \tag{6.11}$$

$$= \min\left[1, \frac{\mathcal{N}^{QM}(\boldsymbol{q}_f)}{\mathcal{N}^{QM}(\boldsymbol{q}_i)} \frac{\pi^{MM}(\boldsymbol{q}_f \to \boldsymbol{q}_i)}{\pi^{MM}(\boldsymbol{q}_i \to \boldsymbol{q}_f)}\right] \tag{6.12}$$

$$= \min\left[1, \frac{\mathcal{N}^{QM}(\boldsymbol{q}_f)}{\mathcal{N}^{QM}(\boldsymbol{q}_i)} \frac{\alpha^{MM}(\boldsymbol{q}_f \to \boldsymbol{q}_i)\theta^{MM}(\boldsymbol{q}_f \to \boldsymbol{q}_i)}{\alpha^{MM}(\boldsymbol{q}_i \to \boldsymbol{q}_f)\theta^{MM}(\boldsymbol{q}_i \to \boldsymbol{q}_f)}\right] \tag{6.13}$$

which can be used to generate an ensemble that is distributed according to the QM distribution. Alternatively, recalling the relation given by equation (3.30), equation (6.13) can be rewritten as

$$\theta^{QM}(\boldsymbol{q}_i \to \boldsymbol{q}_f) = \min\left[1, \frac{\mathcal{N}^{QM}(\boldsymbol{q}_f)}{\mathcal{N}^{QM}(\boldsymbol{q}_i)} \frac{\mathcal{N}^{MM}(\boldsymbol{q}_i)}{\mathcal{N}^{MM}(\boldsymbol{q}_f)}\right] \tag{6.14}$$

As previously mentioned, the expressions of the probability densities appearing in equation (6.14) are ensemble-dependent. For example, if one is interested in sampling from an MM NVT distribution to generate a QM NVT ensemble, the canonical probability density function given by equation (3.14) must be used, leading to the following acceptance criterion

$$\theta(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) = \min\left\{1, \exp\left[-\beta'_U \Delta U^{QM}(\boldsymbol{q}_i, \boldsymbol{q}_f) + \beta_U \Delta U^{MM}(\boldsymbol{q}_i, \boldsymbol{q}_f)\right]\right\} \quad (6.15)$$

which can be recognised as the expression previously stated in equation (6.9). On the other hand, if one is interested in sampling from an MM NVE distribution to generate a QM NVT ensemble, it must be ensured that the MD run of the hMC algorithm conserves the energy of the system. In practical terms, this means that the time step used must be sufficiently small so that the numerical errors are negligible. Whenever this condition is ensured, hMC steps are accepted with unit probability, *i.e.*, $\theta^{MM}(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) = \theta^{MM}(\boldsymbol{q}_f \rightarrow \boldsymbol{q}_i) = 1$, reducing equation (6.13) to

$$\theta(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) = \min\left\{1, \exp\left[-\beta'_U \Delta U^{QM}(\boldsymbol{q}_i, \boldsymbol{q}_f) + \beta_K \Delta K\right]\right\} \quad (6.16)$$

where the results of the transition probabilities given by equations (6.3) and (6.4) were employed. Note that equation (6.16) can be rewritten as equation (6.15) since, under energy conservation conditions, $\Delta K = \Delta U$. Hence, overall we conclude that the same acceptance criterion can be used to sample from the MM chain to the QM chain, whether sampling is performed from the MM NVE or the MM NVT ensemble. However, while the former sampling option requires using the hMC acceptance criterion of equation (6.1) to first generate the MM NVT ensemble, the latter option avoids this step, permitting a direct switch between the MM and QM Markov chains. Lastly, note that although in this study the hMC algorithm was used to generate trial states for the QM Markov chain, in

practice any sampling method, such as, *e.g.*, standard MC simulations, parallel tempering MC, or umbrella sampling schemes, can be used for that purpose.

6.2.3 Force field reparameterisation

The key requirement for convergence of the nMC-MC algorithm is ensuring a favourable distribution overlap between the FF and the QM level of theory. To fulfill this condition, we generated low-level models that were closer to the target high-level of theory than the original GAFF.[182] The aim of this procedure was to improve the overlap between the energy difference distributions of the MM and QM levels. The increased overlap was attained through reparameterisation of GAFF-like FFs, a step performed to make the MM models more QM-like. To perform the optimisation of the FF parameters, we resorted to the methodologies implemented in ParaMol, in which an FF is fitted to a target level of theory through minimisation of the following objective function

$$X(\boldsymbol{p}) = X_F(\boldsymbol{p}) + X_U(\boldsymbol{p}) + \Theta(\boldsymbol{p}) \tag{6.17}$$

where X_F is given by equation (5.3), corresponding to the term of the objective function by which every component of the MM atomic forces is fitted to QM data; X_U is given by equation (5.5), amounting to the fitting of energies to reference QM data; and $\Theta(\boldsymbol{p})$ is given by equation (5.13), corresponding to an optional L2 regularisation included to prevent overfitting.

6.2.4 Phase space overlap metrics

As a means of establishing the similarity between the MM and QM levels, we evaluated their phase space overlap using two different metrics. The first metric resorts to the idea that the phase space overlap can be calculated as the overlap between the distributions of the total energy difference between the two considered levels of theory.[248] These distributions can be obtained by performing

MD simulations using both the MM and the QM Hamiltonians, whereupon the differences $\Delta E_{MM}^{MM\to QM} = E_{MM}^{QM} - E_{MM}^{MM}$ and $\Delta E_{QM}^{QM\to MM} = E_{QM}^{MM} - E_{QM}^{QM}$ are evaluated for the trajectories obtained utilising the MM and QM Hamiltonians, respectively, where the subscript indicates the level of theory used for sampling, and the superscript indicates the level of theory used to evaluate the potential energy. The corresponding histograms of the calculated ΔEs are then approximated by assuming Gaussian-shaped distributions, in such a way that the energy difference distribution of the QM Hamiltonian is given by

$$\mathcal{N}_{QM}(\Delta E_{QM}^{QM\to MM}) = \sqrt{\frac{1}{2\pi\sigma_{QM}^2}} \exp\left[-\frac{\left(\Delta E_{QM}^{QM\to MM} - \langle \Delta E_{QM}^{QM\to MM} \rangle\right)^2}{\sigma_{QM}^2}\right] \tag{6.18}$$

where σ_{QM} is the standard deviation of the $\Delta E_{QM}^{QM\to MM}$ values. The Gaussian representation of the energy difference distribution of the MM Hamiltonian, $\mathcal{N}_{MM}$, can be written analogously. Hence, it possible to measure the overlap, Ω, between the $\mathcal{N}_{QM}$ and $\mathcal{N}_{MM}$ distributions by using the following equation

$$\Omega = \frac{\langle \mathcal{N}_{QM}, \mathcal{N}_{MM} \rangle}{\max\left[\langle \mathcal{N}_{QM}, \mathcal{N}_{QM} \rangle, \langle \mathcal{N}_{MM}, \mathcal{N}_{MM} \rangle\right]} \tag{6.19}$$

where $\langle f, g \rangle = \int dx\, f(x) \cdot g(x)$ is the inner product or overlap integral between the f and g functions. The integration of this overlap integral was performed numerically using SciPy's integrate.quad function with default settings.[264]

We also used the descriptors of the phase space overlap between two states that were developed by Wu and Kofke.[248,326,327] In particular, we used a metric based on the overlap of total energy distributions that reads

$$\Sigma_{MM,QM} = 2 \int_{-\infty}^{+\infty} dE_{QM}\, \rho_{QM}^{QM}(E_{QM}) \int_{-\infty}^{E_{QM}} dE'_{QM}\, \rho_{MM}^{QM}(E'_{QM}) \tag{6.20}$$

where ρ_A^A and ρ_B^A are the probability distributions of state A energies observed within simulations of states A and B, respectively. Similarly, the expression for $\Sigma_{QM,MM}$ can be written as

$$\Sigma_{QM,MM} = 2 \int_{-\infty}^{+\infty} dE_{MM}\, \rho_{MM}^{MM}(E_{MM}) \int_{-\infty}^{U_{MM}} dE'_{MM}\, \rho_{QM}^{MM}(E'_{MM}) \tag{6.21}$$

The value of $\Sigma_{B,A}$ varies from 0 to 2 and indicates the offset of ρ_B^A relative to ρ_A^A. If ρ_B^A is centred left with respect to ρ_A^A, then $1 < \Sigma_{B,A} \leq 2$. Otherwise, if ρ_B^A is centred right with respect to ρ_A^A, then $0 \leq \Sigma_{B,A} < 1$. The integration of the double integrals of equations (6.20) and (6.21) was performed numerically using SciPy's integrate.dblquad function with default settings.

6.2.5 Numerical experiments protocol

The numerical experiments presented in this study relied on refined low-level models that attempted to approximate a high level of theory. The high level of theory used was the DFTB+[273,274] implementation of SCC-DFTB,[1] including the D3 dispersion correction[278] and with Becke-Johnson damping.[279] This choice was based on the evidence that SCC-DFTB-D3 performs quite well in determining conformations of druglike molecules and respective energies.[79–81] SCC-DFTB-D3 is also computationally cheap, allowing for extensive testing of various compounds. As low-level models, optimally tuned FFs were generated to improve the overlap between the energy difference distributions of the low and high levels of theory. These FFs were reparameterised to reproduce the SCC-DFTB-D3 level of theory. They consisted of refined versions of GAFF, for which the functional form is given by equation (4.25). The optimally tuned FFs were designed systematically such that, for every molecule shown in Figure 6.2, the following set of reparameterised FFs was generated

- B FF - bond force constants (K_b) and equilibrium values (r_{eq}) were optimised.

- BA FF - bond and angle force constants (K_θ) and equilibrium values (θ_{eq}) were optimised.
- BAT FF - bond, angle and dihedral force constants (K_b, K_θ, and V_n), bond and angle equilibrium values (r_{eq}, and θ_{eq}), and dihedral phase constants (γ_n) were optimised.
- BAT-LJ FF - in addition to the parameters optimised in the BAT FF, the σ and ϵ LJ 12-6 parameters were also optimised.
- BAT-Q FF - in addition to the parameters optimised in the BAT FF, the atomic charges (q) were also optimised (under the constraint of total molecular charge conservation).
- BAT-LJQ FF - in addition to the parameters optimised in the BAT FF, the σ and ϵ LJ 12-6 parameters were also optimised, as well as the atomic charges (q) under the constraint of total molecular charge conservation.

The optimisation of the parameters was performed using ParaMol with the SciPy's SLSQP optimiser.[328] The optimisations were deemed to be converged whenever the objective function between two successive iterations did not change by more than 10^{-6}, *i.e.*, $X_{n+1} - X_n < 10^{-6}$. The original GAFF parameters were used as the initial guess for the optimisations. These were obtained by initially parameterising the druglike molecules using Antechamber packages, which are part of AmberTools.[329] AM1-BCC charges[225,226] were calculated after the geometry was optimised at the SCC-DFTB-D3 level of theory. The topology and coordinates files used as inputs to ParaMol were created using LEaP. All FF modifications given by the frcmod file created by parmchk2 were included. No atom-type symmetries were preserved during the reparameterisation. Consequently, the results presented are close to the limits of accuracy that the GAFF functional form can achieve.

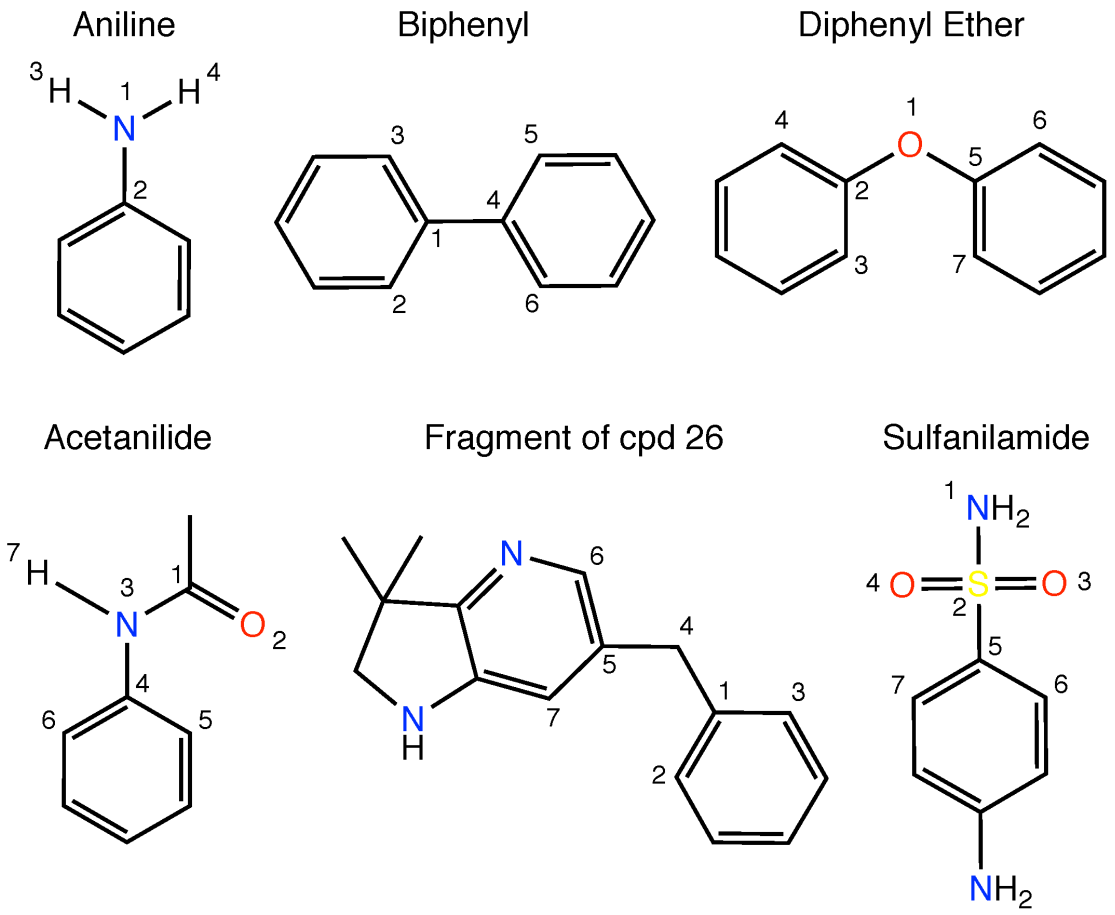

Figure 6.2: Molecular structures of the test molecules used in this study.

The objective function minimised in the optimisation procedure included as targets both forces and energies, as shown in equation (6.17). The reparameterisations of the FFs used either the uniform or the non-Boltzmann weighting (weighting temperature of 300 K) schemes available in ParaMol (see the discussion presented in Section 5.2.4). They were conducted applying either no regularisation or L2 regularisation. The prior widths used for the regularisation term are reported in Table 5.1. The value of the scaling factor used in the regularisation term was $\alpha = 1/N_p$, where N_p is the number of parameters being optimised. The training data sets consisted of configurational ensembles generated at the SCC-DFTB-D3 level of theory. They were obtained by performing gas-phase Langevin dynamics at a temperature of 500 K (time step of 1 fs, and friction coefficient of 2 ps^{-1}). We chose to simulate at a high temperature to ensure a thorough exploration of the SCC-DFTB-D3 conformational space.

Snapshots of the simulations were collected every 1 ps, resulting in a final data set of 10000 configurations.

The nMC-MC calculations performed to estimate the acceptance rates used a time step of 1 fs (100 MD steps per hMC run). Velocities were sampled from the Maxwell-Boltzmann distribution at 300 K. This temperature was also used for the MM and QM chains. No fine-tuning of these hyperparameters was attempted, as these tend to be molecule-specific. A total of 4 independent samplers were run for each molecule, all starting from the same initial structure but using different random seeds.

6.3 Results and discussion

6.3.1 nMC-MC acceptance rates

The most direct metrics one can obtain from nMC-MC simulations are the acceptance rates. There are two of these: the hMC acceptance rate shown in equation (6.1), and the MM to QM switching step acceptance rate shown in equation (6.9). The hMC acceptance rate gives information about the stability of the NVE MD runs. It is useful to identify energy conservation issues. The MM to QM switching step gives information about the similarity between the MM and QM levels of theory. This switching step acceptance rate is the focus of this work because, as discussed later, it is highly correlated with phase space overlap metrics. Consequently, this acceptance rate is a valuable metric of how close to the QM level of theory the MM FFs sample, as it measures the MM $\rightarrow$ QM overlap. Note, however, that this is a unidirectional relation, because the acceptance rate does not give details about how close to the MM level of theory the QM Hamiltonian samples. Measuring the QM $\rightarrow$ MM overlap would require performing nMC-MC calculations using the QM Hamiltonian in the lower chain. This is computationally expensive and, in many cases, unfeasible owing to the requirement of performing *ab initio* MD. Hence, we cannot exclude the possibility

of predicting high acceptance rates for FFs that do not completely represent the QM level. This situation occurs whenever a FF explores only a subset of QM configurations that are well described at the MM level. Note that, however, this situation never occurred for the optimally tuned FFs generated in this study. Incidentally, as the phase space overlap metrics reveal, the MM configurational distributions generated at a given temperature were of a similar extent to (or broader than) their QM counterparts. This observation corroborates that the switching step acceptance rate is a robust metric of the similarity between the MM and QM levels of theory.

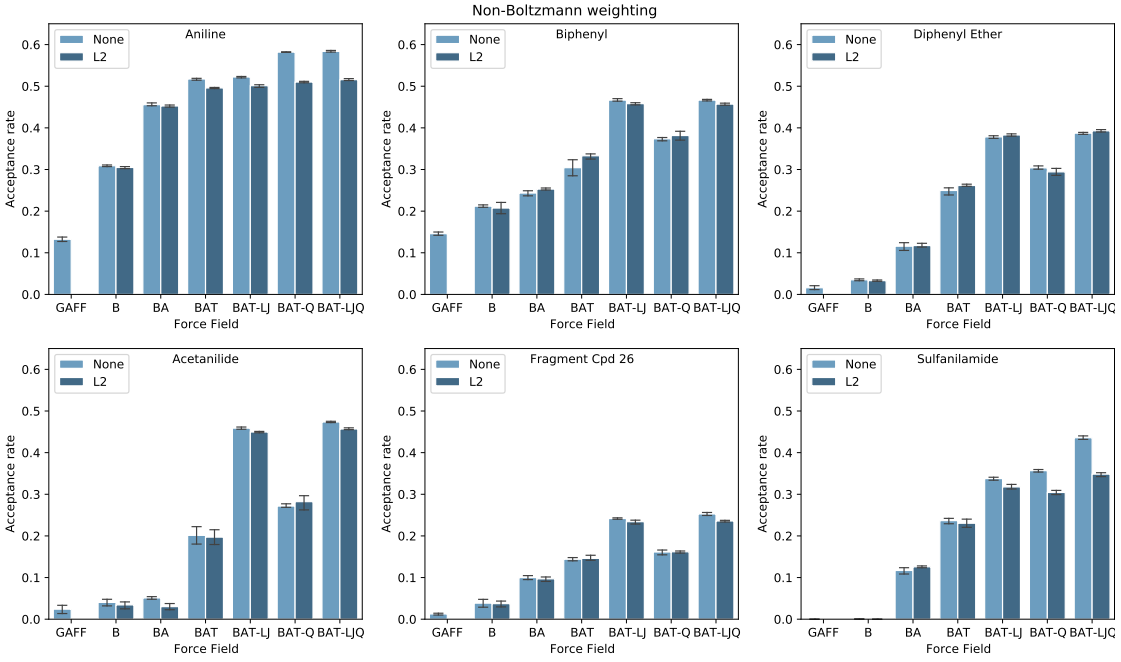

Figure 6.3: nMC-MC acceptance rates for the set of molecules represented in Figure 6.2. The FFs used to calculate the acceptance rates were derived employing non-Boltzmann weighting with (dark blue) or without (light blue) L2 regularisation. The training data set contained configurations sampled at 500 K. The errors bars correspond to the standard deviation of the results of 4 different nMC-MC samplers. Each sampler performed a total of 2×10^5 nMC-MC sweeps.

The acceptance rates obtained when using non-Boltzmann weighting in the reparameterisations are shown in Figure 6.3. Through the analysis of this figure, we conclude that the variations observed in the acceptance rates are in line with what we would expect from a systematical reparameterisation of FFs: for the nonregularised FFs, the more classes of FF parameters optimised, the higher

the acceptance rates obtained. In general, they follow the trend B $<$ BA $<$ BAT $<$ BAT-Q $<$ BAT-LJ $<$ BAT-LJQ. The only disparity is observed for aniline and sulfanilamide, for which the BAT-Q FF performed better than the BAT-LJ. This reveals that the optimisation of charges is more important to accurate modelling of the aniline scaffold, which both molecules share, than the optimisation of the LJ parameters.

The optimisation of the nonbonded part of the FFs (charges, and LJ parameters) was here used as an *ad hoc* workaround to accelerate the sampling of the target SCC-DFTB-D3 distribution. It must be stressed, however, that this approach is only viable for gas-phase situations like those presented in these examples. As these parameters affect the intermolecular interactions, the BAT-LJ, BAT-Q, and BAT-LJQ FFs may have a limited applicability when applied to solutes in solution. Since solute-solvent interactions influence the solute's configurational ensemble, and the nonbonded parameters influence the energetics of intermolecular interactions, a training data set including interactions with solvent needs to be considered if the LJ parameters and partial atomic charges are to be optimised.

In comparison to the original GAFF, the improvements obtained for the acceptance rates were substantial (Figure 6.3). This indicates that the optimally tuned FFs increased their similarity with respect to the SCC-DFTB-D3 Hamiltonian. For all test cases, except aniline and biphenyl (for which GAFF acceptance rates of *ca.* 13-15% were obtained), the GAFF acceptance rates were lower than 3%, being virtually 0% for sulfanilamide. These observations support the idea that MM FFs struggle to correctly model sulfonamides.[330]

The properties of non-Boltzmann weighting help to understand the previous observations. During the FF optimisation, this weighting scheme gives larger weights to conformations in which the MM energy is underestimated ($U^{MM} - U^{QM} < 0$) than to conformations in which the MM energy is overestimated ($U^{MM} - U^{QM} > 0$) with respect to the QM energy. Consequently, non-Boltzmann weighting mitigates the creation of spurious minima and drives the errors towards high-energy regions, thus overestimating transition-state energies and

underestimating fluctuations.[4,263] This is a desirable property because if a perfect fit to the QM PES is unattainable, it is preferable to have a mismatch that makes ΔU^{MM} greater than ΔU^{QM}. Driving the errors towards high-energy regions maximises acceptance into the QM chain (see equation (6.9)) and leads to the stable and systematic improvements that are observed in Figure 6.3.

The results obtained for the L2-regularised FFs follow similar trends to their non-regularised counterparts (Figure 6.3). For the FFs in which bonded parameters (B, BA, and BAT) were optimised, the acceptance rates of the regularised and nonregularised FFs were identical, presenting variations that are not statistically significant. However, for the regularised FFs in which nonbonded parameters were also optimised (BAT-LJ, BAT-Q, and BAT-LJQ), there was a noticeable decrease in the acceptance rates for some molecules (*e.g.*, aniline and sulfanilamide). The prior widths used to constrain the charges and LJ parameters may be the source of this decrease in acceptance rates, as the prior widths may have not allowed the nonbonded parameters to stray too far away from their initial guesses. Consequently, poorer fits were obtained, leading to lower acceptance rates. For the remaining molecules, the regularised FFs performed equally or slightly better than the nonregularised FFs. The small differences observed are, in most cases, not statistically significant.

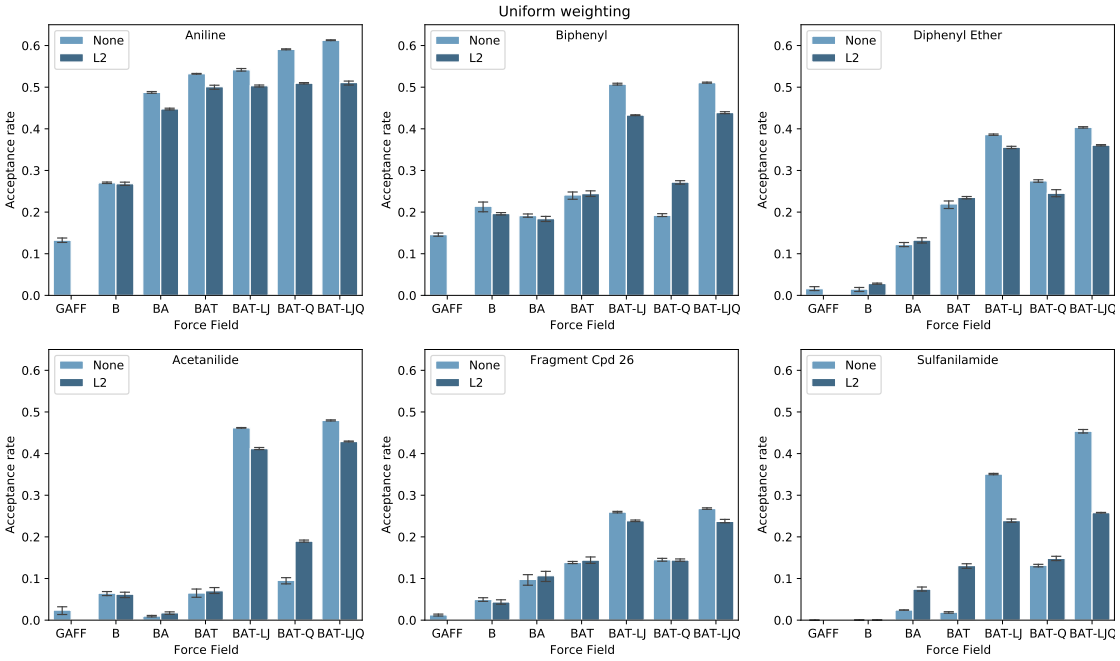

Figure 6.4: nMC-MC acceptance rates for the set of molecules represented in Figure 6.2. The FFs used to calculate the acceptance rates were derived employing uniform weighting with (dark blue) or without (light blue) L2 regularisation. The training data set contained configurations sampled at 500 K. The errors bars correspond to the standard deviation of the results of 4 different nMC-MC samplers. Each sampler performed a total of 2×10^5 nMC-MC sweeps.

Before analysing the results obtained when employing uniform weighting in the reparameterisations (Figure 6.4), it is important to discuss the consequences of the asymmetries that might be imposed on the MM PES by equally allowing positive and negative errors in the fittings. This feature of uniform weighting has already been reported in previous studies.[4,263] Let us consider the diagrams of Figure 6.5, which show hypothetical MM PES fittings that can be obtained when employing uniform weighting. Assuming that the data set used in the reparameterisations comprises only the structures at configurations $\boldsymbol{q}_i$ and $\boldsymbol{q}_f$, all the represented uniform-weighted fittings have equal squared errors of the energy with respect to the QM PES, *viz.* $\sum_i \left(U_i^{QM} - U_i^{MM}\right)^2 = 2U^2$. Despite this, each case would lead to a different behaviour if the corresponding FF were used in the nMC-MC algorithm. Firstly, it is important to note that case A truly corresponds to a perfect fitting, as the uniform-weighted MM PES can be superimposed with QM PES by a simple translation (the difference between them is only a constant). Therefore, in what follows, we exclude

this situation from the discussion. Furthermore, for case B, $\Delta\Delta U(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) = \Delta U^{QM}(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) - \Delta U^{MM}(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) = -2U$ for uniform weighting, leading to a fitting that maximises acceptance into the QM chain. On the other hand, case C minimises the probability of accepting structures into the QM chain since $\Delta\Delta U(\boldsymbol{q}_i \rightarrow \boldsymbol{q}_f) = 2U$. Hence, the uniform weighting scheme is prone to creating series of FFs that show unpredictable, non-systematic behaviour since positive and negative $U^{MM} - U^{QM}$ differences are equally probable. This leads us to advocate for the use of non-Boltzmann weighting if the aim is to use FFs in the nMC-MC algorithm. Incidentally, non-Boltzmann weighting also tends to be superior to uniform weighting for general-purpose applications.[4,263] Note that the squared error of the energy with respect to the QM for the non-Boltzmann-weighted fitting is not equal to $2U^2$ as it is in the uniform-weighted fittings. Nevertheless, the point was to illustrate the possible asymmetries that uniform weighting may impose that can negatively impact the acceptance rates.

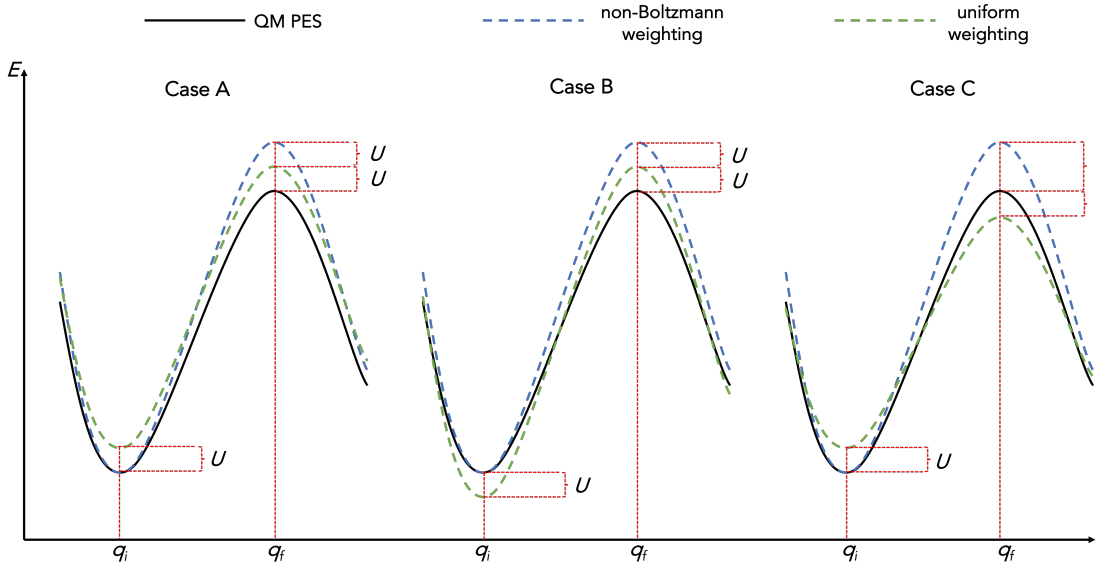

Figure 6.5: Diagram illustrating typical possible fittings that can be obtained when employing either the uniform and non-Boltzmann weighting schemes. All the represented uniform-weighted fittings have equal squared errors of the energy with respect to the QM PES, *viz.*, $\sum_i \left(U_i^{QM} - U_i^{MM}\right)^2 = 2U^2$, but they behave differently when used in the nMC-MC algorithm.

For the FFs derived employing uniform weighting, systematic reparameterisation only led to systematically higher acceptance rates for aniline, diphenyl

ether, and the fragment of cpd 26 (Figure 6.4). This is observed for both the L2 and nonregularised versions of the FFs. On the other hand, for biphenyl and acetanilide, the BA FF resulted in lower acceptance rates than the B FF, and the BAT FF only performed slightly better than the B FF. The acceptance rates obtained for sulfanilamide are unexpected, as the BA and BAT FFs gave very low acceptance into the QM chain (*ca. 2%*). Interestingly, higher acceptance rates were obtained for their L2-regularised counterparts, as they increased from *ca.* 2% to *ca.* 7-8% and 13% for the BA and BAT FFs, respectively. These results can be understood by inspecting the optimised angle force constants, K_θ, and angle equilibrium values, θ_{eq}, of the nonregularised BA and BAT FFs of sulfanilamide (see Figure 6.6). From these plots, it is clear that the nonregularised optimisations drove the parameters towards nonphysical values. For K_θ, close to zero or even negative values were obtained, whereas for θ_{eq}, values close to 0° or 360° were obtained, meaning that bent angles became practically linear. These artifacts, created by the optimisation to minimise the objective function, ultimately had a strong impact on the acceptance rates, as they led to poor dynamics and energy prediction. These large and unphysical variations were lessened by applying L2 regularisation, resulting in FFs that had optimised parameters with values in physically-sensible ranges. The L2-regularised FFs were also superior in terms of QM similarity in regards to their nonregularised counterparts and led to higher acceptance rates. Besides sulfanilamide, the only test case for which a similar behaviour occurred was acetanilide, in which the nonregularised uniform-weighted BA and BAT FFs also contained unphysical parameter values. This event did not occur for any non-Boltzmann-weighted or L2-regularised FF.

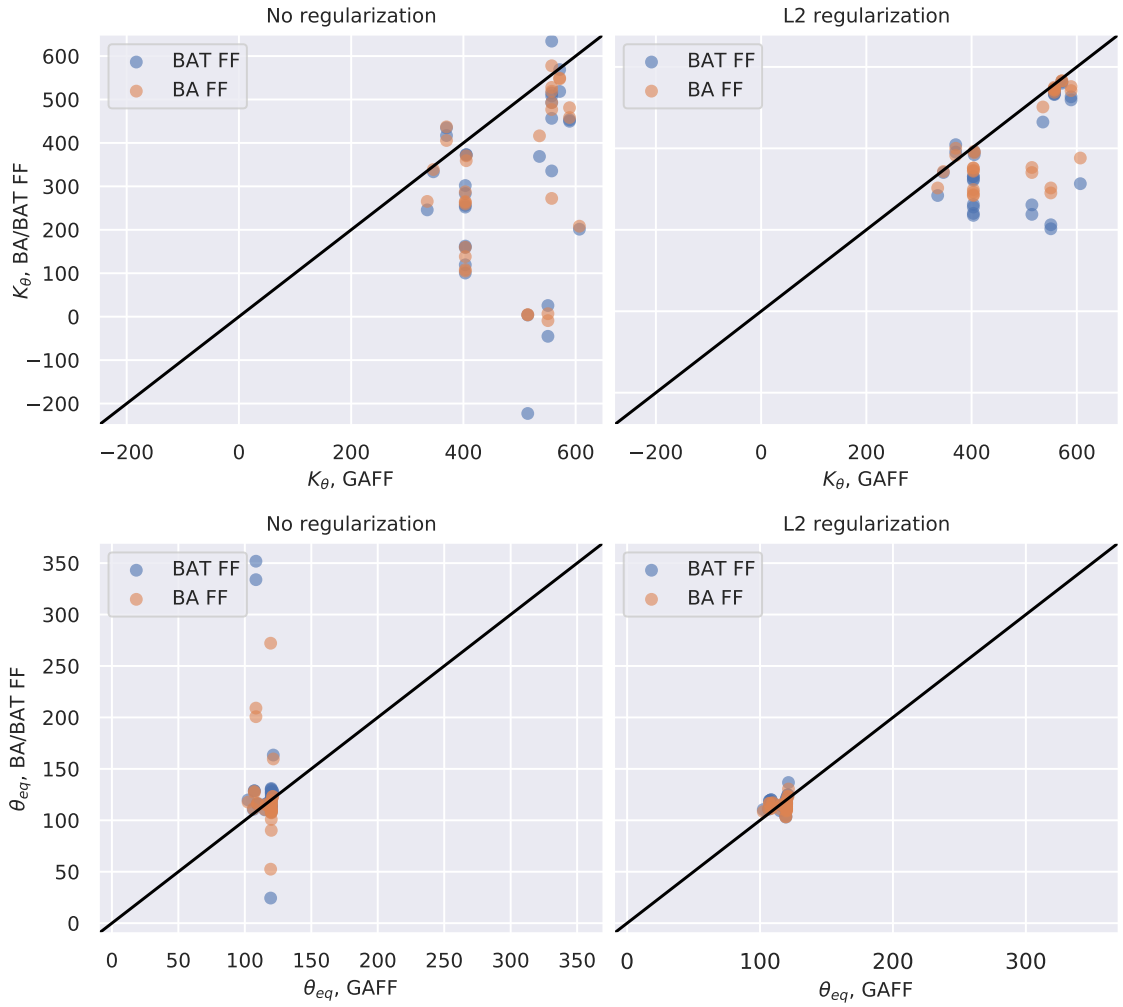

Figure 6.6: Sulfanilamide parameters before (GAFF; x axis) and after reparameterisation (uniform-weighted BA/BAT FFs; y axis). The parameters represented are angle force constants (top panels) and angle equilibrium values (lower panels).

Two synergistic factors may have contributed to the deviations from physically sensible values observed for the nonregularised uniform-weighted FFs. Firstly, some of the angle terms may have been used to compensate for deficiencies in other parts of the FF, since when nonbonded parameters were concomitantly optimised, these large deviations were not observed. Incidentally, as previously mentioned, the optimisation of the nonbonded terms performed in this study was, in some sense, a workaround to make up for possible limitations in the FF functional form. Secondly, the fact that the training data sets consisted of NVT ensembles sampled at a temperature of 500 K also contributed to nonphysical parameters. This can be observed by comparing the acceptance rates obtained

for the training data sets generated at 500 K with those obtained for the training data sets generated at 300 K (Figure 6.7; no regularisation, uniform weighting). At the QM level, especially when sampling at very high temperatures, bonds and angles oscillate anharmonically. This phenomenon may impact the FF reparameterisation owing to limitations of the GAFF functional form that, by design, imposes harmonicity in these DOFs (see equation 4.25). The consequence is that the optimisers tend to pull the angle force constants towards lower values to generate wider potentials that better fit the sampled anharmonicities. This is the response observed in Figure 6.6. Angles are particularly prone to straying away from physical-sensible values, as they have the lowest force constants of the hard DOFs. Consequently, they may suffer more from using data sets containing configurations in the anharmonic regime.

Although the training data sets at 500 K did not lead to nonphysical bond parameters, it is clear in Figure 6.6 that, for some molecules, they led to lower average acceptance rates in comparison to the FFs derived using data sets at 300 K. Interestingly, the differences in acceptance rates, which initially became apparent in the B or BA FF, were then somewhat propagated into the FFs for which parameters of nonbonded terms were also optimised (BAT-LJ, BAT-Q, and BAT-LJQ). Optimisation of nonbonded parameters either mitigated these contrasts or worsened them owing to the presence of high-energy structures in the high-temperature ensemble that were unimportant to the low-temperature ensemble. Despite these observations, we still opted to use the high-temperature data sets for most of the reparameterisations in this study, as they ensured a thorough exploration of the conformational space of the molecules being studied. Ideally, one would employ enhanced-sampling methods to have the best of both ensembles: the extensive conformational sampling of the high-temperature data set, and the harmonic behaviour of the low-temperature data set.

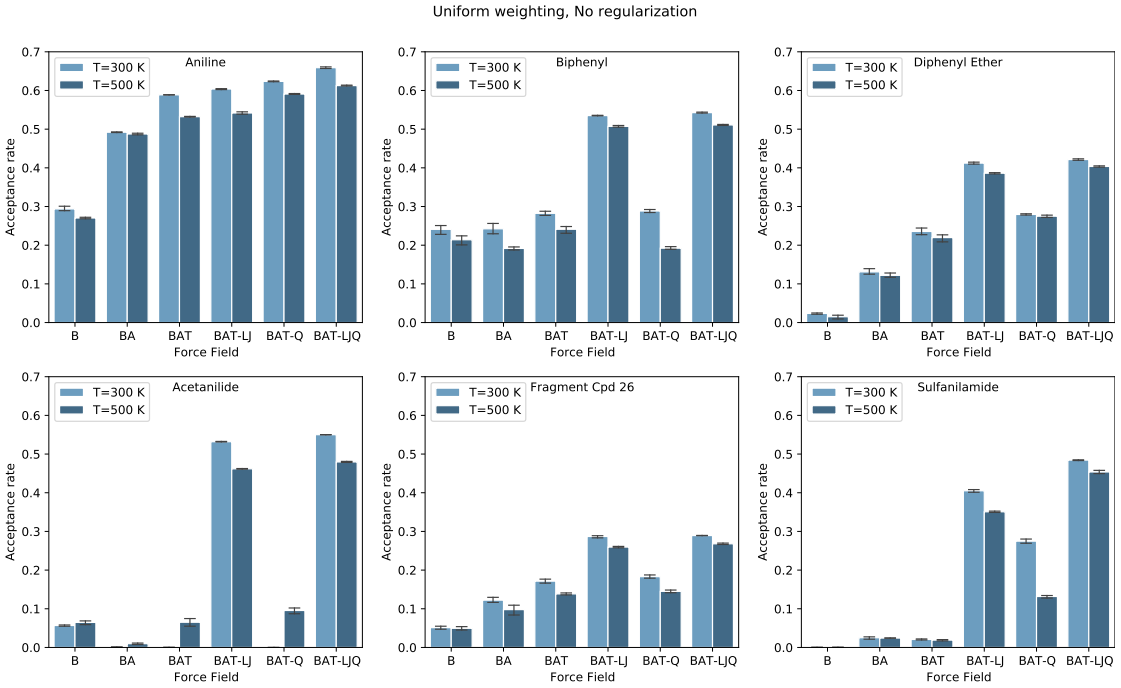

Figure 6.7: Comparison between the nMC-MC acceptance rates obtained for FFs reparameterised using data sets containing structure sampled at either 300 or 500 K. The FFs used to calculate the acceptance rates were derived employing uniform weighting without regularisation. The error bars correspond to the standard deviation of the results of 4 different nMC-MC samplers. Each sampler performed a total of 2×10^5 nMC-MC sweeps.

Overall, FF reparameterisation proves to be an efficient strategy to increase the acceptance rate of the switching step from the MM to the QM level of theory. It permits accelerating convergence of the sampling of the target QM configurational distribution, otherwise impractical owing to very low acceptance rates. The best acceptance rates were obtained for aniline (*ca.* 65%), whereas the molecule with the lowest acceptance rate was the fragment of cpd 26. This is expected given that the latter is the largest and most complex molecule of the test set. We expect that both molecular size and chemical complexity have an impact on the acceptance rates: the former because small differences between the MM and QM Hamiltonians accumulate as the number of DOFs increases; the latter due to the challenge that some functional groups pose to the functional form of the GAFF. Furthermore, uniform-weighted FFs, especially without regularisation, are to be avoided. They are prone to generate nonphysical parameters to obtain the best possible fit. On the other hand, L2-regularised and

non-Boltzmann-weighted FFs tend to perform the best and usually exhibit stable behaviour. Hard DOFs, such as bonds and angles, are crucial to be reparameterised to increase the acceptance rates. As hard DOFs have large force constants, small differences in their values lead to large changes in energy. Consequently, poor parameters for the hard DOFs may considerably impact the switching efficiency from the MM to the QM chain.

6.3.2 NH_2 inversion in aniline

As a first example of how nMC-MC allows recovery of the exact quantum configurational distribution using an approximate FF, let us consider the inversion of the NH_2 group in aniline. It is well established, both experimentally[331,332] and theoretically,[333,334] that the primary amine of aniline has a pyramidal geometry and that interconversion between two equally stable conformations occurs through nitrogen inversion. Nevertheless, although simple, this is a clear instance of a functional group for which GAFF fails to predict the correct conformational dynamics.

Through the analysis of the configurational distributions represented in Figure 6.8, it can be seen that GAFF generated NVT configurational distributions at 300 K that differ substantially from those generated by SCC-DFTB-D3. Specifically, GAFF (lower panel) predicted that the NH_2 group assumes a trigonal planar geometry, hence failing to reproduce the interconversion between the two local minima. On the other hand, SCC-DFTB-D3 (top panel) predicted the expected conformational behaviour. Furthermore, the reparameterised BAT FF distribution was much closer to the SCC-DFTB-D3 distribution than the original GAFF, and the nMC-MC distribution successfully reproduced the SCC-DFTB-D3 distribution when sampling was performed using the BAT FF. The agreement obtained is excellent, as there is negligible loss of accuracy.

The fast recovery of the target SCC-DFTB-D3 distribution through the nMC-MC algorithm was only possible due to the increased acceptance rates that were

achieved after reparameterisation of the original FF. GAFF gave acceptance rates of *ca.* 12-13%. Even though these acceptance rates are high in comparison with other test cases, they are still much lower than the acceptance rates of *ca.* 49-50% that were obtained for the non-Boltzmann-weighted L2-regularised BAT FF. This high acceptance rate enabled recovery of the target SCC-DFTB-D3 in only 2×10^5 nMC-MC sweeps (hMC runs of 100 steps with a 1 fs time step). No attempt was made to optimise the length of these calculations, as our main goal was to prove the implementation and principles of the methodology and not to optimise the protocol in itself.

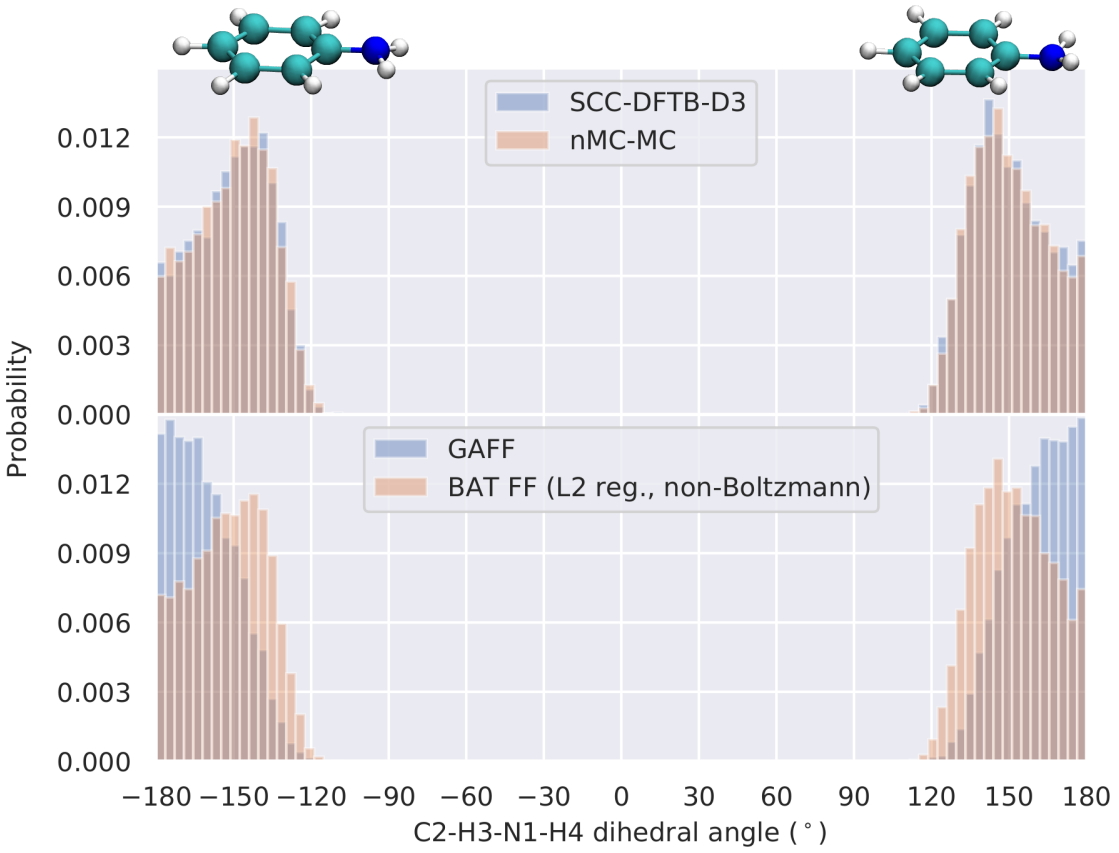

Figure 6.8: Top panel: Distribution of the C2-H3-N1-H4 improper dihedral of aniline as obtained in the SCC-DFTB-D3 MD and nMC-MC simulations. Lower panel: Distribution of the C2-H3-N1-H4 improper dihedral of aniline as obtained in MD simulations using the original GAFF and the non-Boltzmann-weighted L2-regularised BAT FF. The SCC-DFTB-D3, GAFF, and BAT MD calculations were performed during 10 ns (snapshots collected every 1 ps), and the nMC-MC sampler performed a total of 2×10^5 MC sweeps. The temperature of the simulations was 300 K.

6.3.3 Fragment of cpd 26

Let us now discuss the results obtained when applying the nMC-MC algorithm to the fragment of cpd 26 shown in Figure 6.2. Cpd 26 is a nonpeptidic, orally bioavailable, and efficacious low nM antagonist of the inhibitor of apoptosis proteins cIAP1 and XIAP.[15] Therefore, this test case aims to mimic the application of the presented methods to a recently designed and relevant druglike molecule.

By analysing the SCC-DFTB-D3 configurational distribution represented in Figure 6.9, we were able to identify 8 conformations for this molecule. Their molecular structures, which map to the red stars in the plot, are represented in Figure 6.10. These conformations essentially arise from the different relative positions that the phenyl group can assume relative to the azaindoline ring. This conformational dynamics is in line with what has been observed experimentally.[15] Surprisingly, although GAFF gave very low nMC-MC acceptance rates (close to 1%) and predicted incorrect relative abundances, it still fairly described the global features of the configurational distribution. The BAT (L2-regularised, non-Boltzmann-weighted) optimisation of GAFF led to a much closer distribution to the target SCC-DFTB-D3, demonstrating the quality of this FF. This observation is further supported by the increase in acceptance rate to 23-24%. Finally, as expected, the nMC-MC flawlessly reproduced the SCC-DFTB-D3 when using the BAT FF (L2-regularised, non-Boltzmann-weighted) in the low-level Markov chain, allowing recovery of the target distribution in 3×10^6 nMC-MC sweeps (hMC runs of 250 steps with a 1 fs time step). To accelerate sampling in the MM chain, a temperature of 400 K was used for the T_K and T_U values entering in equation (6.1), while the temperature of the target QM NVT ensemble was kept at 300 K.

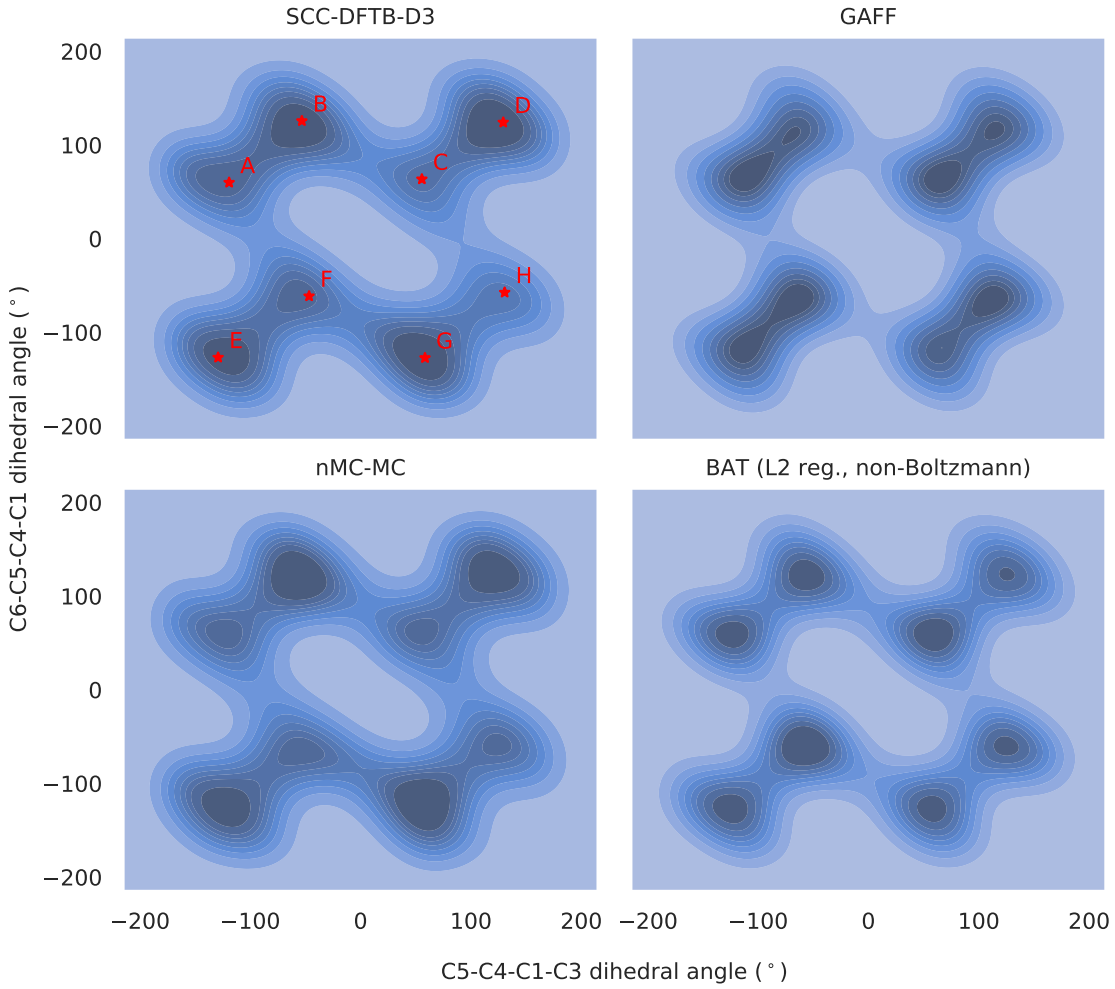

Figure 6.9: Configurational distributions of the C5-C4-C1-C3 *vs.* C6-C5-C4-C1 dihedrals for the fragment of cpd 26. The SCC-DFTB-D3 MD was simulated during 10 ns (snapshots collected every 1 ps), and the GAFF and BAT MD were simulated during 1 μs (snapshots collected every 100 ps). The nMC-MC sampler performed a total of 3×10^6 MC sweeps. The temperature of the simulations was 300 K. The conformations identified on the top left plot are shown in Figure 6.10.

Despite the success of the results obtained when applying the nMC-MC algorithm to the fragment of cpd 26, the main pitfall of this methodology became apparent in this test case. As mentioned before, the larger and more complex the molecule, the more difficult it is to reparameterise to the QM level of theory owing to accumulations of errors that are unavoidable and usually related to FF functional form constraints. There are two possible solutions to this bottleneck if a two-chain nMC-MC algorithm is to be kept. The simpler approach consists of artificially broadening the MM distribution by manipulating the thermodynamic

variables of the MM chain. This would involve, *e.g.*, increasing the temperature of the MM chain such that its energy distribution becomes wider and overlaps to a greater extent with the QM energy distribution. This strategy was successfully applied in past studies,[306,307] though success is not guaranteed if the mismatch between the energy distributions is too large. On the other hand, a more complex but perhaps more reliable method involves developing and employing more accurate low-level models. In this regard, there are different classes of FFs of increasing complexity that can be applied and are still computationally cheap in comparison with the QM calculations. ML potentials are also an attractive option, especially owing to their blindness to functional forms, which make them potentially more accurate than MM FFs.[335,336] Nevertheless, in principle, in an nMC-MC context ML potentials would have to be used alongside a high-level of theory similar to that they were trained to reproduce. It is also possible to use intermediate levels of theory to bridge the gap between the low-level and high-level models. Unfortunately, this solution becomes computationally expensive, especially if hybrid energy models such as $\lambda_i U^{MM} + (1 - \lambda_i) U^{QM}$, where λ_i controls the weight of each energy component in the i-th chain, are used, as they still require high-level calculations to be performed.

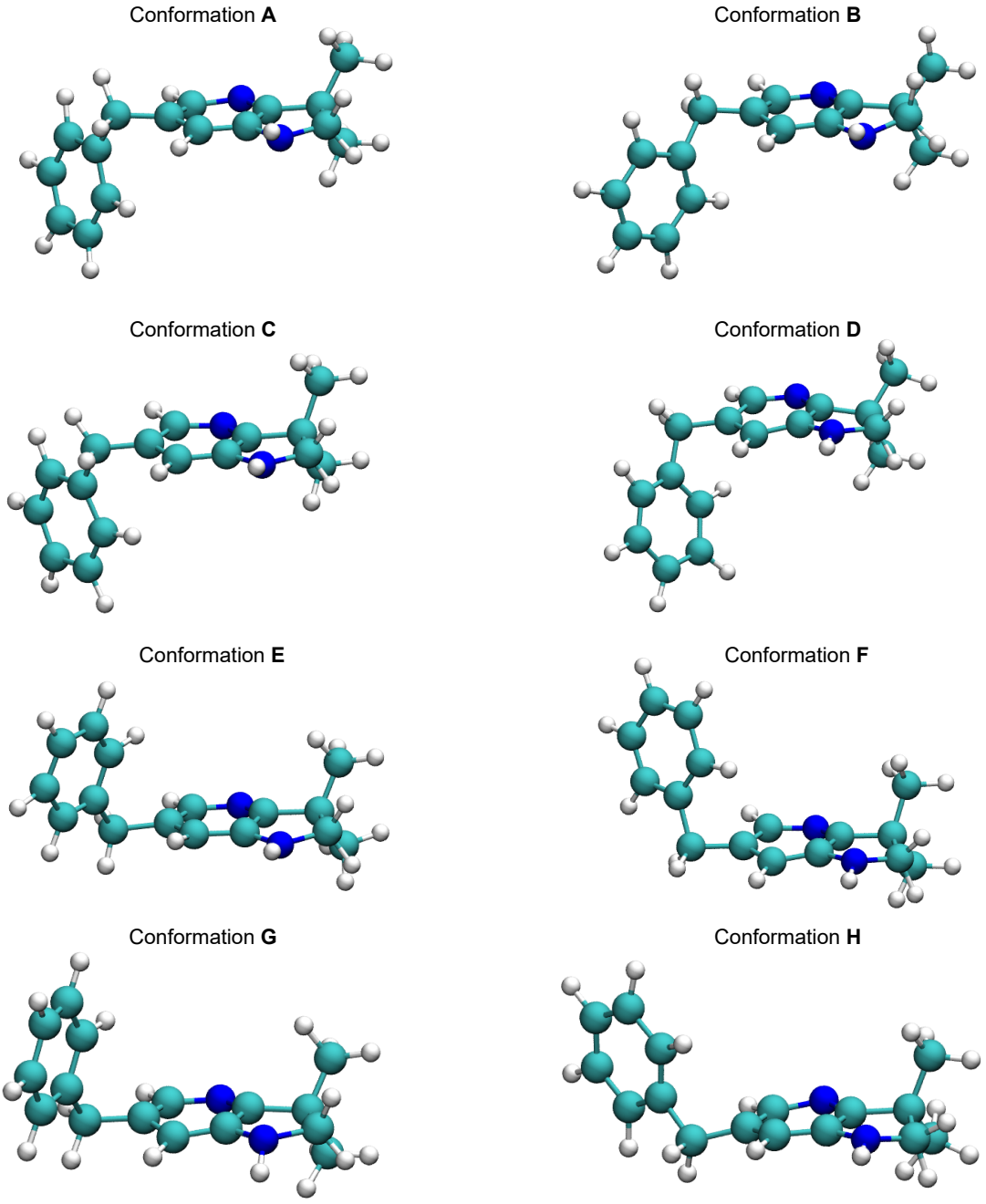

Figure 6.10: Main conformations of the fragment of cpd 26 identified in Figure 6.9.

6.3.4 Analysis of the phase space overlap

To verify the variations in the acceptance rates observed when systematically reparameterising the FFs, let us now turn our discussion to the evaluation of the

phase space overlap between the ensembles generated using the MM FFs and the SCC-DFTB-D3 level of theory. In the following results, all the testing data sets contained 10000 configurations sampled from MM and QM MD simulations. These were performed with the same settings applied when generating the training data sets, except for the temperature, which was set to 300 K to make the results directly comparable to the nMC-MC acceptance rates. Note that phase space is here employed as a synonym of configuration space as we only consider situations that compare total energy distributions at the same temperature, thus making the momentum coordinates irrelevant (see proof in the Appendix B, Section B.1). Owing to this, even though these metrics were initially presented as depending on the total energy, in practical terms only the potential energy was used in their calculation.

The most direct and robust metric to measure the phase space overlap between ensembles obtained using different levels of theory is given by equation (6.19). In the context of this study, it required performing MD simulations with the MM and QM Hamiltonians and, subsequently, evaluating the energies of each ensemble at both the MM and QM levels of theory. The potential energy difference between the QM and MM levels for structures sampled using SCC-DFTB-D3, $\Delta U_{QM}^{QM\rightarrow MM} = U_{QM}^{MM} - U_{QM}^{QM}$, and minus the potential energy difference between the MM FF and the QM level of theory for structures sampled using the FF, $-\Delta U_{MM}^{MM\rightarrow QM} = -\left(U_{MM}^{QM} - U_{MM}^{MM}\right)$, were then calculated, and the corresponding histograms determined. The resulting probability distributions were translated along the ΔU axis so that $\Delta U = 0$ was the midpoint between the two distribution means. Each histogram was fitted to a Gaussian function, and the overlap obtained between the Gaussians was evaluated numerically and used as an estimation of the phase space overlap. Any structure for which the absolute difference of its energy relative to the average energy of the respective distribution was larger than 100 kJ mol^{-1}, *i.e.* $|U - \langle U \rangle| > 100$, was removed from the data set.

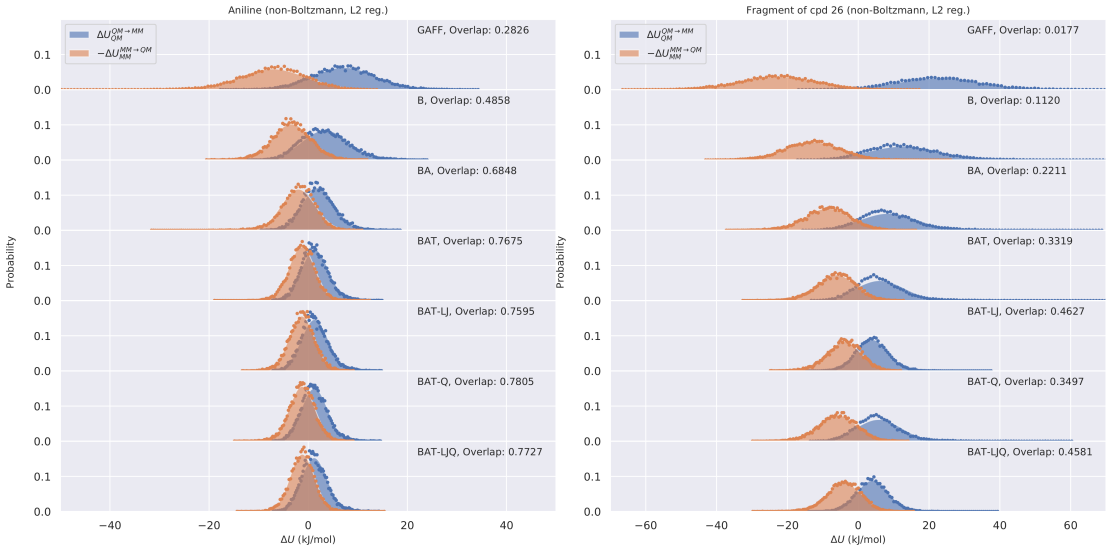

Figure 6.11: Energy difference histograms of MM→QM and QM→MM for aniline (left) and the fragment of cpd 26 (right). The distributions were translated along the ΔU axis so that $\Delta U = 0$ was the midpoint between the means of the two distributions.

The energy difference histograms and phase space overlaps for the two molecules in which we have primarily focused our discussion thus far, *viz.*, aniline and the fragment of cpd 26, are shown in Figure 6.11 (non-Boltzmann weighting, L2 regularisation). In these plots, it can be seen that the energy difference distributions are well approximated by Gaussian functions, and that the phase space overlap increased from the B FF to the BAT-LJQ FF, as observed in the acceptance rates of Figure 6.3.

Using the nMC-MC switching step acceptance rate as a metric of the similarity between the MM and QM levels of theory necessarily requires establishing a strong correlation between the acceptance rates of equation (6.9), θ, and the phase space overlap of equation (6.19), Ω. To assess the degree of correlation between both measurements, we computed the linear regressions of four sets of data: uniform-weighted data, non-Boltzmann-weighted data, non-Boltzmann-weighted L2-regularised data, and all data. From the results shown in Figure 6.12, it is clear that there is a high degree of correlation between θ and Ω. Although these linear fittings are only an approximation of the true correlation between both measurements, we consider that the observed correlations are

close enough to linear behaviour to be considered as so. Owing to this, a direct comparison between R^2 of different fittings is avoided. Nevertheless, for the uniform-weighted data set, there is an outlier located at $(\theta \approx 0.20, \Omega \approx 0)$, for which the nMC-MC acceptance rate was significant, but the phase space overlap was estimated to be practically 0. A close inspection revealed that this data point corresponds to the BAT-Q uniform-weighted L2-regularised FF of acetanilide. This observation is in line with the results obtained in Chapter 5, which reports situations in which reparameterisation of charges using uniform weighting also led to a decrease in the FF quality.[248] Interestingly, the nMC-MC algorithm still allowed high acceptance rates to be obtained for problematic FFs with nonphysical parameters if the hMC runs were short enough to prevent the molecules from converting to the spuriously stable, nonphysical geometries in which they got trapped in regular MD simulations. Overall, the results show that the nMC-MC acceptance rate is a robust metric of the phase space overlap. Therefore, it can be employed to evaluate the similarity between the levels of theory used in the low- and high-level chains of the nMC-MC algorithm

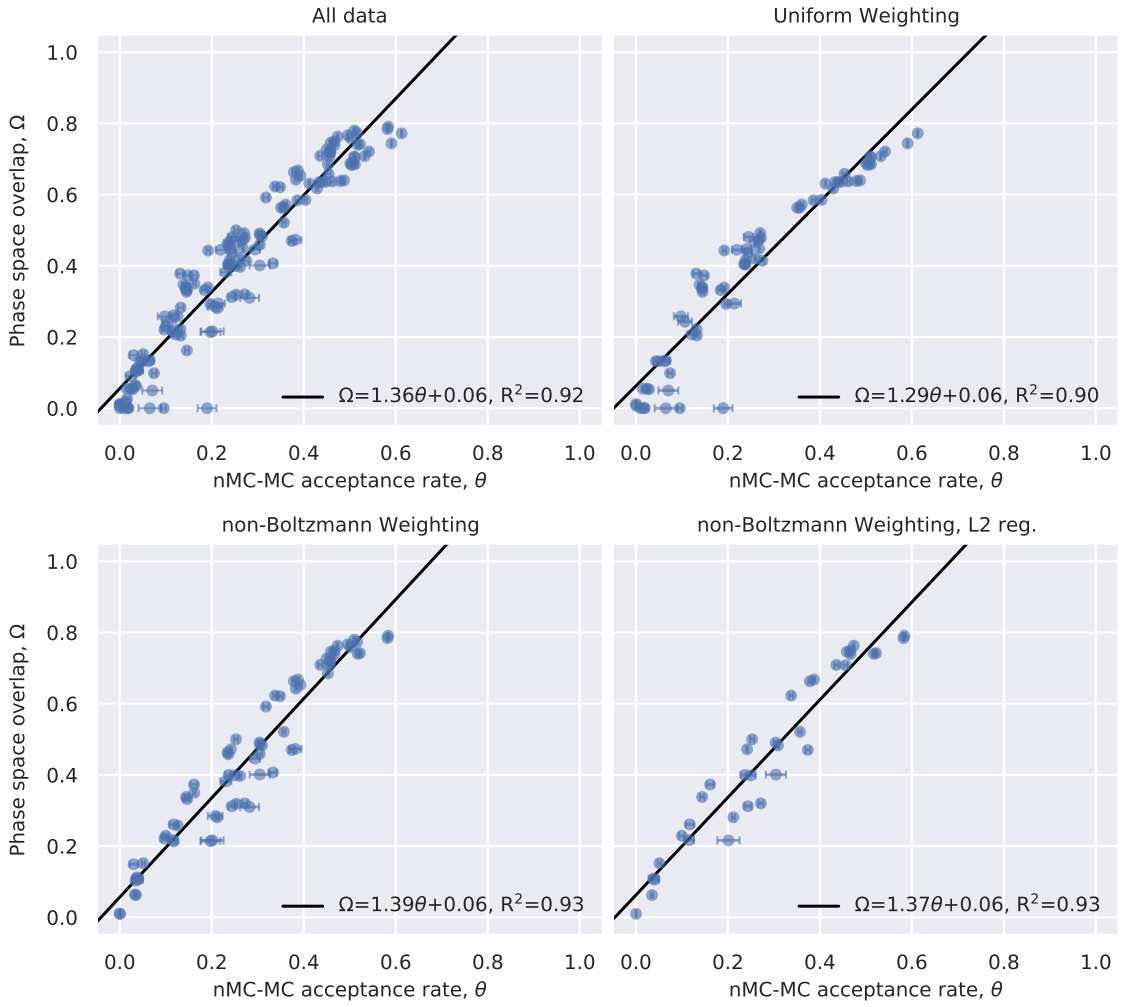

Figure 6.12: Correlation between the nMC-MC acceptance rate, θ, as given by equation (6.9), and the phase space overlap, Ω, as given by equation (6.19), for 4 different data sets: all data (top left), not-regularised data (top right), L2-regularised data (lower left), and non-Boltzmann-weighted L2-regularised data (lower right). The GAFF data points are included in the "all data" data set.

Before analysing the results obtained for the Wu and Kofke metrics given by equations (6.20) and (6.21), it is useful to explain their physical meaning. Since these metrics are a measure of the offset of an energy distribution with respect to another, they provide insights into how much the MM-sampled phase space lies inside the QM-sampled phase space. Firstly, it is important to notice that when $0 \leq \Sigma_{QM,MM} < 1$, the probability distribution of the MM energies observed for a simulation performed using the QM Hamiltonian, ρ_{QM}^{MM}, is centred right relative to the probability distribution of the MM energies observed for a simulation performed using the MM FF, ρ_{MM}^{MM}. This means that the QM high-energy

structures that lay above ρ_{MM}^{MM} are unimportant to the MM FF or that the MM FF low-energy structures that lay below ρ_{QM}^{MM} are undersampled by the QM Hamiltonian. On the other hand, when $1 < \Sigma_{QM,MM} \leq 2$, the probability distribution of the MM energies calculated for a simulation performed using the QM Hamiltonian, ρ_{QM}^{MM}, is centred left relative to the probability distribution of the MM energies calculated for a simulation performed using the MM FF, ρ_{MM}^{MM}. This means that the QM Hamiltonian preferentially accesses a small set of structures that are either not sampled by the MM FF or, if energetically favourable, are entropically disfavoured, or that the MM FF samples high-energy structures that are unimportant to the QM Hamiltonian. Identical reasoning can be applied to $\Sigma_{MM,QM}$.

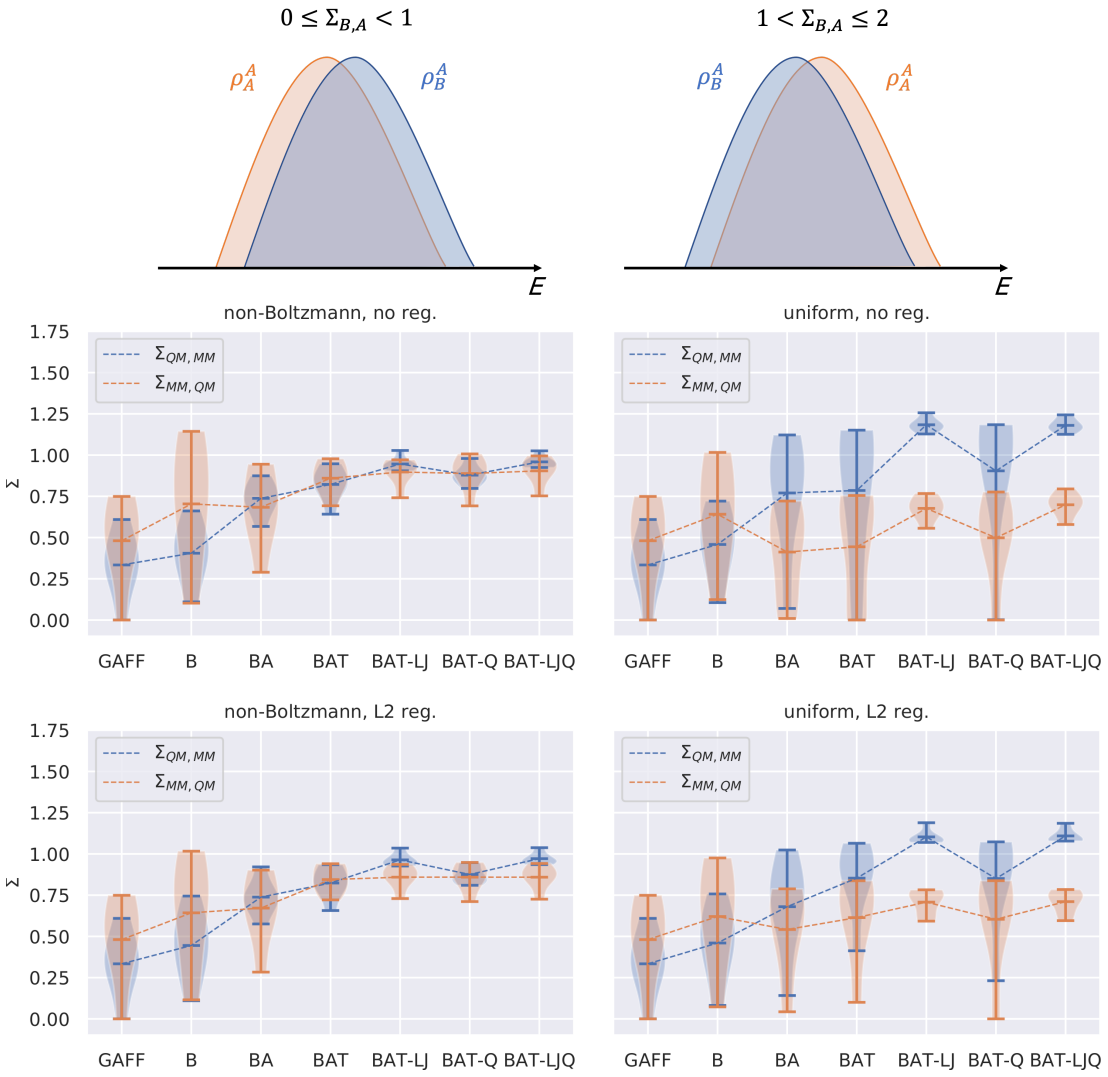

Figure 6.13: Violin plots showing the distribution of the Wu and Kofke overlap metrics between the MM and QM levels of theory, as given by equations (6.20) and (6.21), for all molecules represented in Figure 6.2. The solid lines indicate the mean and the extrema of the distribution for each type of reparameterised FF, and the dashed lines connect their mean values. Four data sets are represented: non-Boltzmann-weighted not-regularised data (top left), uniform-weighted not-regularised data (top right), non-Boltzmann-weighted L2-regularised data (lower left), and uniform-weighted L2-regularised data (lower right).

From the results presented in Figure 6.13, which shows the average behaviour of $\Sigma_{QM,MM}$ and $\Sigma_{MM,QM}$ for all molecules in the test set, it can be seen that for the non-Boltzmann-weighted FFs, $\Sigma_{QM,MM}$ starts with a value close to 0.4 (B FF), likely meaning that the energy of high-energy QM structures was overestimated in the MM FF or that the MM FF sampled spurious minima. Nevertheless, this mismatch progressively diminishes with systematic reparameterisation, since

$\Sigma_{QM,MM}$ gets closer to 1, showing how well the most refined MM FFs predicted the energy of structures generated through an MD performed using the QM Hamiltonian. A similar trend is observed for $\Sigma_{MM,QM}$, though this quantity reaches a plateau at $\Sigma_{MM,QM} \approx 0.8$ for the most refined FFs (BAT, BAT-LJ, BAT-Q, and BAT-LJQ). This combination of $\Sigma_{QM,MM}$ and $\Sigma_{MM,QM}$ values indicates that systematic reparameterisation of the FFs turned an overlap relation into a subset relation between the phase space distributions explored by the MM and QM levels of theory, supporting the idea that the MM distributions of the most refined FFs were somewhat broader than their QM counterparts since the former sampled high-energy structures that were unimportant to the QM FF.

An identical interpretation follows for the plots concerning the uniform-weighted FFs. Worthy of note are the BAT-LJ and BAT-LJQ FFs, for which $1 < \Sigma_{QM,MM} \leq 2$ and $0 \leq \Sigma_{MM,QM} < 1$. This is a case of special concern because it means that these FFs either sampled spurious high-energy structures or undersampled the QM minima. Finally, even though not observed in this study, $0 \leq \Sigma_{QM,MM} < 1$ and $1 < \Sigma_{MM,QM} \leq 2$ is a hugely undesirable case, as it indicates that an FF either sampled spurious minima or overstabilised the QM minima, situations that can potentially lead to trapping of MD simulations in overstabilised basins. Likewise, the case in which $1 < \Sigma_{QM,MM} \leq 2$ and $1 < \Sigma_{MM,QM} \leq 2$ is also of concern, as it suggests that an FF not only undersampled the true QM minima but also sampled other spurious minimina.

6.3.5 Self-parameterising nMC-MC

As proof of principle, we tested the self-parameterising methodology that iteratively couples the nMC-MC algorithm with a parameterisation step. This algorithm allows on-the-fly derivation of optimally tuned FFs owing to its capability of performing sampling of relevant configurations and subsequent optimisation of the FF parameters, all in one scheme. Specifically, we used the nMC-MC algorithm to sample QM configurations in such a way that a configuration belonging to the QM ensemble was added to the FF training data set for

every 5 configurations accepted into this QM ensemble. Configurations belonging to the MM ensemble or rejected configurations of any kind were discarded for reparameterisation purposes. Despite this choice, in some situations they can be informative for the optimisations and, inclusively, accelerate the convergence of the self-parameterising procedure.[312] Furthermore, every time 500 new QM structures were added to the existing training data set, a new reparameterisation of the FF parameters was performed using the total training data. The temperature used for the Markov chains and the Maxwell-Boltzmann distribution was 300 K. The self-parameterising nMC-MC procedure was deemed to be converged when the RMSD of the FF parameters between two successive iterations was less than 10^{-4}.

The molecule used in this application was octahydrotetracene (see molecular structure in Figure 6.14). This choice relied on the fact that this scaffold was previously identified as a challenging case for FFs, which struggle in reproducing its QM energies.[330] All bonded parameters were optimised in every optimisation, such that the vector of optimisable parameters was given by $\boldsymbol{p} = \left(\boldsymbol{K_b}, \boldsymbol{r_{eq}}, \boldsymbol{K_\theta}, \boldsymbol{\theta_{eq}}, \boldsymbol{V_n}, \boldsymbol{\gamma_n}\right)$. The total number of optimisable parameters was 72. Using the nomenclature of the previous examples, this corresponded to generating a BAT-type FF. The objective function included energy, force, and regularisation terms, as given by equations (5.3), (5.5), and (5.13), respectively. Uniform weighting was applied to weight the conformations. In contrast to the previous applications, AMBER atom-type symmetries were preserved, and the prior widths used in the regularisation were estimated from the arithmetic mean for each class of parameters, a feature included in ParaMol.

The results obtained for the self-parameterising nMC-MC calculation of octahydrotetracene are shown in Figure 6.14, in which we see that the acceptance rate increased smoothly and monotonically, progressively stabilising as the nMC-MC sweeps increased. The behaviour of the RMSD plot is somewhat more irregular, with sudden jumps that are explained by the necessity of the optimisation to adapt the parameters to a new set of configurations. Although this was not problematic in this example, in other applications large variations of the RMSD

may lead to a premature ending of the self-parameterising procedure, which is something that may be resolved by employing tighter thresholds. The algorithm took in total 11 iterations to converge, resulting in a final training data set of 5500 QM structures. The convergence of the procedure is confirmed by analysing the components of the objective function as a function of the iteration number (see Appendix B, Figure B.1).

Finally, to test the quality of the derived FF, we generated a BAT-type FF using a training data set containing a total of 10000 QM structures. This FF was derived following the same philosophy applied in the previous examples: firstly, we built the training data set using Langevin dynamics at a temperature of 300 K; afterwards, we optimised the FF. By comparing the blue and green lines of the top panel of Figure 6.14, we conclude that both approaches led to identical acceptance rates of *ca.* 15-16%, strongly indicating the robustness of the self-parameterising nMC-MC procedure. Overall, this self-parameterising algorithm is quite appealing since, by combining sampling and parameterisation in one scheme, it does not require *a priori* generation of a training data set of unknown size, therefore limiting the computational work to what is strictly necessary.

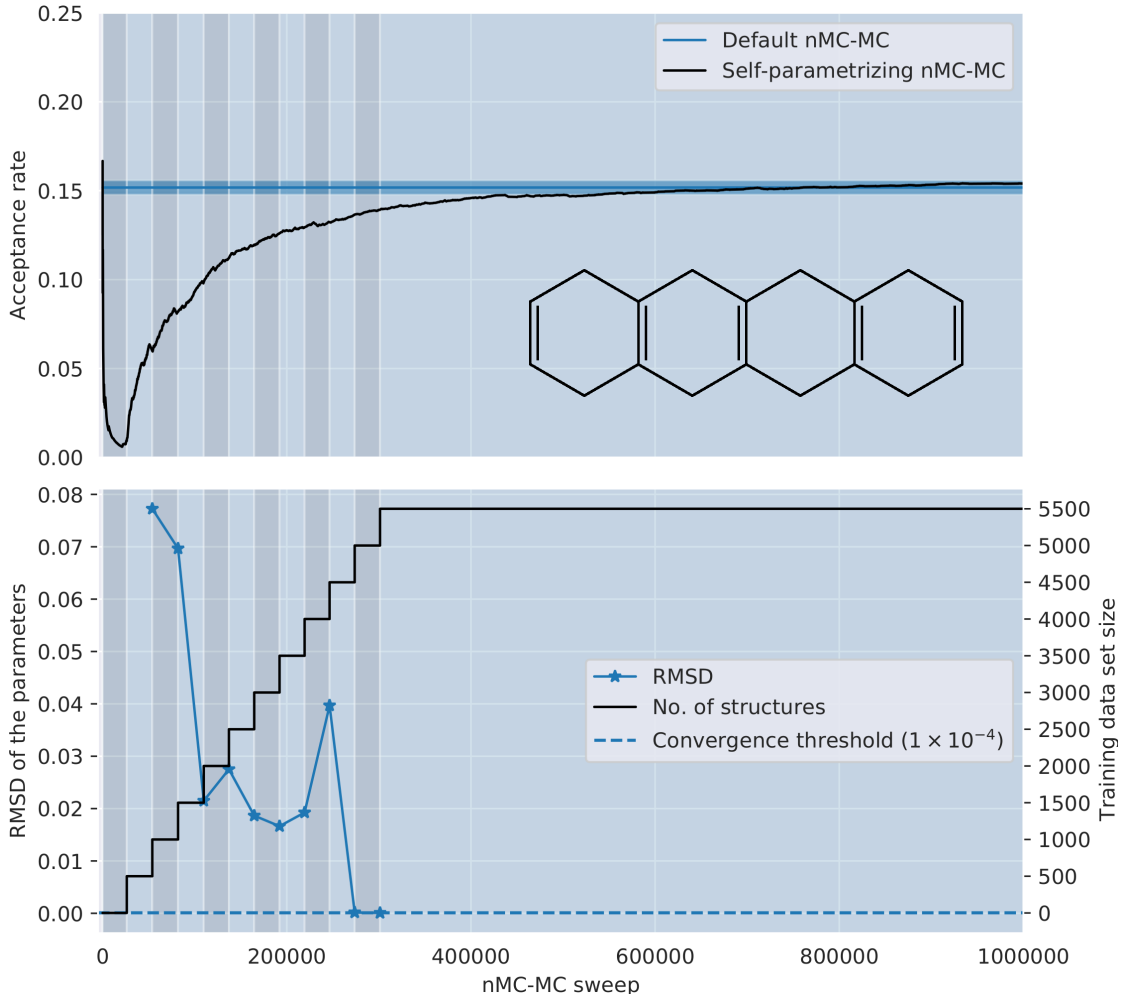

Figure 6.14: Top panel: Acceptance rates of the self-parameterising nMC-MC procedure as a function of the nMC-MC sweep for octahydrotetracene. The nMC-MC acceptance rate and standard deviation of the FF derived following the same philosophy applied for the test cases of Figure 6.2 are also shown. The background shading indicates different iterations of the procedure. Bottom panel: Plot of the RMSD of the FF parameters (left axis) and of the total number of structures in the training data set (right axis) as a function of the nMC-MC sweep.

6.3.6 nMC-MC sampling into a QM/MM Hamiltonian

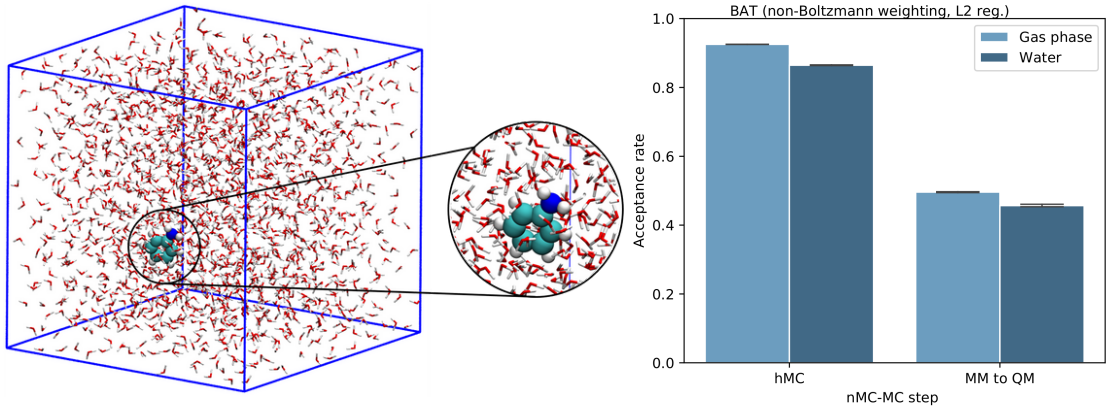

Figure 6.15: Left: Snapshot of the nMC-MC simulation of aniline in water. Right: Comparison of the hMC and switching step acceptance rates obtained for aniline in the gas phase and aqueous solution.

As a final application, we immersed the aniline molecule in a TIP3P water box (the total system had 5327 atoms) and equilibrated the system in the NPT ensemble for 1 ns at 300 K using the Langevin integrator[337] (time step of 1 fs and friction coefficient of 2 ps^{-1}). The pressure of the system was maintained at 1 bar using the Monte Carlo barostat[135,338] implemented in OpenMM. Periodic boundary conditions were applied and long-range electrostatics were handled by the PME method.[221,224] The cutoff applied to all nonbonded interactions was 12 Å. The final configuration of the equilibration run, which had a box size of $37.38 \times 37.26 \times 38.55$ $Å^3$, was subsequently used as the starting point for a set of 4 NVT nMC-MC simulations in which the MM system was used in the low-level chain and a QM/MM model was employed in the high-level chain. The MM model used for aniline was the previously derived non-Boltzmann-weighted L2-regularised BAT FF.

In a system composed of a ligand in solution, the energy of the total MM system is given by

$$U^{MM}(\boldsymbol{q}^s, \boldsymbol{q}^l, \boldsymbol{q}^{s-l}) = U^{MM}_{sol}(\boldsymbol{q}^s) + U^{MM}_{lig}(\boldsymbol{q}^l) + U^{MM}_{lig-sol}(\boldsymbol{q}^{s-l}) \tag{6.22}$$

where U_{sol}^{MM} is the energy of the solvent (TIP3P waters), U_{lig}^{MM} the energy of the ligand (aniline), $U_{lig-sol}^{MM}$ is the ligand-solvent (aniline-TIP3P waters) interaction energy, and $\boldsymbol{q}^{s}$, $\boldsymbol{q}^{l}$ and $\boldsymbol{q}^{s-l}$ are the DOFs of the solvent, ligand, and the interaction between them. The QM/MM energy of a system in which only the ligand is included in the QM region and there are no covalent bonds between the ligand and solvent reads[339,340]

$$U^{QM/MM}(\boldsymbol{q}^{s},\boldsymbol{q}^{l},\boldsymbol{q}^{s-l}) = U_{sol}^{MM}(\boldsymbol{q}^{s}) + U_{lig}^{QM}(\boldsymbol{q}^{l}) + U_{lig-sol}^{MM}(\boldsymbol{q}^{s-l}) \tag{6.23}$$

where the only difference with respect to equation (6.22) is that now the potential energy of the ligand, U_{lig}, is calculated at the QM level. Note that the interaction between the MM and QM regions, $U_{lig-sol}^{MM}$, is still calculated at the MM level. For the present test case, the point charges used to calculate this interaction term did not change during the simulation in the QM/MM Hamiltonian of equation (6.23). This corresponds to a mechanical embedding model with fixed-point charges in the QM region.[341] Consequently, by combining equations (6.22) and (6.23) we obtain that the $\Delta\Delta U$ term that has to be introduced in the nMC-MC acceptance rate of equation (6.9) is given by

$$\begin{aligned}\Delta\Delta U(\boldsymbol{q}_i,\boldsymbol{q}_f) &= \left[U^{QM/MM}(\boldsymbol{q}_f) - U^{MM}(\boldsymbol{q}_f)\right] - \left[U^{QM/MM}(\boldsymbol{q}_i) - U^{MM}(\boldsymbol{q}_i)\right] \\ &= \left[U_{lig}^{QM}(\boldsymbol{q}_f^{l}) - U_{lig}^{MM}(\boldsymbol{q}_f^{l})\right] - \left[U_{lig}^{QM}(\boldsymbol{q}_i^{l}) - U_{lig}^{MM}(\boldsymbol{q}_i^{l})\right]\end{aligned} \tag{6.24}$$

from which we conclude that the switching step from the MM to the QM/MM Hamiltonian only requires the calculation of the energies of the ligand at the MM and QM levels. Equation (6.24) was employed in this test case to sample the QM/MM distribution of aniline in a box of TIP3P waters.

The acceptance rates obtained are shown in Figure 6.15. As expected, the hMC acceptance rate decreased when going from the gas phase to water solvent because

of the increase in system size. There was also a small but significant decrease of *ca.* 4% in the switching step acceptance rate. The successful application of the gas-phase-derived BAT FF is attributed to the fact that the conformational preferences of aniline in aqueous solution (see Figure 6.16) were well captured by the nonbonded interactions. If this were not the case, a possible solution would involve reparameterising GAFF using a training data set consisting of conformations extracted from explicit solution simulations.

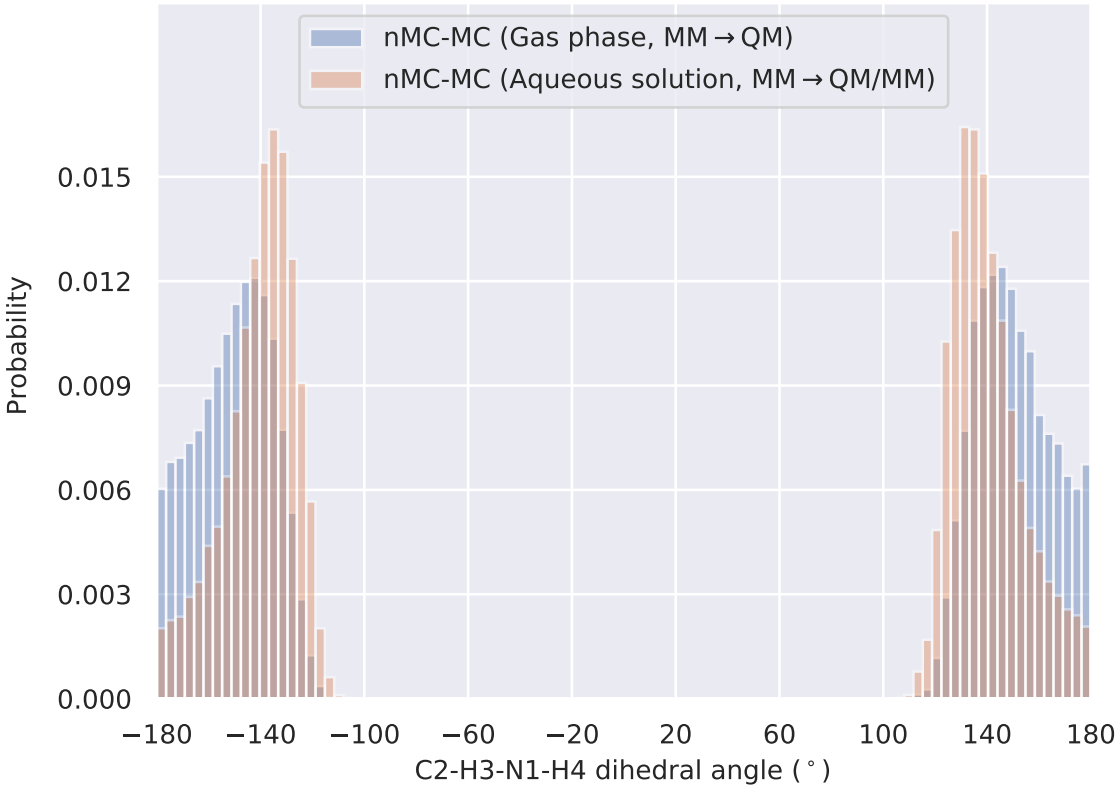

Figure 6.16: Comparison of the nMC-MC-sampled configurational distributions of aniline in the gas phase and in aqueous solution.

In this application example, we used a fixed set of charges for the QM region, which not only simplified the acceptance rate equation but also increased the similarity between the MM and the QM/MM models. Nevertheless, in QM/MM calculations, it is common to consider a QM region with varying partial charges, usually derived using a least-squares fitting to the QM electrostatic potential.[211,255,256] Although ParaMol provides the tools to parameterise nonbonded parameters in solution, from our experience, even in a mechanical embedding context, in which the electrostatic coupling between the MM and

QM regions is calculated at the MM level, a ligand with a varying set of MM partial charges is already a challenging system for the nMC-MC algorithm, as the term $U_{lig-sol}^{MM'}(\boldsymbol{q}^{s-l}) - U_{lig-sol}^{MM}(\boldsymbol{q}^{s-l})$, where MM' is the FF with on-the-fly fitted charges and MM is the FF with original charges, does not vanish as it does in equation (6.24). Unfortunately, since energy is an extensive property, the larger the simulation box, the larger the mismatch between the QM/MM and MM levels of theory. Therefore, this QM/MM framework commonly leads to large energy differences in the ligand-solvent interaction energy, which make the nMC-MC algorithm unviable for all but the simplest cases.

It is also important to stress that, besides conformational changes, there are important nonadditive electrostatic effects that cannot be properly described by using a single set of charges. For example, the electrostatic embedding scheme, in which the electrostatic contribution from the MM subsystem is included in the QM Hamiltonian, poses additional difficulties that worsen the mismatch between the MM and QM representations of the ligand due to polarisation by the solvent. Hence, owing to the dynamic nature of the electrostatic cloud, these are typical applications for which fixed point-charges FFs are unsuitable. FF models that include descriptions of polarisation and hyperconjugation may prove useful for these applications, such as, *e.g.*, AMOEBA,[201,202] polarisable CHARMM[203,204] or fluc-q.[205–207] The "electron spill-out" problem[342,343] is also a well-known pitfall of the electrostatic embedding scheme that may artificially distort the electron density of the QM region, thus increasing the mismatch between the energy of the QM/MM and MM models. So far, our research has set the mechanical-embedding fixed-point charge QM/MM model used above as the limiting case for successful sampling using the nMC-MC algorithm. In general, further complexity of the high-level models seems to be unsuitable to be reproduced by simple fixed point-charge FFs.

6.4 Conclusions

In this chapter, we have presented a multilevel procedure that allows estimation of quantum configurational ensembles while keeping the computational cost at a minimum. This work is of paramount importance for conformational analysis because it combines the feasibility of computationally cheap methods, such as MM FFs, with the accuracy of the more expensive QM level. The algorithms presented are implemented and made available in the ParaMol, free software that aims to ease the process of parameterisation of MM FFs.[4] The code can be found at `https://github.com/JMorado/ParaMol`, and examples of how to use it are available through ParaMol's website, `https://paramol.readthedocs.io`.

The presented methodology involves coupling the hMC algorithm with a switching step between two Markov chains, the latter as formalized by Gelb. In the context of this work, the low-level Markov chain corresponded to a GAFF-like MM FF in which sampling of configurations was performed, whereas the high-level Markov chain was the SCC-DFTB-D3 level of theory to which conformations were periodically attempted to be sampled. Owing to the low energetics similarity between GAFF and SCC-DFTB-D3, a straightforward application of the methodology led to very slow convergence of the target configurational distributions due to low acceptance rates. Therefore, we resorted to FF reparameterisation as a means of ensuring sufficient overlap between the MM and QM levels of theory. We demonstrated this to be a successful strategy of generating more QM-like FFs and, consequently, increasing the nMC-MC switching step acceptance rates, thus accelerating the convergence of the sampling of the target quantum configurational distribution.

Overall, systematic reparameterisation of FFs proved to be an efficient strategy to increase the acceptance rates of the switching step from the MM to the QM level of theory. The best acceptance rates were obtained for aniline (*ca.* 65%), whereas the molecule with the lowest possible acceptance rate was the fragment of cpd 26. This is expected since both molecular size and chemical complexity have an impact on the acceptance rates. Moreover, we determined

that the optimal reparameterisation recipe involves employing non-Boltzmann weighting alongside L2 regularisation. Uniform-weighted FFs, especially without regularisation, are to be avoided as they easily sacrifice physicality in the FF parameters to obtain the best possible fit. These observations are in line with the conclusions of the study presented in Chapter 5. The systematic parameterisation also showed that hard DOFs, such as bonds and angles, are crucial to be reparameterised to increase the acceptance rates, mainly due to their large force constants. Reparameterisation of charges with the uniform weighting scheme seems to be deleterious to the quality of the FFs. The acceptance rates data were supported by information obtained from various phase space overlap metrics. These metrics revealed further insights into the features of the weighting methods, leading us to suggest the switching step acceptance rates as a robust metric of phase space overlap.

We also presented a self-parameterising algorithm that combines sampling and FF parameterisation in one scheme. This method does not require *a priori* generation of a training data set of unknown size, thus limiting the computational work to the strictly necessary. We illustrated its *modus operandi* and showed that it gives identical results to the standard approach.

Finally, we also applied the nMC-MC algorithm to generate the QM/MM distribution of a ligand in aqueous solution. We proved that within a fixed-point charge mechanical embedding framework, the nMC-MC algorithm is a viable methodology that permits recovery of the target QM/MM configurational ensemble. This application example also provided useful guidelines for future research efforts because it illustrates the limitations of using a GAFF-like MM FF as the low-level model. Since this FF has fixed-point charges, it appears to be generally unsuitable for application in contexts involving varying solute charges, which may occur either due to conformational changes or polarisation originating from electrostatic embedding. A possible solution for this bottleneck may involve resorting to polarisable FFs or machine-learning models.

The next chapter attempts to answer the question: ”does a machine-learnt potential perform better than an optimally tuned traditional force field?”. Having developed a method to parameterise druglike molecules in Chapter 5, and an algorithm to generate quantum configurational ensembles at a low computational cost in this chapter, we apply these techniques to derive optimally tuned FFs, which are tested against an ML potential. We then evaluate the performance of a standard FF, an optimally tuned FF, and an ML potential in the modelling of a set of γ-fluorohydrins. We assess the performance of each model by comparing its predictions to those obtained from QM methods and experiments. The current strengths and shortcomings of each model are analysed, from which guidelines for improvement are drawn.

Chapter 7

Does a Machine-Learnt Potential Perform Better Than an Optimally Tuned Traditional Force Field? A Case Study on Fluorohydrins

In this chapter, we present a comparative study that evaluates the performance of a ML potential (ANI-2x), a conventional FF (GAFF), and an optimally tuned GAFF-like FF in the modelling of a set of 10 γ-fluorohydrins that exhibit a complex interplay between intra- and intermolecular interactions in determining conformer stability. To benchmark the performance of each molecular model, we evaluated their energetic, geometric, and sampling accuracy relative to QM data. This benchmark involved conformational analysis both in the gas phase and chloroform solution. We also assessed the performance of the aforementioned molecular models in estimating nuclear spin-spin coupling constants by comparing their predictions to experimental data available in chloroform. The results and discussion presented in this study highlight the strengths and weaknesses of each model, providing guidelines for future development of force fields and machine learning potentials.

7.1 Introduction

Over the years, many methods have been developed to describe the potential energy of small organic molecules. Researchers have been attempting to find a compromise between accuracy and computational cost, and a balance between the time scale that simulations can achieve and the size of the systems being simulated. Despite many efforts and advances towards this equilibrium, in general there is still a positive correlation between computational cost and accuracy, and these two properties are often negatively correlated with system size and simulation time scale. The gold-standard method in ligand modelling remains to be quantum mechanics. QM methods approximate the Schrödinger equation using wavefunction-based methods or DFT. From the wavefunction-based methods, the coupled-cluster methods[45,344] are thus far the most accurate for quantum chemistry applications. Unfortunately, owing to the high computational cost of QM methods, their use in *ab initio* simulations is limited to all but the simplest systems, despite recent advances that attempt to combine the sampling efficiency of cheap, approximate potentials with the accuracy of the quantum level.[285,298–300,345] A less computational demanding and widely used alternative are the MM FFs. FFs resort to classical, empirical functions to describe the potential energy of systems. The popularity of FFs resides in their ability to simulate systems containing thousands of atoms on time scales that can go up to the millisecond.[204,346] The main drawback of FFs comes from the same feature that confers them their strength: the simplicity of their functional form, while computationally attractive, is often unsuited to model complex chemistry and challenging chemical interactions, a constraint that is further aggravated by the requirement of a sometimes unknown set of FF parameters. Although FF parameters are usually available for many classes of molecules at a satisfactory degree of accuracy, novel chemical entities, of which ligands are a striking example, frequently demand derivation of new FF parameters.[4,229,330,347] FF parameterisation, however, is neither trivial nor straightforward in many applications. Recently, neural network potentials (NNP) have emerged as a promising

ML alternative to FFs.[348–350] NNPs learn the QM energy of an atom in its surrounding chemical environment, requiring neither a FF functional form nor FF parameters to work. For these reasons, NNPs are generally readily transferable to classes of compounds similar to those included in the training data set. The accuracy of NNPs when applied to chemical environments outside the training data set, however, is unpredictable and should be evaluated beforehand. Furthermore, although NNPs computational cost is much smaller than that of the QM methods (*ca.* 10^6 times), even with GPU-accelerated computing they are still considerably more (up to 100 times) computationally expensive than conventional FFs, limiting the size of the systems that can be simulated and the simulation time scales that can be achieved.

In this chapter, we evaluate the accuracy of an NNP, a conventional class I FF, and an optimally tuned FF in the modelling of a set of 10 γ-fluorohydrins (Figure 7.1). In 2015, Linclau *et al.* used these molecules to demonstrate for the first time the occurrence of OH-F intramolecular hydrogen bonds (IMHBs) in acyclic saturated γ-fluorohydrins.[351] This set of molecules exhibits a complex interplay between intra- and intermolecular interactions in determining conformer stability, making it an interesting test case to study. We benchmarked the performance of the aforementioned molecular models both in the gas phase and chloroform solution by comparing their predictions to both experimental (NMR J-coupling constants) and theoretical data (energies and geometries).

The NNP we tested in this study was ANI-2x,[352] a model from the ANI family[336,348,353–355] that has been trained to reproduce the ωb97X[70]/6-31G* level of theory. Specifically, ANI-2x was derived using a data set comprising 8.9 million molecular conformations, and it includes active learning refinements to torsional profiles, nonbonded interactions, and bulk water behaviour. This NNP has already been successfully applied in binding free energy calculations,[356] in the description of the torsional[357,358] and bond potential energy surfaces,[358] and integrated into fast FF parameterisation protocols as the reference level of theory.[359] Furthermore, as our traditional FF we used the GAFF,[182] which is commonly employed in the modelling of druglike molecules.[360] Finally, our

optimally tuned FF was an optimised version of the GAFF in which bonded parameters were optimised to the same level of theory that ANI-2x was trained to reproduce (ωb97X/6-31G*).

MD was used to sample the molecular conformations. Since the NNP was used only to model the intramolecular interactions of the ligand, the simulations in chloroform employed a model that embeds the NNP inside a conventional MM FF. In this approach, the ligand intramolecular interactions are treated at the ML level, whereas the ligand-solvent intermolecular interactions and solvent-solvent interactions are treated using MM. This hybrid NNP/MM strategy excludes any polarisation of the solute by the solvent and thus corresponds to a mechanical embedding model[341] with fixed-point charges in the ML region. It has already been applied in past studies,[356,357,361–363] being close in philosophy to that of QM/MM models, but with the NNP in place of the QM method.

This chapter is structured as follows: we first describe the basic theory and methods underlying the present study, *viz.*, the ANI-2x NNP, the FF reparameterisation protocol, the hybrid NNP/MM model, the details regarding the MD simulations, and the procedure used to calculate the populations of conformers and the NMR J-coupling constants. We then present a thorough analysis of the performance of the tested molecular models both in the gas phase and chloroform solution and, finally, conclude with some final remarks and future work.

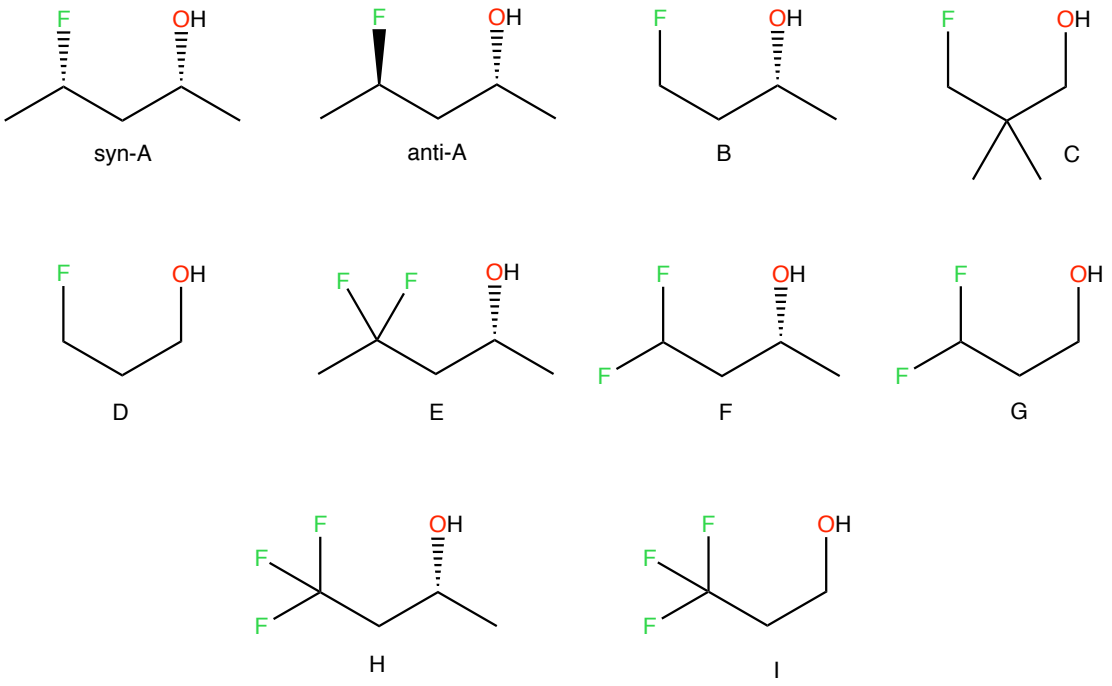

Figure 7.1: List of γ-fluorohydrins studied in this study.

7.2 Theory and methods

7.2.1 The ANI-2x neural network potential

NNPs are currently one of the most promising potentials to be used in place of MM FFs to model the intramolecular interactions of ligands. There are currently two models of the ANI family that may be applied to a broad spectrum of problems in chemical sciences. ANI-1ccx, which is trained to reproduce CCSD(T*)/CBS, is the ANI NNP with the highest level of accuracy,[357,361,362,364] although it can only simulate organic molecules containing elements H, C, N, and O. ANI-2x, on the other hand, is trained to reproduce ωb97X/6-31G* and has also shown promising results in various applications.[356,357] ANI-2x has the advantage of extending the chemical space covered by ANI-1ccx to organic molecules also containing elements F, S, and Cl. Owing to this, ANI-2x covers a chemical space that encompasses 90% of druglike molecules[352] and is the only

ANI model that can be employed to simulate the γ-fluorohydrins considered in this study due to the presence of the fluorine atoms.

The ANI-2x training data set was composed of molecules obtained from different sources, *viz.*, the GDB-11[365,366] and ChEMBL[367] databases, the s66x8 benchmark,[368] as well as sulfur-containing amino acids and dipeptides randomly generated using RDKit.[277] In total, 8.9 million molecular conformations were used. These geometries were sampled through active learning algorithms, designed to sample the relevant chemical space, refine torsions and nonbonded interactions, and improve bulk water behaviour. In these active learning procedures, non-equilibrium conformations were generated using dimer sampling, normal-mode sampling, ensembles of MD simulations, and ML-driven torsion sampling.

ANI NNPs overcome the requirement of an analytical FF functional form and a set of FF parameters by learning the QM energy of an atom i, U_i, in its surrounding chemical environment. The sum of the individual atomic energies yields the total potential energy, U, of a given molecular species, *i.e.*

$$U(\boldsymbol{R}) = \sum_{i}^{N_a} U_i(\boldsymbol{R}) \tag{7.1}$$

where N_a is the number of atoms of the system, and $\boldsymbol{R}$ is a vector that maps a molecule into a certain mathematical representation, ideally invariant to translation and rotation. In terms of performance, the most encouraging feature of many ML models is that, once they have been trained, they can be applied to a myriad of systems without demanding the calculation of additional QM data. This yields the (near-)linear scaling attributed to the ANI methods, which is bounded by the molecular featurisation method employed to generate the descriptors that capture the atomic local environment.[369] As can be noted from equation (7.1), a molecular descriptor is the only input required by the NNPs to output the atomic energy. There are various flavours of descriptors,[369–371] such as, *e.g.*, Coulomb matrices,[372–374] bag of bonds,[375] bispectrum components,[376]

or smooth overlap of atomic positions.[377,378] Specifically, ANI-2x uses the same atom-centred symmetry function as previous ANI models,[348] *viz.*, a form of Behler and Parrinello-type descriptors[379] with a modified symmetry function for the angular part. Hence, for each ith atom of the molecule with atomic number μ, an atomic environment vector, $\boldsymbol{G_i^{\mu}} = \{G_1, G_2, G_3, ..., G_M\}$, is generated, which captures features of each atomic local environment in radial and angular terms. The radial descriptors are given by

$$G_m^R = \sum_{i \neq j}^{N} \exp\left[-\eta(R_{ij} - R_s)^2\right] f_c(R_{ij}) \tag{7.2}$$

in which R_{ij} is the distance between atoms i and j, and the index m is over a set of parameters η and R_s that change the width of the Gaussian distributions and shift the centres of their peaks, respectively. Furthermore, f_c is a piecewise cutoff function that sets the local environment approximation and is given by

$$f_c(R_{ij}) = \begin{cases} 0.5\cos\left(\frac{\pi R_{ij}}{R_c}\right) + 0.5 & \text{for } R_{ij} \leq R_c \\ 0 & \text{for } R_{ij} > R_c \end{cases} \tag{7.3}$$

For the angular symmetry functions, ANI uses a modified version of the Behler and Parrinello descriptor that reads

$$G_m^A = 2^{1-\xi} \sum_{i \neq j,k}^{N} \left[1 + \cos(\theta_{ijk} - \theta_s)\right]^{\xi} \exp\left[-\eta\left(\frac{R_{ij} - R_{ik}}{2} - R_s\right]^2\right) f_c(R_{ij}) f_c(R_{ik}) \tag{7.4}$$

where θ_{ijk} is the angle between atoms i, j and k, and ξ and θ_s serve similar purposes as η and R_s. The local environment approximation of equation (7.3) is thus imposed by using short cutoff values for the radial (4.6 Å) and angular (3.1 Å) descriptors. These short values highlight the local, short-range nature of the ANI models, which are unable to explicitly capture long-range effects. In this regard, a recent study has shown that the poor long-range electrostatic

description of ANI-2x has a deleterious effect on the prediction of water bulk properties, such as, *e.g.*, the internal pressure, even though this situation can be artificially compensated through the use of high external pressure values.[380]

Finally, it is also worth mentioning that ANI-2x was trained to minimise the following objective function

$$X(U^{ANI}, \boldsymbol{F}^{ANI}) = \frac{1}{N_s} \sum_{i=1}^{N_s} \left[\left(U_i^{ANI} - U_i^{DFT} \right)^2 + \frac{\alpha}{N_a} \sum_{j=1}^{N_a} \left(\boldsymbol{F}_{ij}^{ANI} - \boldsymbol{F}_{ij}^{DFT} \right)^2 \right] \quad (7.5)$$

which involves fitting of both forces ($\boldsymbol{F}$) and potential energies (U). In equation 7.5, $\alpha = 0.1$ is a constant used to balance the force and energy fitting terms during the NNP training, and the sum over i runs over N_s systems with N_a atoms per system.

7.2.2 Force field reparameterisation

The conventional MM FF we used in this study was the GAFF,[182] for which the functional is given by equation (4.25). The GAFF partial charges were derived using the multiconformational RESP method.[211,255,256] The gas-phase QM ESPs entering the RESP-fitting procedure were calculated at HF/6-31G*[211] from gas-phase geometries optimised at ωb97X/6-31G*. The conformations used in these calculations were the major conformations found at the MP2/6-311++G(2d,p) level for each γ-fluorohydrin, as reported in Ref. 351. Two stages were performed in this charge-derivation process:[211,381,382] first, all atoms were allowed to vary their charges while applying hyperbolic regularisation with a scaling factor of 0.01; second, only the symmetry-equivalent H and F atoms were allowed to vary their charges, and these charges were constrained to have the same value within a given symmetry group (in this stage, the scaling factor used for the hyperbolic regularisation was 0.001).

The GAFF-like optimally tuned FFs, herein called GAFF.MOD, used the set of RESP charges previously derived, but their bonded parameters were further optimised to reproduce a training data set at ωb97X/6-31G*, the same level of theory that ANI-2x aims to reproduce. This training data set was composed of structures sampled from the DFT ensemble using the nMC-MC algorithm[295,296,345] implemented in ParaMol,[4] interfaced with Psi4[383] for the QM calculations. nMC-MC combines sampling at an approximate potential with periodic switching attempts to the QM level, enabling recovery of the exact quantum DFT ensemble.[345] Owing to this feature, this method was used to generate high-quality structures representative of our reference level of theory. For each γ-fluorohydrin, a variable number of nMC-MC samplers were spawned, each starting from the major conformers reported in Ref. 351 (the same conformers used to calculate the ESPs for the RESP procedure). In total, each nMC-MC sampler performed 2.5×10^4 sweeps, with hMC[292] runs of 100 steps carried out using a 1 fs time step. To accelerate sampling in the low-level chain (ANI-2x), a temperature of 350 K was used for its kinetic and potential energy terms, while the temperature of the target ωb97X/6-31G* NVT ensemble was 300 K. The collected nMC-MC data for each molecule were then merged, and from these a training (1×10^4 structures) and testing (3×10^4 structures) data sets were generated by randomly collecting structures from the final DFT ensembles. The reparameterisation of GAFF was finally conducted by concomitantly optimising all bond, angle, and dihedral parameters (except the dihedral phases) in equation (4.25). Exceptionally for molecule B, the dihedral phases also entered in the optimisation, as molecules containing chiral centres sometimes require optimisation of the dihedral phases to obtain a closer fit to the QM PES.[259,260] Although some of the molecules represented in Figure 7.1 contain chiral centres, optimisation of their dihedral phases only improved the GAFF.MOD performance for molecule B. For the remaining molecules, optimisation of the dihedral phases led to decreased FF accuracy or broke important molecular symmetries. The parameters optimisation was attained by minimising the following objective function

$$X(\boldsymbol{p}) = X_U(\boldsymbol{p}) + \Theta(\boldsymbol{p}) \tag{7.6}$$

where $\boldsymbol{p}$ is the vector of parameters entering the optimisation. Furthermore, X_U is given by equation (5.5) and amounts to the fitting of energies to reference QM data, and $\Theta(\boldsymbol{p})$ is given by equation (5.13) and corresponds to an L2 regularisation term included to prevent overfitting. The values used for the prior widths can be found in Table 5.1. Both the RESP-fitting and reparameterisation procedures were performed in ParaMol.[4]

7.2.3 The hybrid neural network potential/molecular mechanics model

To determine the conformational dynamics of the set of the γ-fluorohydrins considered in this study, we performed MD simulations both in the gas phase and chloroform solution. For the simulations in chloroform that described the γ-fluorohydrins using FFs, the MM energy of the system reads

$$U^{MM}(\boldsymbol{q}^{\boldsymbol{s}}, \boldsymbol{q}^{\boldsymbol{l}}, \boldsymbol{q}^{\boldsymbol{s-l}}) = U^{MM}_{sol}(\boldsymbol{q}^{\boldsymbol{s}}) + U^{MM}_{lig}(\boldsymbol{q}^{\boldsymbol{l}}) + U^{MM}_{lig-sol}(\boldsymbol{q}^{\boldsymbol{s-l}}) \tag{7.7}$$

where U^{MM}_{sol} is the energy of the solvent, which in this study were $CHCl_3$ molecules described by the MM model by Caldwell *et al.*;[384] U^{MM}_{lig} corresponds to the energy of the ligand (γ-fluorohydrin), herein described using either GAFF or GAFF.MOD; $U^{MM}_{lig-sol}$ is the ligand-solvent (γ-fluorohydrin-$CHCl_3$) interaction energy, a term that depends on the LJ parameters of GAFF (the same as those used by the traditional AMBER FF) and on the system partial charges; finally, $\boldsymbol{q}^{\boldsymbol{s}}$, $\boldsymbol{q}^{\boldsymbol{l}}$ and $\boldsymbol{q}^{\boldsymbol{s-l}}$ are the degrees of freedom of the solvent, ligand, and interactions between them, respectively.

As an attempt to improve the accuracy of the pure MM model represented in equation (7.7), several studies[356,357,361–363] have been conducted in which a ML

model was employed to represent the ligand term, $U_{lig}^{MM}(\boldsymbol{q}^{l})$. Besides having the advantage of avoiding the parameterisation of individual ligands, this hybrid model has led, in general, to higher accuracy in simulations. Owing to the similarity of this hybrid scheme to the QM/MM model, it has come to be known as the NNP/MM model. The NNP/MM energy of a system in which only the ligand is included in the ML region and there are no covalent bonds between the ligand and the solvent reads

$$U^{NNP/MM}(\boldsymbol{q}^{s},\boldsymbol{q}^{l},\boldsymbol{q}^{s-l}) = U_{sol}^{MM}(\boldsymbol{q}^{s}) + U_{lig}^{NNP}(\boldsymbol{q}^{l}) + U_{lig-sol}^{MM}(\boldsymbol{q}^{s-l}) \tag{7.8}$$

Note that the only change in equation (7.8) relative to equation (7.7) is that the intramolecular representation of the ligand (γ-fluorohydrin) is now made by the ANI-2x NNP. Hence, since the ligand-solvent (γ-fluorohydrin-$CHCl_3$) and solvent-solvent ($CHCl_3$-$CHCl_3$) interactions are still treated at the MM level, this hybrid model corresponds to a mechanical-embedding scheme[341] in which the partial charges of the ML region are kept fixed. As pointed out by Lahey *et al.*,[361] FFs are parameterised in an internally consistent manner. Consequently, there is a chance that the MM parameters used to described the ligand-solvent nonbonded interactions are not optimal for the NNP/MM potential. The degree to which these non-optimal nonbonded parameters may cause an imbalance between different parts of the model (in our case, between the MM ligand-solvent intermolecular interactions and the NNP ligand intramolecular interactions) is unknown *a priori* and must be investigated.

7.2.4 Molecular dynamics simulations

The gas-phase MD simulations were performed in the NVT ensemble using a Langevin integrator with a temperature of 298.15 K and a friction coefficient of 2 ps^{-1}. These simulations ran for 100 ns with a time step of 1 fs. They were performed in triplicate for GAFF and GAFF.MOD, whereas for ANI-2x only one simulation per molecule was run for reasons of computational cost. For the

simulations in chloroform, which used the same Langevin integrator settings as the gas-phase simulations, the chloroform box was created by adding $CHCl_3$ molecules around the γ-fluorohydrins for 20 Å in the positive and negative x, y, and z directions. The solvated systems were then equilibrated in the NPT ensemble during 1 ns using the MC barostat[135,136] to fix the pressure at 1 bar. The LJ cutoff was set at a distance of 12 Å with a switching distance of 10 Å. Long-range electrostatic interactions were handled using the PME method.[221,224] The final NVT production runs were performed in duplicate (ANI-2x) or triplicate (GAFF and GAFF.MOD) during 100 ns. Snapshots of the trajectories were saved every picosecond. All MD simulations were run in OpenMM,[385] and those that used ML models used the openmm-ml plugin that can be found at `https://github.com/openmm/openmm-ml`. The initial topology and coordinate files used as inputs to OpenMM and ParaMol were generated using LEaP.[386]

7.2.5 Populations of conformers and spin-spin coupling constants

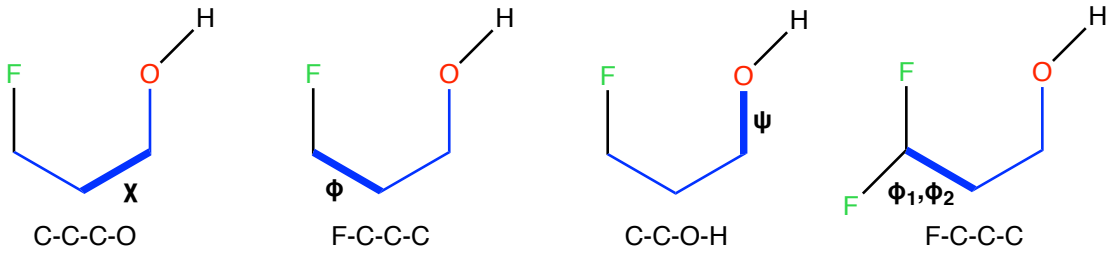

Figure 7.2: Dihedral angles used to identify the conformers of the γ-fluorohydrins.

The terminology used to identify the conformers of the γ-fluorohydrins follows that commonly employed to characterise the rotamers of protein side-chains. Hence, the conformers arising from the rotation of the three threefold torsional barriers identified in Figure 7.2 are labelled according to the following definitions:[351,387,388]

- $0^\circ \leq \chi, \phi, \psi < 120^\circ \implies$ g+

- $120^\circ \leq \chi, \phi, \psi < 240^\circ \implies$ t
- $-120^\circ \leq \chi, \phi, \psi < 0^\circ \implies$ g-

Using these labelling rules, the populations of the conformers obtained from MD simulations were estimated by clustering every individual frame of the final trajectories into the respective conformer. For the monofluoroderivatives, the conformers were identified by the sequence $\chi\phi(\psi)$; for the difluoroderivatives, the conformers were identified by the sequences $\chi\phi_1\phi_2(\psi)$ or $\chi\phi_2\phi_1(\psi)$; and for trifluoderivatives, the conformers were identified by the sequence $\chi(\psi)$. Furthermore, to have estimations of populations that are independent of the potential models being tested and that can thus be used as reference values, we calculated populations at various QM levels for the identified energetic minima of each γ-fluorohydrin. To do this, geometry optimisations and frequency calculations were carried out at the ωb97X/6-31G* and MP2/6-311++G(2d,p) levels of theory using Gaussian 09[389] interfaced with ASE.[275] Whenever required, solvent effects ($CHCl_3$) were introduced through the polarisable continuum model (PCM). The Boltzmann populations of the conformers were then estimated from the calculated relative standard Gibbs free energies in the harmonic approximation.

The J-coupling constants were computed from optimised geometries at the PCM-ωB97X/6-311++G(2d,p) level of theory using the gauge-invariant atomic orbital (GIAO) method.[390–392] In these calculations, the hybrid B97-2 functional[393] and the pcJ-2 basis set,[394] which exhibit good performance in the calculation of these NMR parameters,[351,395,396] were used. Again, solvent effects were included through the PCM model. The calculated J-coupling constants were finally averaged over all conformers, according to their relative populations in chloroform at 298.15 K, using the following equation

$$J_\lambda^\nu = \sum_i^{N_{conf}} P_i^\nu J_{\lambda,i} \tag{7.9}$$

where λ identifies the J-coupling constants being considered, and ν refers to the QM level or molecular model from which the populations are estimated.

7.3 Results and discussion

7.3.1 nMC-MC acceptance rates

We first present the results obtained in the nMC-MC simulations. The nMC-MC simulations aimed at generating ensembles of configurations representative of the ωb97X/6-31G* ensemble by using ANI-2x as the approximate potential. The acceptance rates that the nMC-MC simulations produce give important information about the ANI-2x performance in the gas phase.[345] On the one hand, the hMC acceptance rate measures the stability of the short MD runs performed in the nMC-MC algorithm, and it is positively correlated with energy conservation during the MD run. Since we obtained hMC acceptance rates $\geq$ 90% for all molecules in the test set (Figure 7.3), we conclude that ANI-2x is a viable model to use in MD simulations. These high hMC acceptance rates are comparable in magnitude to those we have obtained in the study presented in Chapter 6 for molecules of similar size modelled using GAFF-like FFs.[345] On the other hand, the ANI-2x to DFT acceptance rate measures the similarity between the ANI-2x potential and the ωb97X/6-31G* level of theory. We obtained acceptance rates in the ANI-2x to DFT step $>$ 60% (Figure 7.3). Compared to the results of the previous study,[345] these are very high acceptance rates, indicating excellent agreement between ANI-2x and ωb97X/6-31G*.

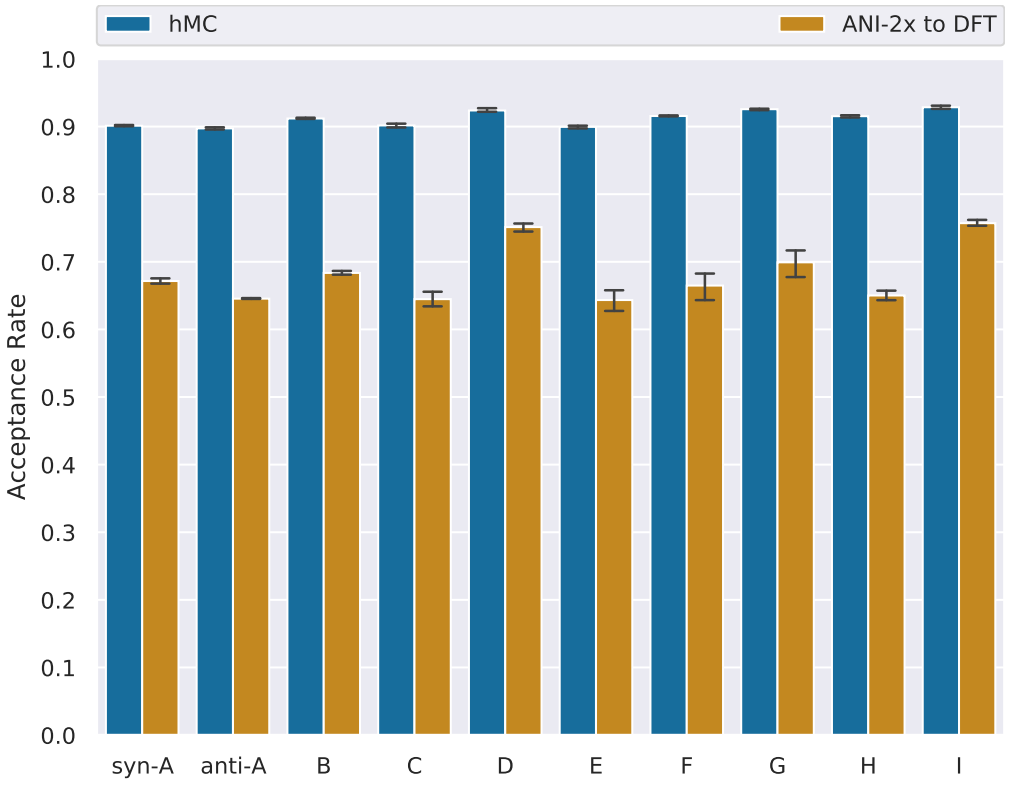

Figure 7.3: Acceptance rates obtained in the nMC-MC simulations for each γ-fluorohydrin. Only the 3 samplers that gave the lowest acceptance rates were included in the calculation of the mean and standard deviation of each bar.

7.3.2 Energetic and geometric agreement in the gas phase

Comparing the performance of different ligand models in the absence of experimental data is not straightforward owing to the lack of an absolute reference. FFs and ML potentials, however, are often fitted to reproduced gas-phase energies and geometries at specific QM levels of theory. This is the case for GAFF, fitted to reproduce experimental, MP2/6-31G* (equilibrium bonds and angles) and MP4/6-311G(d,p)//MP2/6-31G* (dihedral parameters) data,[182] and for GAFF.MOD and ANI-2x, fitted to reproduce ωb97X/6-31G* data.[352] The natural way of benchmarking the accuracy of molecular models in the gas phase is, therefore, to compare their performance to that of the QM level against which they were fitted. Here, we present and discuss our gas-phase results by comparing the performance of GAFF, GAFF.MOD, and ANI-2x to reference QM data.

To assess the energetic agreement between the different molecular models and a given QM reference, we calculated relative energy differences using the following equation[330]

$$\Delta\Delta E = \left(E^X - E_0^X\right) - \left(E^{QM} - E_0^{QM}\right) \tag{7.10}$$

where X denotes the molecular model used (GAFF, GAFF.MOD, or ANI-2x), and the 0 subscript identifies the conformer with the lowest QM energy for the given molecule. Ideally, a model should give $\Delta\Delta E$ values close to 0 kJ mol^{-1}, indicating good agreement to the QM level. Broader distributions indicate larger deviations with respect to the QM level. A negative $\Delta\Delta E$ value indicates that the model relative energy is underestimated compared to the QM relative energy, whereas a positive $\Delta\Delta E$ value indicates that the model relative energy is overestimated compared to the QM relative energy.

For each γ-fluorohydrin, we calculated the relative energy differences for the 3 molecular models relative to ωb97X/6-31G* using 3×10^4 structures extracted from the nMC-MC simulations. The distributions of the relative energy differences show an unequivocal trend: the performance of the models tested decreases as ANI-2x $>$ GAFF.MOD $>$ GAFF (Figure 7.4). This presents evidence that ANI-2x excels in reproducing the level of theory it has been trained to reproduce, with RMSEs for all molecules below the chemical accuracy of 4.184 kJ mol^{-1} (1 kcal mol^{-1}). GAFF.MOD, the optimally tuned FF optimised using ωb97X/6-31G* data, underperforms relative to ANI-2x, though still showing notable improvements relative to the original GAFF. Importantly, the GAFF.MOD distributions invariably show increased precision (narrower distributions) and, for most molecules, increased accuracy (mean of the distributions closer to zero) than GAFF, confirming that reparameterisation improved the FF agreement with ωb97X/6-31G*.

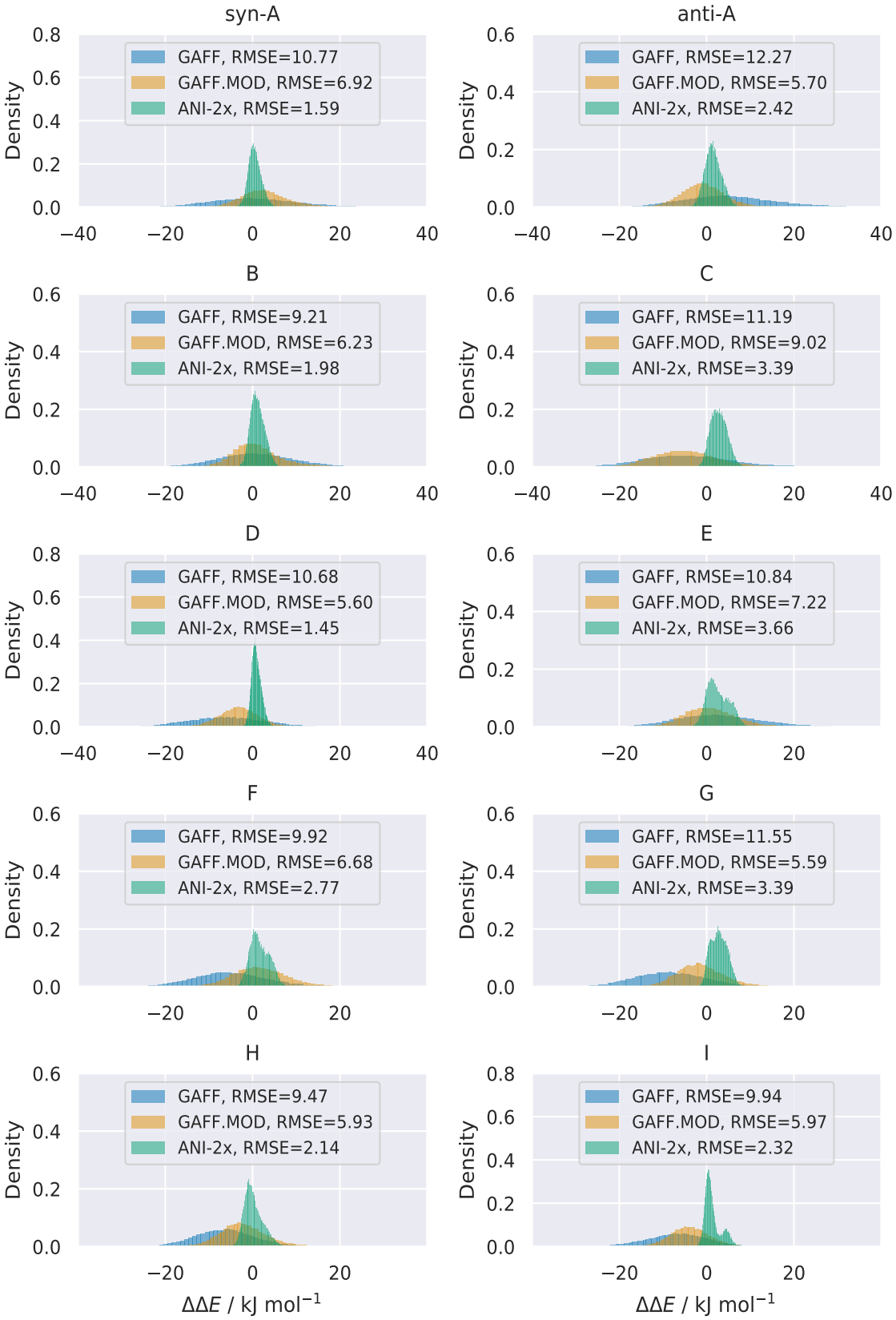

Figure 7.4: Distributions of the relative energy differences ($\Delta\Delta E$) for GAFF, GAFF.MOD, and ANI-2x with respect to the ωb97X/6-31G* level of theory. The testing data set was composed of 3×10^4 structures extracted from the nMC-MC simulations. The molecular structure used as a reference was removed from the histograms.

Several confounding factors may have contributed to preventing GAFF.MOD from achieving the same level of accuracy as ANI-2x. First, since GAFF.MOD is a GAFF-like model, it is constrained by the FF functional form, which may not have been ideal to energetically represent some configurations sampled in the ωb97X/6-31G* ensemble. Second, to derive GAFF.MOD, only the GAFF bonded parameters were optimised, leaving the nonbonded parameters untouched. It is well-known that ωb97X lacks dispersion interactions since the semi-local correlation functionals cannot capture long-range correlation effects.[380,397–400] This physical artifact may not have been entirely captured by only optimising the bonded part of GAFF, as dispersion physics is modelled by the LJ 12-6 potential. Third, as the objective function (equation 7.6)) included a regularisation term that depends on the initial FF parameters, the solutions of the optimisation problem were dependent on this initial guess (in the present work, the GAFF parameters). We cannot exclude the possibility of obtaining higher-quality FF parameters if another initial guess were used, though this also poses the non-trivial problem of determining alternative initial guesses. Fourth, the completeness of the training data set used in the reparameterisation procedure may also have impacted the quality of the optimised FF parameters. The nMC-MC simulations, however, generated representative ωb97X/6-31G* ensembles. We thus believe the completeness of the training data set did not have a significant impact on the quality of the reparameterisation procedure. These four issues could be addressed by using either more advanced FF functional forms or by changing the nature of the optimisation procedure, work which is outside the scope of the current study.

Two types of error are present when molecular models are used to perform energy measurements: random errors and systematic errors. Both originate from functional form constraints and/or inadequate model parameters. Random errors are related to the precision of the molecular model and are normally distributed around the true value. Systematic errors are related to the accuracy of the molecular model and cause the mean of the error distribution to deviate from the true value and/or the error distribution to be non-Gaussian. The distributions of the relative energy differences (Figure 7.4) show that GAFF

has a bias towards negative relative energy differences for most molecules (C, D, F, G, H, and I). On the other hand, a bias towards positive relative energy differences is seen for molecules anti-A and E, whereas for molecules syn-A and B random errors dominate. The systematic errors for the GAFF model are mostly offset errors, as they manifest themselves as deviations from the true value (the distributions remain approximately Gaussian). Furthermore, the GAFF.MOD distributions tend to have a mean closer to the true value (smaller offset errors) than GAFF, while also showing smaller random errors. Finally, although the magnitude of the random errors of ANI-2x are small, they present significant systematic errors that lead to very pronounced non-normally distributed relative energy differences. The distribution of molecule I is of particular concern due to its bimodal shape. This bimodal shape occurs because conformations g-(g+) and g+(g-) of molecule I have a systematic bias towards negative relative energy differences, whereas the remainder conformations have a systematic bias towards positive relative energy differences (Figure 7.5). As we shall see later in the discussion, systematic errors with conformation-dependent signs and/or magnitudes scale the relative energy between conformers, consequently changing their relative populations, an undesirable situation.

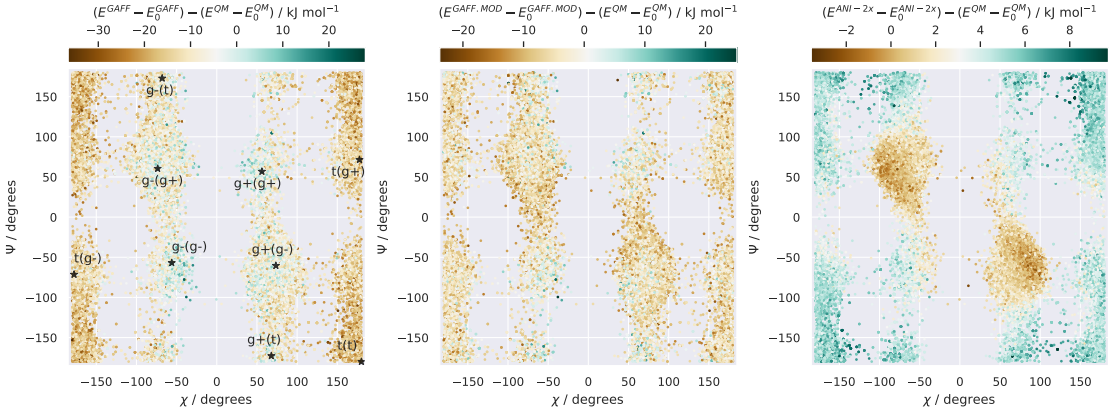

Figure 7.5: Distributions of the χ and Ψ dihedral angles (see definitions in Figure 7.2) of molecule I for configurations sampled using 3 nMC-MC simulations. The color of each point gives the relative energy difference ($\Delta\Delta E$) between the model (GAFF, left; GAFF.MOD, middle; ANI-2x, right) and ωb97X/6-31G*. The black stars locate the QM minima calculated using ωb97X/6-31G*.

We now discuss the performance of the models in reproducing the energies

and geometries of the QM minima. An ideal model should yield optimised geometries similar to QM, and the relative energies of those minima should agree between the models and QM.[330] To assess performance in these two categories, we performed geometry optimisations using GAFF, GAFF.MOD, and ANI-2x, starting from all QM minima within 12.552 kJ mol^{-1} (3 kcal mol^{-1}) from the global minimum. The GAFF and GAFF.MOD geometry optimisations were performed with the L-BFGS algorithm of OpenMM, and the ANI-2x geometry optimisation were performed using the L-BFGS algorithm of ASE. The RMSDs of the relative energy differences and the average RMSD of the atomic positions are shown in Table 7.1. The results obtained agree with the findings presented so far, indicating that ANI-2x is the model that best reproduces the energies and geometries of the ωb97X/6-31G* minima, followed by GAFF.MOD and then by GAFF (Figure 7.6). Interestingly, we observe that ANI-2x tends to predict positive relative energy differences, meaning that the relative energies between the local and global minima tend to be overestimated. This observation is cause for concern, as it indicates that the ANI-2x global minima tend to be systematically overstabilised. Furthermore, GAFF tends to underestimate the relative energy differences, whereas GAFF.MOD errors were mostly random, though still presenting a non-negligible tendency to underestimate the relative energies of some minima relative to QM. As expected, the inverse trend is observed when the QM reference is MP2/6-311++G(2d,p), indicating that GAFF is the model that gives the best energetic agreement with this QM level, followed by GAFF.MOD and ANI-2x (see Appendix C, Figure C.1). Moreover, the average RMSD of the atomic positions is lower for GAFF.MOD and ANI-2x than for GAFF, though these results are heavily influenced by outliers. Removing all points with an RMSD greater than 0.3 Å leads to equal trends for the energetic and geometric agreement, as expected.

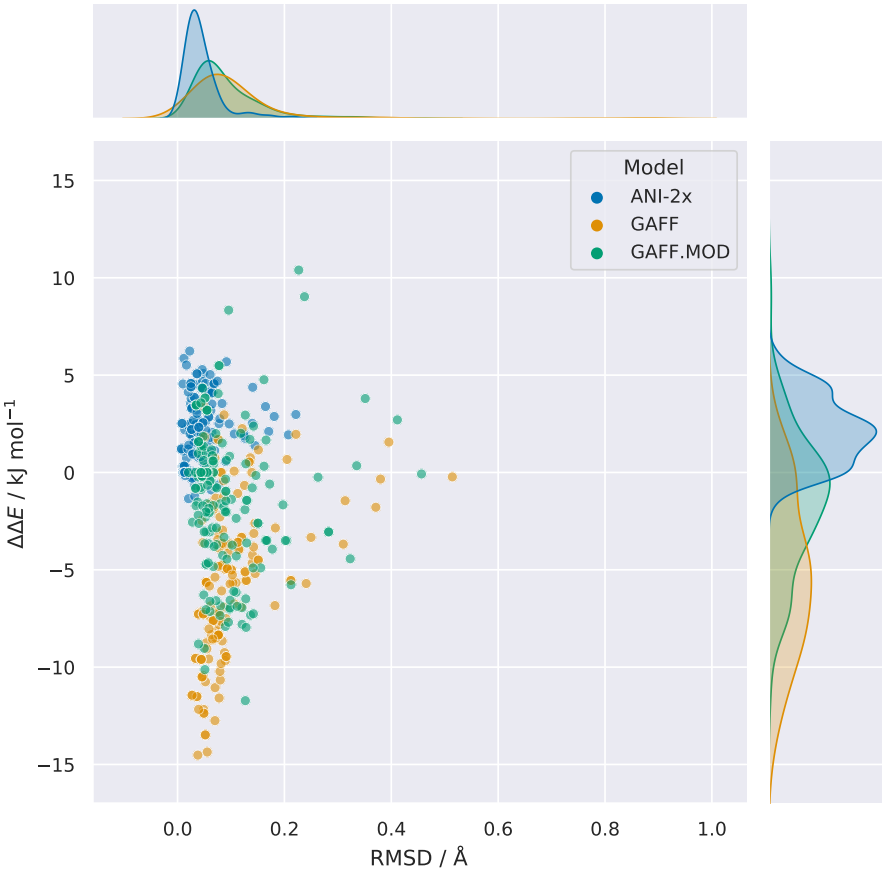

Figure 7.6: Scatter plots of the relative conformer energies ($\Delta\Delta E$) versus the RMSD of atomic positions. Each point was obtained by performing a geometry optimisation using GAFF, GAFF.MOD, or ANI-2x, starting from all QM minima within 12.552 kJ mol^{-1} (3 kcal mol^{-1}) from the global minimum. The QM reference is the ωb97X/6-31G* level of theory.

Table 7.1: RMSDs of the relative energy differences ($\Delta\Delta E$) and average RMSDs of atomic positions for the scatters depicted in Figure 7.6. The molecular structures used as a reference were excluded from the calculation of the RMSDs of the relative energy differences. The QM references are MP2/6-311++G(2d,p) (MP2) and ωB97X/6-31G* (ωB97X).

	$\Delta\Delta E$		atomic positions	
	MP2	ωB97X	MP2	ωB97X
GAFF	3.36	7.45	0.087	0.110
GAFF.MOD	4.26	4.05	0.079	0.095
ANI-2x	5.12	2.80	0.080	0.046

7.3.3 Sampling accuracy in the gas phase

We also assessed the sampling accuracy of each model in the gas phase by comparing the populations of the γ-fluorohydrins, as predicted by MD, to their Boltzmann populations calculated at ωb97X/6-31G* and MP2/6-311++G(2d,p). To estimate the QM populations, the electronic energies were converted into Gibbs free energies in the harmonic approximation using standard thermodynamic corrections obtained from frequency calculations.[351] This approach introduces two approximations: it considers non-interacting particles and assumes that the first and higher electronic excited states are entirely inaccessible.[401] While the latter assumption should not pose a problem for the set of γ-fluorohydrins considered in this study, the former may introduce some error, depending on how much the systems deviate from ideal behaviour. Whenever possible, the QM populations should be estimated by performing MD or MC simulations. However, owing to the prohibitive computational cost of *ab initio* simulations, we believe our approach is sufficiently reasonable to warrant investigation. The metric we used to evaluate the sampling accuracy was the sum of the absolute error of the populations (SAEP), calculated as the absolute difference between the populations predicted by the model X and the QM level, such that

$$SAEP = \sum_{i}^{N_{conf}} |p_i^X - p_i^{QM}| \tag{7.11}$$

where p_i is the population of the ith conformer, and N_{conf} denotes the total number of conformers. When using ωb97X/6-31G* as the reference, it is not possible to determine the model that predicts best sampling accuracy because mixed results were obtained (Figure 7.7, top panel). For example, ANI-2x exhibits significantly higher sampling accuracy than GAFF.MOD for molecules C, D, and E, but GAFF.MOD exhibits significantly higher sampling accuracy than ANI-2x for molecules F, G, H, and I. There are also some molecules (syn-A, anti-A, and B) for which GAFF.MOD and ANI-2x perform similarly. GAFF, however, stands out as the model that worst reproduces the populations at ωb97X/6-31G*.

This trend is inverted when using MP2/6-311++G(2d,p) as the reference (Figure 7.7, bottom panel), indicating that there is a significant disagreement between the predictions of MP2/6-311++G(2d,p) and ωb97X/6-31G*. This is, however, expected behaviour: while ANI-2x was trained to reproduce the ωb97X/6-31G* level of theory, GAFF was derived using experimental, MP2, and MP4 data.[182] Hence, it is natural that GAFF is the model that overall best reproduces the populations at MP2/6-311++G(2d,p). Interestingly, GAFF.MOD FFs still perform reasonably well when MP2/6-311++G(2d,p) is the reference, and they actually outperform GAFF for the trifluoroderivates H and I because the populations for these molecules are similar in both QM references. The good performance of the GAFF.MOD FFs is attributed to the fact that their parameters, while optimised to reproduce ωb97X/6-31G*, retain some "memory" of the original GAFF parameters due to the regularisation applied during the optimisation procedure.

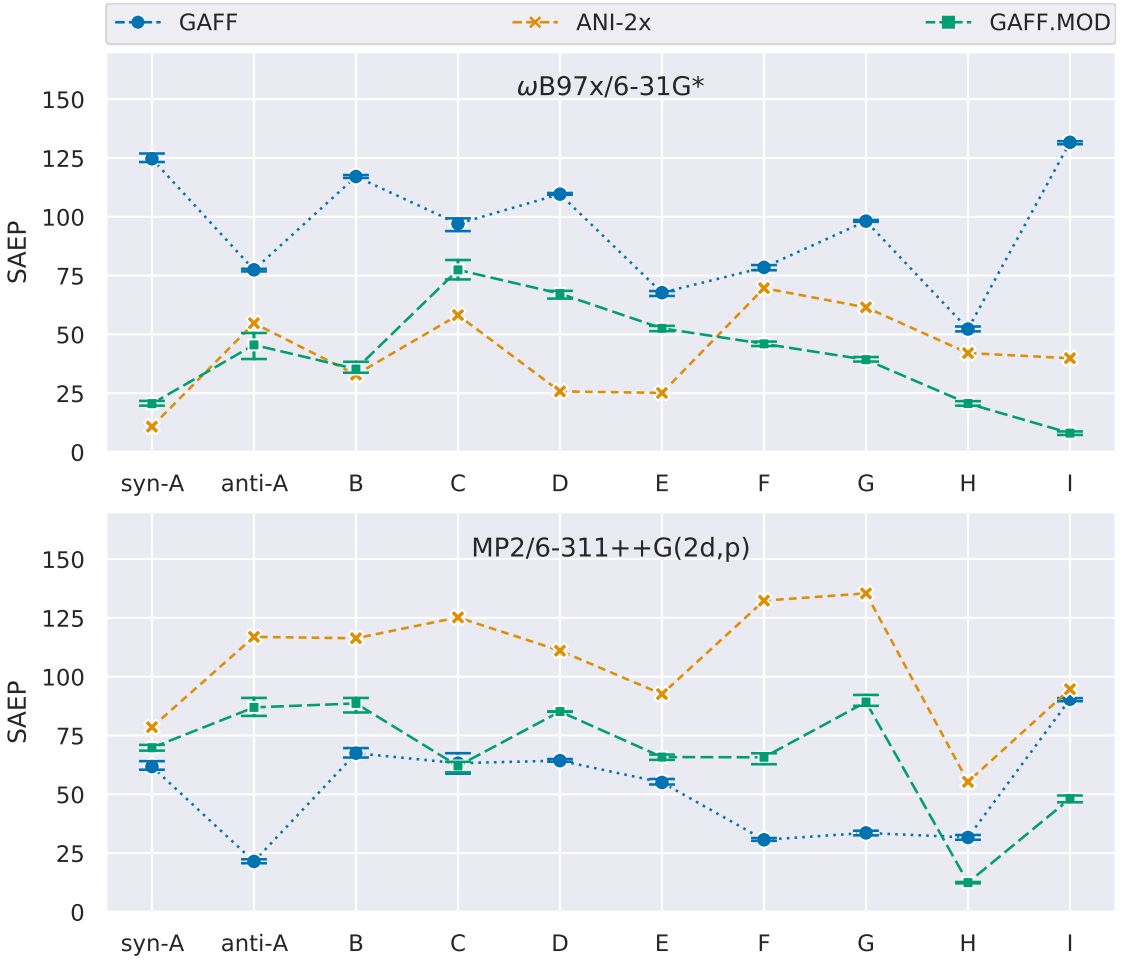

Figure 7.7: Sum of the absolute error of the populations (SAEP), calculated as the absolute difference between the populations predicted by the models (GAFF, GAFF.MOD, and ANI-2x) and the QM level. The QM references are ωb97X/6-31G* (top plot) and MP2/6-311++G(2d,p) (bottom plot).

Two reasons may have contributed to GAFF.MOD predicting a significantly lower SAEP for some molecules (F, G, H, and I) than ANI-2x, even though the GAFF.MOD energy landscape agreement with ωb97X/6-31G* is lower than that of ANI-2x (Figure 7.4). The first reason is energetic: energetic errors that affect the relative energies of conformers change their relative populations. For example, if the relative energy between two conformers increases, the conformer with lower energy becomes more populated. The second reason is entropic: populations depend not only on conformational energies but also on conformational entropies.[402] Broader wells are associated with more configurations than narrower wells, therefore being more entropically favorable. Since for ANI-2x we observe conformation-dependent directions of bias (Figure 7.5) and

a tendency to overstabilise the global minima, we believe these were the factors that, for some molecules, caused the relative energies of GAFF.MOD to be closer to ωb97X/6-31G* than those of ANI-2x (see, *e.g.*, Figure 7.8). Hence, we think GAFF.MOD gave lower SAEPs than ANI-2x for some molecules mainly because of energetic factors. Note that the configurations sampled by each model for a given conformer may differ in degrees of freedom other than those used for conformer assignment (*e.g.*, bonds and angles).

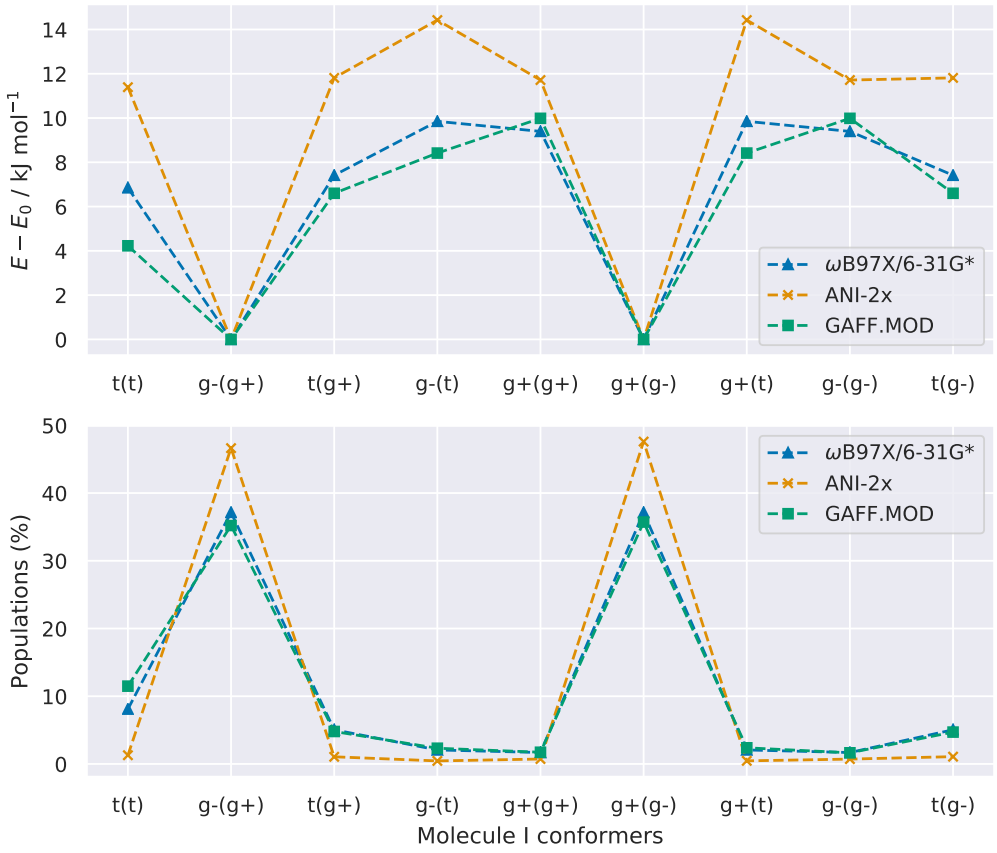

Figure 7.8: Top plot: Relative energies of the conformers of molecule I (optimised geometries), calculated using ωb97X/6-31G*, ANI-2x, and GAFF.MOD. Bottom plot: Populations of each conformer of molecule I, as predicted by ωb97X/6-31G*, ANI-2x, and GAFF.MOD.

In summary, the results obtained in the gas-phase benchmark demonstrate that models tend to behave similarly to the QM levels to which they have been fitted. ANI-2x exhibits levels of accuracy that class I FFs (GAFF and GAFF.MOD) cannot achieve when using ωb97X/6-31G* as the reference (Figure 7.4). Interestingly, the high accuracy of ANI-2x in reproducing the ωb97X/6-31G* energy landscape

does not always translate into high sampling accuracy (Figure 7.7). For some molecules, ANI-2x exhibits systematic errors that cause significant overestimation of the relative energies between the local and global minima, leading to an overpopulation of the lower energy conformers (Figures 7.6 and 7.8). This observation raises concerns on whether to use ANI-2x over a conventional FF to sample the conformational landscape, as the computational cost of ANI-2x is up to 100 times greater than that of a class I FF. These concerns are further aggravated by realising that ANI-2x performs the worst in reproducing MP2/6-311++G(2d,p) (Figure 7.7), which in principle is more accurate than ωb97X/6-31G*. This difference suggests that ωb97X/6-31G* and MP2/6-311++G(2d,p) predict different physical behaviour for the γ-fluorohydrins considered in this study. In the next sections, we evaluate the performance of the models in chloroform solution and then proceed to assess which QM level best reproduces experimentally-determined J-coupling constants.

7.3.4 Sampling accuracy in chloroform solution

We begin the discussion of the results obtained in chloroform solution by assessing the sampling accuracy of each model. To do this, we follow the procedure presented in the previous section and compare the populations of the γ-fluorohydrins, as predicted by MD, to their Boltzmann populations calculated at ωb97X/6-31G*/PCM and MP2/6-311++G(2d,p)/PCM. This approach assumes that these QM levels and solvent model represent the standard against which we compare. GAFF.MOD-RESP/$CHCl_3$ is the model that overall predicts better sampling accuracy when ωb97X/6-31G*/PCM is used as the reference (Figure 7.9). Specifically, GAFF.MOD-RESP/$CHCl_3$ best reproduces the populations of molecules syn-A, anti-A, B, D, E, F, G, and I, whereas the populations of molecules C and H are best reproduced by ANI-2x-RESP/$CHCl_3$ and GAFF-RESP/$CHCl_3$, respectively. Compared to the gas-phase results (Figure 7.7), in which ANI-2x performed slightly better than GAFF.MOD, GAFF.MOD

seems to be better suited for condensed phase simulations. Since in the ANI-2x-RESP/$CHCl_3$ simulations the LJ 12-6 parameters were taken from GAFF, these results suggest that there is an imbalance between the ligand-solvent intermolecular and the ligand intramolecular interactions. Hence, for some systems, the practice[356,357,361–363] of directly combining LJ 12-6 parameters with NNPs may decrease the NNP performance in the condensed phase. This observation strongly indicates that sets of LJ 12-6 parameters consistent with ANI-2x should be developed in the future so that the NNP gas-phase accuracy does not decrease in the condensed phase. The optimally tuned GAFF.MOD FFs, on the other hand, despite having been optimised to reproduce ωb97X/6-31G*, present a better balance between the ligand-solvent intermolecular and the ligand intramolecular interactions. The GAFF.MOD-RESP/$CHCl_3$ are models with greater internal consistency than ANI-2x-RESP/$CHCl_3$ because their parameters are closer to those of GAFF owing to the regularisation applied during the optimisation procedure.

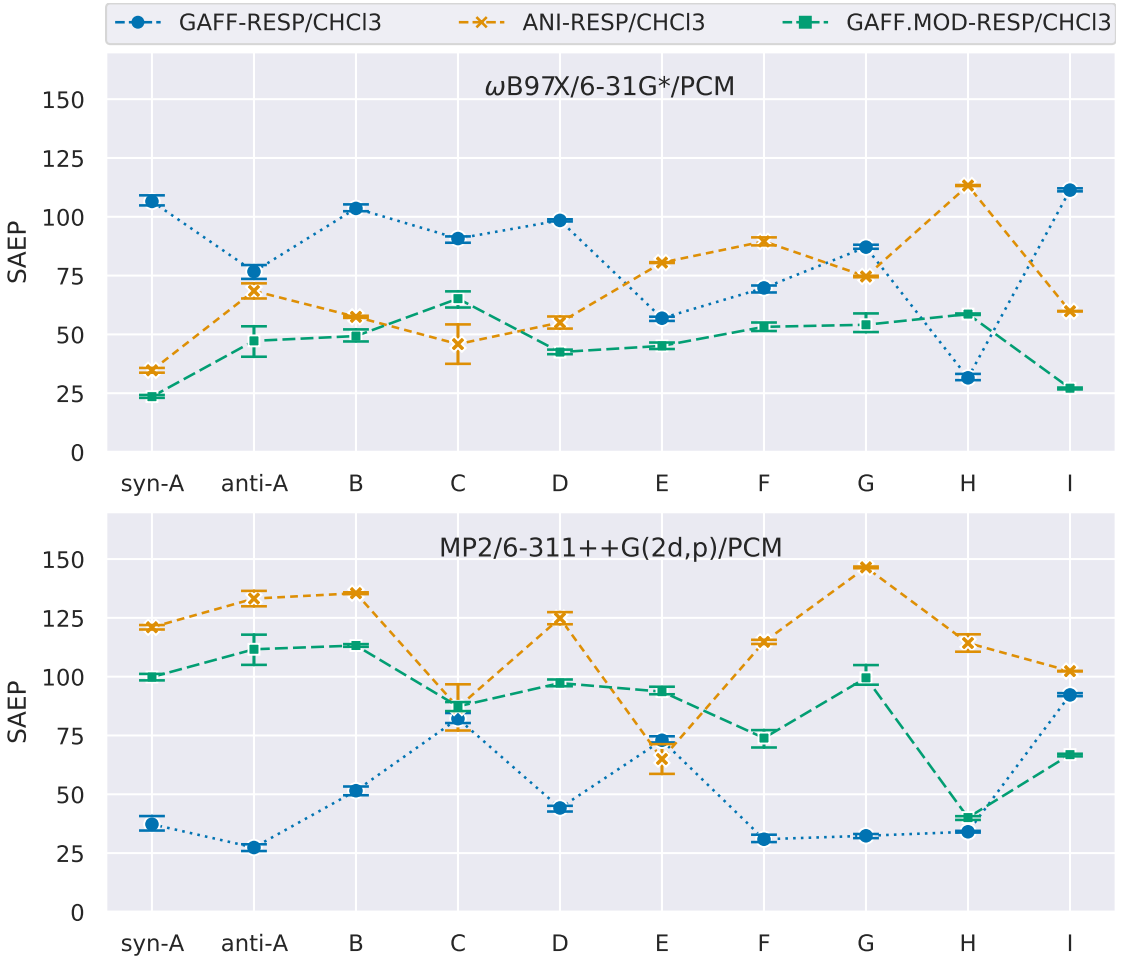

Figure 7.9: Sum of the absolute error of the populations (SAEP), calculated as the absolute difference between the populations predicted by the models (GAFF-RESP/$CHCl_3$, GAFF.MOD-RESP/$CHCl_3$, and ANI-2x-RESP/$CHCl_3$) and the QM level. The QM references are ωb97X/6-31G*/PCM (top plot) and MP2/6-311++G(2d,p)/PCM (bottom plot).

When the reference is MP2/6-311++G(2d,p)/PCM, GAFF-RESP/$CHCl_3$ is the model that overall predicts better sampling accuracy. Exceptions occur for molecule C, for which all models perform similarly; molecule E, for which ANI-2x-RESP/$CHCl_3$ and GAFF-RESP/$CHCl_3$ perform similarly; and molecule I, for which GAFF.MOD-RESP/$CHCl_3$ performs the best. Following the previously presented discussion for the gas-phase benchmark, this is expected behaviour since GAFF was fitted to a QM level similar to MP2/6-311++G(2d,p). Again, the differences in results obtained when using different QM references indicate that the QM references predict different physical behaviour. To understand how this relates to the sampled conformers, we determined the populations of the

conformers with IMHBs for each model and QM reference (Figure 7.10). From these results, we see that ωb97X/6-31G*/PCM tends to overestimate the populations of the conformers with IMHBs relative to MP2/6-311++G(2d,p)/PCM. This overestimation trend is even more pronounced for the ANI-2x-RESP/$CHCl_3$ model. As has been shown previously, ANI-2x tends to overestimate the relative energies of the local minima relative to the global minima (Figures 7.6 and 7.8). Since for this set of γ-fluorohydrins the global minima are mostly the conformers with IMHBs, it follows that the overpopulation of these conformers relative to ωb97X/6-31G*/PCM is a consequence of this energetic error of the ANI-2x model. Similar results were obtained in the gas-phase (see Appendix C, Figure C.2), indicating that this problem cannot be entirely attributed to the imbalance of the hybrid NNP/MM scheme here employed, as part of it is a consequence of the energetic errors intrinsic to ANI-2x.

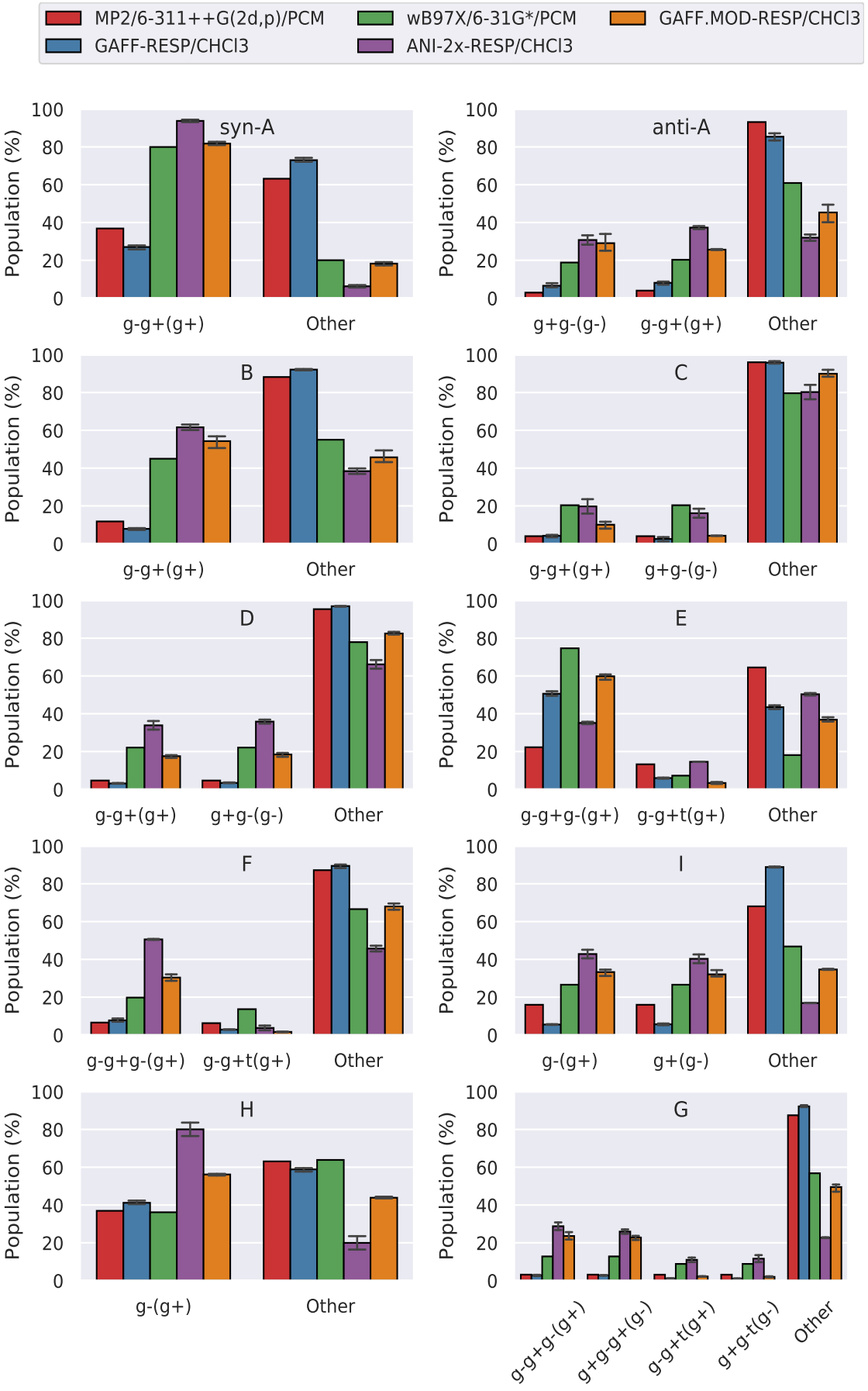

Figure 7.10: Populations in chloroform solution of the conformers with IMHBs.

It is well-known that ωb97X/6-31G* lacks dispersion interactions, and thus it is expected that ANI-2x also suffers from this physical artifact. Incidentally, a previous study has shown that the absence of dispersion interactions in ANI-2x negatively affects the modelling of bulk water and peptides, as it led to stronger-than-expected hydrogen bonds (HBs).[380] For our simulations in chloroform solution, we also observe this phenomenon, as ANI-2x-RESP/$CHCl_3$ and ωb97X/6-31G*/PCM predicted shorter HBs than GAFF-RESP/$CHCl_3$ and MP2/6-311++G(2d,p)/PCM (see Appendix C, Figure C.3). The results obtained could potentially be improved by including a dispersion correction, such as D3.[278] However, it is unclear whether this correction would mitigate the systematic errors in relative energies observed for ANI-2x. Alternatively, for molecules only containing elements H, C, N, and O, the ANI-1cxx NNP could be applied as, in principle, it properly captures dispersion interactions. It would also be possible to run MD simulations in which both the solute and solvent were described at the NNP level. However, our attempts to simulate bulk chloroform using ANI-2x led to radial distribution functions (RDFs) that indicate an overstructuring tendency relative to the experimental RDFs (see Appendix C, Figure C.4). Again, this overstructuring is likely caused by the lack of dispersion interactions, which causes permanent dipole-dipole interactions to dominate. Note that ANI-2x was not trained to reproduce bulk chloroform. Nevertheless, ANI-2x was trained to reproduce bulk water, and the overstructuring tendency is still observed.[364] Additionally, the short-range nature of ANI-2x requires bulk NNP simulations to use high-pressure values (*ca.* 1540 bar for our bulk chloroform simulation) for the barostat so that bulk densities comparable with the experiment are obtained. The absence of long-range electrostatic interactions in ANI-2x naturally suggests the use of hybrid NNP/MM schemes, as long-range interactions are easily computed at the MM level. Unfortunately, NNP/MM hybrid models pose additional and still unsolved problems because, to obtain high levels of accuracy, they require sets of nonbonded parameters consistent with the NNP. In the absence of these parameters, optimally tuned FFs seem to be the most viable alternative to use for cases in which the original FF performs poorly, as for our

test set optimally tuned FFs led to higher sampling accuracy than ANI-2x, with the advantage of having a much lower computational cost.

7.3.5 NMR J-couplings

In the last section of the discussion, we present the results obtained for the J-coupling constants ($^{h1}J_{OH\cdots F}$). We resort to the experimental NMR data available in chloroform solution[351] to determine which molecular model or QM level gives the populations that best reproduce the true behaviour of the molecules considered in this study. To estimate the theoretical J-couplings, we used equation (7.9) to average the calculated $^{h1}J_{OH\cdots F}$ values at ωB97X/6-311++G(2d,p)/PCM//B97-2/pcJ-2/PCM over the conformer populations predicted by each model or QM level.

By analysing the results given in Table 7.2, we conclude that MP2/6-311++G(2d,p)/PCM is the QM level of theory that best reproduces the experimental J-couplings, as it gave the highest squared Pearson correlation coefficient ($R^2 = 0.96$) and lowest RMSE (0.41 Hz). ωb97X/6-31G*/PCM gave the second-best R^2 value (0.87), indicating a strong correlation between theoretical and experimental data. The R^2 values, however, do not reflect systematic errors, which are high for this DFT functional, as can be seen from its RMSE value (4.80 Hz). Concerning the molecular models, GAFF-RESP/$CHCl_3$ gave a lower RMSE (1.28 Hz) than GAFF.MOD-RESP/$CHCl_3$ (4.77 Hz), though with smaller R^2 value (0.68 *vs.* 0.75). As low RMSE values indicate both high precision and accuracy, we consider RMSE to be a better metric than R^2 to measure the agreement with the experiment. Under this assumption, GAFF-RESP/$CHCl_3$ is the model that best reproduces the experimental data, justified by its similarity to the MP2 level. Finally, ANI-2x-RESP/$CHCl_3$ exhibits the greatest disagreement with the experiment, presenting the lowest R^2 (0.58) and highest RMSE (6.78 Hz) values.

Two possible sources of error can impact the accuracy of the calculated J-coupling constants: the error in the populations, which is directly related to the model or QM level of theory used to estimate them, and the error of the QM method used to calculate the J-coupling values for each conformer. We attempted to determine the error associated with the QM method employed to calculate the J-couplings by determining the $^{h1}J_{OH\cdots F}$ value for a conformationally-restricted cyclohexane that predominantly assumes only one conformation (compound 2 in Ref. 351). By doing so, we virtually eliminated the error that comes from the estimation of the populations. For this compound, we obtained a theoretical value (-16.5 Hz) that deviates considerably in magnitude from experiment (12.1 Hz),[403] resulting in a relative error of 16.36%. As we did not find any correlation between the percentage of conformers with IMHBs and the error in the J-couplings for our set of γ-fluorohydrins, which if found could indicate an inability of the method to accurately calculate J-couplings for conformers with IMHBs, this result is surprising. Future work will focus on unraveling the source of this mismatch. Despite this, the excellent agreement obtained for the MP2/6-311++G(2d,p)/PCM data set leads us to believe that the protocol used to compute the J-coupling constants is sufficiently accurate to warrant a fair comparison between different data sets.

All in all, the NMR results here presented lead us to recommend that ANI-2x be used carefully in hybrid models for condensed-phase applications, especially for the modelling of compounds that have chemical interactions poorly described by ωb97X/6-31G* (*e.g.*, HBs). This conclusion is further supported by determining the R^2 (0.86 *vs.* 0.70) and RMSE (1.79 *vs.* 3.42 Hz) values of GAFF.MOD-RESP/$CHCl_3$ and ANI-2x-RESP/$CHCl_3$, respectively, relative to the ωb97X/6-31G*/PCM NMR data. These results corroborate the findings regarding the sampling accuracy in chloroform solution (Figures 7.9 and 7.10), which indicate that GAFF.MOD-RESP/$CHCl_3$ reproduces ωb97X/6-31G*/PCM better than ANI-2x-RESP/$CHCl_3$.

Table 7.2: Experimental and computed J-couplings ($^{h1}J_{OH\cdots F}$) obtained in $CDCl_3$.

Molecule	Experimental[a]	MP2[b]	ωb97X[c]	GAFF[d]	GAFF.MOD[e]	ANI-2x[f]
syn-A	6.6	-7.6	-16.5	-5.6 $\pm$ 0.2	-16.1 $\pm$ 0.4	-19.5
anti-A	1.9	-1.1	-7.4	-2.6 $\pm$ 0.4	-9.9 $\pm$ 0.6	-12.7
B	2.2	-2.2	-9.2	-1.6 $\pm$ 0.2	-10.8 $\pm$ 0.2	-13.7
C	1.7	-1.2	-6.6	-1.1 $\pm$ 0.2	-2.2 $\pm$ 0.3	-6.1
D	1.4	-1.3	-6.7	-0.85 $\pm$ 0.03	-5.4 $\pm$ 0.1	-10.5
E	3.5	-3.2	-11.3	-7.7 $\pm$ 0.2	-9.0 $\pm$ 0.2	-5.2
	1.4	-1.7	-1.0	-0.7 $\pm$ 0.1	-0.6 $\pm$ 0.4	-3.2
F	0.6	-0.6	-1.9	-0.25 $\pm$ 0.03	-0.4 $\pm$ 0.1	-0.4
	0.6	-0.7	-2.7	-0.97 $\pm$ 0.03	-4.1 $\pm$ 0.2	-7.9
G	0.4	-0.1	-1.1	-0.06 $\pm$ 0.05	-2.2 $\pm$ 0.2	-2.5
	0.4	-0.3	-2.0	-0.11 $\pm$ 0.02	-2.3 $\pm$ 0.1	-3.6
H	0.7(q)[g]	-0.8	-0.8	-0.92 $\pm$ 0.03	-1.46 $\pm$ 0.01	-2.7
I	0.3(q)[g]	-0.4	-1.0	0.39 $\pm$ 0.01	-1.29 $\pm$ 0.01	-1.8
R^2		0.96	0.87	0.68	0.75	0.58
RMSE		0.41	4.80	1.28	4.77	6.78

[a]Sign not determined; [b]MP2/6-311++G(2d,p)/PCM; [c]ωb97X/6-31G*/PCM; [d]GAFF-RESP/$CHCl_3$; [e]GAFF.MOD-RESP/$CHCl_3$; [f]ANI-2x-RESP/$CHCl_3$; [g]quartet

7.4 Conclusions

We have presented a comparative study that evaluates the performance of an NNP (ANI-2x), a conventional FF (GAFF), and an optimally tuned FF (GAFF.MOD) relative to experimental and QM data. To this end, for a set of γ-fluorohydrins, we assessed the energetic and geometric agreement in the gas phase, the sampling accuracy in the gas phase and chloroform solution, and the accuracy of the estimates of the J-coupling constants relative to experimental data. The results and discussions presented highlight the strengths and weaknesses of each model, providing guidelines for future development of FFs and ML potentials.

We believe this study may have implications in different areas of chemistry and biology, especially for those interested in applications involving modelling of small organic compounds, which is very important for the drug design community.

The acceptance rates obtained in the nMC-MC simulations, which used ANI-2x as the approximate potential, indicate high similarity between ANI-2x and ωb97X/6-31G*. These nMC-MC results also confirm that ANI-2x can produce stable MD simulations, with numerical stability comparable to that of GAFF-like FFs. The high similarity between ANI-2x and ωb97X/6-31G* was further confirmed by analysing their energetic agreement. Overall, in the gas phase, ANI-2x is the model that best reproduces the ωb97X/6-31G* energy landscapes, followed by GAFF.MOD and GAFF. Optimisation of GAFF to GAFF.MOD, however, led to a significant improvement in the energetic performance, demonstrating the power of bespoke reparameterisation. ANI-2x also proved to be the model that best reproduces the energies and geometries of the gas-phase ωb97X/6-31G* minima, followed by GAFF.MOD and GAFF. Despite this high performance, ANI-2x tends to overstabilise global minima, a feature that scales the relative energies of conformers and, consequently, impacts the relative populations.

In the gas phase, the superior accuracy of ANI-2x in reproducing the ωb97X/6-31G* energy landscape does not always translate into higher sampling accuracy than GAFF.MOD, as the energetic errors of ANI-2x negatively impact its performance in this regard. Surprisingly, GAFF.MOD shows similar performance to ANI-2x in terms of sampling accuracy, raising questions of, given its costs, whether ANI-2x should be used over a FF to sample the conformational landscape of small organic molecules. Hence, while GAFF.MOD performs poorly in describing the minutiae of the ωb97X/6-31G* energy landscape, GAFF.MOD performs reasonably well in terms of relative energy, thus achieving similar performance as ANI-2x. When MP2/6-311++G(2d,p) is the QM reference, GAFF stands out as the best performing model. This difference suggests that ωb97X/6-31G* and MP2/6-311++G(2d,p) predict considerably different physical behaviour for the set of γ-fluorohydrins considered in this study.

In chloroform solution, GAFF.MOD-RESP/$CHCl_3$ is the model that predicts better sampling accuracy when ωb97X/6-31G*/PCM is used as the reference, significantly outperforming ANI-2x-RESP/$CHCl_3$. The decrease in performance of the ANI-2x potential when used in a hybrid NNP/MM model suggests significant imbalances between the ligand-solvent intermolecular and the ligand intramolecular interactions. Hence, combining available LJ 12-6 parameters with an NNP should be done with caution because of the potential decrease in the NNP performance. ANI-2x also tends to overestimate the populations of conformers with IMHBs and predicts stronger hydrogen bonding than expected. These physical artifacts may be caused by the lack of dispersion interactions in the hybrid ωb97X/6-31G* functional and can be potentially mitigated by the use of dispersion corrections.

The NMR analysis also leads us to reinforce the caution of using ANI-2x for condensed-phase applications, especially for the modelling of compounds that have chemical interactions poorly handled by ωb97X/6-31G*. In terms of performance, MP2/6-311++G(2d,p)/PCM is the level of theory that best reproduces the experimental data, followed by GAFF-RESP/$CHCl_3$. These results support the idea that MP2/6-311++G(2d,p) is closer to the experiment than ωb97X/6-31G*, and that GAFF is the model that best reproduces the experimental data. Furthermore, GAFF-RESP/$CHCl_3$ is the model that best reproduces the MP2/6-311++G(2d,p)/PCM data, and GAFF.MOD-RESP/$CHCl_3$ the model that best reproduces the ωb97X/6-31G*/PCM data. All in all, these observations corroborate the findings of the sampling accuracy in chloroform solution, as they indicate that currently FFs are more suited to be used in hybrid models than ANI-2x.

It is indisputable that ANI-2x has its merits, especially when it comes to modelling molecules in the gas phase. The merits of ANI-2x lie in a generally good description of the PES of small organic compounds. However, this study shows some issues with using ANI-2x may have been overlooked in many applications. ANI-2x has a tendency to predict stronger-than-expected hydrogen bonding, the tendency to overstabilise global minima, and cannot properly capture dispersion

interactions. These are observations that may limit the widespread use of ANI-2x in the long term, suggesting that an improved version of this NNP would be welcome. Furthermore, the use of ANI-2x in an NNP/MM framework should be undertaken with caution due to the potential inconsistencies that may arise. For some systems, directly combining ANI-2x with readily-available FF parameters may lead to imbalances between different parts of the hybrid model, resulting in a significant decrease in performance. Owing to their internal consistency, conventional and optimally tuned FFs remain the best models available for simulating condensed-phase systems. FFs also have the advantage of being computationally cheaper than NNPs. For NNP/MM models to become routinely used in condensed-phase simulations, sets of nonbonded MM parameters consistent with NNPs need to be derived; otherwise, the accuracy of NNP/MM models will always be compromised to some extent.

Finally, we must stress that all our conclusions are based on the particular set of γ-flurohydrins considered in this study. We cannot exclude that the observed performance may vary for other systems and that our conclusions may not be, therefore, always extrapolatable. In future work, it would be interesting to use the multilevel MC method presented in Chapter 6 to generate QM/MM ensembles of structures against which we could compare the performance of the NNP/MM data. Despite the significant computational cost that such calculations would entail, QM/MM structures could provide a better reference than the current QM/PCM level of theory used, which, for example, does not consider intermolecular interactions between the ligand and the (implicit) solvent. Moreover, to better understand the origin of the ANI-2x deficits observed, it would also be valuable to train an ANI-like model using QM energies and forces. In principle, this optimally tuned ANI-like model should outperform both ANI-2x and GAFF.MOD, thus proving the superiority of NNPs for modelling small organic molecules.

In the next chapter, we summarise the main conclusions of the thesis and provide suggestions to guide future research efforts.

Chapter 8

Conclusions

The work presented in this thesis aimed to develop, apply, and benchmark molecular models and simulations methods used for the computational modelling of small organic molecules. Ligands were the main class of compounds covered in this work, as they show a chemical diversity and conformational flexibility that requires continual improvement of the conformational analysis methods available to study them. Ligands were also molecules of special interest for Astex and AstraZeneca, pharmaceutical companies that collaborated in this work.

The research projects presented in this thesis focused on three main strands. The first was the development, implementation, and validation of a software to parameterise FFs by fitting to QM data. This software came to be known as ParaMol, and it is available to be used by the scientific community. ParaMol has proved to be an efficient and robust tool to parameterise FFs, requiring as little user intervention as possible to derive optimal FF parameters. Although various parameterisation tools were available when we started developing ParaMol, most of them were hard to use, required mastering of specific software, and were limited to certain functional forms. ParaMol aimed to overcome these limitations by presenting a framework with various built-in parameterisation protocols that can be used by anyone with minimal Python knowledge. ParaMol was developed with flexibility in mind so that users can design parameterisation

protocols specifically tailored to their needs. Besides that, ParaMol was designed to be easily extendable, and developers with some proficiency in Python can implement routines that cover currently unavailable functional forms and parameterisation protocols. Although this philosophy of software usability and extendability has become commonplace for some FF parameterisation tools, it was uncommon - if not absent - when we started developing ParaMol. This rapid paradigm shift is a good example of the importance of FF parameterisation, which remains an incompletely solved problem despite the most recent advances in theory and methods.

The second strand aimed to bridge the efficiency of FFs with the accuracy of QM methods. This research resulted in the development, implementation, and validation of an nMC-MC algorithm that allows quantum configurational ensembles to be generated by performing sampling using approximate potentials. The use of approximate potentials for sampling, despite their potential accuracy issues, presents indisputable advantages relative to QM methods because it allows for extensive sampling of the conformational landscape of molecules. QM methods, while highly accurate, are a far cry from becoming the standard method of choice to simulate molecules due to their computational cost. The nMC-MC algorithm we developed proved to be an efficient and robust method to bridge the gap between cheap approximate potentials and expensive but accurate QM levels. The nMC-MC algorithm was implemented in ParaMol and is available to be used by the scientific community. Importantly, our implementation of the nMC-MC algorithm is agnostic to the used approximate potential and QM level, providing users total flexibility of choice regarding the models that best suit their systems of interest. This means that, for example, besides FFs, NNPs can also be used as approximate potentials, which is very useful considering the similarity to the QM level that NNPs can attain. Our implementation of the nMC-MC algorithm also allows for hybrid QM/MM or NNP/MM models to be used, paving the way towards the extensive sampling of large and realistic systems at high levels of accuracy.

The third and last strand of this thesis focused on benchmarking the current plethora of molecular models available to model small organic molecules. While FFs have been around for decades and have proved to be suitable for many applications, the advent of ML potentials has been a conspicuous game-changer that has undoubtedly revolutionised the way computational scientists approach molecular modelling. Scientific shifts of paradigm, however, do not come without uncertainty, as the emergence of new methodologies raises many unanswered and fundamental questions. In order to evaluate the current performance of ML potentials and FFs, we presented a comparative study that evaluates the performance of an NNP (ANI-2x), a conventional FF (GAFF), and an optimally tuned FF (GAFF.MOD) relative to experimental and QM data. To this end, for a set of γ-fluorohydrins, we assessed the energetic and geometric agreement in the gas phase, the sampling accuracy in the gas phase and chloroform solution, and the accuracy of the estimates of the $^{h1}J_{OH\cdots F}$ coupling constants relative to experimental data. The results and discussions presented highlight the strengths and weaknesses of each model, providing guidelines for future development of FFs and ML potentials.

In regards to the project presented in Chapter 5, which presents ParaMol and establishes the best practices to follow when employing specific parameterisation routes, the following conclusions were drawn:

- Parameterisation using the analytical LLS solver should be preferred over parameterisation using non-linear iterative optimisers. Although we obtained identical results using either method, non-linear iterative optimisers are prone to become trapped in local minima in parameter space, whereas the analytical LLS solver is deterministic and ensures the global minimum is obtained. Non-linear iterative optimisers are advantageous in situations in which it is desirable to find a specific local minimum, such as when the goal is to produce the right helical propensity or orientation of a drug molecule in a protein binding site.

- Dihedrals scans should be performed using the MM-relaxed approach, as fittings using the QM-relaxed approach are critically dependent on the intramolecular FF parameters, which may lead to biased optimisations.

- Non-Boltzmann weighting proved to be the most reliable weighting scheme, despite its tendency to overestimate transition-state energies and underestimate fluctuations. Unless these are undesirable features for a particular application, non-Boltzmann weighting is the recommended weighting scheme for routine parameterisations.

- Boltzmann weighting, which emphasises the description of QM minima, tends to overfit low energy regions of the PES at the cost of poorly describing the remainder of the energy landscape. Boltzmann weighting requires strong regularisation to produce FFs that can be potentially used in MM modeling.

- As uniform weighting equally allows for positive and negative $E^{MM} - E^{QM}$ values, it is prone to creating asymmetries in the PES, which often lead to spurious minima due to artificially large thermodynamics weights and poor description of underrepresented configurations (*e.g.*, transition states). Uniform weighting requires strong regularisation to mitigate some of these undesirable features.

- Using high weighting temperatures (greater than 500 K) in the non-Boltzmann and Boltzmann weightings schemes leads to results that become very similar to those obtained when using uniform weighting. The specific features of these methods are, therefore, more noticeable when using low weighting temperatures.

- Depending on the set of FF parameters that is aimed to be optimised, this parameterisation can be done using either dihedral scans or configurational ensembles as the fitting data sets. Parameterisations using configurational ensembles are, however, more sensitive to the weighting method than those that use dihedral scans. Furthermore, regardless of the data set

type and weighting scheme employed, regularisation should be applied in any situation, as it prevents FF parameters from straying away from physically-sensible values.

- Adaptive self-parameterisation is an attractive and useful way to derive optimal FF parameters, as it combines self-consistent sampling and parameter optimisation in a single protocol.

Overall, owing to its features and robustness, we believe ParaMol is a useful tool for the scientific community and that it has the potential to be used in various applications that require derivation of optimal FF parameters. Despite the many functionalities available in ParaMol, there is still much room for improvement in the software. For example, ParaMol would greatly benefit from allowing experimental data to be used for parameterisation purposes. The implementation of routines that allow parameterisation of FF functional forms belonging to classes II and III would also be of great use. Finally, implementation of Bayesian methods for parameter estimation would be welcome.[240–243]

In regards to the project presented in Chapter 6, which introduces the nMC-MC algorithm, a multilevel Monte Carlo method that allows estimation of quantum configurational while keeping the computational cost at a minimum, the following conclusions were drawn:

- Direct application of the nMC-MC algorithm using traditional FFs, such as GAFF, as the approximate potentials leads to very slow convergence of the target quantum configurational distributions due to low acceptance rates caused by poor phase space overlap between the MM and QM levels.

- FF reparameterisation proved to be an efficient strategy to increase the acceptance rates of the switching step from the MM to the QM level of theory, thus accelerating the sampling convergence of the target quantum configurational distribution.

- Both molecular size and chemical complexity are negatively correlated with the nMC-MC acceptance rates. The best acceptance rates were obtained for aniline, whereas the molecule with the lowest possible acceptance rate was the fragment of cpd 26.

- Hard DOFs, such as bonds and angles, are crucial to be reparameterised to increase the acceptance rates due to their large force constants.

- There is a strong, positive correlation between the nMC-MC switching step acceptance rates and the phase space overlap between the QM and MM levels. This correlation was confirmed by data obtained from various phase space overlap metrics, leading us to suggest the nMC-MC switching step acceptance rates as a robust metric of phase space overlap.

- The nMC-MC version of the adaptive self-parameterising algorithm, which combines sampling at the QM level and FF parameterisation in one scheme, is an efficient parameterisation method that limits the computational work to the strictly necessary and does not require *a priori* generation of a training data set of unknown size.

- Within a fixed point charge mechanical embedding framework, the nMC-MC algorithm is a viable methodology that permits recovery of the target QM/MM configurational ensemble. Currently, this is the limiting case, in terms of model complexity, that still allows reasonable acceptance rates to be obtained in the nMC-MC algorithm when using class I FFs.

All in all, the nMC-MC algorithm along with FF reparameterisation proved to be an efficient strategy to generate quantum configurational ensembles while keeping the computational cost to a minimum. The main drawback of the method is the severe negative impact that system size and target Hamiltonian complexity have on the nMC-MC acceptance rates, which ultimately limit its use to simple systems. A possible solution for this bottleneck may involve resorting to ML models or FFs with more advanced functional forms than those of class

I FFs, which may be able to capture physical phenomena such as anharmonic behaviour, couplings between DOFs, and non-additive electrostatic effects.

In regards to the project presented in Chapter 7, which, using a set of γ-fluorohydrins, presents a comparative study that evaluates the performance of an NNP (ANI-2x), a conventional FF (GAFF), and an optimally tuned FF (GAFF.MOD) relative to experimental and QM data, the following conclusions were drawn:

- nMC-MC simulations performed using ANI-2x as the approximate potential indicate a high similarity between this NNP and the level of theory it was trained to reproduce, ωb97X/6-31G*. These nMC-MC results also show that ANI-2x can produce stable MD simulations, with numerical stability comparable to that of GAFF-like FFs.

- In the gas phase, ANI-2x is the model that best reproduces the ωb97X/6-31G* energy landscape, followed by GAFF.MOD and GAFF. ANI-2x is also the model that best reproduces the energies and geometries of the gas-phase ωb97X/6-31G* minima, followed by GAFF.MOD and GAFF.

- The superior accuracy of ANI-2x in reproducing the ωb97X/6-31G* energy landscape does not always translate into higher sampling accuracy than GAFF.MOD, since ANI-2x shows similar performance to GAFF.MOD in this regard. Hence, while GAFF.MOD performs poorly in describing the minutiae of the ωb97X/6-31G* energy landscape, GAFF.MOD performs reasonably well in terms of relative energies, thus achieving similar performance as ANI-2x.

- ANI-2x excels in describing the minutiae of the ωb97X/6-31G* PES, especially in the gas phase. Nonetheless, ANI-2x also shows a tendency to predict stronger-than-expected hydrogen bonding, a tendency to overstabilise global minima, and cannot properly capture dispersion interactions.

These problems lead to energetic errors that mainly impact the relative energies between conformers, preventing ANI-2x from excelling in sampling accuracy.

- When MP2/6-311++G(2d,p) is the QM reference, GAFF stands out as the best performing model both in terms of sampling accuracy and energetic agreement, followed by GAFF.MOD and GAFF.

- In chloroform solution, GAFF.MOD-RESP/$CHCl_3$ is the model that predicts better sampling accuracy when ωb97X/6-31G*/PCM is used as the reference, significantly outperforming ANI-2x-RESP/$CHCl_3$. The decrease in performance of the ANI-2x potential when used in a hybrid NNP/MM model suggests significant imbalances between the ligand-solvent intermolecular and the ligand intramolecular interactions. Hence, the use of ANI-2x in an NNP/MM framework should be undertaken with caution due to the potential inconsistencies that may arise by directly combining available LJ 12-6 parameters with this NNP. Owing to their internal consistency, conventional and optimally tuned FFs remain the best models available for simulating condensed-phase systems.

- The NMR analysis revealed that MP2/6-311++G(2d,p)/PCM is the level of theory that best reproduces the experimental $^{h1}J_{OH\cdots F}$ coupling constants, followed by GAFF-RESP/$CHCl_3$. These results suggest that MP2/6-311++G(2d,p) is closer to the experiment than ωb97X/6-31G*, and that GAFF is the model that best reproduces the experimental data. Furthermore, GAFF-RESP/$CHCl_3$ is the model that best reproduces the MP2/6-311++G(2d,p)/PCM data, and GAFF.MOD-RESP/$CHCl_3$ the model that best reproduces the ωb97X/6-31G*/PCM data.

Although all our conclusions regarding the work of Chapter 7 are based on the particular set of γ-flurohydrins considered in it, they still provide some valid guidelines for future development of FFs and ML potentials. For example, an improved version of ANI-2x that addresses some of its current weaknesses

would be welcome. Furthermore, in terms of future work, it is clear that for NNP/MM models to become routinely used in condensed-phase simulations, sets of nonbonded MM parameters consistent with NNPs need to be derived so as not to compromise the accuracy of this hybrid model.

Given the complexity and scope of the problems addressed in this thesis, the results presented here are unable to provide complete or definitive solutions. Even so, we believe the outcomes of this project may have implications in different areas of chemistry and biology, especially for those interested in the modelling of small organic molecules in the gas phase and solution. The methods and software developed are available to be used and further developed by the scientific community. Moreover, the data, discussions, and conclusions may also provide insights to guide future research efforts that attempt to develop methods for efficient and accurate simulation of the conformational landscape of ligands.

Appendix A

Appendix of Chapter 5

A.1 Norfloxacin

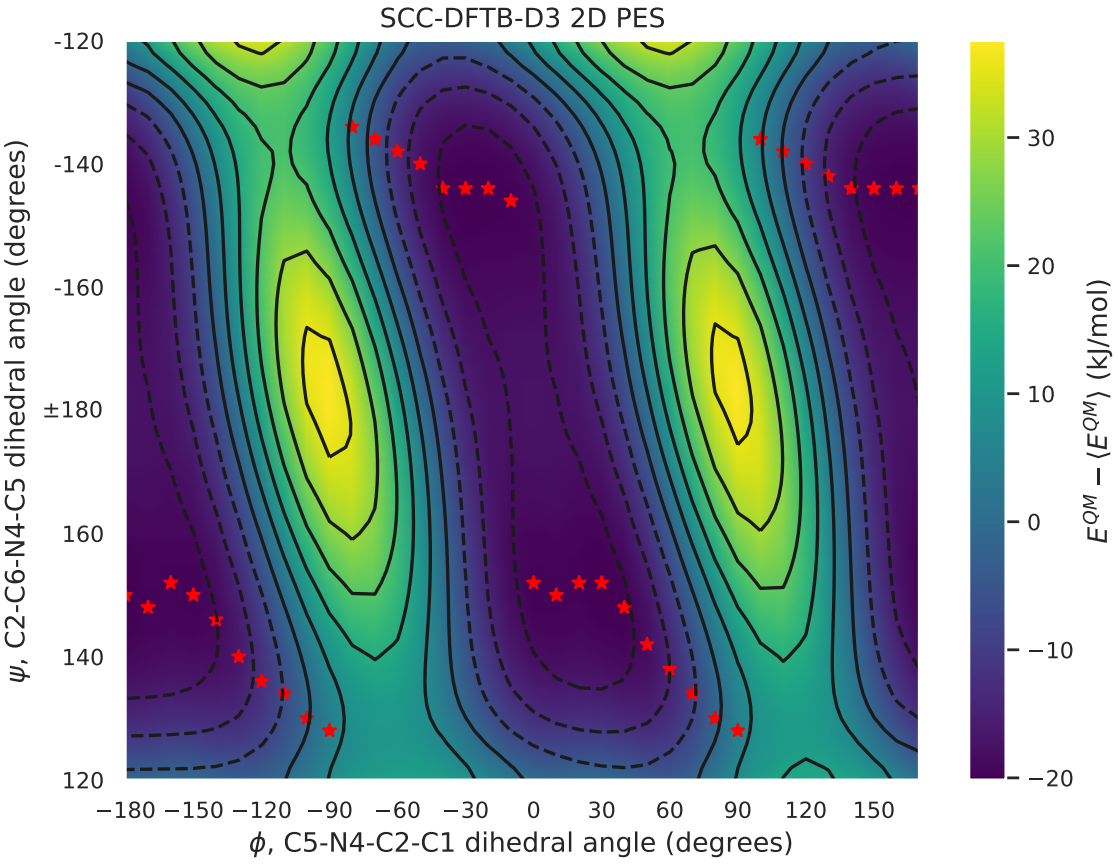

Figure A.1: SCC-DFTB-D3 PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) 2D dihedral scan for the norfloxacin analogue fragment. The red stars correspond to the minimum energy structure for a given ϕ dihedral angle value.

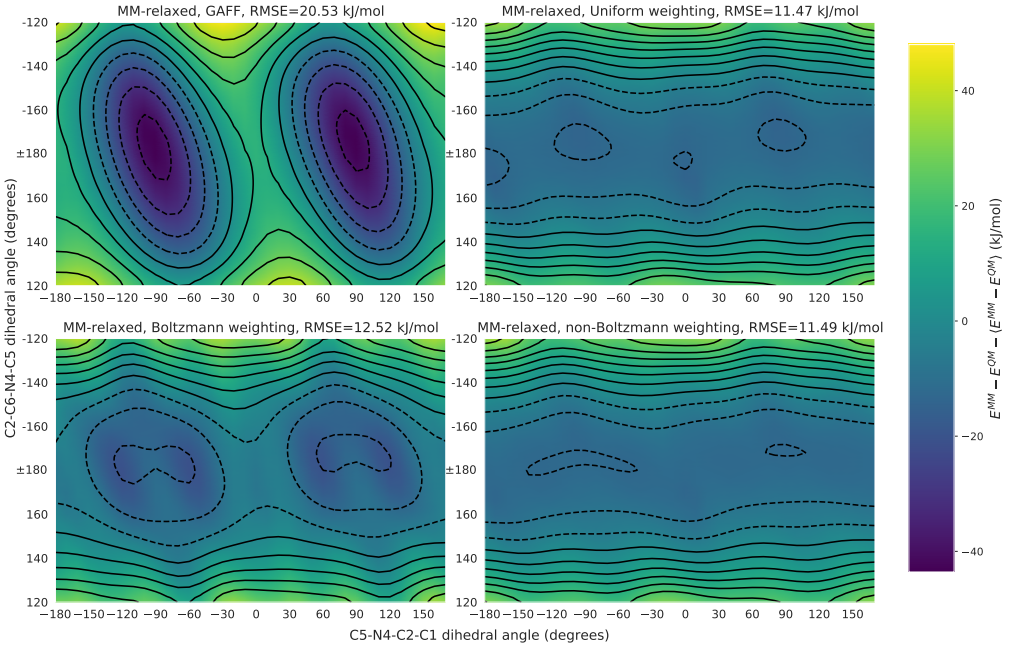

Figure A.2: Relative errors of the MM FFs (GAFF, uniform, Boltzmann and non-Boltzmann weightings) with respect to the target (SCC-DFTB-D3) PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) 2D dihedral scan. The MM-relaxed approach was employed to optimise the FFs.

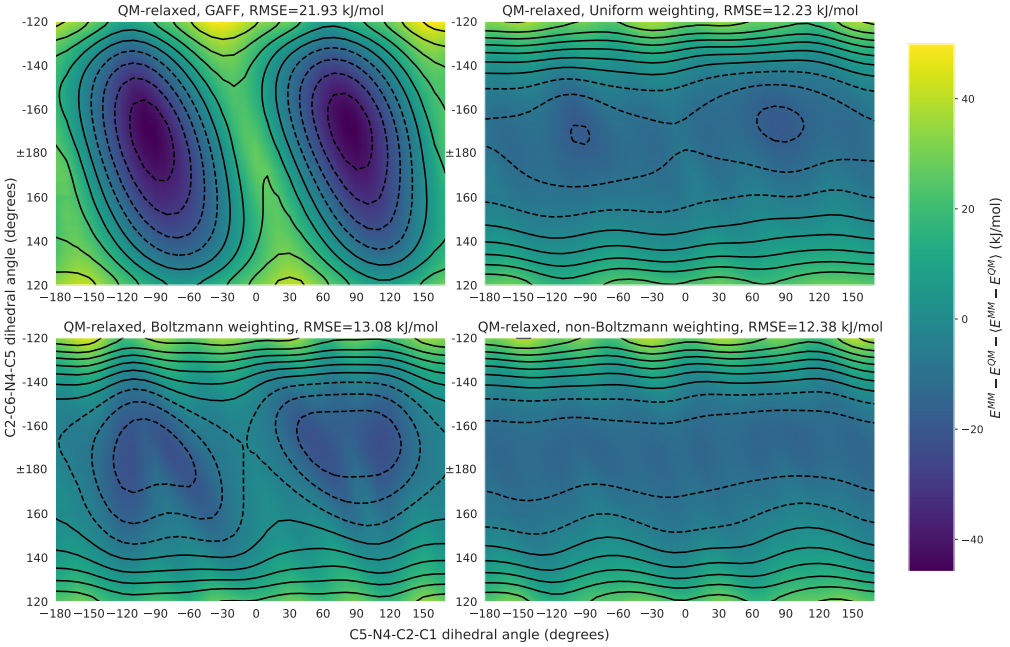

Figure A.3: Relative errors of the MM FFs (GAFF, uniform, Boltzmann and non-Boltzmann weightings) with respect to the target (SCC-DFTB-D3) PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) 2D dihedral scan. The QM-relaxed approach was employed to optimise the FFs

	GAFF	Uniform	Boltzmann	non-Boltzmann
			SciPy SLSQP solution	
V_1	0.00	-0.89 / -2.31	-0.61 / -1.47	-0.56 / -2.33
V_2	17.57	11.56 / 12.04	9.97 / 10.75	11.54 / 12.83
V_3	0.00	-4.40 / -2.24	-2.73 / -6.71	-4.21 / 0.15
V_4	0.00	-0.47 / 1.14	-1.48 / -0.25	-0.82 / 0.87
V_5	0.00	0.49 / -0.28	-1.45 / -4.00	0.49 / 0.24
V_6	0.00	0.30 / 0.13	1.37 / 1.24	0.31 / 0.62
			LLS solution	
V_1	0.00	-0.89 /-2.31	-0.61 / -1.48	-
V_2	17.57	11.56 /12.04	9.97 / 10.75	-
V_3	0.00	-4.40 /-2.24	-2.72 / -6.71	-
V_4	0.00	-0.47 / 1.14	-1.48 / -0.26	-
V_5	0.00	0.49 /-0.28	-1.45 / -4.00	-
V_6	0.00	0.30 / 0.13	1.37 / 1.24	-

Table A.1: Dihedral force constants (kJ mol^{-1}) derived using the MM-relaxed/QM-relaxed approach. The fittings were performed using the SCC-DFTB-D3 PES.

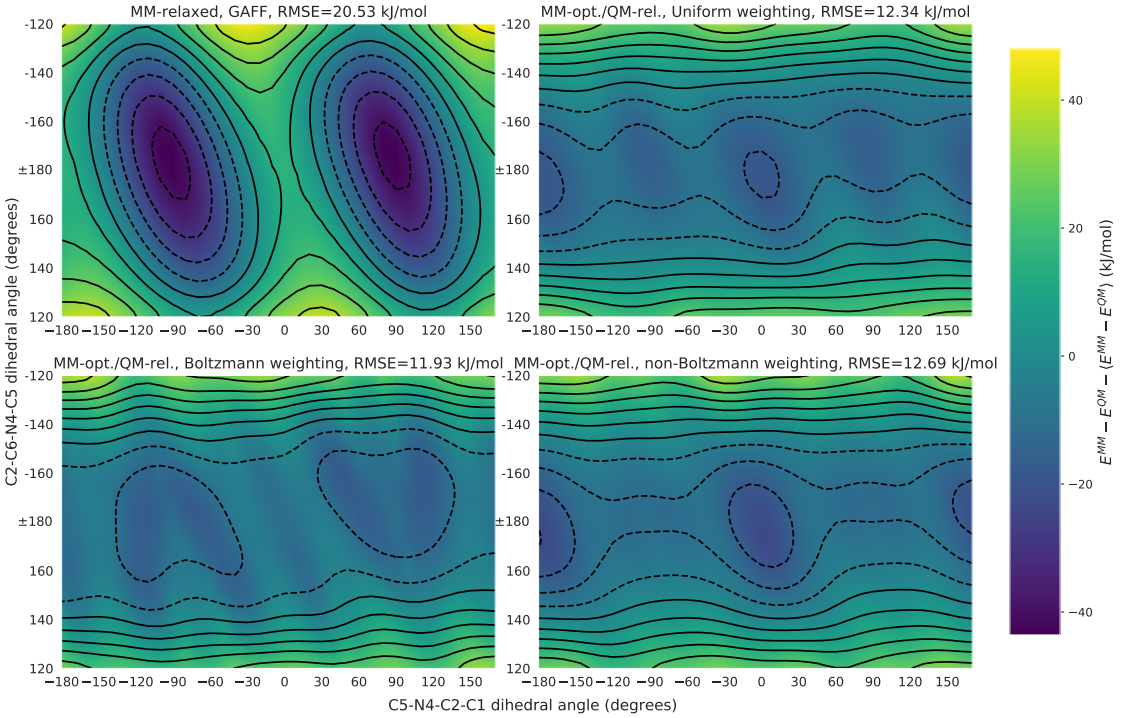

Figure A.4: Relative errors of the MM FFs (GAFF, uniform, Boltzmann and non-Boltzmann weightings) with respect to the target (SCC-DFTB-D3) PES of the C5-N4-C2-C1 (ϕ) *vs.* C2-C6-N4-C5 (ψ) 2D dihedral scan. The MM PESs used to calculate the relative errors were obtained by MM optimisation of the QM-relaxed PES of figure A.3.

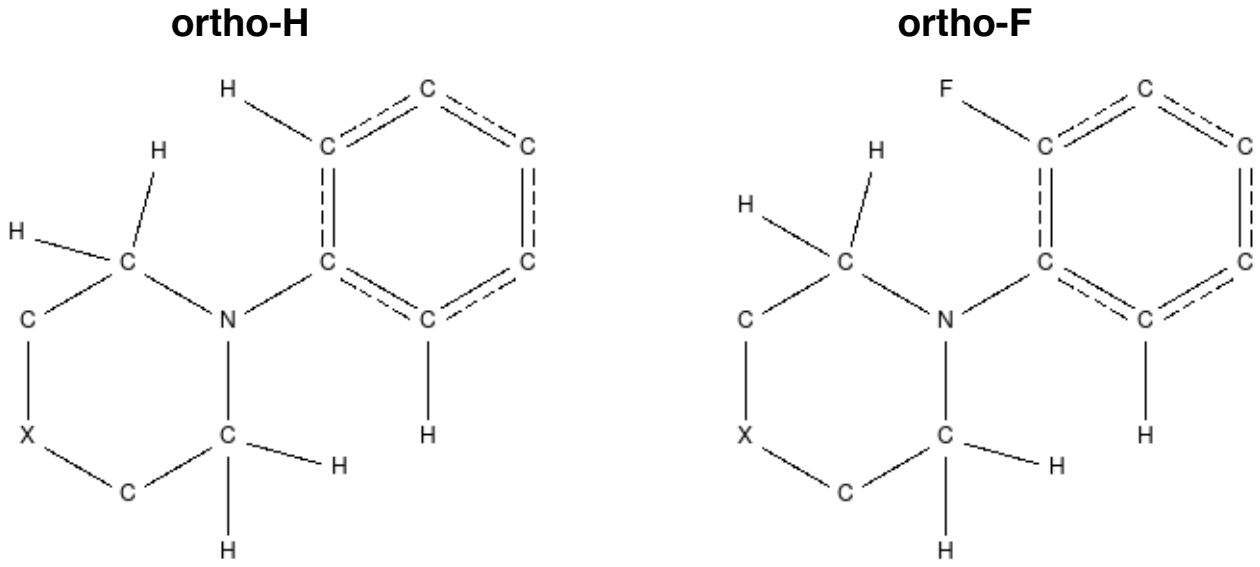

Figure A.5: Templates used to perform the search for crystal structures in ConQuest.

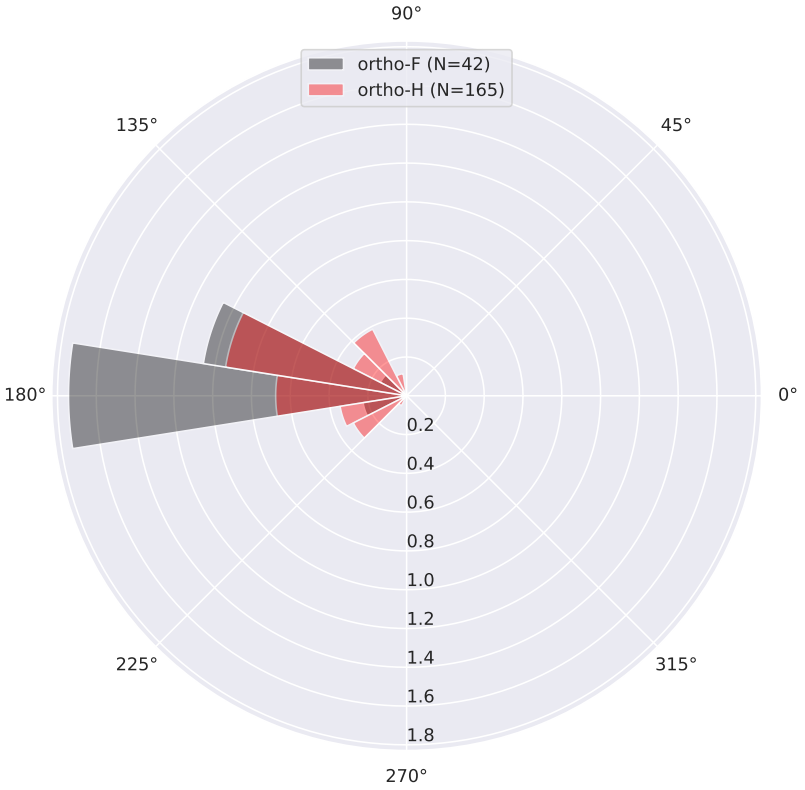

Figure A.6: Polar histograms (in frequency) for the C5-N4-C2-C1 dihedral of norfloxacin (ortho-F) and of the norfloxacin analogue (ortho-H) used in the paper. The crystal structures used in this plot were obtained from the CSD.[2] N corresponds to the number of hits obtained in ConQuest after pruning all structures that were not published in peer-reviewed journals. Ortho-F and ortho-H have similar torsional preferences.

A.2 Aspirin

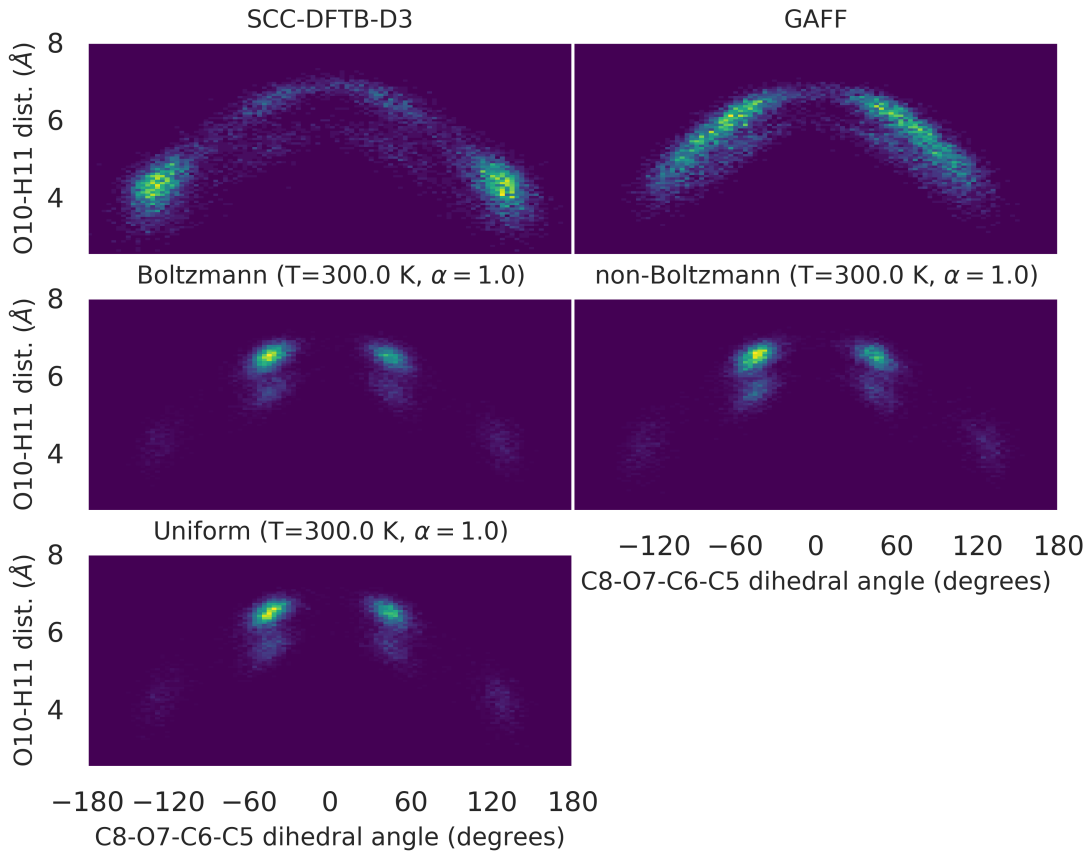

Figure A.7: Configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle of aspirin obtained from MD simulations using SCC-DFTB-D3, the GAFF, and the GAFF.MOD FFs. The latter were derived employing Boltzmann, non-Boltzmann, and uniform weighting, with a weighting temperature of 300 K, and a regularisation strength of $\alpha = 1.0$. No symmetry breaking of the pair of dihedrals C5-C6-O7-C8 and C4-C6-O7-C8 was enforced during the optimisation. The data set used in the reparameterisations was the SCC-DFTB-D3 configurational ensemble. All represented distributions contain 10000 configurations.

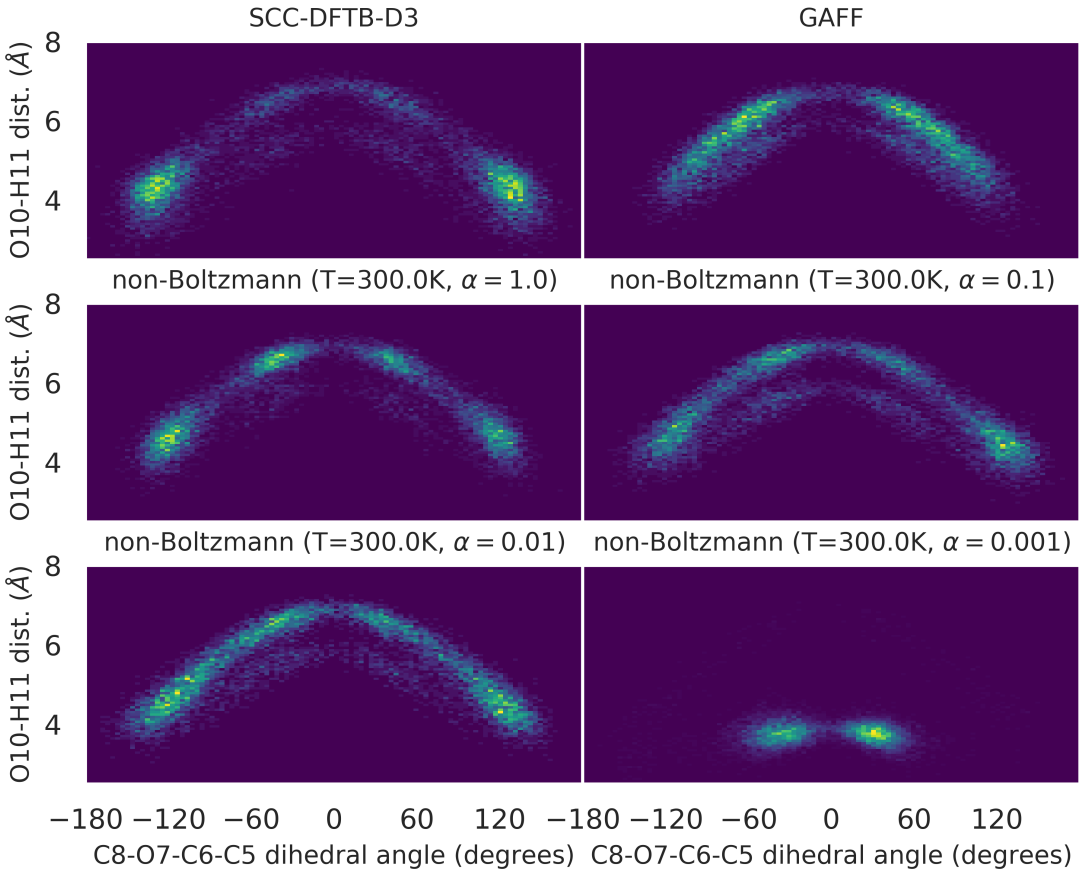

Figure A.8: Configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle of aspirin obtained from MD simulations using SCC-DFTB-D3, the GAFF, and the GAFF.MOD FFs. The latter were derived employing non-Boltzmann weighting, with a weighting temperature of 300 K and using different regularisation strengths ($\alpha = (1.0, 0.1, 0.01, 0.001)$). The data set used in the reparameterisation was the SCC-DFTB-D3 configurational ensemble. All represented distributions contain 10000 configurations.

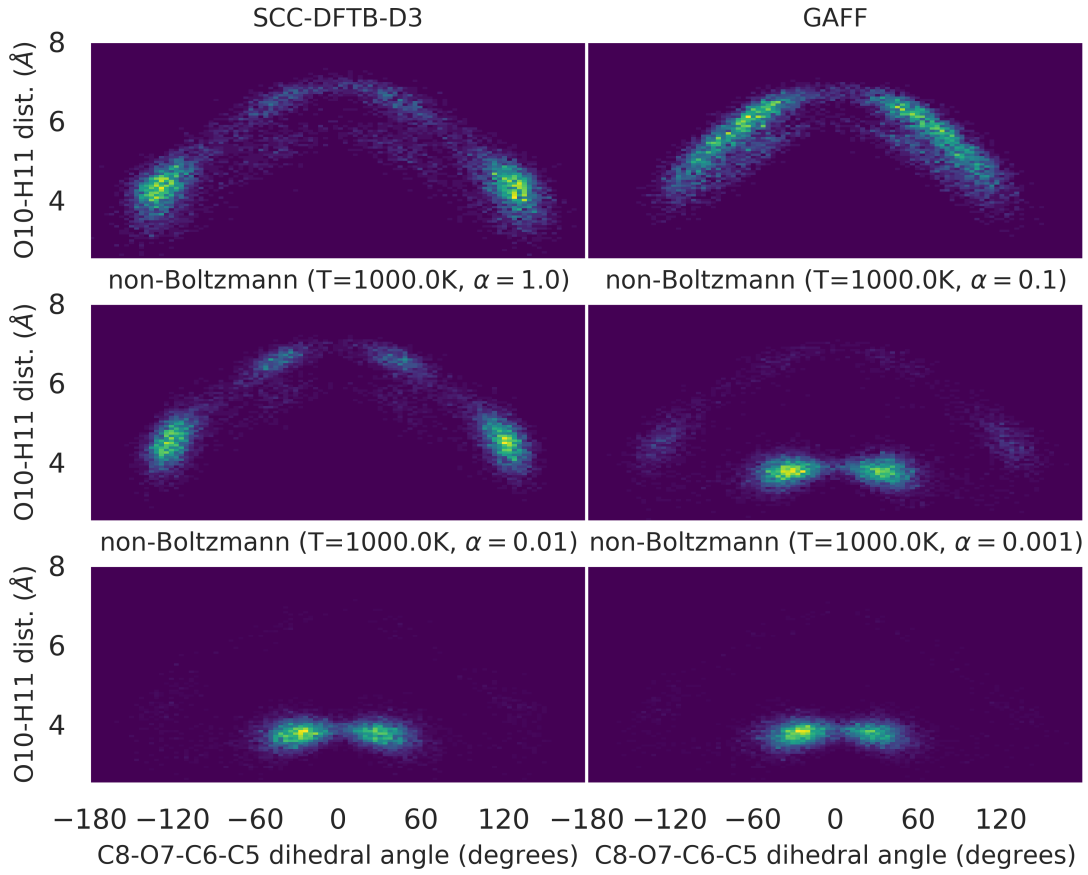

Figure A.9: Configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle of aspirin obtained from MD simulations using SCC-DFTB-D3, the GAFF, and the GAFF.MOD FFs. The latter were derived employing non-Boltzmann weighting, with a weighting temperature of 1000 K and using different regularisation strengths ($\alpha = (1.0, 0.1, 0.01, 0.001)$). The data set used in the reparameterisation was the SCC-DFTB-D3 configurational ensemble. All represented distributions contain 10000 configurations.

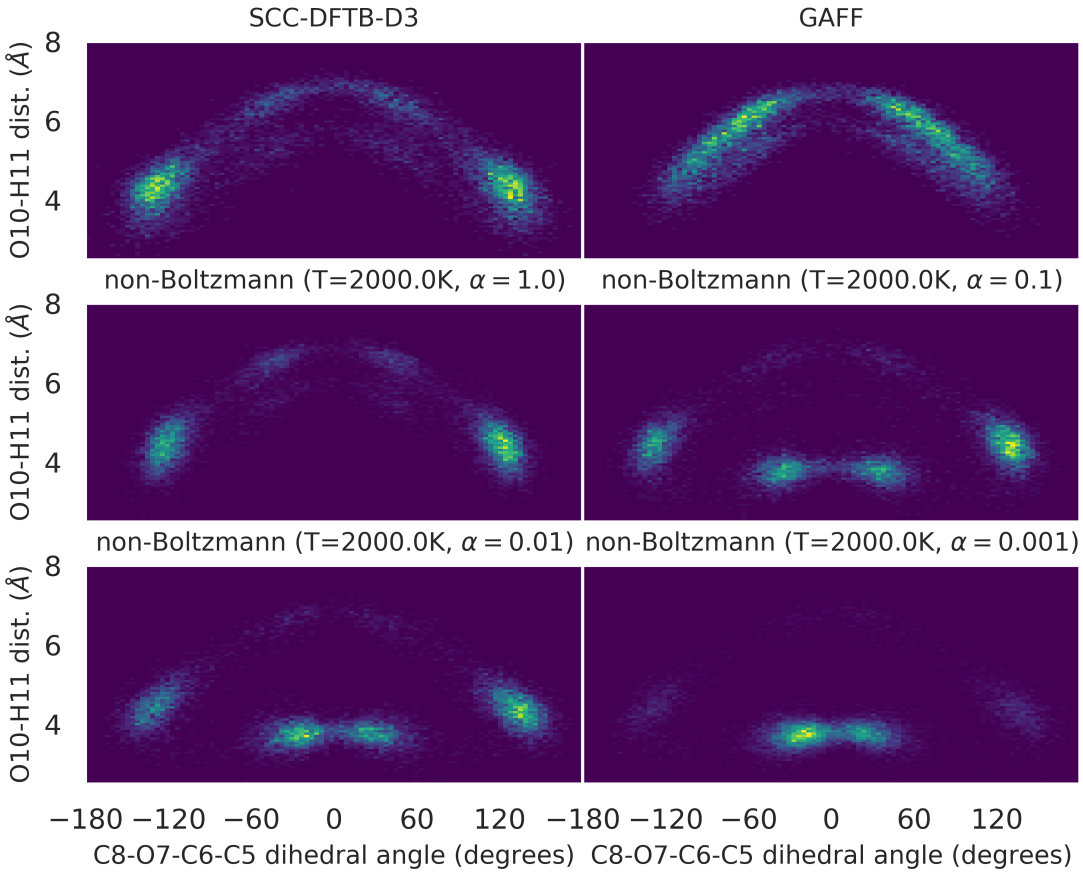

Figure A.10: Configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle of aspirin obtained from MD simulations using SCC-DFTB-D3, the GAFF, and the GAFF.MOD FFs. The latter were derived employing non-Boltzmann weighting, with a weighting temperature of 2000 K and using different regularisation strengths ($\alpha = (1.0, 0.1, 0.01, 0.001)$). The data set used in the reparameterisation was the SCC-DFTB-D3 configurational ensemble. All represented distributions contain 10000 configurations.

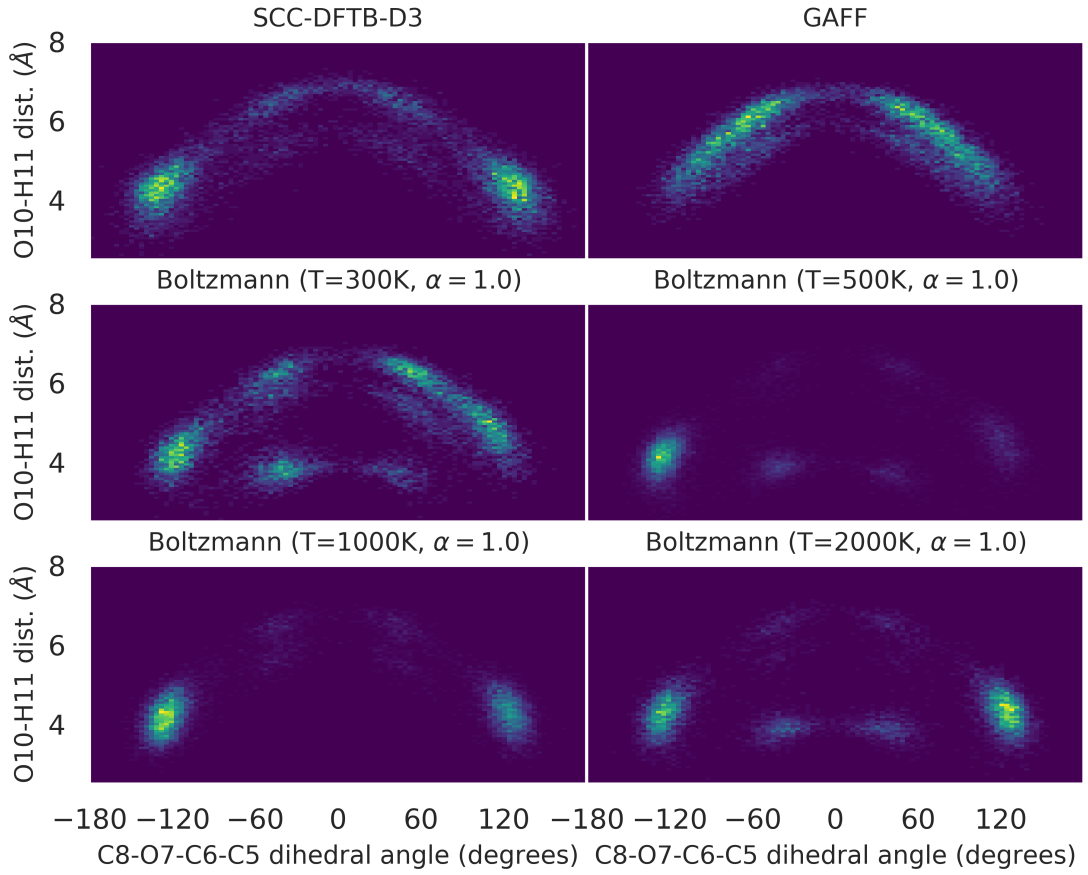

Figure A.11: Configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle of aspirin obtained from MD simulations using SCC-DFTB-D3, the GAFF, and the GAFF.MOD FFs. The latter were derived employing Boltzmann weighting, with weighting temperatures of 300, 500, 1000, and 2000 K and using a regularisation strength of $\alpha = 1.0$. The data set used in the reparameterisation was the SCC-DFTB-D3 configurational ensemble. All represented distributions contain 10000 configurations.

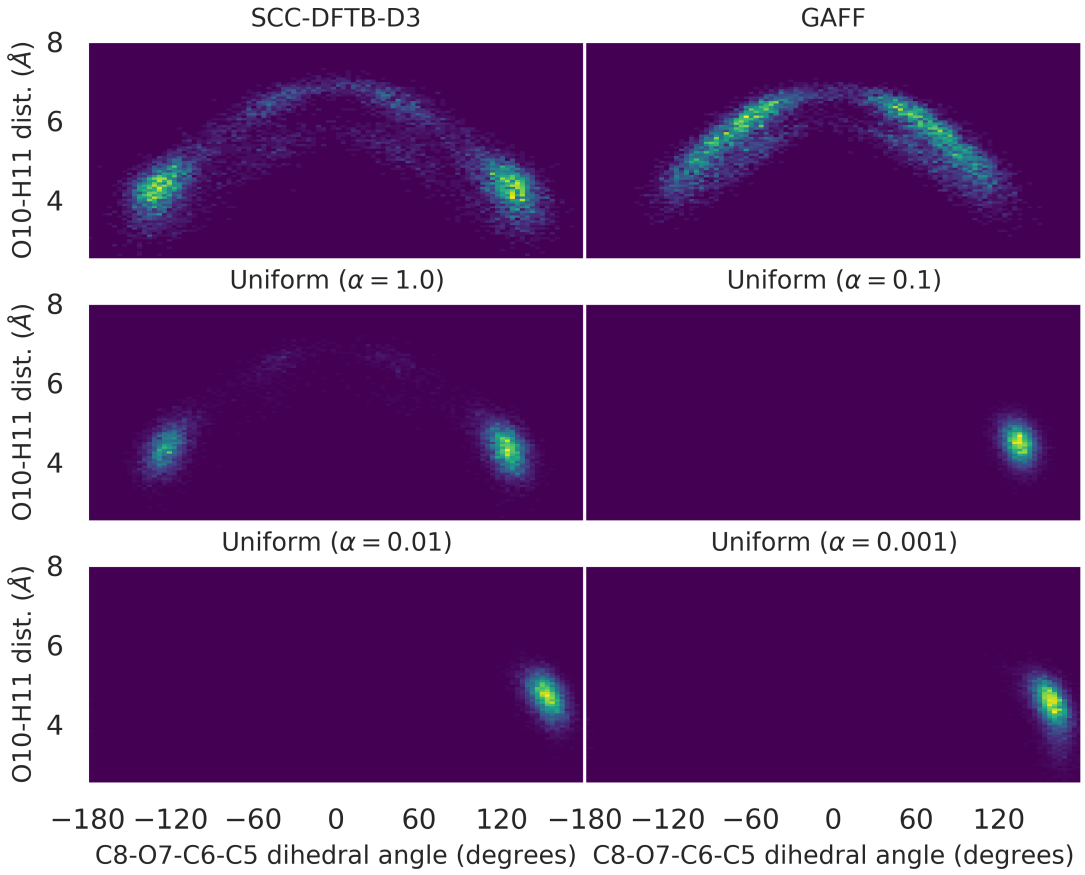

Figure A.12: Configurational distributions of the O10-H11 distance *vs.* the C5-C6-O7-C8 dihedral angle of aspirin obtained from MD simulations using SCC-DFTB-D3, the GAFF, and the GAFF.MOD FFs. The latter were derived employing uniform weighting, with different regularisation strengths ($\alpha = (1.0, 0.1, 0.01, 0.001)$). The data set used in the reparameterisation was the SCC-DFTB-D3 configurational ensemble. All represented distributions contain 10000 configurations.

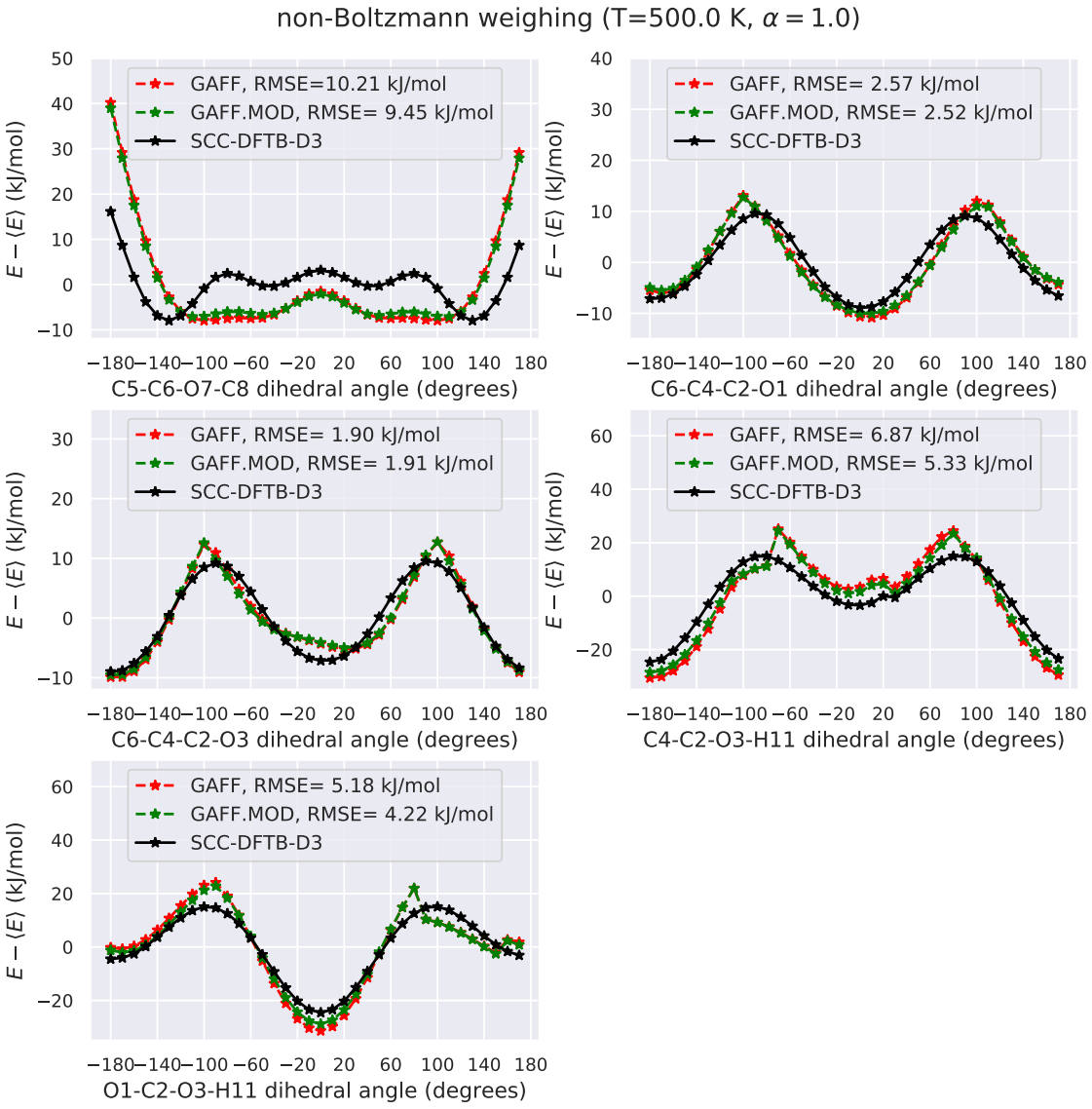

Figure A.13: Comparison of the SCC-DFTB-D3, GAFF, and GAFF.MOD (reparameterised FF) dihedral energy profiles for the C5-C6-O7-C8, C6-C4-C2-O1, C6-C4-C2-O3, C4-C2-O3-H11, and O1-C2-O3-H11 dihedral angles. The GAFF curves correspond to MM-relaxed energy profiles. The GAFF.MOD FF was obtained by employing the MM-relaxed approach with non-Boltzmann weighting (T=500.0 K, $\alpha = 1.0$). The parameters of the dihedrals represented in this Figure were concomitantly optimised along those of the C4-C6-O7-C8 dihedral using the ParaMol's automatic soft dihedral parameterisation task.

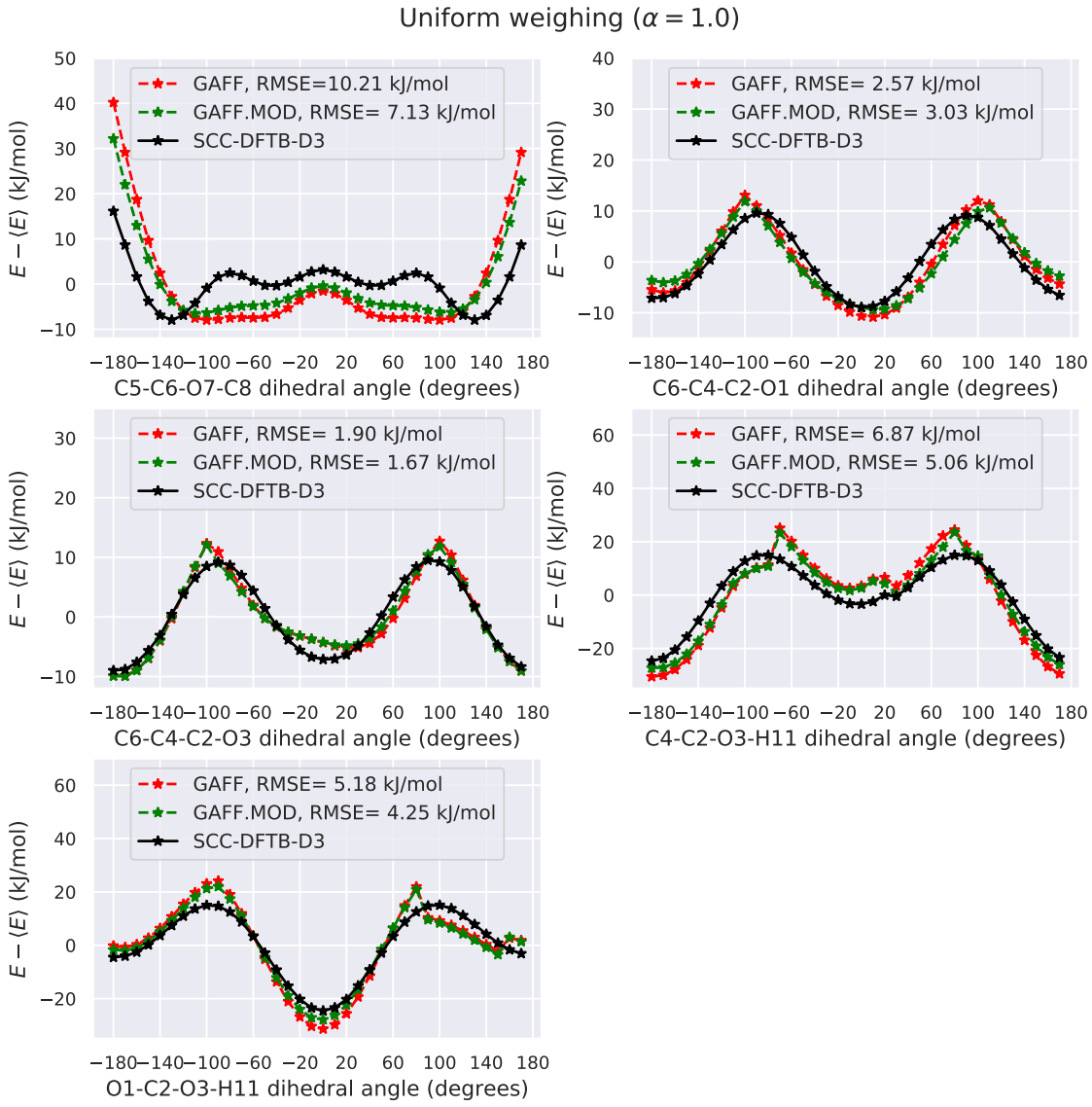

Figure A.14: Comparison of the SCC-DFTB-D3, GAFF, and GAFF.MOD (reparameterised FF) dihedral energy profiles for the C5-C6-O7-C8, C6-C4-C2-O1, C6-C4-C2-O3, C4-C2-O3-H11, and O1-C2-O3-H11 dihedral angles. The GAFF curves correspond to MM-relaxed energy profiles. The GAFF.MOD FF was obtained by employing the MM-relaxed approach with uniform weighting ($\alpha = 1.0$). The parameters of the dihedrals represented in this Figure were concomitantly optimised along those of the C4-C6-O7-C8 dihedral using the ParaMol's automatic soft dihedral parameterisation task.

A.3 Caffeine

In every iteration of the adaptive parameterisation of caffeine to the SCC-DFTB-D3 level of theory, 100 new configurations separated 0.5 ps from each other were generated and added to the previous ones. These configurations were obtained

using Langevin dynamics with a friction coefficient of 2 ps^{-1}, a time-step of 1 fs, and a temperature of 300 K. The adaptive parameterisation procedure was deemed to be converged when the root-mean-square deviation of the parameters between two successive iterations was less than 10^{-4}. The adaptive parameterisation performed 22 iterations, corresponding to a total of 2200 structures in the last iteration.

The plots of the RMSD of the parameters and the components of the objective function as a function of the iteration number are shown in figure A.15. The plot that shows the correlation between the SCC-DFTB-D3 energies and the MM energies is represented in figure A.16. The atomic forces errors are shown in the molecular structures of figure A.17.

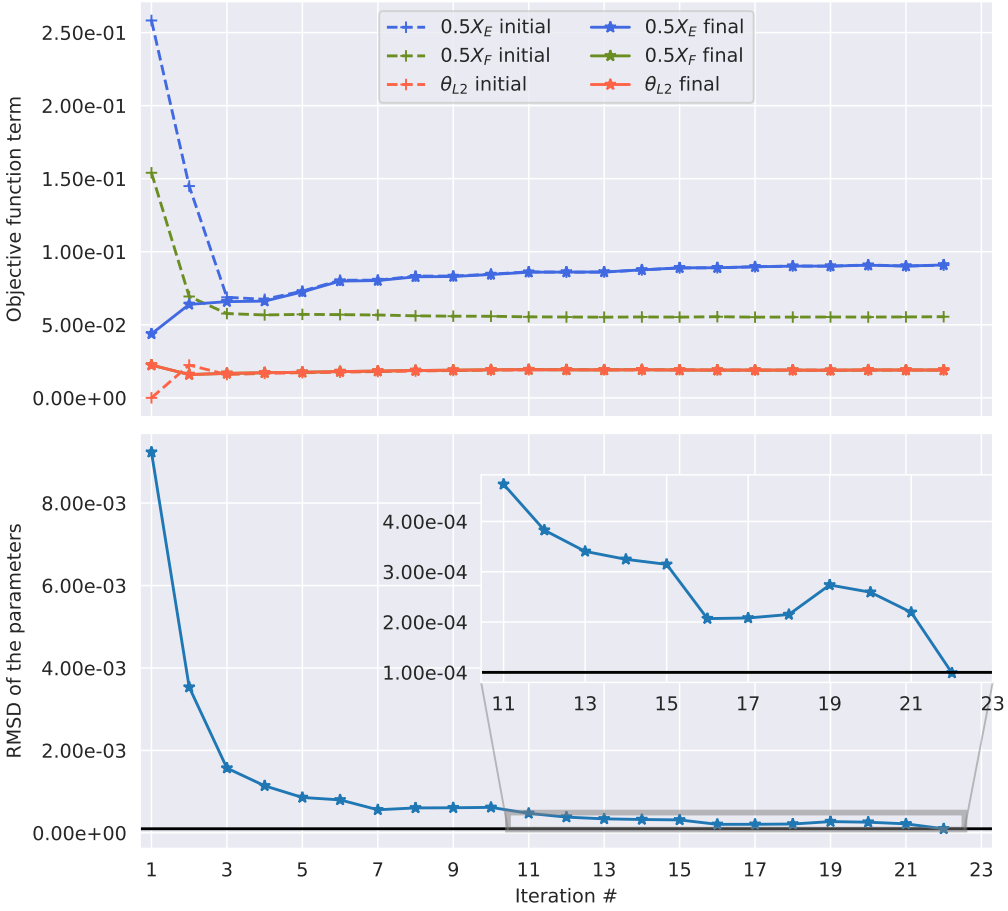

Figure A.15: Top panel: Plot of the values of each term included in the objective function at the beginning (dashed lines) and end (solid lines) of each iteration. Bottom panel: Plot of the RMSD of the parameters as a function of the iteration number.

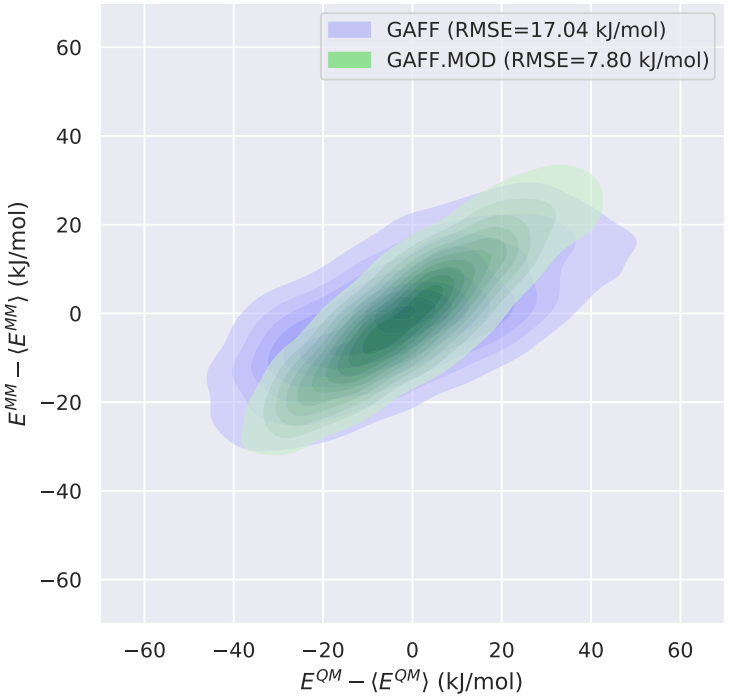

Figure A.16: Correlation between the QM energies and the MM energies of caffeine before and after the adaptive reparameterisation to the SCC-DFTB-D3 level of theory. Each data sets consists of 1000 configurations generated though a short MD simulation that used the respective FF. The RMSE of the energy improved from 17.04 kJ mol^{-1} (GAFF) to 7.80 kJ mol^{-1} after reparameterisation (GAFF.MOD).

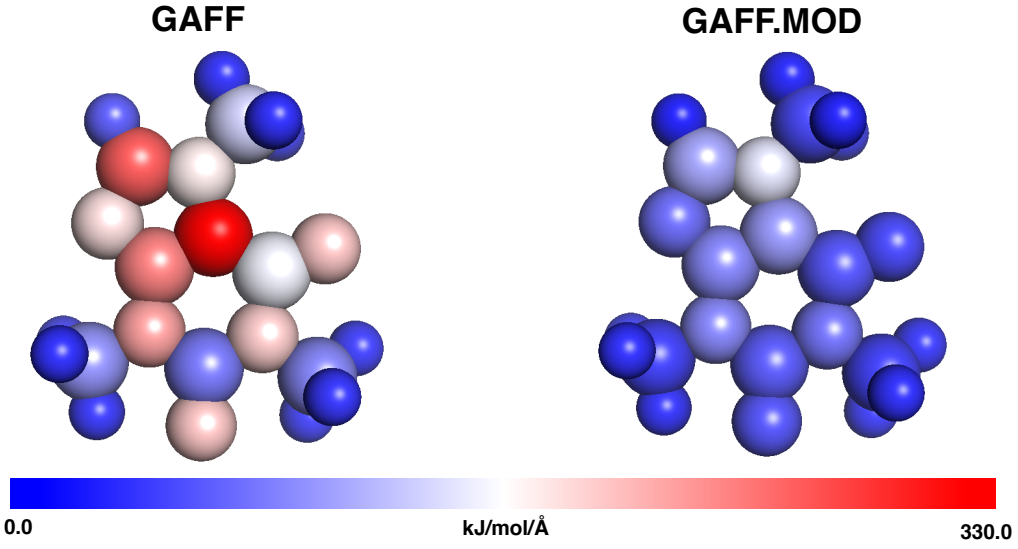

Figure A.17: Atomic force errors before (GAFF, left) and after (GAFF.MOD, right) reparameterisation to the SCC-DFTB-D3 level of theory. The average RMSE of the atomic forces improved from 124.67 kJ mol^{-1} $Å^{-1}$ $atom^{-1}$ (GAFF) to 57.86 kJ mol^{-1} $Å^{-1}$ $atom^{-1}$ after reparameterisation (GAFF.MOD).

A.4 ParaMol's soft dihedral parameterisation algorithm

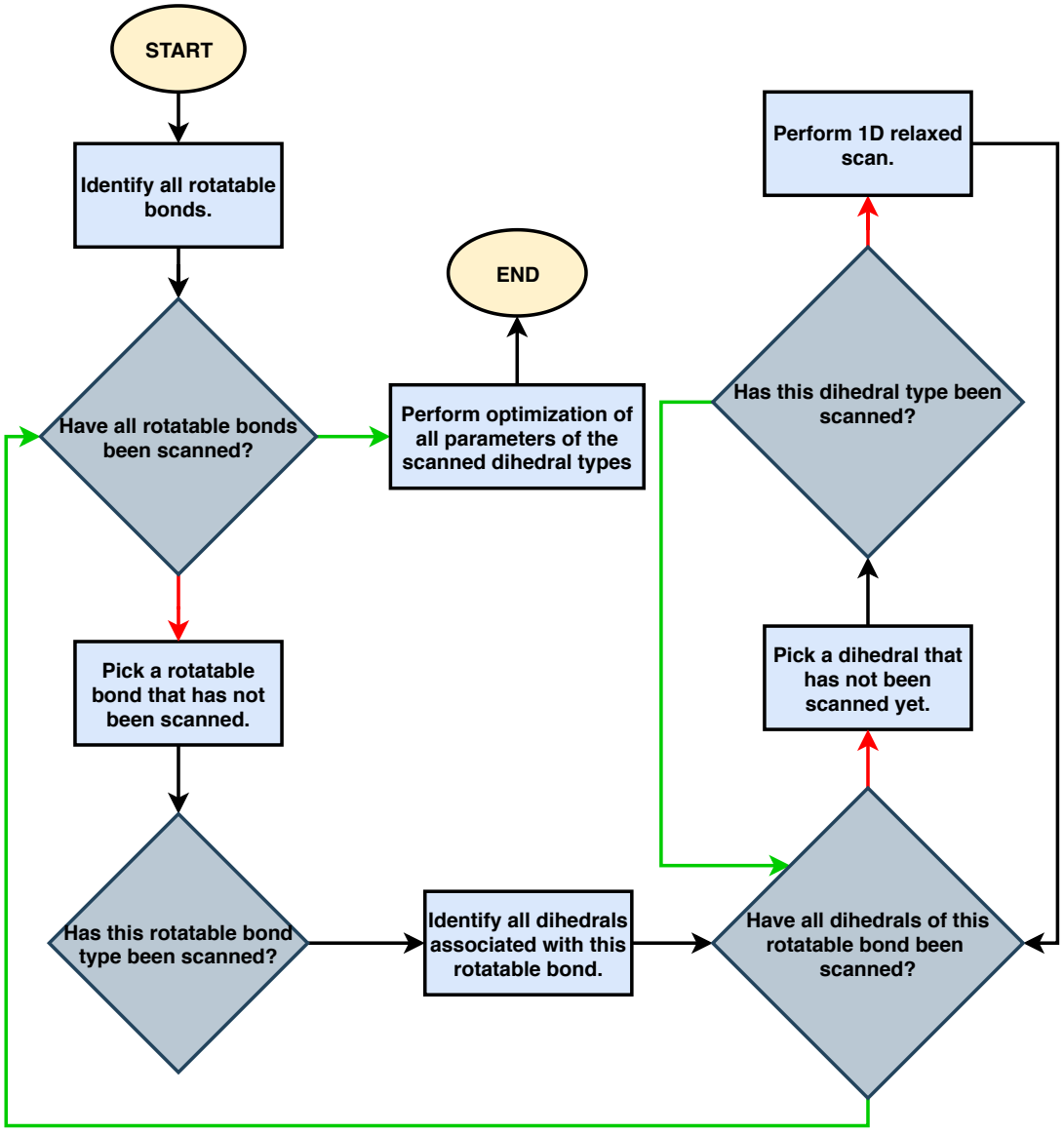

Figure A.18: Flowchart representing the workflow of the ParaMol's built-in task that automatically identifies and optimises soft dihedrals. The green arrows denote conditionals for which the evaluated condition is true, whereas the red arrows denote conditionals for which the evaluated condition is false.

Appendix B

Appendix of Chapter 6

B.1 Phase space overlap metrics

In the phase space overlap calculations of Chapter 6, phase space is employed as a synonym of configuration space as only situations that compare total energy distributions at the same temperature are considered, making the momentum coordinates irrelevant. To confirm that the momentum coordinates do not affect these calculations, let us first consider the metric defined in equation (6.19). Ω is independent of the kinetic energy because

$$\begin{aligned} \Delta E_{QM}^{QM\rightarrow MM} &= E_{QM}^{MM} - E_{QM}^{QM} \\ &= U_{QM}^{MM} + K_{QM} - U_{QM}^{QM} - K_{QM} \\ &= U_{QM}^{MM} - U_{QM}^{QM} \end{aligned} \tag{B.1}$$

in which K_{QM} was included in E_{QM}^{MM} and E_{QM}^{QM} as both total energies consider frames sampled during a QM MD simulation, though with potential energies evaluated at the QM and MM levels of theory, respectively (an analogous derivation can be done for $E_{MM}^{MM\rightarrow QM}$).

Furthermore, the metric defined in equation (6.20) is also independent of the kinetic energy whenever comparing total energy distributions of NVT ensembles

at the same temperature. This is so because, by splitting the total energy (E) into its kinetic energy (K) and potential energy (U) components, we obtain

$$\begin{aligned}
\Sigma_{MM,QM} &= 2\int_{-\infty}^{+\infty} dE_{QM}\, \rho_{QM}^{QM}(E_{QM}) \int_{-\infty}^{E_{QM}} dE'_{QM}\, \rho_{MM}^{QM}(E'_{QM}) \\
&= 2\int_{-\infty}^{+\infty} dU_{QM} dK_{QM}\, \rho_{QM}^{QM}(U_{QM})\rho_{QM}^{QM}(K_{QM}) \times \cdots \\
&\cdots \times \int_{-\infty}^{U_{QM}+K_{MM}} dU'_{QM} dK'_{MM}\, \rho_{MM}^{QM}(U'_{QM})\rho_{MM}^{QM}(K'_{MM}) \\
&= 2\int_{-\infty}^{+\infty} dK_{QM}\, \rho_{QM}^{QM}(K_{QM}) \int_{-\infty}^{K_{MM}} dK'_{MM}\, \rho_{MM}^{QM}(K'_{MM}) \times \cdots \\
&\cdots \times \int_{-\infty}^{+\infty} dU_{QM}\, \rho_{QM}^{QM}(U_{QM}) \int_{-\infty}^{U_{QM}} dU'_{QM}\, \rho_{MM}^{QM}(U'_{QM})
\end{aligned} \tag{B.2}$$

in which $\int_{-\infty}^{+\infty} dK_{QM}\, \rho_{QM}^{QM}(K_{QM}) \int_{-\infty}^{K_{MM}} dK'_{MM}\, \rho_{MM}^{QM}(K'_{MM}) = 1$ because two kinetic energy distributions of the same system at the same temperature totally overlap (they are independent of the level of theory used for the potential energy, depending only on the temperature if the same system is regarded). Therefore, we have that

$$\Sigma_{MM,QM} = 2\int_{-\infty}^{+\infty} dU_{QM}\, \rho_{QM}^{QM}(U_{QM}) \int_{-\infty}^{U_{QM}} dU'_{QM}\, \rho_{MM}^{QM}(U'_{QM}) \tag{B.3}$$

which demonstrates the independence of this metric on the kinetic energy.

B.2 Self-parameterising nMC-MC

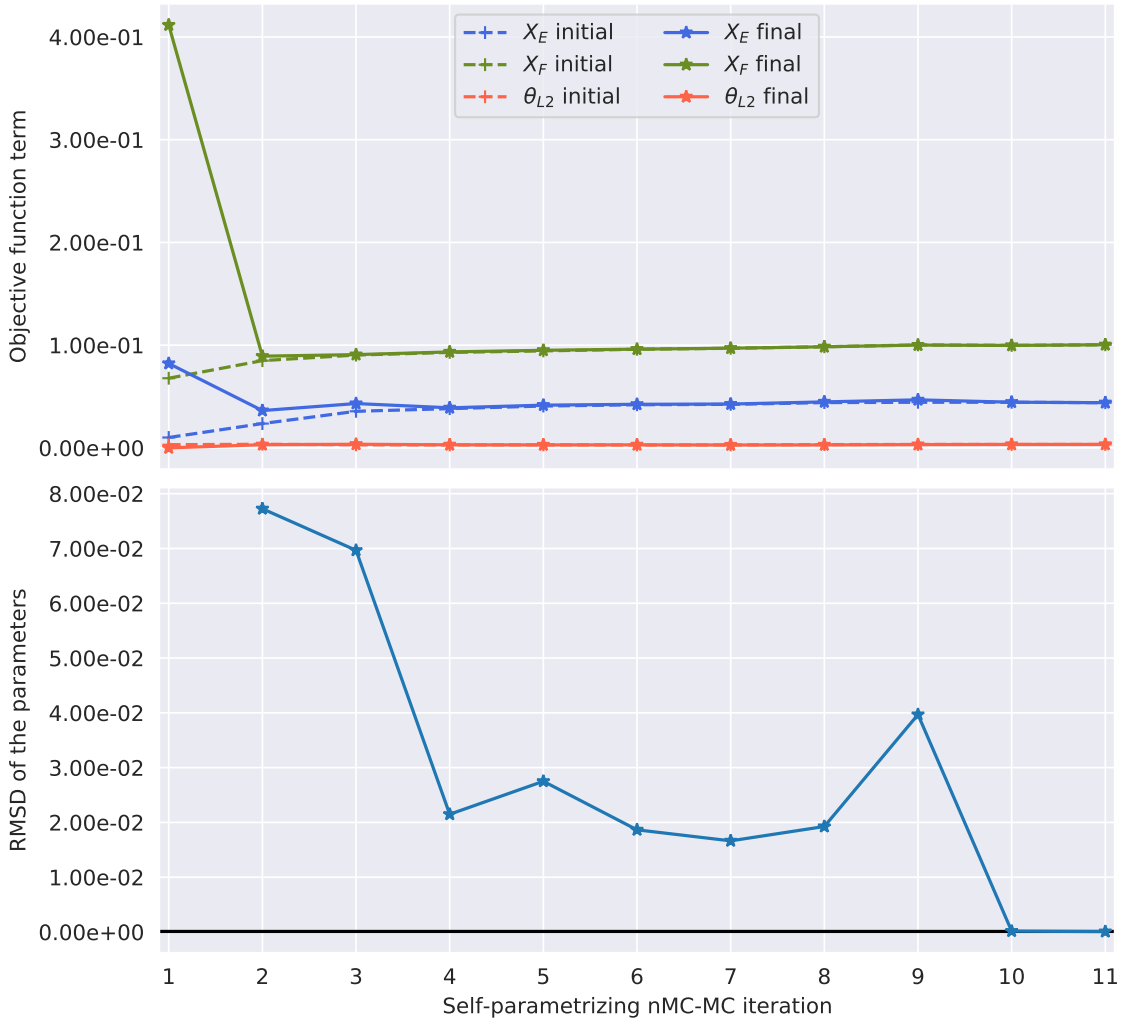

Figure B.1: Convergence of the self-parameterising nMC-MC calculation for octahydrotetracene. Top panel: Plot of the values of each term included in the objective function at the beginning (dashed lines) and end (solid lines) of each iteration. X_E corresponds to the potential energy term, X_F to the forces term, and θ_{L2} to the regularisation term. Bottom panel: Plot of the RMSD of the parameters as a function of the iteration number.

B.3 Acceptance rates

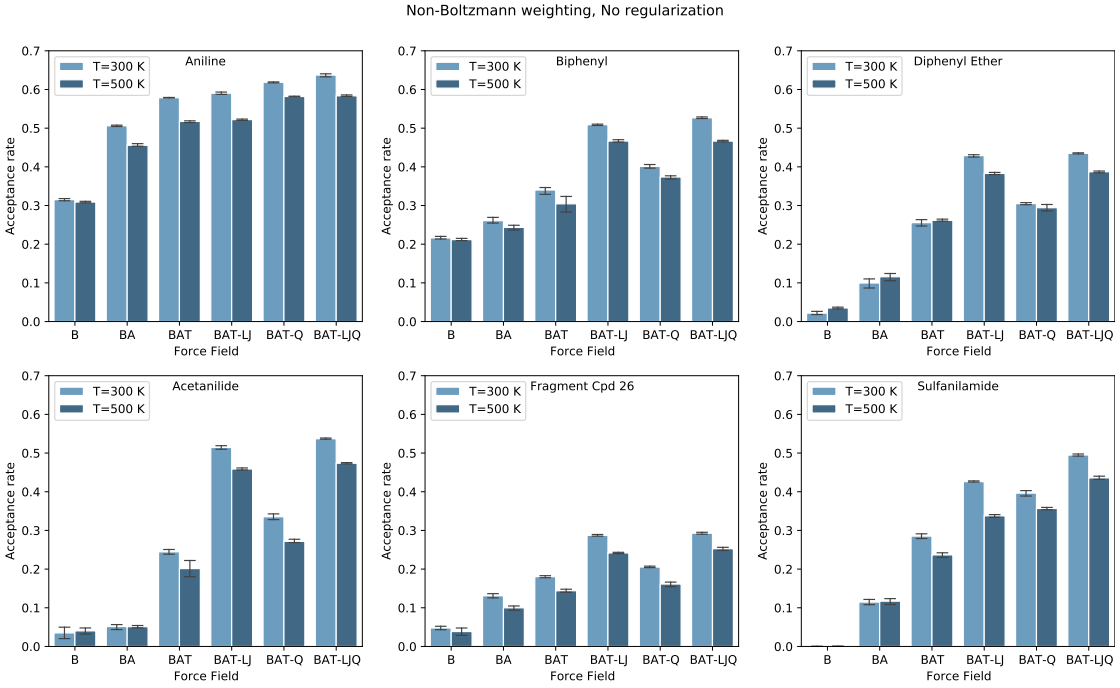

Figure B.2: Comparison between the nMC-MC acceptance rates obtained for FFs reparameterised using data sets containing structure sampled at either 300 K or 500 K. The FFs used to calculate the acceptance rates were derived employing non-Boltzmann weighting without any regularisation. The error bars correspond to the standard deviation of the results of 4 different nMC-MC samplers. Each sampler performed a total of 2×10^5 nMC-MC sweeps.

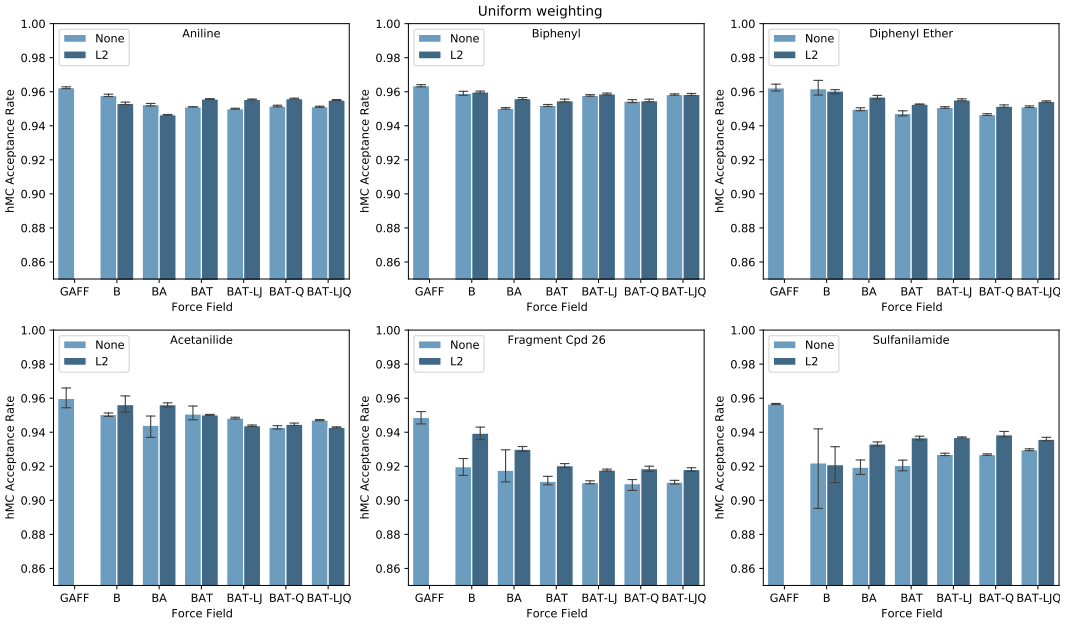

Figure B.3: hMC acceptance rates for the set of molecules used in Chapter 6. The FFs were derived employing uniform weighting with (dark blue) or without (light blue) L2 regularisation. The training data set contained configurations sampled at 500 K. The errors bars correspond to the standard deviation of the results of 4 different nMC-MC samplers. Each sampler performed a total of 2×10^5 nMC-MC sweeps.

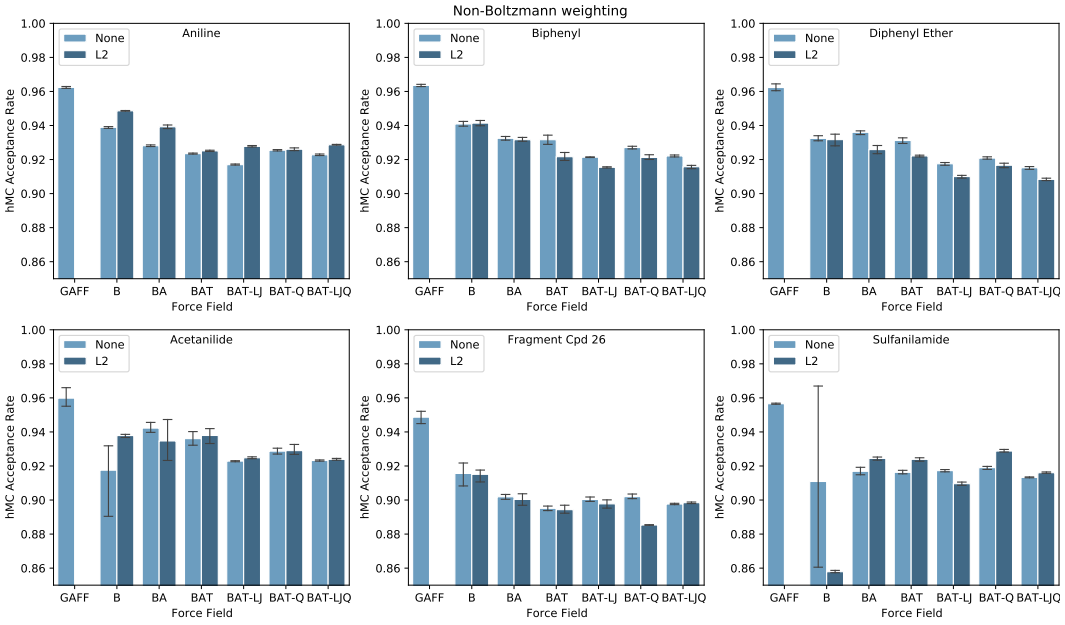

Figure B.4: hMC acceptance rates for the set of molecules used in Chapter 6. The FFs were derived employing non-Boltzmann weighting with (dark blue) or without (light blue) L2 regularisation. The training data set contained configurations sampled at 500 K. The errors bars correspond to the standard deviation of the results of 4 different nMC-MC samplers. Each sampler performed a total of 2×10^5 nMC-MC sweeps.

B.4 Configurational distributions

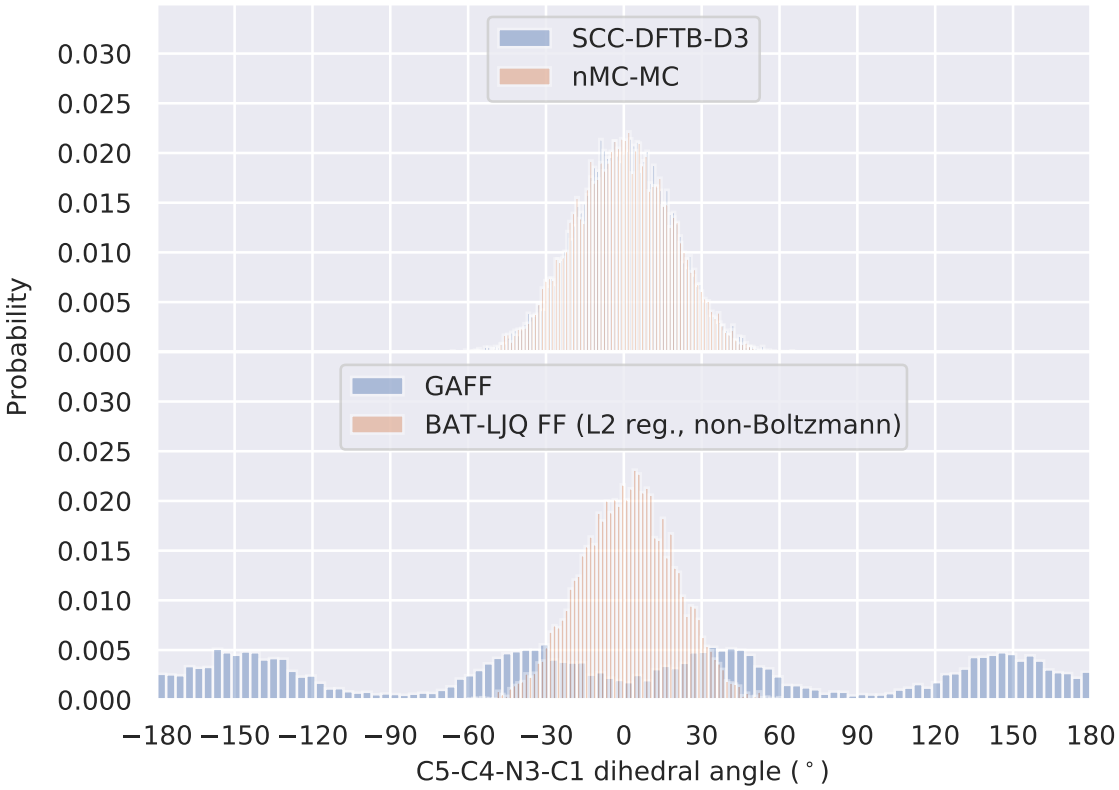

Figure B.5: Top panel: Distribution of the C5-C4-N3-C1 dihedral of acetanilide as obtained in SCC-DFTB-D3 MD and nMC-MC simulations. Lower panel: Distribution of the C5-C4-N3-C1 dihedral of acetanilide as obtained in MD simulations using the original GAFF and the non-Boltzmann-weighted L2-regularised BAT-LJQ FF. The SCC-DFTB-D3, GAFF, and BAT-LJQ MD were simulated during 10 ns (snapshots collected every 1 ps), and the nMC-MC sampler performed a total of 2×10^6 MC sweeps. The temperature of the simulations was 300 K.

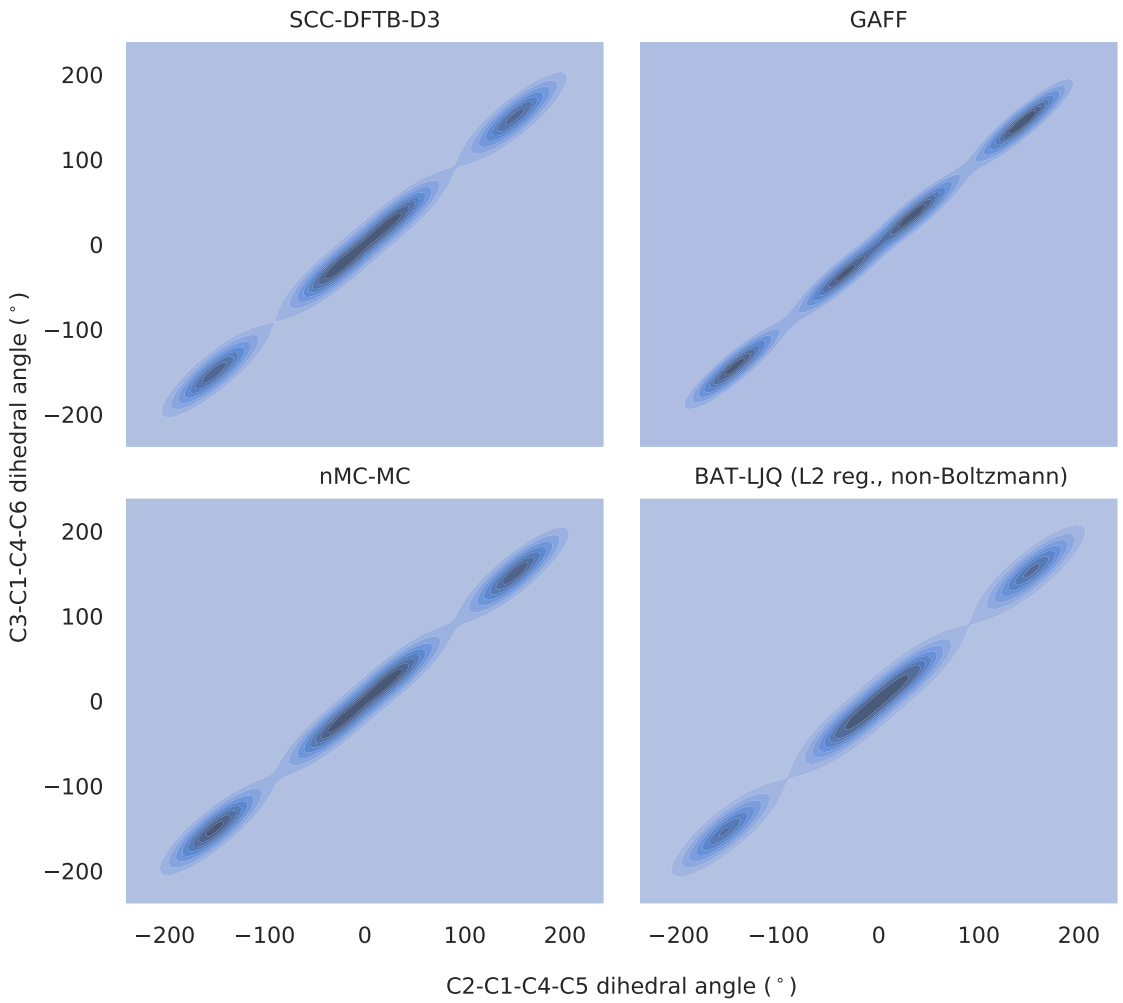

Figure B.6: Configurational distributions of the C2-C1-C4-C5 *vs.* C3-C1-C4-C6 dihedrals for biphenyl. The SCC-DFTB-D3 MD was simulated during 10 ns (snapshots collected every 1 ps), and the GAFF and BAT-LJQ MD were simulated during 100 ns (snapshots collected every 10 ps). The nMC-MC sampler performed a total of 4×10^6 MC sweeps. The temperature of the simulations was 300 K.

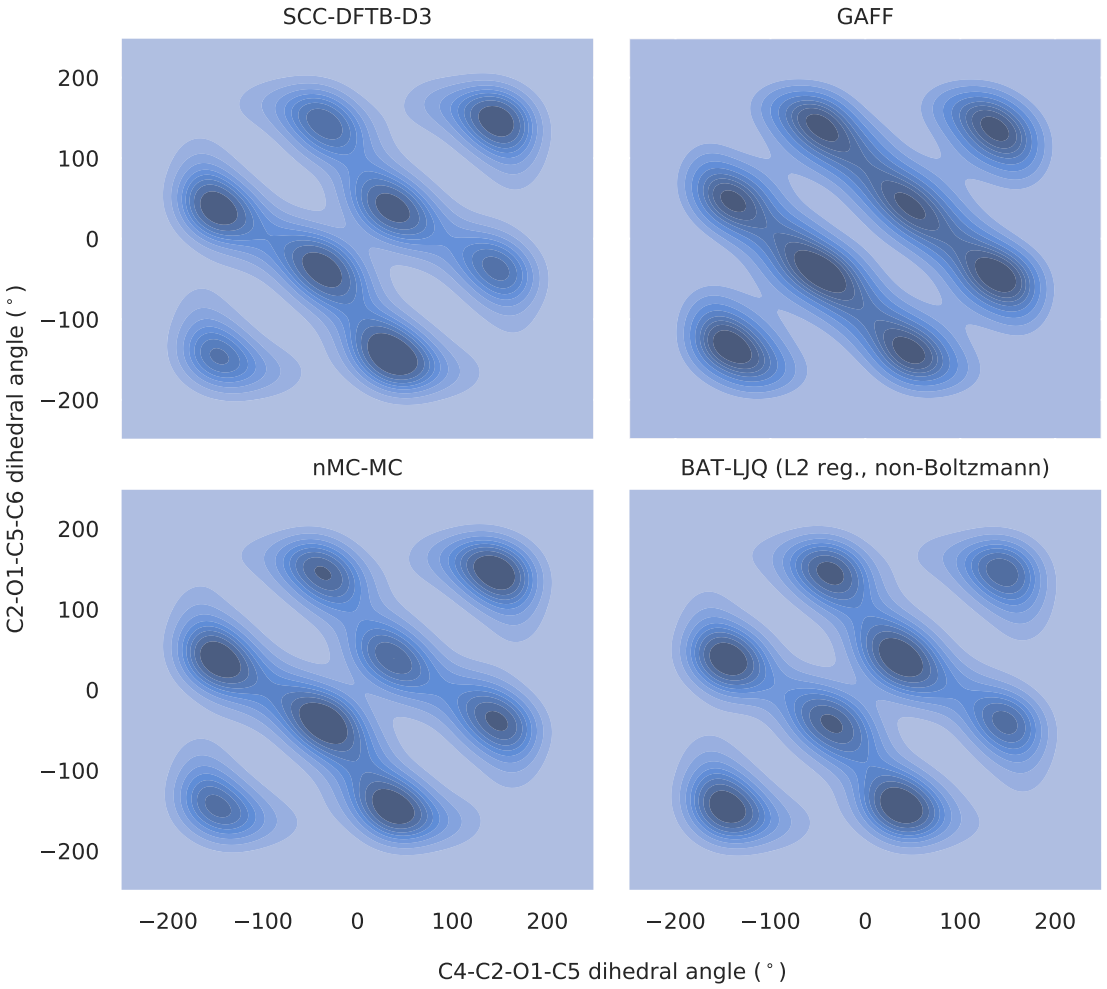

Figure B.7: Configurational distributions of the C4-C2-O1-C5 *vs.* C2-O1-C5-C6 dihedrals for diphenyl ether. The SCC-DFTB-D3 MD was simulated during 10 ns (snapshots collected every 1 ps), and the GAFF and BAT-LJQ MD were simulated during 1 μs (snapshots collected every 100 ps). The nMC-MC sampler performed a total of 2×10^6 MC sweeps. The temperature of the simulations was 500 K. The distributions at 300 K are not show as they were very far from convergence.

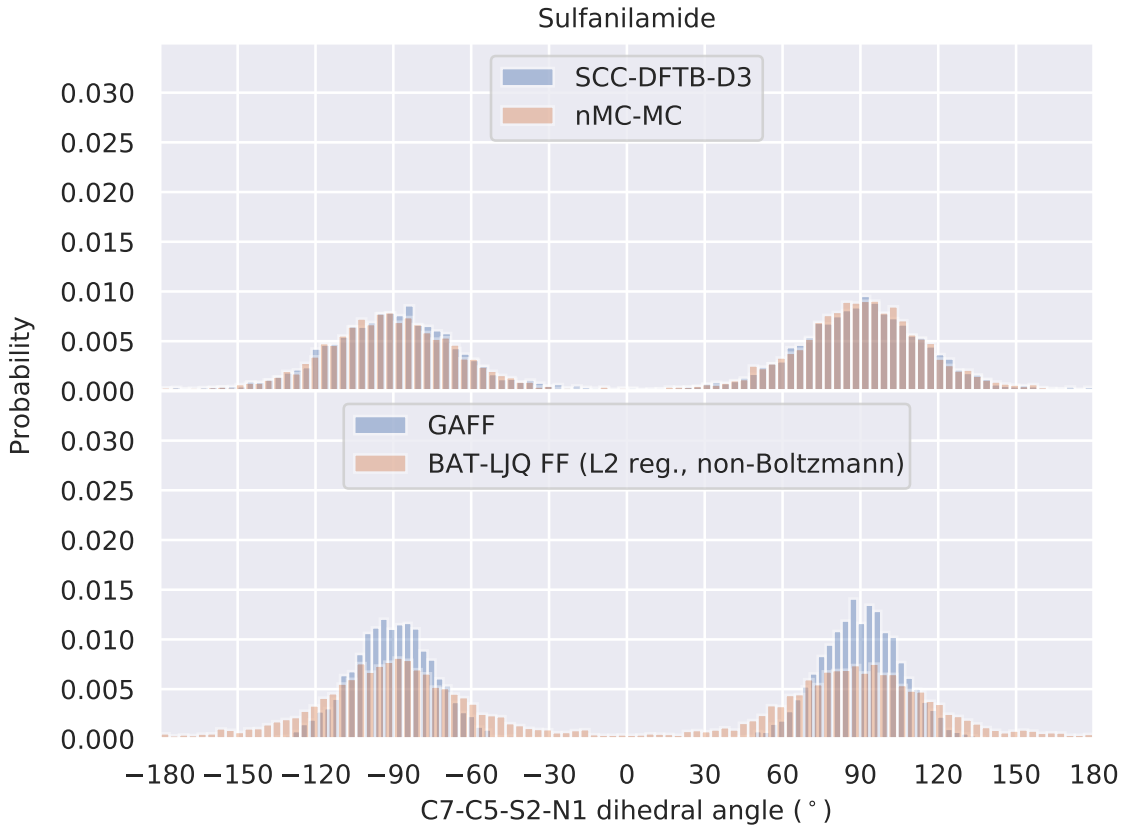

Figure B.8: Top panel: Distribution of the C7-C5-S2-N1 dihedral of sulfanilamide as obtained in SCC-DFTB-D3 MD and nMC-MC simulations. Lower panel: Distribution of the C7-C5-S2-N1 dihedral of sulfanilamide as obtained in MD simulations using the original GAFF and the non-Boltzmann-weighted L2-regularised BAT-LJQ FF. The SCC-DFTB-D3 and BAT-LJQ MD were simulated during 10 ns (snapshots collected every 1 ps), and the GAFF MD was simulated during 1 μs (snapshots collected every 100 ps). The nMC-MC sampler performed a total of 2923640 MC sweeps. The temperature of the simulations was 300 K.

Appendix C

Appendix of Chapter 7

C.1 Supporting Figures

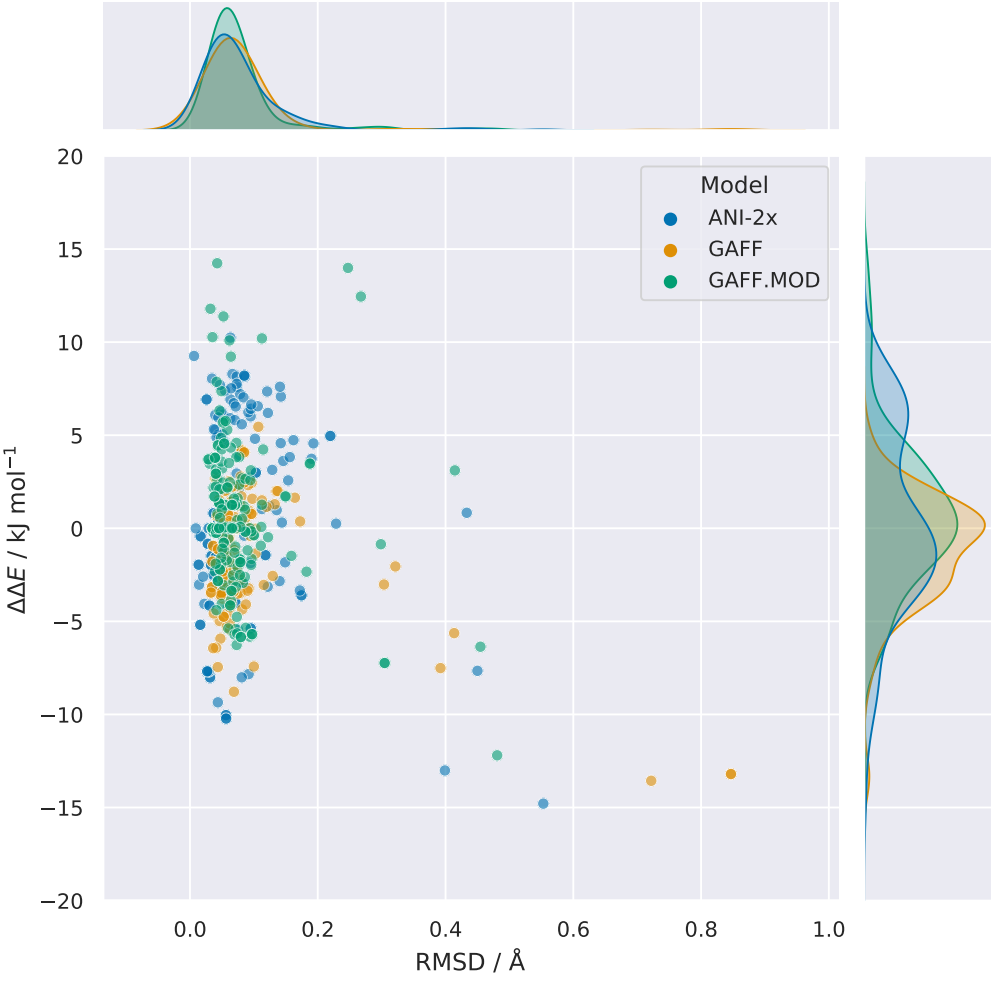

Figure C.1: Scatter plots of the relative conformer energies ($\Delta\Delta E$) versus the RMSD of atomic positions. Each point was obtained by performing a geometry optimisation using GAFF, GAFF.MOD, or ANI-2x, starting from all QM minima within 12.552 kJ mol^{-1} (3 kcal mol^{-1}) from the global minimum. The QM reference is the MP2/6-311++G(2d,p) level of theory.

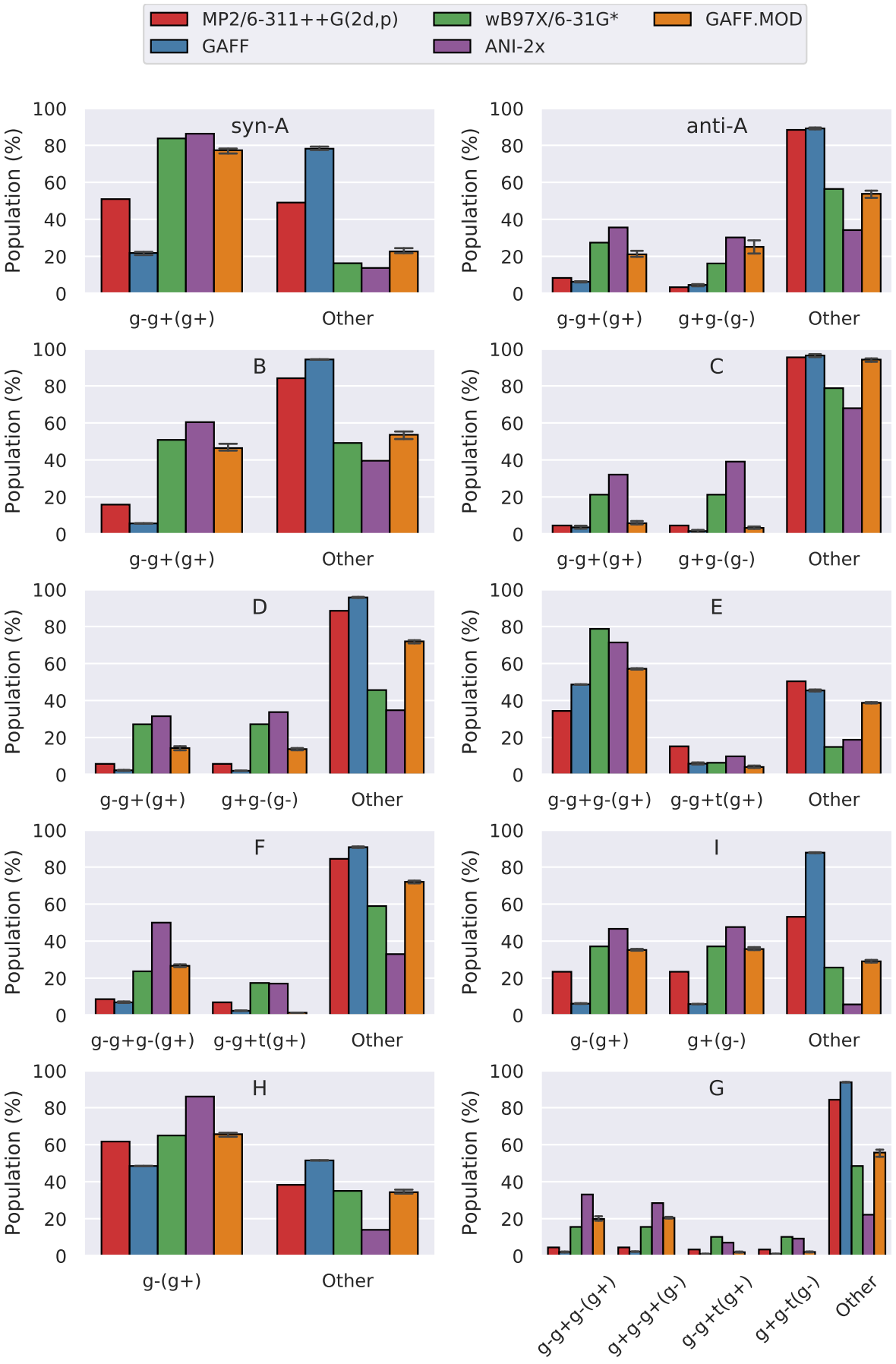

Figure C.2: Populations in the gas phase of the conformers with IMHBs.

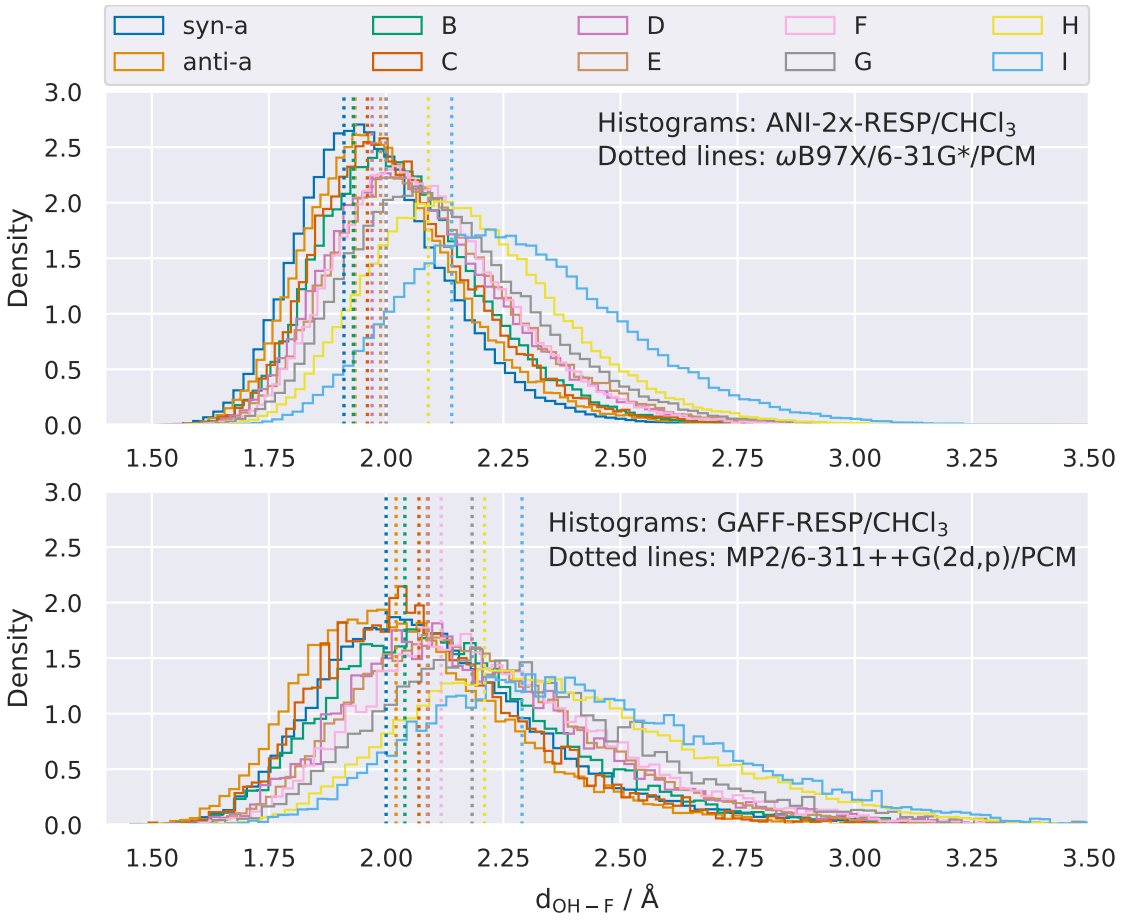

Figure C.3: Top panel: Distributions of the hydrogen bond (HB) lengths as obtained from the ANI-2x-RESP/$CHCl_3$ MD simulations (solid lines), and HB lengths of the geometries optimised at ωB97X/6-31G*/PCM (dashed lines). Bottom panel: Distributions of the hydrogen bond (HB) lengths as obtained from the GAFF-RESP/$CHCl_3$ MD simulations (solid lines), and HB lengths of the geometries optimised at MP2/6-311++G(2d,p)/PCM (dashed lines). Only conformers with IMHBs are represented.

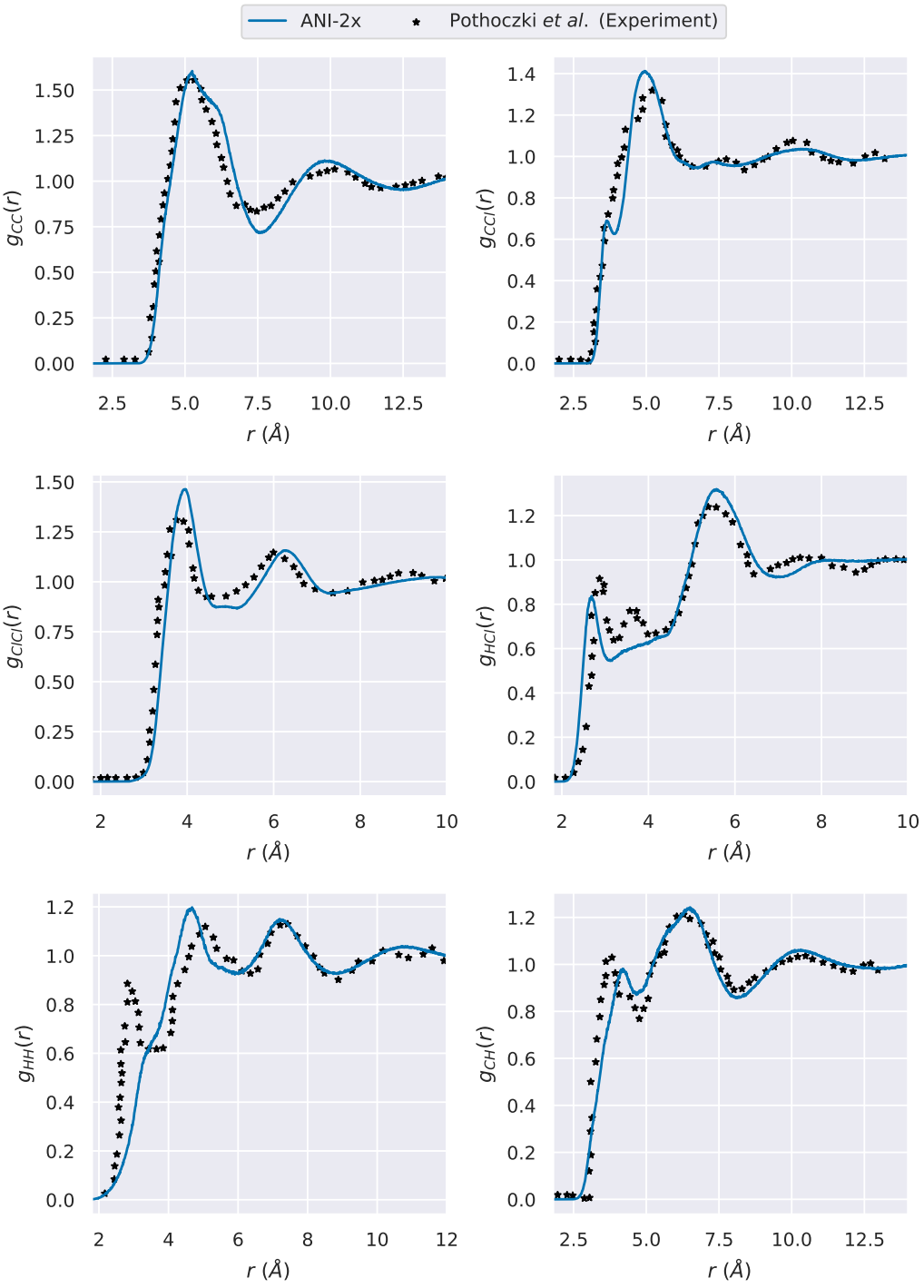

Figure C.4: Experimental and ANI-2x radial distribution functions (RDFs) of bulk chloroform. The experimental data is reproduced from Refs. 5 and 6.

References

[1] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai and G. Seifert, *Phys. Rev. B*, 1998, **58**, 7260–7268.

[2] C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2016, **72**, 171–179.

[3] J. B. Dressman, A. Nair, B. Abrahamsson, D. M. Barends, D. Groot, S. Kopp, P. Langguth, J. E. Polli, V. P. Shah and M. Zimmer, *J. Pharm. Sci.*, 2012, **101**, 2653–2667.

[4] J. Morado, P. N. Mortenson, M. L. Verdonk, R. A. Ward, J. W. Essex and C.-K. Skylaris, *J. Chem. Inf. Model.*, 2021, **61**, 2026–2047.

[5] S. Pothoczki, L. Temleitner, S. Kohara, P. Jóvári and L. Pusztai, *J. Phys.: Condens. Matter*, 2010, **22**, 404211.

[6] C.-C. Yin, A. H.-T. Li and S. D. Chao, *J. Chem. Phys.*, 2013, **139**, 194501.

[7] D. J. Huggins, P. C. Biggin, M. A. Dämgen, J. W. Essex, S. A. Harris, R. H. Henchman, S. Khalid, A. Kuzmanic, C. A. Laughton, J. Michel, A. J. Mulholland, E. Rosta, M. S. P. Sansom and M. W. van der Kamp, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2019, **9**, e1393.

[8] D. F. Hahn, C. I. Bayly, H. E. B. Macdonald, J. D. Chodera, V. Gapsys, A. S. J. S. Mey, D. L. Mobley, L. P. Benito, C. E. M. Schindler, G. Tresadern and G. L. Warren, *arXiv.org, e-Print Arch., Quant. Biol.*, 2021.

[9] S. Kashefolgheta, M. P. Oliveira, S. R. Rieder, B. A. C. Horta, W. E. Acree and P. H. Hünenberger, *J. Chem. Theory Comput.*, 2020, **16**, 7556–7580.

[10] E. B. Miller, R. B. Murphy, D. Sindhikara, K. W. Borrelli, M. J. Grisewood, F. Ranalli, S. L. Dixon, S. Jerome, N. A. Boyles, T. Day, P. Ghanakota, S. Mondal, S. B. Rafi, D. M. Troast, R. Abel and R. A. Friesner, *J. Chem. Theory Comput.*, 2021, **17**, 2630–2639.

[11] S.-Y. Huang and X. Zou, *Int. J. Mol. Sci.*, 2010, **11**, 3016–3034.

[12] N. Foloppe and I.-J. Chen, *Curr. Med. Chem.*, 2009, **16**, 3381–3413.

[13] N. Foloppe and I.-J. Chen, *Bioorg. Med. Chem.*, 2016, **24**, 2159–2189.

[14] R. J. Hall, P. N. Mortenson and C. W. Murray, *Prog. Biophys. Mol. Biol.*, 2014, **116**, 82–91.

[15] E. Tamanini, I. M. Buck, G. Chessari, E. Chiarparin, J. E. H. Day, M. Frederickson, C. M. Griffiths-Jones, K. Hearn, T. D. Heightman, A. Iqbal, C. N. Johnson, E. J. Lewis, V. Martins, T. Peakman, M. Reader, S. J. Rich, G. A. Ward, P. A. Williams and N. E. Wilsher, *J. Med. Chem.*, 2017, **60**, 4611–4625.

[16] A. Bruno, G. Costantino, L. Sartori and M. Radi, *Curr. Med. Chem.*, 2019, **26**, 3838–3873.

[17] K. Williams, E. Bilsland, A. Sparkes, W. Aubrey, M. Young, L. N. Soldatova, K. De Grave, J. Ramon, M. de Clare, W. Sirawaraporn and et al., *J. R. Soc., Interface*, 2015, **12**, 20141289.

[18] E. Ferrero, S. Brachat, J. L. Jenkins, P. Marc, P. Skewes-Cox, R. C. Altshuler, C. Gubser Keller, A. Kauffmann, E. K. Sassaman, J. M. Laramie and et al., *PLoS Comput. Biol.*, 2020, **16**, e1008126.

[19] J. Tong and S. Zhao, *J. Chem. Inf. Model.*, 2021, **61**, 1180–1192.

[20] E. Chiarparin, M. J. Packer and D. M. Wilson, *Future Med. Chem.*, 2019, **11**, 79–82.

[21] I.-J. Chen and N. Foloppe, *Bioorg. Med. Chem.*, 2013, **21**, 7898–7920.

[22] C. D. Blundell, T. Nowak and M. J. Watson, *Prog. Med. Chem.*, 2016, 45–147.

[23] D. Cui, B. W. Zhang, N. Matubayasi and R. M. Levy, *J. Chem. Theory Comput.*, 2018, **14**, 512–526.

[24] Z. Li and T. Lazaridis, *Phys. Chem. Chem. Phys.*, 2007, **9**, 573–581.

[25] T. Young, R. Abel, B. Kim, B. J. Berne and R. A. Friesner, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 808–813.

[26] Y. Murakami, S. Omori and K. Kinoshita, *J. Struct. Funct. Genomics*, 2016, **17**, 101–110.

[27] H. M. Berman, *Nucleic Acids Res.*, 2000, **28**, 235–242.

[28] A. Y. S. Balazs, R. J. Carbajo, N. L. Davies, Y. Dong, A. W. Hird, J. W. Johannes, M. L. Lamb, W. McCoull, P. Raubo, G. R. Robb, M. J. Packer and E. Chiarparin, *J. Med. Chem.*, 2019, **62**, 9418–9437.

[29] A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 12562–12566.

[30] A. Barducci, G. Bussi and M. Parrinello, *Phys. Rev. Lett.*, 2008, **100**, 020603.

[31] L. Wang, R. A. Friesner and B. J. Berne, *J. Phys. Chem. B*, 2011, **115**, 9431–9438.

[32] R. H. Swendsen and J.-S. Wang, *Phys. Rev. Lett.*, 1986, **57**, 2607–2609.

[33] K. Hukushima and K. Nemoto, *J. Phys. Soc. Jpn.*, 1996, **65**, 1604–1608.

[34] G. Torrie and J. Valleau, *J. Comput. Phys.*, 1977, **23**, 187–199.

[35] P. Atkins and R. Friedman, *Molecular Quantum Mechanics*, Oxford University Press, 2011.

[36] G. Schatz and M. Ratner, *Quantum Mechanics in Chemistry*, Dover Publications, 2002.

[37] M. Born and R. Oppenheimer, *Ann. Phys. (Berlin, Ger.)*, 1927, **389**, 457–484.

[38] D. R. Hartree, *Math. Proc. Cambridge Philos. Soc.*, 1928, **24**, 89–110.

[39] V. Fock, *Z. Phys.*, 1930, **61**, 126–148.

[40] J. C. Slater, *Phys. Rev.*, 1930, 210–211.

[41] J. C. Slater, *Phys. Rev.*, 1951, **81**, 385–390.

[42] A. Szabo and N. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, Dover Publications, 1996.

[43] C. Møller and M. S. Plesset, *Phys. Rev.*, 1934, **46**, 618–622.

[44] B. Roos, *Chem. Phys. Lett.*, 1972, **15**, 153–159.

[45] J. Čížek, *J. Chem. Phys.*, 1966, **45**, 4256–4266.

[46] P. E. M. Siegbahn, J. Almlöf, A. Heiberg and B. O. Roos, *J. Chem. Phys.*, 1981, **74**, 2384–2396.

[47] B. O. Roos, P. Linse, P. E. Siegbahn and M. R. Blomberg, *Chem. Phys.*, 1982, **66**, 197–207.

[48] J. Olsen, *Int. J. Quantum Chem.*, 2011, **111**, 3267–3272.

[49] P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, **136**, B864–B871.

[50] W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, **140**, A1133–A1138.

[51] J. C. A. Prentice, J. Aarons, J. C. Womack, A. E. A. Allen, L. Andrinopoulos, L. Anton, R. A. Bell, A. Bhandari, G. A. Bramley, R. J. Charlton and et al., *J. Chem. Phys.*, 2020, **152**, 174111.

[52] L. H. Thomas, *Math. Proc. Cambridge Philos. Soc.*, 1927, **23**, 542–548.

[53] E. Fermi, *Rend. Accad. Naz. Lincei*, 1927, **6**, 32.

[54] P. A. M. Dirac, *Math. Proc. Cambridge Philos. Soc.*, 1930, **26**, 376–385.

[55] S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200–1211.

[56] J. P. Perdew and Y. Wang, *Phys. Rev. B*, 1992, **45**, 13244–13249.

[57] D. M. Ceperley and B. J. Alder, *Phys. Rev. Lett.*, 1980, **45**, 566–569.

[58] J. P. Perdew and A. Zunger, *Phys. Rev. B*, 1981, **23**, 5048–5079.

[59] J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

[60] J. Perdew, *Electronic Structure Theory of Solids*, Akademie Verlag, 1991, p. 11.

[61] A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.

[62] C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.

[63] J. Tao, J. P. Perdew, V. N. Staroverov and G. E. Scuseria, *Phys. Rev. Lett.*, 2003, **91**, 146401.

[64] Y. Zhao and D. G. Truhlar, *J. Chem. Phys.*, 2006, **125**, 194101.

[65] A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.

[66] C. Adamo and V. Barone, *Chem. Phys. Lett.*, 1997, **274**, 242–250.

[67] C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.

[68] R. Sabatini, T. Gorni and S. de Gironcoli, *Phys. Rev. B*, 2013, **87**, 041108.

[69] O. A. Vydrov and T. Van Voorhis, *J. Chem. Phys.*, 2010, **133**, 244103.

[70] J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615.

[71] S. Grimme and F. Neese, *J. Chem. Phys.*, 2007, **127**, 154116.

[72] E. Brémond, I. Ciofini, J. C. Sancho-García and C. Adamo, *Acc. Chem. Res.*, 2016, **49**, 1503–1513.

[73] S. Grimme, *J. Chem. Phys.*, 2006, **124**, 034108.

[74] E. Brémond and C. Adamo, *J. Chem. Phys.*, 2011, **135**, 024106.

[75] J. P. Perdew, AIP Conference Proceedings, 2001, p. 1–20.

[76] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert and R. Kaschner, *Phys. Rev. B*, 1995, **51**, 12947–12957.

[77] G. Seifert, D. Porezag and T. Frauenheim, *Int. J. Quantum Chem.*, 1996, **58**, 185–192.

[78] M. Gaus, Q. Cui and M. Elstner, *J. Chem. Theory Comput.*, 2011, **7**, 931–948.

[79] M. Elstner, P. Hobza, T. Frauenheim, S. Suhai and E. Kaxiras, *J. Chem. Phys.*, 2001, **114**, 5149–5155.

[80] M. Elstner, K. J. Jalkanen, M. Knapp-Mohammady, T. Frauenheim and S. Suhai, *Chem. Phys.*, 2001, **263**, 203–219.

[81] M. Elstner, *Theor. Chem. Acc.*, 2006, **116**, 316–325.

[82] W. Hujo and S. Grimme, *J. Chem. Theory Comput.*, 2011, **7**, 3866–3871.

[83] S. Lehtola, *Int. J. Quantum Chem.*, 2019, **119**, e25968.

[84] J. C. Slater, *Phys. Rev.*, 1930, **36**, 57–64.

[85] E. R. Davidson and D. Feller, *Chem. Rev*, 1986, **86**, 681–696.

[86] S. F. Boys, *Proc. R. Soc. London, Ser. A*, 1950, **200**, 542–554.

[87] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, 2017, vol. 1.

[88] Y. I. Yang, Q. Shao, J. Zhang, L. Yang and Y. Q. Gao, *J. Chem. Phys.*, 2019, **151**, 070902.

[89] G. J. Martyna, M. L. Klein and M. Tuckerman, *J. Chem. Phys.*, 1992, **97**, 2635–2643.

[90] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Academic Press, 2nd edn., 2002.

[91] K. Huang, *Introduction to Statistical Physics*, CRC Press, 2nd edn., 2010.

[92] M. E. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation*, Oxford University Press, 2010.

[93] R. K. Pathria and P. D. Beale, *Statistical Mechanics*, Elsevier/Academic Press, 3rd edn., 2011.

[94] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids: With Applications to Soft Matter*, Academic Press, 2013.

[95] R. L. Harrison, *AIP Conf. Proc.*, 2010, **1204**, 17–21.

[96] J. H. Halton, *SIAM Rev.*, 1970, **12**, 1–63.

[97] N. Metropolis, *Los Alamos Sci.*, 1987, **15**, 125–130.

[98] D. N. Theodorou, *Ind. Eng. Chem. Res.*, 2010, **49**, 3047–3058.

[99] D. Meimaroglou and C. Kiparissides, *Ind. Eng. Chem. Res.*, 2014, **53**, 8963–8979.

[100] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *J. Chem. Phys.*, 1953, **21**, 1087–1092.

[101] F. A. Escobedo and J. J. de Pablo, *J. Chem. Phys*, 1996, **105**, 4391–4394.

[102] J. S. Liu and R. Chen, *J. Am. Stat. Assoc.*, 1998, **93**, 1032–1044.

[103] Y. Iba, *Int. J. Mod. Phys. C*, 2001, **12**, 623–656.

[104] L. D. Gelb, *J. Chem. Phys.*, 2003, **118**, 7747–7750.

[105] H. M. Cuppen, L. J. Karssemeijer and T. Lamberts, *Chem. Rev. (Washington, DC, U. S.)*, 2013, **113**, 8840–8871.

[106] J. P. Nilmeier, G. E. Crooks, D. D. L. Minh and J. D. Chodera, *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, E1009–E1018.

[107] D. N. Theodorou, *Ind. Eng. Chem. Res.*, 2010, **49**, 3047–3058.

[108] K. I. Tenekedjiev, N. D. Nikolova and K. Kolev, *Applications of Monte Carlo Simulation in Modelling of Biochemical Processes*, InTech, 2011.

[109] M. J. Yu, *J. Cheminf.*, 2012, **4**, 32.

[110] D. Meimaroglou and C. Kiparissides, *Ind. Eng. Chem. Res.*, 2014, **53**, 8963–8979.

[111] B. Kelly and W. R. Smith, *Mol. Phys.*, 2019, **117**, 2778–2785.

[112] D. J. C. Mackay, in *Introduction to Monte Carlo Methods*, ed. M. I. Jordan, Springer Netherlands, 1998, p. 175–204.

[113] C. Lemieux, *Monte Carlo and Quasi-Monte Carlo Sampling*, Springer New York, 1st edn., 2009.

[114] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics*, Cambridge University Press, 4th edn., 2014.

[115] K. Binder and D. W. Heermann, *Monte Carlo Simulation in Statistical Physics: An Introduction*, Springer International Publishing, 6th edn., 2019.

[116] W. K. Hastings, *Biometrika*, 1970, 14.

[117] A. A. Barker, *Aust. J. Phys.*, 1965, **18**, 119–134.

[118] B. J. Alder and T. E. Wainwright, *J. Chem. Phys.*, 1957, **27**, 1208–1209.

[119] F. H. Stillinger and A. Rahman, *J. Chem. Phys.*, 1974, **60**, 1545–1557.

[120] J. A. McCammon, B. R. Gelin and M. Karplus, *Nature*, 1977, **267**, 585–590.

[121] A. Jász, A. Rák, I. Ladjánszki and G. Cserey, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2020, **10**, e1444.

[122] R. W. Hockney, *Methods Comput. Phys.*, 1970, **9**, 135–211.

[123] L. Verlet, *Phys. Rev.*, 1967, **159**, 98–103.

[124] P. Schofield, *Comput. Phys. Commun.*, 1973, **5**, 17–23.

[125] D. Beeman, *J. Comput. Phys.*, 1976, **20**, 130–139.

[126] W. C. Swope, H. C. Andersen, P. H. Berens and K. R. Wilson, *J. Chem. Phys.*, 1982, **76**, 637–649.

[127] L. Woodcock, *Chem. Phys. Lett.*, 1971, **10**, 257–261.

[128] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

[129] E. Braun, S. M. Moosavi and B. Smit, *J. Chem. Theory Comput.*, 2018, **14**, 5262–5272.

[130] G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.

[131] S. Nosé, *J. Chem. Phys.*, 1984, **81**, 511–519.

[132] W. G. Hoover, *Phys. Rev. A*, 1985, **31**, 1695–1697.

[133] H. C. Andersen, *J. Chem. Phys.*, 1980, **72**, 2384–2393.

[134] T. Schneider and E. Stoll, *Phys. Rev. B*, 1978, **17**, 1302–1322.

[135] J. Åqvist, P. Wennerström, M. Nervall, S. Bjelic and B. O. Brandsdal, *Chem. Phys. Lett.*, 2004, **384**, 288–294.

[136] K.-H. Chow and D. M. Ferguson, *Comput. Phys. Commun.*, 1995, **91**, 283–289.

[137] M. Bernetti and G. Bussi, *J. Chem. Phys.*, 2020, **153**, 114107.

[138] M. Parrinello and A. Rahman, *Phys. Rev. Lett.*, 1980, **45**, 1196–1199.

[139] M. Parrinello and A. Rahman, *J. Appl. Phys. (Melville, NY, U. S.)*, 1981, **52**, 7182–7190.

[140] M. Parrinello and A. Rahman, *J. Chem. Phys.*, 1982, **76**, 2662–2666.

[141] W. G. Hoover, *Phys. Rev. A*, 1986, **34**, 2499–2500.

[142] G. J. Martyna, D. J. Tobias and M. L. Klein, *J. Chem. Phys.*, 1994, **101**, 4177–4189.

[143] G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, 014101.

[144] A. Kavalur, V. Guduguntla and W. K. Kim, *Mol. Simul.*, 2020, **46**, 911–922.

[145] R. Erban, *Proc. R. Soc. A*, 2014, **470**, 20140036.

[146] B. Leimkuhler and C. Matthews, *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*, Springer, 2015.

[147] S. Vandenhaute, S. M. J. Rogge and V. Van Speybroeck, *Front. Chem. (Lausanne, Switz.)*, 2021, **9**, 718920.

[148] J. Jung, W. Nishima, M. Daniels, G. Bascom, C. Kobayashi, A. Adedoyin, M. Wall, A. Lappala, D. Phillips, W. Fischer and et al., *J. Comput. Chem.*, 2019, **40**, 1919–1930.

[149] J. D. Durrant, S. E. Kochanek, L. Casalino, P. U. Ieong, A. C. Dommer and R. E. Amaro, *ACS Cent. Sci.*, 2020, **6**, 189–196.

[150] O. M. H. Salo-Ahen, I. Alanko, R. Bhadane, A. M. J. J. Bonvin, R. V. Honorato, S. Hossain, A. H. Juffer, A. Kabedev, M. Lahtela-Kakkonen, A. S. Larsen and et al., *Processes*, 2020, **9**, 71.

[151] R. Schneider, A. R. Sharma and A. Rai, in *Introduction to Molecular Dynamics*, ed. H. Fehske, R. Schneider and A. Weiße, Springer Berlin Heidelberg, 2008, vol. 739, p. 3–40.

[152] R. Lazim, D. Suh and S. Choi, *Int. J. Mol. Sci.*, 2020, **21**, 6339.

[153] J. Debnath, M. Invernizzi and M. Parrinello, *J. Chem. Theory Comput.*, 2019, **15**, 2454–2459.

[154] O. Valsson, P. Tiwary and M. Parrinello, *Annu. Rev. Phys. Chem.*, 2016, **67**, 159–184.

[155] D. D. Frantz, D. L. Freeman and J. D. Doll, *J. Chem. Phys.*, 1990, **93**, 2769–2784.

[156] E. Darve and A. Pohorille, *J. Chem. Phys.*, 2001, **115**, 9169–9183.

[157] J. Lee, H. Scheraga and S. Rackovsky, *J. Comput. Chem.*, 1997, **18**, 1222–1232.

[158] I. Joung, J. Y. Kim, S. P. Gross, K. Joo and J. Lee, *Computer Physics Communications*, 2018, **223**, 28–33.

[159] Y. Sugita and Y. Okamoto, *Chem. Phys. Lett.*, 1999, **314**, 141–151.

[160] T. Okabe, M. Kawata, Y. Okamoto and M. Mikami, *Chem. Phys. Lett.*, 2001, **335**, 435–439.

[161] P. Liu, B. Kim, R. A. Friesner and B. J. Berne, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 13749–13754.

[162] L. Wang, R. A. Friesner and B. J. Berne, *J. Phys. Chem. B*, 2011, **115**, 9431–9438.

[163] R. C. Bernardi, M. C. Melo and K. Schulten, *Biochim. Biophys. Acta, Gen. Subj.*, 2015, **1850**, 872–877.

[164] A. D. Mackerell, *J. Comput. Chem.*, 2004, **25**, 1584–1604.

[165] E. G. Lewars, in *The Concept of the Potential Energy Surface*, Springer International Publishing, 2016, p. 9–49.

[166] A. Hinchliffe, *Molecular Modelling for Beginners*, Wiley, 2nd edn., 2008.

[167] S. Riniker, *J. Chem. Inf. Model.*, 2018, **58**, 565–578.

[168] Z. Liu, X. Wu and W. Wang, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1096.

[169] S.-W. Chiu, S. A. Pandit, H. L. Scott and E. Jakobsson, *J. Phys. Chem. B*, 2009, **113**, 2748–2763.

[170] X. Periole and S.-J. Marrink, in *The Martini Coarse-Grained Force Field*, ed. L. Monticelli and E. Salonen, Humana Press, 2013, vol. 924, p. 533–565.

[171] J. Barnoud and L. Monticelli, in *Coarse-Grained Force Fields for Molecular Simulations*, ed. A. Kukol, Springer New York, 2015, vol. 1215, p. 125–149.

[172] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid and A. Kolinski, *Chem. Rev. (Washington, DC, U. S.)*, 2016, **116**, 7898–7936.

[173] K. Vanommeslaeghe, O. Guvench and A. D. MacKerell, *Curr. Pharm. Des.*, 2014, **20**, 3281–3292.

[174] A. J. Rzepiela, M. Louhivuori, C. Peter and S. J. Marrink, *Phys. Chem. Chem. Phys.*, 2011, **13**, 10437.

[175] P. Kar and M. Feig, *J. Chem. Theory Comput.*, 2017, **13**, 5753–5765.

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

[177] D. Dubbeldam, K. S. Walton, T. J. H. Vlugt and S. Calero, *Adv. Theory Simul.*, 2019, **2**, 1900135.

[178] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard and W. M. Skiff, *J. Am. Chem. Soc.*, 1992, **114**, 10024–10035.

[179] S. L. Mayo, B. D. Olafson and W. A. Goddard, *J. Phys. Chem.*, 1990, **94**, 8897–8909.

[180] S. Shi, L. Yan, Y. Yang, J. Fisher-Shaulsky and T. Thacher, *J. Comput. Chem.*, 2003, **24**, 1059–1076.

[181] J. W. Ponder and D. A. Case, in *Force Fields for Protein Simulations*, Elsevier, 2003, vol. 66, p. 27–85.

[182] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.

[183] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. Mackerell, *J. Comput. Chem.*, 2009, 671–690.

[184] C. Oostenbrink, A. Villa, A. Mark and W. van Gunsteren, *J. Comput. Chem.*, 2004, **25**, 1656–1676.

[185] E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, J. W. Kaus, D. S. Cerutti, G. Krilov, W. L. Jorgensen, R. Abel and R. A. Friesner, *J. Chem. Theory Comput.*, 2016, **12**, 281–296.

[186] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives and W. L. Jorgensen, *J. Phys. Chem. B*, 2001, **105**, 6474–6487.

[187] W. L. Jorgensen, D. S. Maxwell and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1996, **118**, 11225–11236.

[188] W. L. Jorgensen and J. Tirado-Rives, *J. Am. Chem. Soc.*, 1988, **110**, 1657–1666.

[189] N. L. Allinger, Y. H. Yuh and J. H. Lii, *J. Am. Chem. Soc.*, 1989, **111**, 8551–8566.

[190] J. H. Lii and N. L. Allinger, *J. Am. Chem. Soc.*, 1989, **111**, 8566–8575.

[191] J. H. Lii and N. L. Allinger, *J. Am. Chem. Soc.*, 1989, **111**, 8576–8582.

[192] J. R. Maple, M.-J. Hwang, T. P. Stockfisch, U. Dinur, M. Waldman, C. S. Ewig and A. T. Hagler, *J. Comput. Chem.*, 1994, **15**, 162–182.

[193] S. Ósk Jónsdóttir and K. Rasmussen, *New J. Chem.*, 2000, **24**, 243–247.

[194] H. Sun, *J. Phys. Chem. B*, 1998, **102**, 7338–7364.

[195] H. Sun, P. Ren and J. Fried, *Comput. Theor. Polym. Sci.*, 1998, **8**, 229–246.

[196] T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519.

[197] T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 520–552.

[198] T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 553–586.

[199] T. A. Halgren and R. B. Nachbar, *J. Comput. Chem.*, 1996, **17**, 587–615.

[200] T. Halgren, *J. Comput. Chem.*, 1996, **16**, 616–641.

[201] P. Ren, C. Wu and J. W. Ponder, *J. Chem. Theory Comput.*, 2011, **7**, 3143–3161.

[202] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson and T. Head-Gordon, *J. Phys. Chem. B*, 2010, **114**, 2549–2564.

[203] S. Patel and C. L. Brooks, *J. Comput. Chem.*, 2004, **25**, 1–16.

[204] K. Vanommeslaeghe and A. MacKerell, *Biochim. Biophys. Acta, Gen. Subj.*, 2015, **1850**, 861–871.

[205] S. W. Rick, S. J. Stuart and B. J. Berne, *J. Chem. Phys.*, 1994, **101**, 6141–6156.

[206] S. W. Rick and B. J. Berne, *J. Am. Chem. Soc.*, 1996, **118**, 672–679.

[207] K. Ando, *J. Chem. Phys.*, 2001, **115**, 5228–5237.

[208] P. M. Morse, *Phys. Rev.*, 1929, **34**, 57–64.

[209] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha and et al., *J. Phys. Chem. B*, 1998, **102**, 3586–3616.

[210] M. Buck, S. Bouguet-Bonnet, R. W. Pastor and A. D. MacKerell, *Biophys. J.*, 2006, **90**, L36–L38.

[211] C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, *J. Phys. Chem.*, 1993, **97**, 10269–10280.

[212] M. Schauperl, P. Nerenberg, H. Jang, L.-P. Wang, C. I. Bayly, D. Mobley and M. Gilson, *ChemRxiv*, 2019.

[213] H. A. Lorentz, *Ann. Phys. (Berlin, Ger.)*, 1881, **248**, 127–136.

[214] D. Berthelot, *C. R. Hebd. Seances Acad. Sci.*, 1898, **126**, 1703–1706.

[215] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks and V. S. Pande, *PLoS Comput. Biol.*, 2017, **13**, 1–17.

[216] S. D. Stoddard and J. Ford, *Phys. Rev. A*, 1973, **8**, 1504–1512.

[217] J. Nicolas, K. Gubbins, W. Streett and D. Tildesley, *Mol. Phys.*, 1979, **37**, 1429–1454.

[218] J. Powles, W. Evans and N. Quirke, *Mol. Phys.*, 1982, **46**, 1347–1370.

[219] S. Toxvaerd and J. C. Dyre, *J. Chem. Phys.*, 2011, **134**, 081102.

[220] M. R. Shirts, D. L. Mobley, J. D. Chodera and V. S. Pande, *J. Phys. Chem. B*, 2007, **111**, 13052–13063.

[221] T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.

[222] A. Y. Toukmaji and J. A. Board, *Comput. Phys. Commun.*, 1996, **95**, 73–92.

[223] F. G. J. Longford, J. W. Essex, C.-K. Skylaris and J. G. Frey, *J. Chem. Phys.*, 2018, **148**, 214704.

[224] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *J. Chem. Phys.*, 1995, **103**, 8577–8593.

[225] A. Jakalian, B. L. Bush, D. B. Jack and C. I. Bayly, *J. Comput. Chem.*, 2000, **21**, 132–146.

[226] A. Jakalian, D. B. Jack and C. I. Bayly, *J. Comput. Chem.*, 2002, **23**, 1623–1641.

[227] N. Allinger, in *Calculation of Molecular Structure and Energy by Force-Field Methods*, Elsevier, 1976, vol. 13, p. 1–82.

[228] D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts and et al., *J. Chem. Theory Comput.*, 2018, **14**, 6076–6092.

[229] Y. Qiu, D. G. A. Smith, S. Boothroyd, H. Jang, D. F. Hahn, J. Wagner, C. C. Bannan, T. Gokey, V. T. Lim, C. D. Stern, A. Rizzi, B. Tjanaka, G. Tresadern, X. Lucas, M. R. Shirts, M. K. Gilson, J. D. Chodera, C. I. Bayly, D. L. Mobley and L.-P. Wang, *J. Chem. Theory Comput.*, 2021, **17**, 6262–6280.

[230] L.-P. Wang, T. J. Martinez and V. S. Pande, *J. Phys. Chem. Lett.*, 2014, **5**, 1885–1891.

[231] F. Ercolessi and J. B. Adams, *Europhys. Lett.*, 1994, **26**, 583–588.

[232] F. Leonarski, F. Trovato, V. Tozzini and J. Trylska, in *Genetic Algorithm Optimization of Force Field Parameters: Application to a Coarse-Grained Model of RNA*, ed. C. Pizzuti, M. D. Ritchie and M. Giacobini, Springer Berlin Heidelberg, 2011, vol. 6623, p. 147–152.

[233] M. V. Ivanov, M. R. Talipov and Q. K. Timerghazin, *J. Phys. Chem. A*, 2015, **119**, 1422–1434.

[234] A. Bin Faheem, J.-Y. Kim, S.-E. Bae and K.-K. Lee, *J. Mol. Liq.*, 2021, **337**, 116579.

[235] J. T. Horton, A. E. A. Allen, L. S. Dodda and D. J. Cole, *J. Chem. Inf. Model.*, 2019, **59**, 1366–1381.

[236] J. M. Seminario, *Int. J. Quantum Chem.*, 1996, **60**, 1271–1277.

[237] A. E. A. Allen, M. C. Payne and D. J. Cole, *J. Chem. Theory Comput.*, 2018, **14**, 274–281.

[238] D. J. Cole, J. Z. Vilseck, J. Tirado-Rives, M. C. Payne and W. L. Jorgensen, *J. Chem. Theory Comput.*, 2016, **12**, 2312–2323.

[239] D. J. Cole, I. Cabeza de Vaca and W. L. Jorgensen, *MedChemComm*, 2019, **10**, 1116–1120.

[240] O. C. Madin, S. Boothroyd, R. A. Messerly, J. D. Chodera, J. Fass and M. R. Shirts, *Bayesian Inference-Driven Model Parameterization and Model Selection for 2CLJQ Fluid Models*, 2021.

[241] J. Köfinger and G. Hummer, *ChemRxiv*, 2021.

[242] J. Köfinger and G. Hummer, *Eur. Phys. J. B*, 2021, **94**, 245.

[243] Y. Xie, J. Vandermause, L. Sun, A. Cepellotti and B. Kozinsky, *npj Comput. Mater.*, 2021, **7**, 1–10.

[244] P. Liu, Q. Shi, H. Daumé and G. A. Voth, *J. Chem. Phys.*, 2008, **129**, 214114.

[245] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.

[246] P. S. Hudson, S. Boresch, D. M. Rogers and H. L. Woodcock, *J. Chem. Theory Comput.*, 2018, **14**, 6327–6335.

[247] P. S. Hudson, K. Han, H. L. Woodcock and B. R. Brooks, *J. Comput.-Aided Mol. Des.*, 2018, **32**, 983–999.

[248] T. J. Giese and D. M. York, *J. Chem. Theory Comput.*, 2019, **15**, 5543–5562.

[249] R. M. Betz and R. C. Walker, *J. Comput. Chem.*, 2015, **36**, 79–87.

[250] C. G. Mayne, J. Saam, K. Schulten, E. Tajkhorshid and J. C. Gumbart, *J. Comput. Chem.*, 2013, **34**, 2757–2770.

[251] L. Huang and B. Roux, *J. Chem. Theory Comput.*, 2013, **9**, 3543–3556.

[252] M. Doemer, P. Maurer, P. Campomanes, I. Tavernelli and U. Rothlisberger, *J. Chem. Theory Comput.*, 2014, **10**, 412–422.

[253] D. Shivakumar, J. Williams, Y. Wu, W. Damm, J. Shelley and W. Sherman, *J. Chem. Theory Comput.*, 2010, **6**, 1509–1519.

[254] L.-P. Wang, J. Chen and T. Van Voorhis, *J. Chem. Theory Comput.*, 2013, **9**, 452–460.

[255] B. H. Besler, K. M. Merz and P. A. Kollman, *J. Comput. Chem.*, 1990, **11**, 431–439.

[256] C. A. Reynolds, J. W. Essex and W. G. Richards, *J. Am. Chem. Soc.*, 1992, **114**, 9075–9079.

[257] S. K. Burger, P. W. Ayers and J. Schofield, *J. Comput. Chem.*, 2014, **35**, 1438–1445.

[258] K. Vanommeslaeghe, M. Yang and A. D. MacKerell, *J. Comput. Chem.*, 2015, **36**, 1083–1101.

[259] C. W. Hopkins and A. E. Roitberg, *J. Chem. Inf. Model.*, 2014, **54**, 1978–1986.

[260] O. Guvench and A. D. MacKerell, *J. Mol. Model.*, 2008, **14**, 667–679.

[261] M. Zgarbová, M. Otyepka, J. Šponer, A. Mládek, P. Banáš, T. E. Cheatham and P. Jurečka, *J. Chem. Theory Comput.*, 2011, **7**, 2886–2902.

[262] L.-P. Wang and T. Van Voorhis, *J. Chem. Phys.*, 2010, **133**, 231101.

[263] L.-P. Wang, K. A. McKiernan, J. Gomes, K. A. Beauchamp, T. Head-Gordon, J. E. Rice, W. C. Swope, T. J. Martínez and V. S. Pande, *J. Phys. Chem. B*, 2017, **121**, 4023–4039.

[264] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, *Nat. Methods*, 2020, **17**, 261–272.

[265] M. J. Powell, *Mathematics Today - Bulletin of the Institute of Mathematics and Its Applications*, 2007, **43**, 12.

[266] D. Kraft, *A Software Package for Sequential Quadratic Programming*, Wiss. Berichtswesen d. DFVLR, 1988.

[267] A. R. Conn, N. I. M. Gould and P. L. Toint, *Trust-region methods*, Society for Industrial and Applied Mathematics, 2000.

[268] H. Do and A. Troisi, *Phys. Chem. Chem. Phys.*, 2015, **17**, 25123–25132.

[269] K. Claridge and A. Troisi, *J. Phys. Chem. B*, 2019, **123**, 428–438.

[270] L. Bottou, Proceedings of COMPSTAT'2010, Heidelberg, 2010, pp. 177–186.

[271] D. Bertsimas and J. Tsitsiklis, *Stat. Sci.*, 1993, **8**, 10–15.

[272] S. J. W. Jorge Nocedal, *Numerical Optimization*, Springer New York, 2006.

[273] B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshaye, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. van der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W.-z. Yu and T. Frauenheim, *J. Chem. Phys.*, 2020, **152**, 124101.

[274] B. Aradi, B. Hourahine and T. Frauenheim, *J. Phys. Chem. A*, 2007, **111**, 5678–5684.

[275] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, *J. Phys.: Condens. Matter*, 2017, **29**, 273002.

[276] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del R'ıo, M. Wiebe, P. Peterson, P. G'erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant, *Nature*, 2020, **585**, 357–362.

[277] G. Landrum, *RDKit: Open-source cheminformatics*, Accessed 1st of August, 2020, http://www.rdkit.org.

[278] S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.

[279] S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.

[280] J. C. Womack, N. Mardirossian, M. Head-Gordon and C.-K. Skylaris, *J. Chem. Phys.*, 2016, **145**, 204114.

[281] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi and M. C. Payne, *J. Chem. Phys.*, 2005, **122**, 084119.

[282] R. M. Parrish, L. A. Burns, D. G. A. Smith, A. C. Simmonett, A. E. DePrince, E. G. Hohenstein, U. Bozkaya, A. Y. Sokolov, R. Di Remigio, R. M. Richard, J. F. Gonthier, A. M. James, H. R. McAlexander, A. Kumar, M. Saitow, X. Wang, B. P. Pritchard, P. Verma, H. F. Schaefer, K. Patkowski, R. A. King, E. F. Valeev, F. A. Evangelista, J. M. Turney, T. D. Crawford and C. D. Sherrill, *J. Chem. Theory Comput.*, 2017, **13**, 3185–3197.

[283] G. A. Jeffrey, *An Introduction to Hydrogen Bonding*, Oxford University Press, 1997.

[284] Y. Li and C. Kang, *Molecules*, 2017, **22**, 1399.

[285] M. Rosa, M. Micciarelli, A. Laio and S. Baroni, *J. Chem. Theory Comput.*, 2016, **12**, 4385–4389.

[286] H. H. Heenen, J. A. Gauthier, H. H. Kristoffersen, T. Ludwig and K. Chan, *J. Chem. Phys.*, 2020, **152**, 144703.

[287] M. Aminpour, C. Montemagno and J. A. Tuszynski, *Molecules*, 2019, **24**, 1693.

[288] L. Zhang, W. Li, T. Fang and S. Li, *Theor. Chem. Acc.*, 2016, **135**, 34.

[289] M. Dračínský, H. M. Möller and T. E. Exner, *J. Chem. Theory Comput.*, 2013, **9**, 3806–3815.

[290] D. Wei, H. Guo and D. R. Salahub, *Phys. Rev. E*, 2001, **64**, 011907.

[291] S. Zivanovic, F. Colizzi, D. Moreno, A. Hospital, R. Soliva and M. Orozco, *J. Chem. Theory Comput.*, 2020, **16**, 6575–6585.

[292] S. Duane, A. Kennedy, B. J. Pendleton and D. Roweth, *Phys. Lett. B*, 1987, **195**, 216–222.

[293] E. Akhmatskaya, N. Bou-Rabee and S. Reich, *J. Comput. Phys.*, 2009, **228**, 2256–2265.

[294] C. R. Sweet, S. S. Hampton, R. D. Skeel and J. A. Izaguirre, *J. Chem. Phys.*, 2009, **131**, 174106.

[295] L. D. Gelb, *J. Chem. Phys.*, 2003, **118**, 7747–7750.

[296] R. Iftimie, D. Salahub, D. Wei and J. Schofield, *J. Chem. Phys.*, 2000, **113**, 4852.

[297] S. K. Burger, P. W. Ayers and J. Schofield, *J. Comput. Chem.*, 2014, **35**, 1438–1445.

[298] C. Sampson, T. Fox, C. S. Tautermann, C. Woods and C.-K. Skylaris, *J. Phys. Chem. B*, 2015, **119**, 7030–7040.

[299] C. Cave-Ayland, C.-K. Skylaris and J. W. Essex, *J. Chem. Theory Comput.*, 2017, **13**, 415–424.

[300] J. Michel, R. D. Taylor and J. W. Essex, *J. Chem. Theory Comput.*, 2006, **2**, 732–739.

[301] J. Leiding and J. D. Coe, *J. Chem. Phys.*, 2016, **144**, 174109.

[302] A. Mittal, N. Lyle, T. S. Harmon and R. V. Pappu, *J. Chem. Theory Comput.*, 2014, **10**, 3550–3562.

[303] A. Mittal, A. S. Holehouse, M. C. Cohan and R. V. Pappu, *J. Mol. Biol.*, 2018, **430**, 2403–2421.

[304] S. Ito and Q. Cui, *J. Chem. Phys.*, 2020, **153**, 044115.

[305] P. S. Hudson, S. Boresch, D. M. Rogers and H. L. Woodcock, *J. Chem. Theory Comput.*, 2018, **14**, 6327–6335.

[306] J. D. Coe, T. D. Sewell and M. S. Shaw, *J. Chem. Phys.*, 2009, **131**, 074105.

[307] J. D. Coe, T. D. Sewell and M. S. Shaw, *J. Chem. Phys.*, 2009, **130**, 164104.

[308] P. Bandyopadhyay, *Chem. Phys. Lett.*, 2013, **556**, 341–345.

[309] J. Leiding and J. D. Coe, *J. Chem. Phys.*, 2014, **140**, 034106.

[310] I. Andricioaei and J. E. Straub, *Phys. Rev. E*, 1996, **53**, R3055–R3058.

[311] N. E. Jackson, M. A. Webb and J. J. de Pablo, *J. Chem. Phys.*, 2018, **149**, 072326.

[312] Y. Nagai, M. Okumura, K. Kobayashi and M. Shiga, *Phys. Rev. B*, 2020, **102**, 041124.

[313] R. B. Jadrich and J. A. Leiding, *J. Phys. Chem. B*, 2020, **124**, 5488–5497.

[314] P. Bandyopadhyay, *J. Chem. Phys.*, 2005, **122**, 091102.

[315] A. Nakayama, N. Seki and T. Taketsugu, *J. Chem. Phys.*, 2009, **130**, 024107.

[316] S. Bulusu and R. Fournier, *J. Chem. Phys.*, 2012, **136**, 064112.

[317] L. D. Gelb and T. N. Carnahan, *Chem. Phys. Lett.*, 2006, **417**, 283–287.

[318] A. Warshel and M. Levitt, *J. Mol. Biol.*, 1976, **103**, 227–249.

[319] V. Tomar, *J. Appl. Phys. (Melville, NY, U. S.)*, 2007, **101**, 103512.

[320] G. Marsaglia and T. A. Bray, *SIAM Rev.*, 1964, **6**, 260–264.

[321] B. Mehlig, D. W. Heermann and B. M. Forrest, *Phys. Rev. B*, 1992, **45**, 679–685.

[322] W. C. Swope, H. C. Andersen, P. H. Berens and K. R. Wilson, *J. Chem. Phys.*, 1982, **76**, 637–649.

[323] L. Verlet, *Phys. Rev.*, 1967, **159**, 98–103.

[324] Y. Fang, J. M. Sanz-Serna and R. D. Skeel, *J. Chem. Phys.*, 2014, **140**, 174108.

[325] J. A. Izaguirre and S. S. Hampton, *J. Comput. Phys.*, 2004, **200**, 581–604.

[326] D. Wu and D. A. Kofke, *J. Chem. Phys.*, 2005, **123**, 054103.

[327] D. Wu and D. A. Kofke, *J. Chem. Phys.*, 2005, **123**, 084109.

[328] D. Kraft, *A Software Package for Sequential Quadratic Programming*, Wiss. Berichtswesen d. DFVLR, 1988.

[329] J. Wang, W. Wang, P. A. Kollman and D. A. Case, *J. Mol. Graphics Modell.*, 2006, **25**, 247–260.

[330] V. T. Lim, D. F. Hahn, G. Tresadern, C. I. Bayly and D. L. Mobley, *F1000Research*, 2020, **9**, 1390.

[331] J. C. Brand, D. R. Williams and T. J. Cook, *J. Mol. Spectrosc.*, 1966, **20**, 359–380.

[332] M. Mukherjee, B. Bandyopadhyay, P. Biswas and T. Chakraborty, *Indian J. Phys.*, 2012, **86**, 201–208.

[333] C. W. Bock, P. George and M. Trachtman, *Theor. Chim. Acta*, 1986, **69**, 235–245.

[334] K. C. Gross and P. G. Seybold, *Int. J. Quantum Chem.*, 2000, **80**, 1107–1115.

[335] P. Gkeka, G. Stoltz, A. Barati Farimani, Z. Belkacemi, M. Ceriotti, J. D. Chodera, A. R. Dinner, A. L. Ferguson, J.-B. Maillet, H. Minoux, C. Peter, F. Pietrucci, A. Silveira, A. Tkatchenko, Z. Trstanova, R. Wiewiora and T. Lelièvre, *J. Chem. Theory Comput.*, 2020, **16**, 4757–4775.

[336] J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.

[337] J. A. Izaguirre, C. R. Sweet and V. S. Pande, in *Multiscale Dynamics of Macromolecules Using Normal Mode Langevin*, World Scientific, 2009, p. 240–251.

[338] K.-H. Chow and D. M. Ferguson, *Comput. Phys. Commun.*, 1995, **91**, 283–289.

[339] D. Bakowies and W. Thiel, *J. Phys. Chem.*, 1996, **100**, 10580–10594.

[340] S. J. Weiner, U. C. Singh and P. A. Kollman, *J. Am. Chem. Soc.*, 1985, **107**, 2219–2229.

[341] T. Vreven and K. Morokuma, in *Chapter 3 Hybrid Methods: ONIOM(QM:MM) and QM/MM*, Elsevier, 2006, vol. 2, p. 35–51.

[342] A. O. Dohn, *Int. J. Quantum Chem.*, 2020, **120**, e26343.

[343] E. Brunk and U. Rothlisberger, *Chem. Rev. (Washington, DC, U. S.)*, 2015, **115**, 6217–6263.

[344] R. J. Bartlett and M. Musiał, *Rev. Mod. Phys.*, 2007, **79**, 291–352.

[345] J. Morado, P. N. Mortenson, J. W. M. Nissink, M. L. Verdonk, R. A. Ward, J. W. Essex and C.-K. Skylaris, *J. Chem. Theory Comput.*, 2021, **17**, 7021–7042.

[346] L. C. Pierce, R. Salomon-Ferrer, C. Augusto F. de Oliveira, J. A. McCammon and R. C. Walker, *J. Chem. Theory Comput.*, 2012, **8**, 2997–3002.

[347] J. N. Ehrman, V. T. Lim, C. C. Bannan, N. Thi, D. Y. Kyu and D. L. Mobley, *J. Comput.-Aided Mol. Des.*, 2021, **35**, 271–284.

[348] J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.

[349] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *J. Chem. Phys.*, 2018, **148**, 241722.

[350] O. T. Unke and M. Meuwly, *J. Chem. Theory Comput.*, 2019, **15**, 3678–3693.

[351] B. Linclau, F. Peron, E. Bogdan, N. Wells, Z. Wang, G. Compain, C. Q. Fontenelle, N. Galland, J.-Y. Le Questel and J. Graton, *Chem. - Eur. J.*, 2015, **21**, 17808–17816.

[352] C. Devereux, J. S. Smith, K. K. Huddleston, K. Barros, R. Zubatyuk, O. Isayev and A. E. Roitberg, *J. Chem. Theory Comput.*, 2020, **16**, 4192–4202.

[353] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev and A. E. Roitberg, *J. Chem. Phys.*, 2018, **148**, 241733.

[354] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev and A. E. Roitberg, *Nat. Commun.*, 2019, **10**, 2903.

[355] J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev and S. Tretiak, *Sci. Data*, 2020, **7**, 134.

[356] D. A. Rufa, H. E. Bruce Macdonald, J. Fass, M. Wieder, P. B. Grinaway, A. E. Roitberg, O. Isayev and J. D. Chodera, *bioRxiv*, 2020.

[357] S.-L. J. Lahey, T. N. Thien Phuc and C. N. Rowley, *J. Chem. Inf. Model.*, 2020, **60**, 6258–6268.

[358] D. L. Folmsbee, D. R. Koes and G. R. Hutchison, *J. Phys. Chem. A*, 2021, **125**, 1987–1993.

[359] R. Galvelis, S. Doerr, J. M. Damas, M. J. Harvey and G. De Fabritiis, *J. Chem. Inf. Model.*, 2019, **59**, 3485–3493.

[360] S. Zhu, *J. Chem. Inf. Model.*, 2019, **59**, 4239–4247.

[361] S.-L. J. Lahey and C. N. Rowley, *Chem. Sci.*, 2020, **11**, 2362–2368.

[362] J. W. Vant, S.-L. J. Lahey, K. Jana, M. Shekhar, D. Sarkar, B. H. Munk, U. Kleinekathöfer, S. Mittal, C. Rowley and A. Singharoy, *J. Chem. Inf. Model.*, 2020, **60**, 2591–2604.

[363] D. J. Cole, L. Mones and G. Csányi, *Faraday Discuss.*, 2020, **224**, 247–264.

[364] M. Wieder, J. Fass and J. D. Chodera, *Chem. Sci.*, 2021, **12**, 11364–11381.

[365] T. Fink and J.-L. Reymond, *J. Chem. Inf. Model.*, 2007, **47**, 342–353.

[366] T. Fink, H. Bruggesser and J.-L. Reymond, *Angew. Chem., Int. Ed.*, 2005, **44**, 1504–1508.

[367] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos and J. P. Overington, *Nucleic Acids Res.*, 2014, **42**, D1083–D1090.

[368] B. Brauer, M. K. Kesharwani, S. Kozuch and J. M. L. Martin, *Phys. Chem. Chem. Phys.*, 2016, **18**, 20905–20925.

[369] D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge and P. W. Chung, *Sci. Rep.*, 2018, **8**, 9059.

[370] W. Pronobis, A. Tkatchenko and K.-R. Müller, *J. Chem. Theory Comput.*, 2018, **14**, 2991–3003.

[371] M. Pinheiro, F. Ge, N. Ferré, P. O. Dral and M. Barbatti, *Chem. Sci.*, 2021, **12**, 14396–14413.

[372] M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.

[373] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. Lilienfeld and K.-R. Müller, Advances in Neural Information Processing Systems, 2012.

[374] F. Faber, A. Lindmaa, O. A. von Lilienfeld and R. Armiento, *Int. J. Quantum Chem.*, 2015, **115**, 1094–1101.

[375] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2015, **6**, 2326–2331.

[376] A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Phys. Rev. Lett.*, 2010, **104**, 136403.

[377] A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B*, 2013, **87**, 184115.

[378] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi and M. Ceriotti, *Sci. Adv.*, 2017, **3**, e1701816.

[379] J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.

[380] D. Rosenberger, J. S. Smith and A. E. Garcia, *J. Phys. Chem. B*, 2021, **125**, 3598–3612.

[381] W. D. Cornell, P. Cieplak, C. I. Bayly and P. A. Kollman, *J. Am. Chem. Soc.*, 1993, **115**, 9620–9631.

[382] R. Woods and R. Chappelle, *J. Mol. Struct.: THEOCHEM*, 2000, **527**, 149–156.

[383] J. M. Turney, A. C. Simmonett, R. M. Parrish, E. G. Hohenstein, F. A. Evangelista, J. T. Fermann, B. J. Mintz, L. A. Burns, J. J. Wilke, M. L. Abrams, N. J. Russ, M. L. Leininger, C. L. Janssen, E. T. Seidl, W. D. Allen, H. F. Schaefer, R. A. King, E. F. Valeev, C. D. Sherrill and T. D. Crawford, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 556–565.

[384] P. Cieplak, J. Caldwell and P. Kollman, *J. Comput. Chem.*, 2001, **22**, 1048–1057.

[385] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern and et al., *PLoS Comput. Biol.*, 2017, **13**, e1005659.

[386] J. Wang, W. Wang, P. A. Kollman and D. A. Case, *J. Mol. Graphics Modell.*, 2006, **25**, 247–260.

[387] P. J. Flory, *Statistical Mechanics of Chain Molecules*, Interscience Publishers, 1969.

[388] E. Benedetti, G. Morelli, G. Némethy and H. A. Scheraga, *Int. J. Pept. Protein Res.*, 1983, **22**, 1–15.

[389] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi,

M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian 09 Revision D.01*, Gaussian Inc. Wallingford CT 2009.

[390] R. Ditchfield, *Mol. Phys.*, 1974, **27**, 789–807.

[391] K. Wolinski, J. F. Hinton and P. Pulay, *J. Am. Chem. Soc.*, 1990, **112**, 8251–8260.

[392] T. Helgaker, M. Watson and N. C. Handy, *J. Chem. Phys.*, 2000, **113**, 9402–9409.

[393] P. J. Wilson, T. J. Bradley and D. J. Tozer, *J. Chem. Phys.*, 2001, **115**, 9233–9242.

[394] F. Jensen, *J. Chem. Theory Comput.*, 2006, **2**, 1360–1369.

[395] T. W. Keal, T. Helgaker, P. Sałek and D. J. Tozer, *Chem. Phys. Lett.*, 2006, **425**, 163–166.

[396] T. Kupka, M. Nieradka, M. Stachów, T. Pluta, P. Nowak, H. Kjær, J. Kongsted and J. Kaminsky, *J. Phys. Chem. A*, 2012, **116**, 3728–3738.

[397] S. Kristyán and P. Pulay, *Chem. Phys. Lett.*, 1994, **229**, 175–180.

[398] J. F. Dobson, K. McLennan, A. Rubio, J. Wang, T. Gould, H. M. Le and B. P. Dinte, *Aust. J. Chem.*, 2001, **54**, 513.

[399] J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615.

[400] Y.-S. Lin, G.-D. Li, S.-P. Mao and J.-D. Chai, *J. Chem. Theory Comput.*, 2013, **9**, 263–272.

[401] J. W. Ochterski, *Thermochemistry in Gaussian*, 2000, https://gaussian.com/thermo/.

[402] A. R. Leach, in *A Survey of Methods for Searching the Conformational Space of Small and Medium-Sized Molecules*, ed. K. B. Lipkowitz and D. B. Boyd, John Wiley & Sons, Inc., 2007, p. 1–55.

[403] J. Graton, Z. Wang, A.-M. Brossard, D. Gonçalves Monteiro, J.-Y. Le Questel and B. Linclau, *Angew. Chem., Int. Ed.*, 2012, **51**, 6176–6180.