The development of accurate *in silico* multiscale physics models of the interface with biology

Mohamed Ali al-Badri

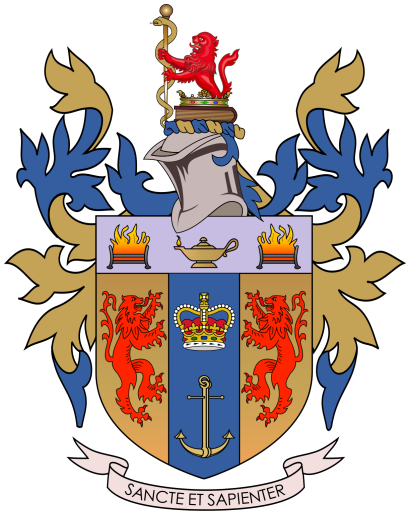
A dissertation submitted in fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

King's College London.



Department of Physics

King's College London

April 14, 2021

I, Mohamed Ali al-Badri, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

There is an increasing demand for accurate modelling of biological processes where conventional experimental and computational methods break down. In particular, the engineering of biotechnology at the nano-scale makes accurate atomistic and dynamic simulations of complex systems at the bio-nano interface indispensable. This thesis explores the accurate modelling of systems within this domain, where accuracy is applied both in structural characterisation as well as defining the resulting chemical properties. Electronic structure theory calculations are applied for the development of bespoke molecular dynamics forcefield parameters. This work illustrates the capacity for costly quantum-mechanical calculations to be extrapolated to classical molecular dynamics simulations of systems composed of hundreds of thousands of atoms while retaining the accuracy observed in both experiment and state-of-the-art *ab initio* methods. Accurate characterisation is studied using different theoretical tools to elucidate both the function and inhibition of proteins. The sensitivity of adsorbed protein denaturing, protein corona formation and the cellular uptake of a protein-nanomaterial complex to nanomaterial functionalisation is studied to explain interfacial interactions that drive unwanted phenomena in biotechnology. Additionally, the accuracy of the dynamic atomistic or electronic character of protein active sites is investigated. In particular, the cessation of proteolytic activity of SARS-CoV-2 is detailed through the disruption of a catalytic dyad in the main protease active site. Finally, the accurate modelling of the strongly correlated electronic ground state of the hemocyanin oxy-

gen transporting protein active site is investigated using a hybrid density functional theory + dynamical mean field theory (DFT+DMFT) quantum-mechanical treatment. These multiscale modelling applications convey the ability to extend preexisting theoretical tools to the burgeoning demand of accurate and large-scale modelling of biological phenomena, both to understand the otherwise impenetrable processes using conventional tools and to inform the development of new interdisciplinary tools in bio-nano engineering.

And say, All praise is due to Allah.
He will show you His signs, and
you will recognise them. And your
Lord is not unaware of what you do.

The Holy Quran, 27:93

To my beloved parents

Acknowledgements

First and foremost, I owe my supervisor Chris Lorenz my gratitude for always encouraging me in my pursuit of research questions, no matter how challenging. I have him to thank for introducing me to computational research; setting a Metropolis Monte Carlo project in my second year as an undergraduate at King's, after which I haven't looked back. Chris has never allowed me to be affected by the challenges of a PhD thanks to his support and open door. I will forever be grateful to him for his guidance and friendship.

The work in this thesis is built on the knowledge and support of my supervisors and collaborators, Khuloud al-Jamal, Cedric Weber, Daniel Cole, Edward Linscott, Paul Smith, Robert Sinclair, Antoine Georges, Khaled Abdel-Maksoud and Jonathan Essex, to all of whom I am grateful. I would also like to thank the London Interdisciplinary Doctoral Programme (LIDo) team for all their hard work to equip us with the requisite skills to confidently tackle fundamental questions in biology, and for moulding a convivial collaborative environment for us to develop in outside our host institutions.

My thanks goes to my mentors at The Alan Turing Institute; Oliver Strickson, Eric Daub and Martin O'Reilly and the rest of the Hut 23 research software engineering team, who, over four months of working on the uncertainty quantification of multi-scale and multi-physics computer models project, welcomed, stimulated and taught me so much and I am indebted to them all for their generosity.

Having spent more than 8 years at the physics department at King's, I am lucky to have met my King's family; especially the members of the Lorenz Lab, Paul Le Long, James French, John Ellis, Dylan Owen, Malcolm Fairbairn, Eva Philippaki, Nashwan Sabti, James Alvey, Bethan Cornell, Sreedevi Varma, Dries Seynaeve, Claudio Zeni, Eloy de Jong, Thomas Helfer and of course those who are no longer with us; Alan Michette and Alessandro De Vita. A special thank you to my friends to whom I could always turn to for support; Adam al-Makroudi, Fred Pedicona, Greg Szep, Nashwan Sabti, Khaled Abdel-Maksoud and Abdulah Fawaz.

I dedicate this thesis to my family. Principally my parents, without whom I would not be enjoying the freedoms I have today; having both been imprisoned, tortured and persecuted out of their homeland, they sacrificed everything to carry a six-month old out of the horrors of fascism, and agonisingly made their way to a safe haven, where I have been welcomed and now call home. My wife Georgina, whose irritating ability to do everything perfectly in life and academia makes her someone I proudly call my role model and my best friend. The light of my life, my daughter Noura; spending the last year of the PhD at home due to the pandemic has been the greatest time of my life, just to see you grow up so quickly while "Baba doing working" in the background. A lot of thanks and love to my best friends; my siblings.

Conferences

During the PhD, I have presented at or attended the following conferences/workshops:

- *Green's function methods: the next generation III*, June 2017, Toulouse, France
- *International Summer School on Computational Quantum Materials*, May 2018, Quebec, Canada
- *The CCP9 Young Researchers & Community Meeting*, July 2018, Cambridge, UK
- *The 6th Annual CCPBioSim Meeting: Molecular Simulations in Drug Discovery and Development*, September 2018, Oxford, UK
- *South West Computational Chemists Meeting*, November 2018, Bath, UK
- *Modeling Metal Nanoparticles: environment and dynamical effects*, December 2018, Grenoble, France
- *Forcefields: Status, Challenges & Vision*, January 2019, Daresbury, UK
- *Computational Molecular Science*, March 2019, Warwick, UK
- *ONETEP Masterclass*, August 2019, Warwick, UK
- *American Physical Society Conference*, March 2020, Denver, CO, USA

Contents

List of Figures

List of Tables

Chapter 1

Introduction

I wish to God these calculations had been executed by steam.

Charles Babbage

The advent of the modern digital computer, as formalised by Alan Turing,[1] ignited the field of computational physics, aided by preexisting theoretical formulations of algorithms. Starting from the first experiments with Monte Carlo (MC) simulations in the 1930s by Fermi and the formulation of the Markov-Chain Monte Carlo (MCMC) technique by Ulam in the 1940s, von Neumann programmed the 18,000 vacuum-tube Electronic Numerical Integrator and Computer (ENIAC) computer to investigate neutron diffusion in fissionable materials.[2] This success paved the way for the integration of Newton's equations of motion to compute the time evolution of a many-body system.

In 1953, Fermi, Pasta, Ulam and Tsingou simulated a 1-dimensional analogue of atoms in a crystal, formed of masses connected by springs that obey Hooke's law and a weak non-linear term on the Mathematical Analyzer Numerical Integrator and Automatic Computer Model I (MANIAC I) computer. Using this, they were able to simulate the time evolution of the energy of a system to understand the origins of irreversibility in nature. This is known as the Fermi–Pasta–Ulam–Tsingou problem.[3]

The first use of molecular dynamics (MD) simulations were implemented by Alder and Wainwright in 1957 using an IBM 704 computer, where they simulated the elastic collisions in a hard sphere fluid composed of a maximum of 108 particles. They outlined the importance of the evolution in time of a system towards a steady-state, which was outside the scope of the preexisting MC techniques and within the grasp of MD. Conscious of the limitations of computational power that was available, they acknowledge that their work is qualitative and indicate that, given better resources, simulations could give quantitative statistical averages of thermodynamic properties at equilibrium. Unlike modern implementations of MD with continuous potentials, Alder and Wainwright modelled their hard sphere liquid using discontinuous potentials where events such as particle collisions define the simulation statistics.[4]

Arguably the first MD simulations that reflected the verifiability of the method as a science, by comparing with experiment, are those of liquid argon by Rahman in 1964 using a Control Data Corporation (CDC) 3600 computer. Similar to modern uses of MD, Rahman's simulations modelled the interactions using a pairwise Lennard-Jones potential, first proposed in 1932,[5] of an 864 particle system. Using these MD simulations, a radial distribution function and the self-diffusion constant were calculated to compare to experiment, where the latter was found to be within 15% of the experimental value. The discrepancy between predicted observables and experiments were in the range of 20%, which spurred Rahman to tweak the repulsive part of the Lennard-Jones potential; forcefield development therefore reared its head as soon as MD came to fruition.[6]

The highlight of the history of molecular simulations is undoubtedly the 2013 Nobel Prize in Chemistry for the development of multiscale mod-

els for complex chemical systems. It recognised the pioneering work of Martin Karplus, Michael Levitt and Arieh Warshel in a range of molecular simulation techniques, from quantum chemistry to MD simulations of biomolecules. Most notable among them is the development of the hybrid quantum mechanics/molecular mechanics (QM/MM) simulation technique by Levitt and Warshel.[7] This method combined the accuracy of quantum mechanical calculations with the efficiency of MD simulations by defining a QM region within an encompassing MM region. In another seminal work, they studied the folding of a model bovine pancreatic trypsin inhibitor protein in 1975, by simplifying the structure of the protein and simulating the system at $T = 1000$ K, to crudely solve for the global minimum through simulated annealing. This paper could be considered the first MD simulation of a biological process, albeit a simplified one.[8] Of his many accomplishments, Martin Karplus and his group created the Chemistry at Harvard Macromolecular Mechanics (CHARMM) MD engine and coordinated its development, as well as formalising and parameterising various iterations of the CHARMM forcefield, which has been one of the primary forcefields used for simulating biological molecules.[9] Karplus proposed the idea of first principles (*ab initio*) molecular dynamics (AIMD) simulations in 1973, where trajectories would describe dynamic organic reactions as in MD, but be defined by quantum mechanical forces instead of semi-empirical mechanical techniques.[10]

1.1 Current state-of-the-art in biomolecular simulations

1.1.1 Groundbreaking simulations

In 2006, MD simulations of an all-atom complete satellite tobacco mosaic virus (Fig. 1.1), in a system composed of a million particles, were conducted using the Nanoscale Molecular Dynamics (NAMD) MD engine for 50 ns. This work demonstrated the capacity for MD to study temporal properties

of matter at such a large scale, which is at the heart of the mechanistic understanding of life itself. One of the major findings in this study was the conditional stability of the viral capsid in the presence of the enclosed ribonucleic acid (RNA).[11] This was a profound result both from an evolutionary perspective and a computational one; to reveal the power of pairwise interaction parameters and a handful of algorithms, to describe the physio-chemical drivers at the heart of the function of biomolecular processes.

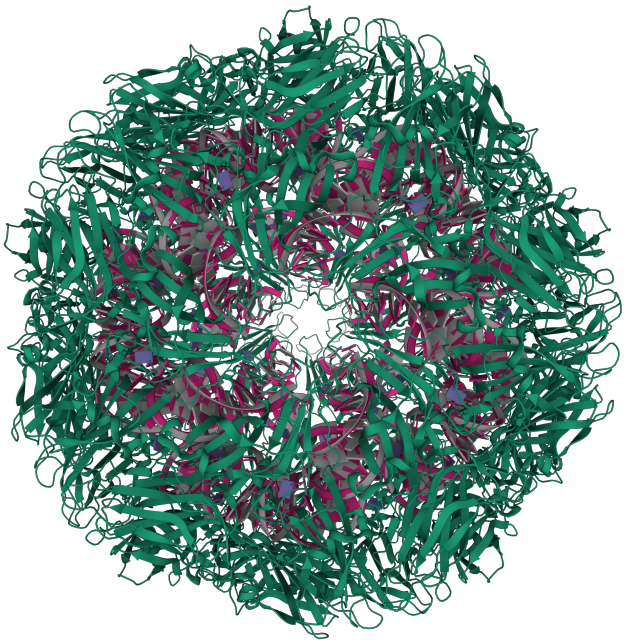


Figure 1.1: The structure of the satellite tobacco mosaic virus (PDB code 4OQ9), showing the protein capsid (green) enveloping the RNA (purple).[12]

Using MD as a tool to investigate the structure and dynamics of systems out of the reach of conventional experimental methods, is therefore an avenue to advancing our understanding of biomolecular processes at the atomic scale. In the understanding of disease, an early breakthrough of MD-guided drug discovery is due to Andrew McCammon and colleagues. In their work, they performed 2 ns MD simulations of HIV integrase where they identified a cryptic trench which became a target for the first FDA ap-

proved HIV integrase inhibitor (raltegravir).[13]

More recently the distributed computing project known as Fold@Home, originally set up by Vijay Pande in 2000, achieved exascale computing using a million personal computers from volunteers worldwide. Using this wealth of computational power, a plethora of MD simulations of protein dynamics were run to develop new therapeutics for a variety of diseases. A highlight of this effort came from the ongoing COVID-19 pandemic. The Fold@Home team ran a 0.1 second MD simulation to study the conformational changes in the SARS-CoV-2 spike protein to understand its role in evading an immune response, and similar to the example of HIV integrase, revealed cryptic binding sites for druggability.[14]

1.1.2 Bespoke forcefield parametrisation

Karplus' proposition to perform MD using QM forces, when correctly implemented, could circumvent the inherent inaccuracies of MD using generalised forcefields and be close to an exact MD. As an approach, AIMD is open to implementing any electronic structure theory method to calculate the atomic forces on-the-fly. The use of density functional theory (DFT) as a basis for calculating coupled electron-ion dynamics for AIMD was actualised by Car and Parrinello in 1985, resulting in a method known as Car-Parrinello MD (CPMD).[15]

This golden standard is — in its modern form — the closest we have come to accurately investigating the dynamics of matter at that scale. It is scale that also defines its limitations; AIMD simulations are restricted to system sizes and simulation times much smaller than the timescales at which important biomolecular processes such as protein conformational changes, ligand binding events and macromolecular interactions take place. Alternatively, more accurate forcefields may be an avenue to achieving a

similar level of accuracy as in AIMD, and therefore provide the desired understanding of biology at the atomistic scale (see Chapter 2.1). For decades, the parametrisation of forcefields has achieved wide coverage through extrapolating parameters from small molecular systems, and where necessary, optimised through experimental observables when a forcefield underperforms.[16] The ability to reproduce the chemical properties of systems is hinged on whether the forcefield training and test sets include corresponding molecules. In the absence of rigorous testing, which is time consuming and certainly not a formal requisite to an MD investigation, the validity of predicted observables is certainly not guaranteed. The advantage of a large international MD community is the rapid identification of errors and the subsequent continuous integration of corrections by the forcefield developers. Importantly however, the persisting divergence from accuracy is almost exclusively limited to molecular systems with exotic chemistries.

1.1.2.1 OpenFF initiative

Accurate MD modelling has commercial potential with significant uptake by the pharmaceutical industry, in the form of the OPLS-3 forcefield,[17] made exclusive to the D. E. Shaw Research (DESRES) developed Desmond MD engine.[18] In contrast, the Open Force Field (OpenFF) initiative is an open-source, collaborative and community-led project for the development of better forcefields.[19] OpenFF is parametrised using the ForceBalance forcefield optimisation software tool,[20] which uses experimental and/or quantum mechanical calculations as reference data to produce forcefield parameters. With sufficient community-contributed data and a strong automated infrastructure, held together by good software development, the OpenFF Parsley forcefield aims to incorporate diverse chemistries for small molecules, and hence improve the accuracy of MD simulations within that niche. Parsley has achieved a similar accuracy for liquid properties when compared to other forcefields, such as the General Amber Force Field (GAFF), while showing improvements in the optimised geometries and con-

formational energetics of small molecules outside its training set.[19] However, it will not rival generalised forcefields on accuracy and wide coverage without the continued integration of data. The uptake of the forcefield may be somewhat retarded by the SMIRKS [21] Native Open Force Field (SMIRNOFF) parameter assignment formalism, which unlike general commonly applied forcefields is based on chemical perception, where the assignment of forcefield parameters is solely based on whether a bond between atoms is a single, double, triple or aromatic bond. This is unlike the assignment of complex atom type labels given to each atom to determine its interaction parameters with other atom types in conventional forcefield formats. The SMIRNOFF formalism is not currently transferable to all MD engines.

1.1.2.2 Gaussian process regression

Other groups have approached the forcefield question from a purely quantum mechanical perspective by learning a non-parametric tabulated forcefield of a system through the use of Gaussian process regression.[22, 23, 24] Such kernel based approaches set out to learn a local configuration relative to each atom in a system and the forces that act upon them from *ab initio* methods such as DFT, on-the-fly. The accuracy of the resulting Gaussian process regression forcefield is determined by its ability to reproduce DFT forces. Gaussian process regression methods for forcefield development have so far been applied successfully in the field of material science, applied to systems such as Nickel nanoclusters,[25] and 10,000 atom systems of stanene.[26] This method has been applied to molecular systems, but in applications where it has been used so far the complexity of the systems have not extended beyond a couple of element species, and system sizes have not extended beyond a handful of atoms.[27, 28] This is most likely due to the training set having factorial scaling with the number of element species.[29]

Another kernel-based approach in the form of so-called gradient-domain machine learning (GDML) employs high-level *ab initio* calculations, such as coupled cluster quantum mechanical calculations (CCSD) to generate MD forcefields for use in small molecular systems. This approach is limited by the scaling capacity of the underlying CCSD calculations, to which the authors give an upper bound of 20 atoms.[30] The advantage of this approach is unparalleled accuracy for such small systems, however it does require a set of a few hundred molecular conformations to train the kernel machine. A potential extension to this limitation, and to achieve large scale quantum accuracy in MD forcefields (where necessary), is to apply a similar approach but using hybrid electronic structure calculations. In Chapter, we apply such a hybrid approach using DFT and dynamical mean field theory (DMFT) in the form of DFT+DMFT, which could scale to hundreds of thousands of atoms.

1.1.2.3 Neural network forcefields

Alternative learning methods in the form of neural networks have been applied to DFT data of organic molecules, to produce MD forcefields that retain accuracy and transferability. The Accurate Neural Network Engine for Molecular Energies (ANAKIN-ME) forcefield and the Neural Network Reactive Forcefield (NNRF) for organic molecular systems have been limited to hydrogen, carbon, nitrogen and oxygen species to date.[31, 32] With the increased availability of electronic structure calculation data, neural network based approaches could help attain a healthy coverage while retaining accuracy from higher level theory. Unfortunately however, they have so far been applied using rudimentary neural network architectures in the form of multilayer perceptrons (MLPs) that utilise no prior knowledge of the data structure.

1.1.2.4 QUBE forcefield

A fully quantum-mechanically derived forcefield bypasses the requirement for experimental data in the forcefield development process. The biggest

advantage of such a protocol is the decentralisation of accurate forcefield efforts; it no longer depends on behemoth electronic structure data sets of a subset of chemical space — which may or may not correlate with the user's molecules of interest — and depends solely on the molecular system in question. Pairwise non-bonded forcefield parameters are derived using an application of electronic structure calculations, where linear-scaling DFT software can be utilised for systems composed of thousands of atoms, outlined in Section (2.2.5).

The Quantum mechanical Bespoke (QUBE) force field model is an example of such a method, which is packaged in code of excellent quality to make the derivation of forcefield parameters for small organic molecules seamless using QUBEKit.[33] Bonded interactions are derived directly from the QM Hessian matrix using a modified Seminario method for the determination of the bond stretching and angle bending forcefield parameters,[34] whereas the dihedral torsional angle forcefield terms are derived using QM torsional scans through constrained geometry optimisation. Software of similar efficiency is required to extrapolate this method to the derivation of forcefield parameters for large biomolecules, the absence of which will limit the use of this approach to specialist groups responsible for its development and restrict its use to a subset of MD engines. With the availability of efficient streamlined software for both small and large organic molecules, the QUBE technique will have a big impact on the accuracy of MD simulations. Protein-ligand simulations will no longer require the mixing of generalised forcefields with bespoke parameterisation of small ligand forcefield parameters, while achieving competitive accuracy in free energies of binding when compared to experiment.[35] Similarly the QUBE forcefield could then be extended to protein-protein interactions, opening a host of exotic application systems where accuracy is paramount, such as metalloproteins.

1.1.2.5 Outlook

The future of MD and forcefield development — where dynamics remain sacrosanct — can greatly benefit from the wealth of method development in the field of deep learning. Instead of using early formulations of neural networks that do not properly reflect the problem at hand, educated architectures have recently been developed to learn to wholly simulate the trajectories of macroscopic particles in the form of Graph Network-based Simulators (GNS).[36] Additionally, with the development of MD engine software where backpropagation through the MD engine itself is possible, entire trajectories are open to be differentiated, opening the doors to interface with machine learning libraries.[37]

Alternatively, where the dynamics of biomolecules are irrelevant and the steady-state solution is of sole interest, a wealth of biological data can give rise to highly accurate predictions. The relationship between amino acid sequence and protein structure was successfully predicted in 24 out of 43 free modelling domains of the Critical Assessment of Protein Structure Prediction experiment,[38] using convolutional residual networks on protein native contact maps computed using the PDB archive.[39] However as the COVID-19 pandemic has clarified, a static protein structure is not as informative for learning to inhibit protein function as it is for protein folding. The challenges of inhibiting protein function with respect to SARS-CoV-2 is studied in Chapter (5).

1.1.3 State of the art architecture

With the inflation of computational power, comes the inflation of system sizes we can study in MD, giving scientists the power to ask questions of the atomistic detail of biomolecular mechanisms on a vast array of scales. Almost all such simulations, as those discussed in section (1.1), have so far been performed on high performance computing (HPC) clusters. These are usually state or corporate-sponsored machines that can cost up to tens of

millions of pounds, used to perform calculations including MD simulations.

1.1.3.1 GPU acceleration

In the last few years, MD engines such as the Groningen Machine for Chemical Simulations (GROMACS) — used for all the MD simulations in this thesis — have modified their software to implement the MD algorithms on graphics processing units (GPUs). GPUs are specialised processors that can simultaneously perform multiple calculations across streams of data. They were originally designed to accelerate graphics rendering but have become the customary computational architecture for scientific applications such as machine learning. Both bonded and non-bonded interaction calculations can be performed on the GPU using the GROMACS MD engine, where the majority of the speedup of using a GPU compared to central processing units (CPU) are the short-ranged non-bonded interactions, where a GPU is inherently more suited to parallelising the problem to reduce the calculation time. The option to perform long-range non-bonded interactions exclusively on the GPU has recently become possible in GROMACS 2020, using the CUDA Fast Fourier Transform library to perform the PME calculations (see section (2.1.7.2)). Furthermore, coordinate updates and constraint calculations can also be performed on the GPU, allowing all MD simulation algorithms to be computed on the GPU. This has a significant impact on accelerating calculations, as it reduces the need for GPU-CPU communication through device-to-host and host-to-device data transfer, which uses lower bandwidth peripheral component interconnect express. Instead, data sharing on the GPU happens through intra-GPU peer-to-peer communication. A great benefit of staying up-to-date with more commercially available architectures gives scientists the ability to study more and more complex systems without the need for exclusive access to HPC facilities.

1.1.3.2 ASIC

An altogether different HPC architecture, built with the special purpose of performing MD simulations for large biomolecular systems is the An-

ton supercomputer by DESRES.[40] Its architecture, which falls under the umbrella of application-specific integrated circuits (ASIC) is therefore inherently efficient at performing MD simulations, achieving microsecond to millisecond timescales. Performing non-bonded interactions via a high-throughput interaction subsystem, which together with its limiting of inter- and intra-chip communication, accelerates MD simulations by two orders of magnitude compared to ordinary CPU architectures.[41]

1.1.3.3 FPGA

Fully integrated field-programmable gate array (FPGA) circuits are architectures that are designed to be configured by the user for any intended use following its manufacture. FPGAs have come to rival ASIC architectures through increased speed, lower cost and the possibility to reconfigure the circuit on-the-fly.[42] MD simulations were first implemented on FPGAs in 2019, where they were simulated on a single Intel Stratix 10 FPGA chip, achieving a throughput similar to that of GPU architectures of 630 ns/day for a 24,000 atom system.[43] With this burgeoning field, MD simulations could find that flexible reconfigurable computing may become better suited than conventional architectures in future.

1.2 Motivation

There are a wealth of bespoke forcefield development methods as outlined in section (1.1.2), and they are by no means the only ones available. As illustrated, accurate forcefield parameterisation has always been important and is now becoming increasingly accessible. In this thesis, I have for the first time extended the development of bespoke forcefield parameters to accelerate the accurate investigation of exotic materials — without undue dependence on tedious experimental fitting procedures — and extrapolate the most rigorous of the aforementioned methods to large systems in a domain of bio-nano interactions. The prime requirements for a bespoke forcefield in this domain are different to those applied for small molecular

organic systems; fundamentally the availability of linear-scaling electronic structure calculations to derive the forcefield parameters of macromolecules such as proteins. Furthermore, it is important to consider the transferability of the generated forcefield to MD engines with competitive acceleration on modern architectures, enabling the investigation of complex problems which are often large in scale.

In the study of graphene-oxide (GO) nanomaterials in Chapter (3), we stress the importance of accurate structure of nanomaterials in reproducing the chemical properties in addition to the forcefield parameters. The semi-ordered structure of GO, composed of inhomogeneous regions of correlated oxidised and unoxidised domains, enforces the complete quantum mechanical treatment of the entire nanosheet. This is unlike the averaging of forcefield non-bonded parameters by atom type, as is the convention in transferable forcefields, but instead ascribes bespoke forcefield parameters to each and every constituent atom. With such a large-scale treatment of materials with exotic electronic properties, it becomes exceedingly complex and costly to derive parameters, but the results reflect accuracy of state-of-the-art AIMD while applicable to simulations of hundreds of thousands of atoms.

In the absence of streamlined software, the extension of this approach to fully quantum forcefields for all the constituent molecules is on hold. The work in Chapter (4) extends the importance of accurate modelling of nanomaterials from a structural perspective to the protein corona problem. The technologically advantageous properties of nanomaterials are lagging behind in their translation to the biotechnology field, constrained by undesired physio-chemical effects including the adsorption of proteins as in the protein corona. In section (1.1), the ability to unpick the natural mechanistic workings of biomolecules with atomic precision has been demonstrated

to be valuable, in Chapter (4) we study the impact of changes in engineering nanotechnology on protein structure, denaturing and binding at the bio-nano interface. The spatio-temporal datasets emerging from MD simulations are exceedingly large and thus complex, requiring efficacious data-processing treatment to return coherent results. We make use of different analysis techniques to develop a pipeline that can accurately shine a light on the consequences of nanomaterial functionalisation on macromolecular structures. Sometimes such results may not be exactly verifiable through preexisting experimental results, simply due to the small-scale and time-dependent dynamic behaviour of complex interfacial interactions. The work employed can however serve as a critical approach to informing biotechnological investigations.

Simulating accurate dynamics of protein interactions, as we see in Chapter (5), use MD and rare-event sampling techniques, together with comprehensive analyses to accelerate the discovery of inhibitors to the SARS-CoV-2 main protease. In this work simulations are carried out on both SARS-CoV-1 main protease interacting with a potent inhibitor, and the SARS-CoV-2 main protease with a contender inhibitor. Benefiting from both proteases having total sequence conservation, the identical active sites and the presence of a pre-identified SARS-CoV-1 main protease inhibitor, we focus on the accurate modelling of the main protease active site. We confirm that the main protease activity is mediated by an allosteric mechanism linked to a catalytic dyad between the His41-Cys145 residues. Using an accurate dynamic model of the active site, we investigate the disruption of this catalytic dyad and its stabilisation in an open conformation through favourable ligand interactions. By analysing the MD trajectories of the protease-ligand interactions, we identify distinct states of the open and closed active site and the free energy difference between the two states, that can be recovered through a stabilising ligand. The work uses Metadynam-

ics techniques to define a collective variable through which the sampling of the catalytic dyad disruption can be biased, to aid the acceleration of computer-aided drug design of protease inhibitors for SARS-CoV-2.

In other cases, the accurate modelling of the active site driving protein function requires a more rigorous description using higher level of theory. In Chapter, the hemocyanin protein core is investigated for its quantum-mechanically-driven function of reversible dioxygen binding. We find that the stabilisation of the open-shell singlet electronic ground state is due to strong electronic correlation effects, where conventional electronic structure methods such as DFT fail to correctly describe the multi-reference quantum mechanics of the hemocyanin active site. Instead, the accurate modelling of the hemocyanin active site is simulated with DFT+DMFT; a hybrid method that treats the copper $3d$-electrons with DMFT and the remaining active site with linear-scaling DFT.

Chapter 2

Theoretical Background

"You must understand, young Hobbit, it takes a long time to say anything in Old Entish. And we never say anything unless it is worth taking a long time to say."

J.R.R. Tolkien

This chapter presents a description of the two computational methods which have been primarily applied throughout the course of this thesis. Certain aspects of the thesis have required the use of other simulation methods, and these methods are outlined in the respective chapters.

2.1 Molecular dynamics

Molecular dynamics (MD) is a simulation method to study the dynamic evolution of a system of particles over time in accordance with their predefined interaction rules. For a macromolecular system, this is a powerful tool that is used to complement experiments to investigate the system to atomic resolution in space and femtosecond resolution in time, to predict structural and thermodynamic properties such as free energies.[44] The structural and thermodynamic behaviour can be compared to experimental data and provide insight to complex processes such as adsorption or ligand binding. An MD engine is a collection of numerical algorithms that defines a set of forces

that act on all particles in a system, and iteratively evolves them along trajectories according to Newton's equations of motion.

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \mathbf{F}_i, \quad i = 1, \ldots, N \tag{2.1}$$

Forces acting on particles $\mathbf{F}_i$ at positions $\mathbf{r}_i$ evolve the system to new positions according to the potential energy function $U(\mathbf{r}_i)$.

$$\mathbf{F}_i = -\frac{\partial U}{\partial \mathbf{r}_i} \tag{2.2}$$

In the case of biomolecular systems, the predefined interaction rules are a potential energy function (forcefield) which are requisite to calculating the forces. A forcefield maps the coordinates of the atoms $\mathbf{r}$ onto the potential energy surface as per the bonded and non-bonded energetic contributions of the macromolecular system components; namely the bonds, angles and dihedrals between atoms (bonded) and van der Waals and electrostatic (non-bonded) interactions.[45] A forcefield is usually expressed as sums of particle interactions and in the case of the *optimised potentials for liquid simulations* (OPLS) forcefield [46, 47] it takes the form expressed in (Eq. 2.3), where (K_b, K_θ, K_l,) are respectively the bond, angle and dihedral force constants for each bonded interaction type and (r_0, θ_0, ϕ_l) are the equilibrium terms. The Lennard-Jones parameters (σ_{ij}, ϵ_{ij}) that describe the interactions between types i and j are determined by taking a geometric average of the self-interaction parameters, i.e. $((\sigma_{ii}\sigma_{jj})^{1/2})$. The final term in (Eq. 2.3) is the Coulombic term that describes the energy between a pair of charges (q_i, q_j).

$$\begin{aligned} U(\mathbf{r}) = &\sum_{bonds} K_b(r - r_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} \sum_{l=1}^{4} \frac{K_l}{2}\left[1 + (-1)^{l+1}\cos(l\phi - \phi_l)\right] \\ &+ \sum_{nonbonded} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i>j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \end{aligned} \tag{2.3}$$

The principle of each bonded and non-bonded forcefield terms are illustrated in (Fig. 2.1). Conventional potential energy functions describe a system with fixed charge for electrostatic interactions, electronic repulsion and dispersive non-bonded interactions through a van der Waals (Lennard-Jones) potential and bond stretching, bending and dihedral angle torsion potentials. More complex formulations of the potential energy function can include polarisation effects in polarisable forcefields,[48] or reactive forcefields that can account for bond order, bond formation and breaking.[49]

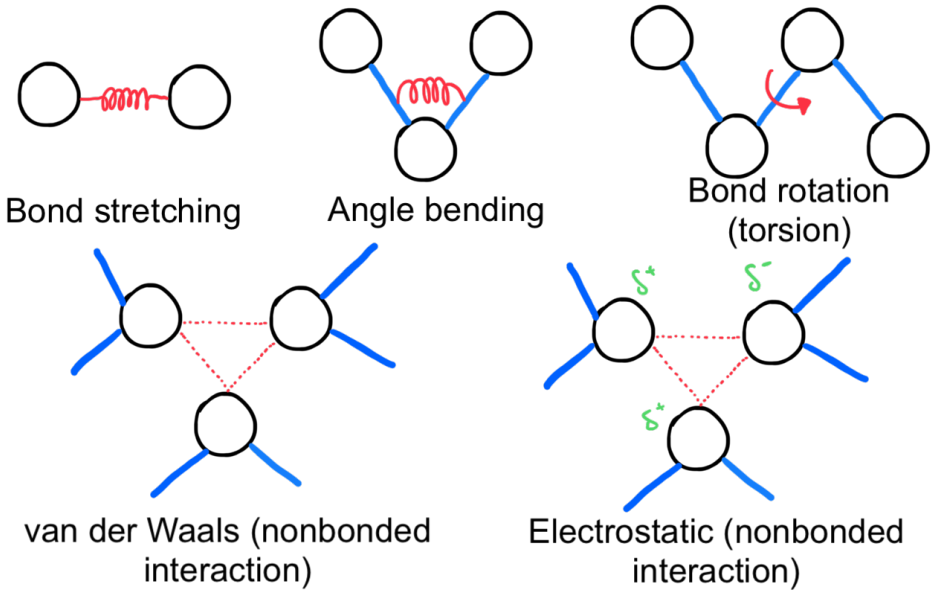


Figure 2.1: The energetic bonded and non-bonded components of a general biomolecular forcefield.

The functional form of a forcefield (Eq. 2.3) is composed of thousands of parameters for different types of atoms, bonds, bond angles, dihedral angles and nonbonded interactions.[17] These parameters are typically derived semi-empirically through *ab initio* quantum calculations on small molecules as well as optimising parameters to reproduce experimental properties of systems such as liquid densities and heats of vaporisation,[50] but retain the ability to accurately extrapolate to macromolecular properties such as the distribution of conformers such as ϕ, ψ angle distributions in proteins.[51] The transferability of forcefields is of course limited to the training and test

sets, using them with an assumption of total coverage is erroneous, therefore special care is required when studying the dynamics of molecules with novel chemistries such as drug structures or exotic nanomaterials.[17] One approach to derive the bespoke forcefield parameters from *ab initio* quantum mechanical calculations is discussed in (2.2.5).

2.1.1 Numerical integrators

To solve the equations of motion (Eq. 2.1) and translate the particle coordinates in the simulation system, one of many possible numerical integration methods is applied to approximate the exact solution to the differential equation. These methods generally start by replacing the derivative with a finite difference approximation of the form (Eq. 2.4).

$$\mathbf{r}'(t) \approx \frac{\mathbf{r}(t+\mathrm{d}t) - \mathbf{r}(t)}{\mathrm{d}t} \tag{2.4}$$

This is rearranged to approximate new particle coordinates according to the force (Eq. 2.2) and potential energy function (Eq. 2.3), with respect to a discretised timestep $\mathrm{d}t$, this is known as the forward Euler algorithm (Eq. 2.5).

$$\mathbf{r}_i(t+\mathrm{d}t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\mathrm{d}t \tag{2.5a}$$

$$\mathbf{v}_i(t+\mathrm{d}t) = \mathbf{v}_i(t) + \frac{\mathbf{F}_i(\{\mathbf{r}_i(t)\})}{m_i}\mathrm{d}t \tag{2.5b}$$

This approach is first order convergent, meaning it has a local truncation error $\epsilon_{t+\mathrm{d}t}$ of order 2; namely the difference between the analytical ($\mathbf{r}(t+\mathrm{d}t)$) and approximated ($\tilde{\mathbf{r}}(t+\mathrm{d}t)$) solutions:

$$\epsilon_{t+\mathrm{d}t} = |\mathbf{r}(t+\mathrm{d}t) - \tilde{\mathbf{r}}(t+\mathrm{d}t)| \tag{2.6}$$

The analytical solution is expressed as a truncated Taylor series expansion:

$$\mathbf{r}(t+\mathrm{d}t) = \mathbf{r}(t) + \mathbf{r}'(t)\mathrm{d}t + \frac{1}{2}\mathbf{r}''(t)\mathrm{d}t^2 \tag{2.7}$$

Such that the truncated error of a single integration step using the forward Euler algorithm is given by:

$$\epsilon_{t+\mathrm{d}t} = |\mathbf{r}(t+\mathrm{d}t) - \tilde{\mathbf{r}}(t+\mathrm{d}t)| \tag{2.8a}$$

$$= \left|\mathbf{r}(t) + \mathbf{r}'(t)\mathrm{d}t + \frac{1}{2}\mathbf{r}''(t)\mathrm{d}t^2 - [\mathbf{r}(t) + \mathbf{v}(t)\mathrm{d}t]\right| \tag{2.8b}$$

$$= \left|\frac{1}{2}\mathbf{r}''(t)\mathrm{d}t^2\right| = \mathcal{O}(\mathrm{d}t^2) \tag{2.8c}$$

As the integrator evolves the system coordinates iteratively, this error accumulates with respect to the size of the time step $\mathrm{d}t$, a scenario where the numerical solution converges to the analytical solution is therefore only possible in the limit of $\mathrm{d}t \to 0$, which is an impractical obstacle that would stand in the way of simulating meaningful (long enough) trajectories. The accumulation of errors results in the gradual change of the total energy of a closed system over time, an artefact known as energy drift. To accurately model the dynamics of a system at equilibrium, integrators are required to be symplectic, have time reversible symmetry and conserve linear momentum, angular momentum and energy. The forward Euler algorithm breaks both time-reversible symmetry i.e. whether the model is invariant to the substitution $\mathrm{d}t \to -\mathrm{d}t$, and symplectic condition i.e. that phase space volume is conserved.

The leapfrog algorithm overcomes these limitations by performing a half velocity update step, followed by a full position update. New positions are used as initial conditions for computing a new set of atomic forces, combined with the final velocity half step this gives the velocity at the full step (Eq. 2.9).

$$\mathbf{v}_i\left(t + \frac{\mathrm{d}t}{2}\right) = \mathbf{v}_i(t) + \frac{\mathbf{F}_i(\{\mathbf{r}_i(t)\})}{m_i}\frac{\mathrm{d}t}{2} \tag{2.9a}$$

$$\mathbf{r}_i(t+\mathrm{d}t) = \mathbf{r}_i(t) + \mathbf{v}_i\left(t + \frac{\mathrm{d}t}{2}\right)\mathrm{d}t \tag{2.9b}$$

$$\mathbf{v}_i(t+\mathrm{d}t) = \mathbf{v}_i\left(t + \frac{\mathrm{d}t}{2}\right) + \frac{\mathbf{F}_i(\{\mathbf{r}_i(t+\mathrm{d}t)\})}{m_i}\frac{\mathrm{d}t}{2} \tag{2.9c}$$

The leapfrog algorithm is second order convergent and therefore has a local truncation error of order 3, following the same approach in (Eq. 2.8):

$$\epsilon_{t+\mathrm{d}t} = |\mathbf{r}(t+\mathrm{d}t) - \tilde{\mathbf{r}}(t+\mathrm{d}t)| \tag{2.10a}$$

$$= \left|\frac{1}{6}\mathbf{r}'''(t)\mathrm{d}t^3\right| = \mathcal{O}(\mathrm{d}t^3) \tag{2.10b}$$

The leapfrog algorithm (Eq. 2.9) meets the conditions outlined above and does not give rise to numerical artefacts for long simulation times as with the forward Euler algorithm.

2.1.2 Initial configuration

The investigation of a system's dynamics using an MD engine — employing the algorithms covered in this chapter — requires the user to identify molecular constituent atom types, atomic coordinates, the topology of the molecules of interest and their velocities. The input system structure is recognised by the employed forcefield to assign parameters in accordance with the functional form of the forcefield, generalised by (Eq. 2.3).

The user defined input coordinates of the system are paramount to correctly modelling the physical body under investigation; whether a protein, ligands or nanomaterials, it can be experimentally derived or inferred but requires pre-processing to ensure no experimental or modelling errors are present. Protein structures are available from X-Ray Diffraction (XRD) experiments, in the form of crystallographic 3-dimensional coordinate data hosted on the protein data bank (PDB) archive.[52] A common experimental artefact from PDB structures is missing residues in the amino acid sequence, these are remedied using homology modelling. MODELLER is such a program, it performs comparative protein structure modelling by satisfaction of spatial restraints, filling in missing residues by aligning the input protein amino acid sequence to known related structures, producing 3-dimensional protein structures.[53, 54]

In the case of small molecules such as ligands, widely available soft-

ware packages can be utilised to define the input structure of molecules by hand using an interactive interface such as Avogadro.[55] Meanwhile, large non-organic materials can be generated by one of the many functionalities of the more complex initial configuration generators such as CHARMM-GUI; a web-hosted graphical user interface with an extensive set of tools ranging from nanomaterial, polymer and lipid membrane builders with multi-forcefield and multi-MD engine support.[56, 57] Another example of a versatile initial configuration builder is Packmol, it is a script-based and locally hosted software package that is widely used for generating complex systems.[58] Packmol affords user flexibility in ways interactive software may not, such as building the initial configuration of systems composed of different constituent molecules.

For entirely bespoke systems the molecular configuration is inferred through experimental techniques that do not accurately translate to 3-dimensional coordinates as is the case with proteins from the PDB. In particular, systems with exotic chemistries are not always readily prepared using the aforementioned initial configuration software, beside standardised protocol that may not reflect the accurate theoretical/phenomenological structure of the system in question. Such systems would require accuracy both in the distribution of atoms within the macromolecule as well as assigning the correct forcefield parameters. For example, some of the work in this thesis required the accurate characterisation of graphitic nanomaterials that required custom software to develop the macromolecular structures, topological properties and their corresponding forcefield parameters.[59] Custom software allows for flexibility in varying nanomaterial structural properties such as shape, size, chemical functionalisation and degree of functional group correlation. Software engineering custom software is an incredibly time consuming project, which is being accelerated by well executed and documented open source software initiatives.

2.1.3 Periodic boundary conditions

Periodic boundary conditions (PBC) are imposed on most MD simulation systems to approximate large systems that reside in a practically infinite environment, by construction. Simply put, particles leaving the unit cell in any of the $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ directions reenter the unit cell from the opposing side in the respective dimension, retaining identical mechanical properties such as particle momenta. Particles in the unit cell interact with the particles in the closest neighbouring unit cell, restricted by the interaction cut-off radius as input by the user. As is commonly the case and has been explained, the initial configuration can raise simulation artefacts and PBCs are no exception. A macromolecule such as a protein can self-interact through the boundary conditions, resulting in distorted dynamics if and when the simulation system is too small and the macromolecule isn't enveloped by sufficient explicit solvent molecules in all of the $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ directions. The recommended protocol is that the simulation system should exceed twice the cut-off radius and have at least 1 nm of explicit solvent surrounding the macromolecule.[60] Charge interactions through PBCs could raise electrostatic artefacts, where a system without a neutral charge could sum to infinite charge through the periodic cells. Instead counter-ions are added to bring the simulation system to neutral charge, usually this sodium and chlorine ions for positive and negative charges, respectively.

2.1.4 Equilibration

The software used to create the initial configuration does not typically consider the energetics of the system, it is instead more suited to successfully geometrically organise a system's 3-dimensional molecular coordinates. As a result, these configurations may — according to the initial configuration method — exist in a high energy state, driven by anything from a high degree of atomic overlap to energetically unfavourable molecular conformations. Studying the system at equilibrium therefore requires a sequential treatment to bring the system's potential energy to a local minimum,

followed by simulating trajectories to bring about thermal and barostatic equilibrium. These are known as the equilibration steps that precede the so called production-run which is used to gather the statistics for the desired system.

2.1.4.1 Energy minimisation

To overcome the high energy of the initial configuration of the system, the MD engine employs the chosen forcefield to tweak the system coordinates until the energy corresponds to a local minimum in the potential energy surface through energy minimisation, instead of employing MD algorithms. A common approach to locate the local minimum of a multivariate and highly dimensional function that cannot be solved analytically is the gradient descent algorithm. A parameter (coordinate) vector ($\mathbf{r}$) is iteratively updated in the negative direction to the first order local gradient of the function ($\nabla V(\mathbf{r})$) being minimised in accordance with (Eq. 2.11) until reaching convergence when the gradient is zero. The gradient is scaled by the step size (η), also known as the learning rate, the choice of which can impact whether the algorithm reaches convergence. Too small a step size could significantly retard the convergence, whereas a large step size could overshoot the local minimum parameter vector.

$$\mathbf{r}_{n+1} = \mathbf{r}_n - \eta_n \nabla V(\mathbf{r}_n) \tag{2.11}$$

Tweaking the step size is therefore paramount to effectively or efficiently locate a local minimum; one such approach is a particular case of gradient descent known as the steepest descent algorithm, where the step size (η) is iteratively chosen to maximise the direction of the negative gradient of the function by minimising the objective function (Eq. 2.12).[61]

$$\eta_n = \arg\min_{\eta} V(\mathbf{r}_n - \eta \nabla V(\mathbf{r}_n)) \tag{2.12}$$

2.1.4.2 Particle velocities

To model the system in equilibrium, the system potential energy should oscillate close to its local minimum due to fluctuations in kinetic energy. Ensuring these oscillations in potential energy are stable and do not diverge due to large fluctuations, initial velocities are assigned to every particle in the system. There exist statistical constraints on the choice of particle velocities to ensure the system is in thermal equilibrium, namely that it abides by the equipartition theorem. To abide by this, all particles are randomly assigned a velocity from the Maxwell-Boltzmann distribution given by (Eq. 2.13), where (v) is the velocity magnitude ($\sqrt{v_x^2 + v_y^2 + v_z^2}$) of the velocity vector ($\mathbf{v} = (v_x, v_y, v_z)$) and ($m$), ($T$) and ($k_B$) are the mass, desired temperature and the Boltzmann constant, respectively.

$$f(v) = \frac{4}{\sqrt{\pi}} \left(\frac{m}{2k_B T} \right)^{3/2} v^2 \exp\left(-\frac{mv^2}{2k_B T} \right) \tag{2.13}$$

Using this, the average square velocity is given by the integral:

$$\langle v^2 \rangle = \int_0^\infty v^2 f(v)\, \mathrm{d}v = \frac{3k_B T}{m} \tag{2.14}$$

2.1.5 Thermostats

Performing iterative integration of Newton's equations of motion (Eq. 2.2) in this way generates a microcanonical ensemble where the simulation system's number of atoms (N), volume (V) and total energy (E) are kept constant. With conserved energy, the system is in a steady state that does not evolve over time, despite all its constituent parts being in motion. The microcanonical (NVE) ensemble assigns equal probability to all the system microstates within this energy E, all microstates outside this range have zero probability and are never sampled in an equilibrium configuration. The alternative approach is to model the system in a canonical (NVT) ensemble, where the system no longer has conserved energy but the average temper-

ature (T) is kept constant. By allowing the simulation system to exchange energy with its environment — implemented through coupling to a heat reservoir — it is able to explore a phase space that includes microstates with different energies.[62] In order to implement the canonical ensemble in MD simulations, in which temperature is conserved instead of energy, the integration of Newton's equations of motion have to be coupled to a thermostat.

A thermostat is an algorithm that introduces a fictitious dynamical variable, which slows down or accelerates particles until the temperature is equal to the desired value through coupling to an external heat bath. One of the simplest implementations of a thermostat is the Berendsen thermostat which rescales the velocities of all particles at each integrator timestep with a nonphysical force (Eq. 2.15) with a scale factor (λ) to control the total mean kinetic energy.[63]

$$\lambda(t) = \left[1 + \frac{\mathrm{d}t}{\tau_T}\left(\frac{T_0}{T(t - \frac{1}{2}\mathrm{d}t) - 1}\right)\right]^{1/2} \tag{2.15}$$

τ is a time constant related to the temperature coupling time constant τ_T by $\tau = 2C_V\tau_T / N_{df}k_B$, where C_V is the total heat capacity of the system, N_{df} the numer of degrees of freedom in the system and k_B is the Boltzmann constant. Deviations of the simulation system temperature T_0 are corrected to the target temperature T according to:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{T_0 - T}{\tau} \tag{2.16}$$

The system kinetic energy (K) is scaled with λ (Eq. 2.15) at each integrator step through:

$$\Delta K = (\lambda - 1)^2 K \tag{2.17}$$

By design (see 2.1.4) the particle velocities are distributed according to the Maxwell-Boltzmann function. Scaling the kinetic energy with the Berendsen thermostat (Eq. 2.17) suppresses the fluctuations in kinetic en-

ergy and particle velocities no longer adopt a Maxwell-Boltzmann distribution. This means that the Berendsen thermostat does not generate a canonical ensemble and will give rise to artefacts and incorrect sampling of the system's phase space. A particular manifestation within a system is that the total kinetic energy is not shared equally among all kinetic degrees of freedom — hence violating the theory of equipartition — where the kinetic energy instead flows from higher energy degrees of freedom to translational and rotational degrees of freedom. This results in a part of the system freezing into a single conformation and travelling through the simulation box with high momentum, this is known as the flying ice cube effect.[64]

To remedy this artefact, a stochastic term ($\mathrm{d}W$) in the form of a Wiener process is added to the Berendsen thermostat (Eq. 2.15) that ensures the correct kinetic energy distribution and that the resulting dynamics produce a correct canonical ensemble.

$$\mathrm{d}K = (K_0 - K)\frac{\mathrm{d}t}{\tau_T} + 2\sqrt{\frac{KK_0}{N_{df}}}\frac{\mathrm{d}W}{\sqrt{\tau_T}} \tag{2.18}$$

This is known as the Bussi–Donadio–Parrinello thermostat (Eq. 2.18).[65]

2.1.6 Barostats

In a similar fashion, an isothermal-isobaric (NPT) ensemble can be generated by a rigorous barostat algorithm by coupling the system to a pressure bath. One approach that successfully generates an isothermal-isobaric ensemble is the Parinello-Rahman barostat.[66] This approach is implemented through coupling the imbalance between the simulation system pressure and the external (bath) tensorial pressure field to changes in the shape and size of the simulation box. The simulation box vectors are represented by the matrix ($\mathbf{b}$) and obey the equation of motion in (Eq. 2.19), defined with respect to the simulation box volume (V), coupling strength matrix ($\mathbf{W}$),

system pressure ($\mathbf{P}$) and reference pressure ($\mathbf{P}_{ref}$).

$$\frac{\mathrm{d}\mathbf{b}^2}{\mathrm{d}t} = V\mathbf{W}^{-1}\mathbf{b}'^{-1}(\mathbf{P} - \mathbf{P}_{ref}) \tag{2.19}$$

The system pressure tensor ($\mathbf{P}$) is calculated from the difference in kinetic energy ($\mathbf{K}$) and the virial tensors (Eq. 2.20), where ($\otimes$) denotes the tensor product.

$$\mathbf{P} = \frac{2}{V}(\mathbf{K} - \Xi) \tag{2.20a}$$

$$\mathbf{K} = \frac{1}{2}\sum_i m_i \mathbf{v}_i \otimes \mathbf{v}_i \tag{2.20b}$$

$$\Xi = -\frac{1}{2}\sum_{i<j} \mathbf{r}_{ij} \otimes \mathbf{F}_{ij} \tag{2.20c}$$

The coupling strength matrix ($\mathbf{W}$) is defined with respect to the isothermal compressibility (β_{ij}), pressure time constant (τ_P) and the largest box matrix element (L) as $(\mathbf{W}^{-1})_{ij} = 4\pi^2\beta_{ij}/3\tau_P^2 L$. Since the implementation of pressure coupling requires a change in both simulation box vector and particle equations of motion, a modification is applied to the system Hamiltonian, where the modified Hamiltonian ($\mathcal{H}_{PR}$) is given by (Eq. 2.21), where (U) and (K) denote the potential and kinetic energies, respectively.

$$\mathcal{H}_{PR} = U + K + \sum_i P_{ii} V + \sum_{ij} \frac{1}{2} W_{ij} \left(\frac{\mathrm{d}\, b_{ij}}{\mathrm{d}t}\right)^2 \tag{2.21}$$

The equations of motion of the system's particles are given by (Eq. 2.22), where the added term takes the form of a friction term. It is however fictitious and solely there to define the particle equations of motion with respect to the simulation box vectors through the term.

$$\frac{\mathrm{d}^2\mathbf{r}_i}{\mathrm{d}t^2} = \frac{\mathbf{F}_i}{m_i} - \Lambda\frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} \tag{2.22a}$$

$$\Lambda = \mathbf{b}^{-1}\left[\mathbf{b}\frac{\mathrm{d}\mathbf{b}'}{\mathrm{d}t} + \frac{\mathrm{d}\mathbf{b}}{\mathrm{d}t}\mathbf{b}'\right]\mathbf{b}'^{-1} \tag{2.22b}$$

2.1.7 Parallelisation

Simulating the dynamics of biomolecular systems requires a large system size to answer questions of meaningful scientific interest. The computational cost of running MD algorithms scales with the number of particles in the simulation system, therefore the implementation of MD algorithms in parallelised standards is necessary. Accelerating MD simulations is achieved by both parallelised standards where the software code is written to specify how tasks are divided between computational resources, and by implementing methods that reduce the complexity of particularly costly calculations involved in MD algorithms, e.g. evaluating a list of distances between particles and calculating long-range interactions.

2.1.7.1 Domain decomposition

The naive approach of calculating the distances between particle coordinates in the simulation system is through a brute-force algorithm, where the distances are computed between all $(N(N-1)/2)$ point pairs and the smallest distance pair is picked. The list of distances is used to evaluate the interaction type and its magnitude, thus its contribution to the total system energy at that point in the simulation. Performing this by brute force requires $\mathcal{O}(N^2)$ time. An alternative approach to calculating particle pair-wise distances and their interactions is known as domain decomposition, in which a data structure is implemented where distances are evaluated only within a predefined cut-off distance, therefore reducing the $\mathcal{O}(N^2)$ time complexity of the closest pair of points problem. The cut-off distance is generally the distance within which short-range non-bonded interactions are present.

This data structure works by decomposing the 3-dimensional simulation domain into cells with a edge length equal to or greater than the interaction cut off distance, where pairwise distances are now computed for particles between neighbouring cells instead of the entire simulation domain. For a 3-dimensional simulation domain, a 3-dimensional cell has 26 neighbouring cells for which pairwise distances need to be calculated. The computational time complexity of this operation is reduced from $\mathcal{O}(N^2)$ to $\mathcal{O}(N\bar{c}) \in \mathcal{O}(N)$, where $\bar{c}$ denotes the average number of particles per cell (N/m), where (m) is the number of cells decomposing the domain.

2.1.7.2 Particle-mesh-Ewald summation

Having reduced the cost of evaluating particle pairwise distances for short-range interactions does not solve the problem of computing pairwise distances to calculate long range non-bonded interactions, such as electrostatics and the Lennard-Jones dispersion $(1/r^6)$ term in (Eq. 2.3). Ewald summation [67] was introduced for the treatment of such long-range interactions, however it scales as $(\mathcal{O}(N^2))$ in time which is impractical for the treatment of large simulation systems. However it does form the foundations of methods such as particle-mesh-Ewald (PME) [68] which scales as $(\mathcal{O}(N \log N))$ in time.

Within Ewald summation formalism, the total electrostatic energy for N particles for a unit cell — defined with respect to periodic images at positions $(\mathbf{r}_j + \mathbf{n})$ — is represented by (Eq. 2.23), where $(\mathbf{n} = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3)$ are the box index vectors and the $(\sum_n^*)$ sum indicates that terms with $(i = j)$ and $(\mathbf{n} = 0)$ should be omitted.

$$E(\mathbf{r}) = \frac{1}{2}\sum_{n}^{*}\sum_{i}\sum_{j}\frac{q_i q_j}{\left|\mathbf{r}_j - \mathbf{r}_i + \mathbf{n}\right|} \tag{2.23}$$

The expression is a slowly converging sum that can be converted to quickly converging expressions for short-range (E_{sr}) and long-range (E_{lr}) electrostatic interactions and a self-interaction correction term (E_{corr}); long-range

interactions are computed in reciprocal space through the use of Fourier transforms. [67, 69]

$$E = E_{sr} + E_{lr} + E_{corr} \tag{2.24a}$$

$$E_{sr} = \frac{1}{2}\sum_{n}^{*}\sum_{i,j}^{N} q_i q_j \frac{\operatorname{erfc}\left(\beta\left|\mathbf{r}_j - \mathbf{r}_i + \mathbf{n}\right|\right)}{\left|\mathbf{r}_j - \mathbf{r}_i + \mathbf{n}\right|} \tag{2.24b}$$

$$E_{lr} = \frac{1}{2\pi V}\sum_{\mathbf{m}\neq 0} \frac{\exp\left(-\pi^2\mathbf{m}^2/\beta^2\right)}{\mathbf{m}^2} S(\mathbf{m})S(-\mathbf{m}) \tag{2.24c}$$

$$E_{corr} = -\frac{1}{2}\sum_{(i,j)\in M} \frac{q_i q_j \operatorname{erf}\left(\beta\left|\mathbf{r}_i - \mathbf{r}_j\right|\right)}{\left|\mathbf{r}_i - \mathbf{r}_j\right|} - \frac{\beta}{\sqrt{\pi}}\sum_{i}^{N} q_i^2 \tag{2.24d}$$

In (Eq. 2.24c), the volume of the unit cell (V) is given by ($\mathbf{a}_1 \cdot \mathbf{a}_2 \times \mathbf{a}_3$), ($\mathbf{m}$) denotes the reciprocal lattice space vectors ($\mathbf{m} = m_1\mathbf{a}_1^* + m_2\mathbf{a}_2^* + m_3\mathbf{a}_3^*$) and ($S(\mathbf{m}) = \sum_j^N q_j \exp\left(2\pi i\,\mathbf{m}\cdot\mathbf{r}\right)$) is the structure factor. The error function ($\operatorname{erf}(x)$) in (Eq. 2.24d) is related to the complementary error function in (Eq. 2.24b) by ($1 - \operatorname{erfc}(x)$). The ($\beta$) parameter in (Eq. 2.24) determines the relative weight of the direct and reciprocal sums and can be used to control the rates of convergence of the respective terms. Algorithmic optimisation of the (β) parameter can reduce the time complexity of Ewald summation from ($\mathcal{O}(N^2)$) to ($\mathcal{O}(N^{3/2})$).[70, 68]

The PME method also separates the short- and long-range interactions, similarly to the Ewald method in (Eq. 2.24), where short-range and long-range interactions are evaluated in real and reciprocal space, respectively. The real space (short-range) sum has a time complexity of ($\mathcal{O}(N)$) by virtue of the pairwise minimum distance computation from domain decomposition, leaving the reciprocal space sum (Eq. 2.24c) as the bottleneck. The treatment of the long-range non-bonded interactions is accelerated to ($\mathcal{O}(N\log N)$) time complexity by assigning the charges (q) to a discrete grid (mesh) using cardinal B-spline interpolation, [69] resulting in a charge density field. The charge density field is defined on a discrete lattice in real space and is Fourier transformed to reciprocal space, enabling the long-

range interaction term to be evaluated using a single sum over the grid in reciprocal ($\mathbf{k}$) space (Eq. 2.25), where ($\tilde{\Phi}_{lr}(\mathbf{k})$) denotes the Fourier transformed potential from Ewald summation — implicit in (Eq. 2.24c) — and $\tilde{\rho}(\mathbf{k})$ the Fourier transformed charge density field.

$$E_{lr} = \sum_{\mathbf{k}} \tilde{\Phi}_{lr}(\mathbf{k}) \left|\tilde{\rho}(\mathbf{k})\right|^2 \tag{2.25}$$

2.2 Density Functional Theory

So far the discussion of molecular dynamics has considered the energetics and the calculation of forces on individual atoms in a macromolecular system, using Newton's laws of motion in classical mechanics (Eq. 2.1). The computation of energies and forces in classical MD is hinged on the validity of a forcefield that is derived from both *ab initio* and phenomenological approaches. *Ab initio* quantum mechanical approaches such as density functional theory (DFT) calculate the electronic ground state of a molecular system using the Schrödinger equation (Eq. 2.26) as a basis for its framework, which involves the treatment of both electrons and nuclei. Within the Schrödinger equation, the Hamiltonian operator ($\mathcal{H}$) acts on the quantum mechanical system; formulated in terms of the many electron wave function (Ψ).[71]

$$\hat{\mathcal{H}} \left|\Psi\right\rangle = E \left|\Psi\right\rangle \tag{2.26}$$

Solving the Schrödinger equation for a many electron system is an intractable and impossible problem to solve analytically, due to the correlated motion of electrons described by the many electron wave function (Ψ), containing $3N$ degrees of freedom, where N is the number of electrons in question. DFT reformulates the Schrödinger equation from a single N-body problem to N single-body problems.

Instead of the wave function, DFT solves the electronic Hamiltonian (Eq. 2.27) by considering the density of the electrons in a system as a fundamental variable. It is expressed with respect to the kinetic energy operators

for each electron ($\hat{T}_e$) and nucleus ($\hat{T}_e$) in the system, as well as the potential energy operators between the electrons and nuclei ($\hat{V}_{ne}$), electron-electron repulsion ($\hat{V}_{ee}$) and nuclei-nuclei repulsion ($\hat{V}_{nn}$).

$$\begin{aligned}\hat{\mathcal{H}} &= \hat{T}_e + \hat{V}_{ne} + \hat{V}_{ee} + \hat{T}_n + \hat{V}_{nn} \\ &= -\frac{\hbar}{2m}\sum_{i}^{N}\nabla_i - e^2\sum_{I=1}^{N}\sum_{i=1}^{N}\frac{Z_I}{|\mathbf{R_I}-\mathbf{r_i}|} + \frac{e^2}{2}\sum_{i\neq j}^{N}\sum_{j\neq i}^{N}\frac{1}{|\mathbf{r_i}-\mathbf{r_j}|} - \frac{\hbar^2}{2}\sum_{I=1}^{P}\frac{\nabla_i^2}{M_I} \\ &\quad + \frac{e^2}{2}\sum_{I=1}^{P}\sum_{J\neq I}^{P}\frac{Z_I Z_J}{|\mathbf{R_I}-\mathbf{R_J}|}\end{aligned} \tag{2.27}$$

The reformulation from wave functions to the density of electrons is based on Hohenberg-Kohn theorem and the Kohn-Sham equations. Fermi and Thomas first proposed this idea and derived a differential equation for the density,[72, 73] therefore negating the need to work in terms of the one-electron orbital. It later required the work of Hohenberg and Kohn to formulate the theory on a strong mathematical basis,[74] giving rise to the Hohenberg-Kohn theorem. The theory was further developed with the Kohn-Sham framework,[75] where the problem of interacting electrons in a static external potential is reduced to non-interacting electrons moving in an effective potential. The Born-Oppenheimer approximation makes solving the Schrödinger equation easier by imposing the separation of nuclear and electronic motion to the wave function ($\Psi(\mathbf{r},\mathbf{R})$).[76] This is justified by the difference in momentum of the nuclei compared to electrons, where the much heavier nuclei must therefore have minuscule velocities.

$$\Psi(\mathbf{r},\mathbf{R}) = \psi(\mathbf{r},\mathbf{R})\phi(\mathbf{R}) \tag{2.28}$$

2.2.1 Hohenberg-Kohn Theorems

The Hohenberg-Kohn theorems are the basis for computing the ground state properties of a system without directly dealing with the highly dimensional many-electron wave function. It instead reformulates the problem in terms

of the electronic density and treats the system of electrons moving under the influence of an external potential, which encompasses all effects external to the electrons themselves such as Coulombic effects from atomic nuclei. As such, and due to the separation of nuclear and electronic motion via the Born-Oppenheimer approximation, the repulsive Coulombic potential between nuclei can be treated separately as an external potential:

$$V_{\text{ext}}(\mathbf{r}) = -\sum_{I} \frac{Z_I}{|\mathbf{r}_I - \mathbf{r}_i|} \tag{2.29}$$

Therefore the entire electronic Hamiltonian (Eq. 2.27) can be expressed as ($\hat{H} = \hat{F} + \hat{V}_{\text{ext}}$), where ($\hat{F}$) is given by (Eq. 2.30) and the external potential operator is given by the sum of external potentials ($\hat{V}_{\text{ext}} = \sum_i V_{\text{ext}}(\mathbf{r}_i)$).

$$\hat{F} = -\frac{1}{2}\sum_{i} \nabla_i^2 + \frac{1}{2}\sum_{i}\sum_{j\neq i} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} = \hat{T}_e + \hat{V}_{ee} \tag{2.30}$$

The expression for ($\hat{F}$) is universal to all systems consisting of N electrons, meaning that the external potential (V_{ext}) and the number of electrons in a given system (N) completely control ($\hat{H}$) and the ground state wave function (Ψ_0). The ground state wave function and the Hamiltonian give rise to the ground state density ($n(\mathbf{r})$:

$$n(\mathbf{r}) = \int \prod_{i=2}^{N} |\Psi_0(\mathbf{r}, \mathbf{r_2}, \mathbf{r_3}, \ldots, \mathbf{r_N})|^2 \mathrm{d}\mathbf{r}_i = \langle \Psi_0 | \hat{n} | \Psi_0 \rangle \tag{2.31}$$

Given this, the electronic density ($n(\mathbf{r})$) and the ground state wave function (Ψ_0) are functionals (a function of a function) of the number of electrons (N) and the external potential (V_{ext}).

The first Hohenberg-Kohn theorem states that the external potential ($V_{\text{ext}}(\mathbf{r})$) is a unique functional of the ground state electronic density ($n(\mathbf{r})$), therefore the total energy is also a functional of the ground state electronic energy. Indeed, it turns out a consequence of this theorem is that all system properties can be determined just by the electronic ground state density.[74]

The proof for this theorem is a simple *reductio ad absurdum* argument. Consider a many-electron Hamiltonian of the form ($H = T + V$) and a different Hamiltonian ($H' = T + V'$), with wave functions (Ψ) and (Ψ') respectively and where ($V - V' \neq$ constant). The electron density is defined by (Eq. 2.31) and we assume that the electron density is equal for (V) and (V'). According to the Rayleigh-Ritz variational theorem, the ground state energy satisfies the inequality (Eq. 2.32).

$$E_0 < \langle \Psi' | H | \Psi' \rangle = \langle \Psi' | H' | \Psi' \rangle + \langle \Psi' | H - H' | \Psi' \rangle \tag{2.32a}$$

$$= E_0' + \int n(\mathbf{r})(V(\mathbf{r}) - V'(\mathbf{r}))d\mathbf{r} \tag{2.32b}$$

The inequality refers to different Hamiltonians, therefore the eigenstates (Ψ) and (Ψ') are different. By swapping the primed and unprimed quantities in (Eq. 2.32) we get the expression (Eq. 2.33).

$$E_0' < \langle \Psi | H' | \Psi \rangle = \langle \Psi | H | \Psi \rangle + \langle \Psi' | H' - H | \Psi' \rangle \tag{2.33a}$$

$$= E_0 + \int n(\mathbf{r})(V'(\mathbf{r}) - V(\mathbf{r}))d\mathbf{r} \tag{2.33b}$$

Summing (Eq. 2.32) and (Eq. 2.33) together results in an absurd statement, namely that ($E_0 + E_0' < E_0' + E_0$), consequently two potentials (V) and (V') can correspond to the same electron density, which is proven not to be the case. Hence, ($V'(\mathbf{r}) \neq V(\mathbf{r})$) for a given ($n(\mathbf{r})$).

The second Hohenberg-Kohn theorem is that the total energy of a system — a functional of the ground state electronic density as by the first Hohenberg-Kohn theorem — is minimised to arrive at the correct ground state energy. Regardless of the external potential of a system (V_{ext}), the functional to be minimised is universal for any system and is given by:

$$E[n(\mathbf{r})] = F[n(\mathbf{r})] + \int n(\mathbf{r}) V_{\text{ext}}(\mathbf{r}) \mathrm{d}\mathbf{r} \tag{2.34}$$

where, as outlined in (Eq. 2.30), the electronic energies are contained within

the universal functional ($F[n(\mathbf{r})]$). The total energy functional (Eq. 2.34) is equivalent to the ground state energy, given by the expectation value with respect to the ground state wave function:

$$E = E[n] = \langle \Psi_0 | \hat{H} | \Psi_0 \rangle \tag{2.35}$$

To show this, consider an electronic density ($n'(\mathbf{r})$) different to the ground state electronic density ($n(\mathbf{r})$) has a different wave function (Ψ'). The energy corresponding to this state (E') has a higher energy than the ground state energy (E), according to the variational theorem as seen in (Eq. 2.32):

$$E = \langle \Psi_0 | \hat{H} | \Psi_0 \rangle > \langle \Psi' | \hat{H} | \Psi' \rangle = E' \tag{2.36}$$

Given this, the total energy functional (Eq. 2.34) gives the exact ground state energy (E) only for the exact ground state electronic density ($n(\mathbf{r})$). However, the difficulty is that ($F[n(\mathbf{r})]$) remains unknown and a requirement for constructing a scheme to calculate the ground state electronic density is addressed by the Kohn-Sham equations.

2.2.2 Kohn-Sham equations

The Kohn-Sham (KS) formulation extended the Hohenberg-Kohn theorems and allowed for the practical implementation of DFT.[75] The non-interacting ground-state electron density ($n(\mathbf{r})$) is represented as a sum of one electron orbitals, known as Kohn-Sham orbitals ($\psi_i(\mathbf{r})$). For doubly occupied orbitals, the electron density is given by:

$$n(\mathbf{r}) = 2 \sum_{i}^{N/2} |\psi_i(\mathbf{r})|^2 \tag{2.37}$$

The Kohn-Sham orbitals are solutions to the Schrödinger equation (Eq. 2.26).

$$\left(-\frac{\hbar^2}{2m_e} \nabla^2 + V_{\text{eff}}(\mathbf{r}) \right) \psi_i(\mathbf{r}) = \epsilon_i \psi_i(\mathbf{r}) \tag{2.38}$$

Where (m_e) is the mass of an electron, (V_{eff}) is the effective (fictitious) potential in which the non-interacting electrons move and (ϵ_i) is the orbital energy of the corresponding Kohn–Sham orbital ($\psi_i(\mathbf{r})$). Given that the KS equations reformulate the problem from interacting electrons in a static external potential to non-interacting electrons moving in an effective potential — by ($V_{\text{eff}}(\mathbf{r}) = V_{\text{ext}}(\mathbf{r}) + V_{\text{xc}}(\mathbf{r})$), where ($V_{\text{xc}}(\mathbf{r})$) is the exchange-correlation potential — the energy functional is given the form:

$$E[n(\mathbf{r})] = T_s[n(\mathbf{r})] + E_H[n(\mathbf{r})] + E_{xc}[n(\mathbf{r})] + \int n(\mathbf{r})V(\mathbf{r})d\mathbf{r} \tag{2.39}$$

where the exchange-correlation energy, $E_{xc}[n(\mathbf{r})]$, is an unknown. The kinetic energy of non-interacting electrons, $T_s[n(\mathbf{r})]$, is given by

$$T_s[n] = -\frac{\hbar^2}{2m_e} 2 \sum_i \langle \psi_i | \nabla^2 | \psi_i \rangle \tag{2.40}$$

The Hartree energy term, $E_H[n(\mathbf{r})]$, is defined as the electrostatic interactions between clouds of charge

$$E_H[n(\mathbf{r})] = \frac{e^2}{2} \int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' \tag{2.41}$$

The KS method builds upon the Hohenberg-Kohn theorem by minimising the functional ($E[n(\mathbf{r})]$) through varying ($n(\mathbf{r})$), such that:

$$\begin{aligned} & \frac{\delta}{\delta n(\mathbf{r})} \left[E[n(\mathbf{r})] - \mu \int n(\mathbf{r}) d\mathbf{r} \right] = 0 \\ \rightarrow \quad & \frac{\delta E[n(\mathbf{r})]}{\delta n(\mathbf{r})} = \mu \end{aligned} \tag{2.42}$$

where (μ) is a Lagrange multiplier. Taking a functional derivative of (Eq. 2.39) minimises the ground state density to give (Eq. 2.43), where the unknown exchange-correlation potential remains defined as a functional

derivative of the exchange-correlation energy.

$$\frac{\delta T_s[n(\mathbf{r})]}{\delta n(\mathbf{r})} + V_{ext}(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' + \frac{\delta E_{xc}[n(\mathbf{r})]}{\delta n(\mathbf{r})} = \mu \tag{2.43}$$

Where (V_{ext}) is the static external potential. The effective potential in which the non-interacting electrons move, $V_{\text{eff}}(\mathbf{r})$, is given by rewriting (Eq. 2.43) as:

$$\frac{\delta T_s[n(\mathbf{r})]}{\delta n(\mathbf{r})} + V_{\text{eff}}(\mathbf{r}) = \mu \tag{2.44}$$

To finally find the ground-state energy (E_0) and density ($n_0(\mathbf{r})$), the Schrödinger equation in (Eq. 2.38) is solved for one electron:

$$\left(-\frac{1}{2}\nabla_i^2 + V_{\text{eff}}(\mathbf{r}) - \epsilon_i\right)\psi_i(\mathbf{r}) = 0 \tag{2.45}$$

2.2.3 Exchange-correlation energy

Explicit functionals of the electron density in the Kohn-Sham equations are the contributions given in terms of the electron density, such as the Hartree term and the interactions between electrons and an effective potential. Other terms such as the kinetic energy of non-interacting electrons (Eq. 2.40) and the exchange energy are known as functionals of the non-interacting orbitals, which are unknown functionals of the density.

Until this point, no approximation has yet been applied in the Kohn-Sham equations. The aforementioned kinetic energy of non-interacting electrons (Eq. 2.40) is not the true kinetic energy, simply because the electrons in a *real* system do interact, the corrections to this are included in the exchange-correlation energy as

$$E_{xc}[n(\mathbf{r})] = T[n(\mathbf{r})] - T_s[n(\mathbf{r})] + E_{ee}[n(\mathbf{r})] - E_H[n(\mathbf{r})] \tag{2.46}$$

where ($T_s[n(\mathbf{r})]$) and ($E_{ee}[n(\mathbf{r})]$) are the exact kinetic energy and electron-electron interaction energies, respectively. It is important to remember that the real form of ($E_{xc}[n(\mathbf{r})]$) is not known, therefore approximate functionals

are employed. Of the many approximations, the oldest and most widely used include the Local Density Approximation (LDA) and the Generalised Gradient Approximation (GGA). The Local Density Approximation was also proposed by Kohn and Sham, [75] and it approximates the functional with a function of the local density ($n(\mathbf{r})$) of a uniform electron gas which has the same density at point ($\mathbf{r}$).

$$E_{xc}[n(\mathbf{r})] = \int \epsilon_{xc}(n(\mathbf{r}))n(\mathbf{r})d\mathbf{r} \tag{2.47}$$

The functional derivative of which is given by:

$$\frac{\delta E[n(\mathbf{r})]}{\delta n(\mathbf{r})} \equiv \mu_{xc}(n(\mathbf{r}))n(\mathbf{r}) = \left(\epsilon_{xc}(n(\mathbf{r})) + n(\mathbf{r})\frac{d\epsilon_{xc}(n(\mathbf{r}))}{dn(\mathbf{r})}\right) \tag{2.48}$$

where ($\epsilon_{xc}(n(\mathbf{r})) = \epsilon_{xc}^{hom}[n(\mathbf{r})]$) is usually based upon Quantum Monte Carlo calculations on homogeneous electron gases at different densities.[77] These were parametrised as interpolation formulae in analytical form.[78] Due to the condition of homogeneity imposed in the LDA functional, it negates the corrections to the exchange-correlation due to inhomogeneities in the electron density at ($\mathbf{r}$). The Generalised Gradient Approximation accounts for inhomogeneities in the electron gas by including a contribution of the electron gradient $\nabla n(\mathbf{r})$ to the exchange-correlation energy.[79]

$$E_{xc}^{GGA}[n(\mathbf{r})] = \int n(\mathbf{r})\epsilon_{xc}^{hom}[n(\mathbf{r})]F_{xc}[n(\mathbf{r}), \nabla n(\mathbf{r})]d\mathbf{r} \tag{2.49}$$

The different variants of the GGA functional vary in their enhancement factor ($F_{xc}[n(\mathbf{r}), \nabla n(\mathbf{r})]$), which is usually written in terms of the Seitz radius (r_s) and the dimensionless reduced density gradient ($s(r)$) in (Eq. 2.50), where the Fermi wave-vector is given by ($k_F = [3\pi^2 n(\mathbf{r})]^{1/3}$).

$$s(\mathbf{r}) = \frac{|\nabla n(\mathbf{r})|}{2k_F(\mathbf{r})n(\mathbf{r})} \tag{2.50}$$

2.2.4 Linear-scaling DFT

In order to implement the KS equations, a representation of the quantum operators and the KS wave function (ψ_i) are chosen and are restricted to a subspace spanned by a finite set of basis functions where the problem is solved. Computer packages build molecular orbitals using the non-orthogonal single particle function (ψ_i), which is usually chosen to enhance the efficiency of performing DFT calculations. A commonly used basis set is a truncated plane-wave basis set:

$$\psi_{i\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}.\mathbf{r}} u_{i\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}.\mathbf{r}} \left[\sum_{\mathbf{G}} c_{i\mathbf{k},\mathbf{G}} e^{i\mathbf{G}.\mathbf{r}} \right] \tag{2.51}$$

where the system is defined as a crystal with lattice vectors ($\mathbf{r}$) and reciprocal lattice vectors ($\mathbf{G}$), this could describe a system of a real periodic crystal or an aperiodic system. The KS wave functions are classified by a Bloch-vector ($\mathbf{k}$) in the Brillouin zone. A plane wave basis set is further defined by:

$$\frac{\hbar^2}{2m} |\mathbf{k} + \mathbf{G}|^2 \leq E_{cut} \tag{2.52}$$

where (E_{cut}) is a cut-off kinetic energy of the plane waves, through which wave vectors are retained in the expansion of the KS wave functions and convergence is controlled. Conventional plane wave basis sets suffer from inefficiency, in that the time required to solve them scales with the cube of the number of electrons in a system, or ($\mathcal{O}(N^3)$) time complexity. Electronic structure calculations are demanding, especially for molecular systems, where the number of atoms can reach thousands of atoms in biomolecular systems. Therefore, given limited computational resources, the only way of efficiently applying DFT to evaluating the electronic ground state of biomolecular systems is through the use of alternative basis sets.

The Order-N Electronic Total Energy Package (ONETEP) package [80, 81] utilises periodic cardinal sin or Psinc functions as basis sets, [82] using

which, the computational expense scales linearly ($\mathcal{O}(N)$) with the number of electrons in a system. These basis sets use the real space representation of truncated plane waves, they retain the orthogonality of (Eq. 2.51) and additionally have the property of spatial localisation that is needed to efficiently expand the spatially-localised support functions. The Psinc basis sets are real linear combinations of plane waves that are highly localised and orthogonal, defined by (Eq. 2.53)

$$D_{\{m\}}(\mathbf{r}) = \prod_{i=1}^{3} \frac{1}{N_i} \sum_{p_i=-J_i}^{J_i} e^{i(p_i \mathbf{b_i}).(\mathbf{r}-\mathbf{r}_{\{m\}})} \tag{2.53}$$

where ($N_i = 2J_i + 1$) are the number of grid points in each lattice direction, the high localised property can be recognised as at points ($\mathbf{r}_m$, $D_{\{m\}}$) are approximations to Dirac-delta functions.

$$D_{\{m\}}(\mathbf{r}) \approx \delta(\mathbf{r} - \mathbf{r}_{\{m\}}) = \frac{\Omega_{cell}}{(2\pi)^3} \int d\mathbf{G} e^{i\mathbf{G}.(\mathbf{r}-\mathbf{r}_{\{m\}})} \tag{2.54}$$

To achieve linear-scaling, it is not efficient to work in terms of the aforementioned electron density ($n(\mathbf{r})$) from the Kohn-Sham equations, and instead formulate the problem in terms of a density matrix defined by (Eq. 2.55), where (f_n) is the occupancy and (ψ) the KS orbitals. It has the property of idempotency ($\rho^2 = \rho$), the electron density can be computed as its diagonal elements ($n(\mathbf{r}) = 2\rho(\mathbf{r}, \mathbf{r}')$) and the total energy of a system is defined as ($E = 2\mathrm{Tr}(\rho H)$).

$$\rho(\mathbf{r}, \mathbf{r}') = \sum_n f_n \psi_n(\mathbf{r}) \psi_n^*(\mathbf{r}') \tag{2.55}$$

Specifically within ONETEP, the density matrix is given in terms of non-orthogonal generalised Wannier functions (NGWFs) that are localised in real space,[82] and defined by (Eq. 2.56). The density kernel ($K^{\alpha\beta}$) is a representation of (f_n) (Eq. 2.55) and must be sparse, so that elements of the

NGWF further than a defined cut-off (r_K) can be truncated.

$$\rho(\mathbf{r},\mathbf{r}') = \sum_{\alpha\beta} \phi_\alpha(\mathbf{r}) K^{\alpha\beta} \phi^*_\beta(\mathbf{r}) \tag{2.56}$$

The ground state energy of a given system is then computed by (Eq. 2.57)

$$E_0 = \min_n E[n] = \min_\rho E[\rho]_{\rho=\rho^2} = \min_{\mathbf{K},\phi} E[\mathbf{K},\phi] \tag{2.57}$$

By (Eq. 2.57), two nested conjugate gradient constrained search loops are used to minimise the total energy. In the outer energy minimisation loop, the density kernel is kept fixed while the energy is minimised with respect to the spatial profile of the NGWFs, which is equivalent to solving a KS eigenvalue problem. In the inner loop, the NGWF expansion is kept fixed while the energy is minimised with respect to the matrix elements of the density kernel. Therefore eventually arriving at a self-consistent Hamiltonian once the right (idempotent) density matrix ($\rho(\mathbf{r},\mathbf{r}')$) is found.

2.2.5 Bespoke non-bonded forcefield parameters

The Tkatchenko-Scheffler (TS) method is an approach to post process the ground state electronic density — produced by the DFT calculations as outlined in (section 2.2.4) — to derive the non-bonded forcefield parameters from (Eq. 2.3) for use in MD simulations.[83] It utilises the charge density dependence of both atomic charges as well as Lennard-Jones parameters, therefore accounting for the impact of the local chemical environment on the van der Waals contributions and point charges of atoms. The TS method requires an atoms-in-molecule method, which uses the ground state DFT electronic density of the molecular system ($n(\mathbf{r})$) to divide into uniform overlapping atomic densities ($n_A(\mathbf{r})$) through a weighting factor $w_A(\mathbf{r})$ (Eq. 2.58).[83, 84]

$$n_A(\mathbf{r}) = w_A(\mathbf{r})\, n_{\text{mol}}(\mathbf{r}) \tag{2.58}$$

The weighting factor is calculated by computing the share of each isolated atom ($n_A^0(\mathbf{r})$) of the (N) atoms that make up a promolecule ($n_{\text{mol}}^0(\mathbf{r})$). A promolecule is an ideal reference system composed of non-interacting atoms fixed at the same positions in the corresponding molecular system.

$$n_{\text{mol}}^0(\mathbf{r}) = \sum_{A=1}^{N} n_A^0(\mathbf{r}) \tag{2.59}$$

where the weighting factors are differently formulated according to the atoms-in-molecule scheme, in the iterative Hirshfeld (IH) scheme,[85] it is given by:

$$w_A^{\text{IH}}(\mathbf{r}) = \frac{n_A^0(\mathbf{r})}{n_{\text{mol}}^0(\mathbf{r})} \tag{2.60}$$

The self-consistency of the IH scheme therefore requires that the share of the isolated atomic densities ($n_A^0(\mathbf{r})$) from the promolecular density ($n_{\text{mol}}^0(\mathbf{r})$) is equivalent to the share of the overlapping atomic density ($n_A(\mathbf{r})$) from the total molecular density ($n_{\text{mol}}(\mathbf{r})$). The atomic electron population (N_A) are derived from the atomic densities ($n_A(\mathbf{r})$) by:

$$N_A = \int n_A(\mathbf{r}) \mathrm{d}\mathbf{r} \tag{2.61}$$

In the IH scheme, the isolated atomic density is composed of the weighted average of the atomic densities with the next lowest integer (lint(N_A)) and next highest integer (uint(N_A)) occupancies.

$$\begin{aligned} n_A^{0,N_A}(\mathbf{r}) = n_A^{0,\text{lint}(N_A)}(\mathbf{r}) \left[\text{uint}\left(N_A\right) - N_A\right] \\ + n_A^{0,\text{uint}(N_A)}(\mathbf{r}) \left[N_A - \text{lint}\left(N_A\right)\right] \end{aligned} \tag{2.62}$$

The weighting factor in the iterated stockholder atoms (ISA) scheme [86] is instead defined with respect to the spherical average of the electronic density around atom (A) by (Eq. 2.63). Unlike IH, the self-consistency in ISA requires that every value of the radius of a sphere around each nucleus (A), the average electron density on the surface of this sphere is the

same in the promolecular atom ($\langle n_A^{0,\mathrm{ISA}}(d)\rangle$) and in the atom in the molecule ($\langle n_A^{\mathrm{ISA}}(d)\rangle$).[87]

$$w_A^{\mathrm{ISA}}(\mathbf{r}) = \frac{\left\langle n_A^{0,\mathrm{ISA}}\left(|\mathbf{r}-\mathbf{R}_A|\right)\right\rangle}{n_{\mathrm{mol}}^{0,\mathrm{ISA}}(\mathbf{r})} \tag{2.63}$$

where ($\langle n_A^{0,\mathrm{ISA}}\left(|\mathbf{r}-\mathbf{R}_A|\right)\rangle$) is the average density over the surface of a sphere with radius (d) for an isolated atom (A). The promolecular density in (Eq. 2.63) is similarly defined as the sum of spherically averaged spheres of radii ($|\mathbf{r}-\mathbf{R}_B|$) around every atom ($B$):

$$n_{\mathrm{mol}}^{0,\mathrm{ISA}}(\mathbf{r}) = \sum_{B=1}^{N}\left\langle n_B^{0,\mathrm{ISA}}\left(|\mathbf{r}-\mathbf{R}_B|\right)\right\rangle \tag{2.64}$$

The density derived electrostatic and chemical electron density partitioning (DDEC) method is one of the atoms-in-molecule electronic density partitioning approaches. DDEC uses a mixture of both IH and ISA methods and iteratively optimises for a converged solution to the weighing factor to resemble the spherical average of the atomic densities ($n_A(\mathbf{r})$) and the density of an isolated reference atom ($n_A^0(\mathbf{r})$). The atomic partial charges (q_A) for MD are derived from DDEC by (Eq. 2.65), where (Z_A) is the nuclear charge of atom A.

$$q_A = Z_A - N_A = Z_A - \int n_A(\mathbf{r})\,\mathrm{d}^3\mathbf{r} \tag{2.65}$$

Whereas the dispersion (B_{ij}) and repulsion (A_{ij}) coefficients are derived from the partitioned electronic density by (Eq. 2.66a) and (Eq. 2.66b).[33]

$$B_i = \left(\frac{V_A^{\mathrm{DDEC}}}{V_A^{\mathrm{free}}}\right)^2 B_i^{\mathrm{free}} \tag{2.66a}$$

$$A_i = \frac{1}{2}B_i\left(2R_A^{\mathrm{DDEC}}\right)^6 \tag{2.66b}$$

The atomic volume (V_A^{DDEC}) is calculated from the electronic density:

$$V_A^{\mathrm{DDEC}} = \int r^3 n_A(\mathbf{r})\,\mathrm{d}^3\mathbf{r} \tag{2.67}$$

Whereas the other terms are derived from alternative methods. Namely, the free dispersion coefficients (B_i^{free}) are computed using time-dependent DFT calculations of free atoms in vacuum,[88] the reference volume (V_A^{free}) is calculated using more accurate *ab initio* approaches with explicit treatment of electron correlation effects and the DDEC effective radius of each atom rescales the experimentally derived reference free atom radius (R_A^{free}) using (Eq. 2.68).[33]

$$R_A^{\text{DDEC}} = \left(\frac{V_A^{\text{AIM}}}{V_A^{\text{free}}} \right)^{1/3} R_A^{\text{free}} \tag{2.68}$$

The dispersion and repulsion coefficients are related to the Lennard-Jones parameters, (ϵ) and (σ), (Eq. 2.3) via ($A_{ij} = 4\epsilon_{ij}\sigma_{ij}^{12}$) and ($B_{ij} = 4\epsilon_{ij}\sigma_{ij}^{6}$). The resulting parameters are an accurate representation of the chemistry of the molecular system under investigation, since they are exclusively derived from *ab initio* quantum mechanical calculations. As such, computing the non-bonded forcefield parameters using DDEC is a basis for bespoke forcefield parametrisation to bypass the use for transferable forcefields. This approach is implemented for small molecules by software such as QUBEKit.[33]

2.3 Dynamical Mean Field Theory

Strongly correlated materials — encompassing many transition metal oxides — typically have incompletely-filled d- or f-shells, where the on-site Coulombic interactions are comparable with the band-gap. Electronic correlations in such systems give rise to complex behaviour that can no longer be described by the treatment of a single electron within a system of non-interacting particles. As such, single-electron theories such as the local-density approximation (LDA) in density functional theory (DFT) or Hartree-Fock theory fail to accurately describe their electronic structure.[89]

Strongly correlated materials embody exotic electronic properties that are technologically advantageous or have evolved to be utilised by nature for biological function.[90, 91, 92] A study of the Hemocyanin protein core

in Chapter requires such a treatment, where extensions to DFT methods are necessary to accurately describe the electronic structure of the strongly correlated metalloprotein active site.[90]

2.3.1 Hubbard model

The Hubbard model is a lattice model that describes the interaction of particles of opposite spin on a lattice with sites (i, j) through on-site electronic repulsion (U) and off-site kinetic hopping (t) terms.[93, 94, 95] The Hubbard model Hamiltonian (Eq. 2.69) is defined with respect to the creation ($c^\dagger$) and annihilation (c) operators, for spin indices (σ) and the occupation of the i-th site (n_i), where $n_{i\sigma} = c^\dagger_{i\sigma} c_{j\sigma}$.[96]

$$\hat{H}_{\text{Hubbard}} = t \sum_{\langle ij \rangle \sigma} c^\dagger_{i\sigma} c_{j\sigma} + U \sum_i n_{i\uparrow} n_{i\downarrow} \tag{2.69}$$

Despite its simplistic formulation, the Hubbard model — unlike conventional approaches — correctly describes the insulating character of Mott insulators by accounting for the strong repulsion between electrons. In the limit of infinite dimensions or lattice coordination number, the Hubbard model is an exact treatment of the local correlations of a strongly correlated system.[97]

2.3.2 Anderson impurity model

The Anderson impurity model (AIM) is a model Hamiltonian that describes magnetic impurities embedded in a metallic system. It is formulated with respect to the intra-site Coulomb repulsion of the impurity energy levels ($\hat{H}_{\text{loc}}$), a bath of conducting electrons ($\hat{H}_{\text{bath}}$) and the coupling term between the impurity and conduction orbitals ($\hat{H}_{\text{mix}}$).[98]

$$H_{\text{AIM}} = \overbrace{\sum_i \epsilon_i a^\dagger_i a_i}^{\hat{H}_{\text{bath}}} + \overbrace{\sum_{i\sigma} \left(V^\sigma_i c^\dagger_\sigma a_{i\sigma} + \text{h.c.} \right)}^{\hat{H}_{\text{mix}}} + \overbrace{U n_\uparrow n_\downarrow - \mu \left(n_\uparrow + n_\downarrow \right)}^{\hat{H}_{\text{loc}}} \tag{2.70}$$

The non-correlated electronic levels (ϵ_i) of the bath are defined with respect to bath creation ($a_i^\dagger$) and annihilation (a_i) operators. The impurity is described by electrons (n) interacting through a Coulomb repulsion (U) and a chemical potential (μ). The coupling between the impurity and bath orbitals is described by the hybridisation term (V_i^σ).

The dynamics of the electrons hopping into and out of the bath are described by the impurity Green's function (G_{imp}) and defined with respect to the hybridisation function ($\Delta(\omega) = \sum_k |V_k|^2 / (\omega + \mu - \epsilon_k)$) and the self energy ($\Sigma(\omega)$) for frequency ($\omega$):

$$G_{\text{imp}}\ (\omega)^{-1} = \omega + \mu - \epsilon_k - \Delta(\omega) - \Sigma_{imp}\,(\omega) \tag{2.71}$$

In dynamical mean-field theory (DMFT),[99] the lattice (Hubbard) model can be mapped onto a single impurity model, where the single correlated impurity orbital is embedded in an uncorrelated bath of conduction-band states. This mapping is self-consistent, requiring the impurity Green's function to reproduce the lattice dynamics, as described by the local lattice Green's function (G_{ii}), such that:

$$G_{\text{imp}}\ (\omega) = G_{\text{ii}}\,(\omega) \tag{2.72}$$

As well as the self-energy:

$$\Sigma_{\text{imp}}\ (\omega) = \Sigma_{\text{ii}}\,(\omega) \tag{2.73}$$

A DMFT calculation is solved through an iterative process where the AIM parameters are chosen such that the model Hamiltonian best describes the physics of the real system in question. As the hybridisation function encapsulates all of the contributions of the bath sites to the physics of the impurity sites, the iterative process converges when the hybridisation function satisfies the self-consistency conditions outlined above.[100]

DMFT is a sophisticated method that includes quantum dynamical effects, and takes into account charge fluctuations, spin fluctuations, and thermal excitations. The DMFT approach is very computationally expensive for studying realistic systems, it is therefore combined with DFT in the form of DFT+DMFT to result in a method that belongs to a broad category of embedding approaches.[99, 101] The DFT+DMFT method accounts for the limitations of DFT by treating the many-body effects of strongly correlated materials explicitly, while limiting this treatment to the correlated subspace of the d- or f-electrons, thereby side-stepping the prohibitive scaling of quantum chemistry methods.

Chapter 3

Accurate large scale modelling of Graphene Oxide

The development of accurate structures and forcefield parameters for exotic materials are paramount to the transferability of molecular dynamics investigations. Using preexisting work that posits the semi-ordered structure of graphene-oxide (GO), implemented using software developed to create semi-ordered rectangular graphene oxide sheets, we develop a bespoke forcefield using density functional theory calculations. We compare the performance of both generalised and bespoke forcefields on the behaviour of GO in solution with respect to experiments. We discover that observables diverge between generalised and bespoke forcefields in interfacial water dynamics and ion adsorption. This work provides an insight to the importance of bespoke forcefield design for combining both accuracy and scale in modelling nanomaterials at the interface. The bespoke forcefield shows strong agreement with AIMD simulations for the interfacial water dynamics and ion adsorption; we conclude that this results in the forcefield having an accuracy near that of AIMD. Contributions for this work are as follows: **Mohamed Ali al-Badri** and **Christian D. Lorenz** conceived and planned the research. **Mohamed Ali al-Badri** and **Robert C. Sinclair** developed the nanomaterial structure software. **Mohamed Ali al-Badri** performed the calculations. **Mohamed Ali al-Badri, Paul Smith** and **Christian D. Lorenz** analysed the data and **Mohamed Ali al-Badri** prepared the final manuscript.

Carbon 174 (2021) 266–275

Contents lists available at ScienceDirect

Carbon

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

Research Article

Accurate large scale modelling of graphene oxide: Ion trapping and chaotropic potential at the interface

Mohamed Ali al-Badri [a, **], Paul Smith [a], Robert C. Sinclair [b], Khuloud T. al-Jamal [c], Christian D. Lorenz [a, *]

[a] Department of Physics, King's College London, London, WC2R 2LS, UK
[b] Centre for Computational Sciences, University College London, London, WC1H 0AJ, UK
[c] Institute of Pharmaceutical Science, King's College London, London, SE1 9NH, UK

ARTICLE INFO

Article history:
Received 21 October 2020
Received in revised form
9 December 2020
Accepted 10 December 2020
Available online 16 December 2020

Keywords:
Molecular dynamics simulations
Bespoke quantum derived forcefield
Graphene oxide
Interfacial phenomena

ABSTRACT

Graphene oxide (GO) shares many novel mechanical and electronic properties with graphene and has been applied extensively for uses in physics, engineering and medicine. Computational simulations of GO have widely neglected accurate characterisation by random functionalisation, forsaking steric strain and abandoning edge functional groups. Here, we show that molecular dynamics forcefield design using electronic structure calculations of hundreds of atoms of GO with accurate functionalisation shows good agreement with state-of-the-art *ab initio* molecular dynamics (AIMD) simulations. We find that the bespoke forcefield shows better agreement with previous AIMD and experimental results in terms of the interfacial water dynamics and ion adsorption. Namely, GO described by the bespoke forcefield is found to disrupt the hydrogen bonding network at the interface by playing a more dynamic role in accepting and donating hydrogen bonds from water. Furthermore, with the bespoke forcefield, we find preferential adsorption of ions to carboxyl functional groups and a similar mean adsorption half-life for Na^+ and Cl^- ions around GO. These findings are critical for future investigations of GO in complex environments in application ranging from desalination to protein adsorption for drug delivery.

1. Introduction

Graphene oxide (GO) is an amorphous 2D material with complex chemistry and a diverse array of potential applications. GO can be difficult to characterise because of its irregular nature, myriad of functional groups, sensitivity to synthesis method, and high reactivity. This has lead to poor understanding in the literature and great difficulty in simulating GO with conflicting reports in experimental papers[1].

There is no precise consensus on the nanostructure of GO. The Lerf-Klinowski model [2] is widely recognised and has formed the basis of much scientific research [3,4]. Lerf and Klinowski identified the most common functional groups in graphene oxide: epoxy and alcohol groups on the surfaces, with alcohol and carboxyl groups around the edges. However, they assume no correlation between the location of functional groups. Correlation between oxidised sites seems chemically intuitive when one considers isolated carbon double bonds are more reactive than conjugated/aromatic systems [5]; indeed, several experiments have shown the presence of oxidised and unoxidised regions[6–8].

Accurately describing the structure of GO is important, but one must also consider the forcefield used to describe it in a molecular dynamics (MD) simulation. Traditionally classical forcefields are parameterised for transferability, and are used to describe a large family of molecules[9–11]. More recently, generalised varieties of the CHARMM and AMBER forcefield have been used to generate, with minimal effort, parameters for an increasingly large number of molecules[11–13]. Previously, MD has been used to study the interactions of GO with ionic solutions[14], water[15–18], gases [19,20], peptides [21,22] and lipid bilayers[23], in which parameters from these traditional families of force fields were used to approximate GO's chemical properties. As we find in our work, water structure is well described by generalised forcefields, requiring only an accurate representation of GO structure and allowing for structural flexibility. Such work has been recently

* Corresponding author.
** Corresponding author.
E-mail addresses: mohamed.al-badri@kcl.ac.uk (M.A. al-Badri), chris.lorenz@kcl.ac.uk (C.D. Lorenz).

https://doi.org/10.1016/j.carbon.2020.12.032
0008-6223/

applied to study the nanopore connectivity and swelling of GO with varying GO membrane separation and degree of oxidation, work that is vital to understand GO for applications in gaseous or aqueous separation[24,25].

The computational cost of electronic structure calculations is a bottleneck to accurate modelling of heterogeneous nanomaterials such as the two-phase or semi-ordered structure of GO. The semi-ordered structure of GO is defined as the inhomogeneous regions of oxidised and unoxidised domains, where amorphous alcohol and epoxy groups make up the oxidised regions. In contrast, random functionalisation assumes an uncorrelated random distribution of oxidised functional groups. Limitations on the size and conformation can be remedied by imposing periodic boundary conditions (thereby simulating an infinite sheet of GO) as well as planarity to the topology. However this neglects both the chemistry of aforementioned edge functional groups and capping hydrogen atoms and the structural deformation induced by steric effects from correlated oxidised sites.

Subasinghege Don et al. [26] and Mouhat et al. [27] have recently studied the structure of water at the interface of GO using state of the art *ab initio* molecular dynamics (AIMD) simulations. Subasinghege Don et al. [26] conducted NVT simulations to compare the performance of classical MD using the Optimized Potentials for Liquid Simulations All-Atom (OPLS-AA) forcefield to AIMD when modelling a single shell of water at the GO interface within the confines of a large cubic simulation box, mainly composed of a vacuum. This sandwiching vacuum biases water molecules towards the surface for adsorption and neglects the role of bulk water on water dynamics at the GO surface. Mouhat et al. [27] present an exceptional comparison of randomly functionalised and more realistic heterogeneous (semi-ordered) models of GO. Their results show that semi-ordered models of GO are the most stable structures in vacuum as well as in liquid water. Despite the accuracy of their realistic semi-ordered model, their work neglects the role of both carboxyl and phenol edge functional groups. We show that, even though they are much less prevalent than functional groups on the GO surface, edge groups play an important role in water and ion dynamics and should not be disregarded. To the best of our knowledge, the behaviour of semi-ordered GO in ionic solution is yet to be investigated using AIMD.

Accurate forcefields are paramount in bridging between experiment and engineering novel applications of graphitic materials, where atomic scale chemistry defines their macroscopic behaviour. In order to achieve this without the limitations on simulation time and system size, we propose an intermediate level of accuracy that advances on generalised forcefields through deriving non-bonded forcefield parameters from *ab initio* Density Functional Theory (DFT) calculations. This approach allows for the transferability of electronic structure calculations to a dynamic picture of GO at biologically relevant time scales (hundreds of nanoseconds) and systems composed of hundreds of thousands of atoms, as is the purview of MD.

In this study, we present novel electronic structure derived forcefield parameters using the Density Derived Electrostatic and Chemical partitioning method (DDEC) from DFT calculations. We use the DDEC derived forcefield to compare water dynamics in both aqueous and ionic solutions with that observed using the generalised OPLS forcefield.

2. Results

2.1. Forcefield parameters

The structure of the GO sheet is given in Fig. 1, where functional groups have been annotated according to the OPLS forcefield naming scheme used herein.

The DDEC and OPLS non-bonded forcefield parameters are presented in Fig. 2. There is a clear deviation from OPLS parameters for some atomic species, particularly significant in the partial charge distributions. The most notable finding in these distributions is the large variance in partial charge of many individual atomic species, including distributions that span either side of a charge of zero. Lennard-Jones parameters are derived from the DDEC charge density using the Tkatchenko-Scheffler relations[28], where ϵ values are mostly in the vicinity of OPLS-AA parameters, σ values however show a large deviation between the two approaches.

The lateral distribution of partial charges across the GO surface illustrates their variation and correlation with functionalised regions in Fig. 3. It displays the disparity between the two approaches, with a continuous DDEC charge distribution – representative of the DFT electrostatic density – in contrast to the constant OPLS charges that are distributed discretely. Furthermore, the DDEC GO surface carbon atoms have a variance in partial charge distribution spanning negative to positive charge, unlike OPLS which restricts all GO surface carbon atoms to either zero or positive partial charge.

2.1.1. Structural fluctuations

Previous studies using reactive forcefields [29], DFT [30] and AIMD [27] have found that the correlated environment of oxygen-bearing (alcohol and epoxy) functional groups in semi-ordered GO stabilise the edifice through intramolecular hydrogen bonds. In solution, the heterogeneous structure of GO induces structural deformation due to the steric effect of correlated epoxy and alcohol groups. We show that the deformation is similar for both DDEC and OPLS forcefields through the distance of each atom to the mean in the orthogonal plane (Fig. 4). Minute structural differences arise within correlated environments, due to the difference in the intramolecular hydrogen bonds between epoxy-epoxy and alcohol-alcohol groups in DDEC and OPLS. A larger number of intramolecular hydrogen bonds stabilise the local chemical environment to a greater degree in DDEC, reducing its perturbation away from the mean in the orthogonal plane.

Using a hydrogen-bond angle cut-off of 150°, the intramolecular hydrogen-bond of O1–O1 (alcohol-alcohol) and O2–O2 (epoxy-epoxy) is double and triple the prevalence in DDEC than in OPLS, respectively (Table 1).

2.2. Water structure

The radial distribution functions (RDFs) of the constituent GO atomic species with respect to the water oxygen atoms are shown in Fig. 5, along with their hydration numbers — the mean water count within the region enveloped by the first RDF minimum. The radial distribution function between atoms a and b is given by Eq. (1), with a radial cumulative distribution function $G_{ab}(r) = \int_0^r \mathrm{d}r'\, 4\pi\, r'^2 g_{ab}$ where the average coordination number of b atoms at radius r is given by $N_{ab}(r) = \rho\, G_{ab}(r)$.

$$g_{ab}(r) = (N_a N_b)^{-1} \sum_{i=1}^{N_a} \sum_{j=1}^{N_b} \langle \delta(|\mathbf{r_i} - \mathbf{r_j}| - r) \rangle \tag{1}$$

The RDFs show that water oxygen atoms are more tightly bound to the GO oxygens than the GO carbon atoms, where the first water shell has a radius of approximately 4 Å and 5 Å, respectively. Water oxygen atoms show weak structuring around GO carbon atoms (Fig. 5B). Aromatic carbon atoms (C1) belonging to pristine (sp^2) graphene regions, show the weakest water structuring of all constituent atom types, noting that the weak structuring of water around the tertiary alkyl carbon (C2) atoms is due to their

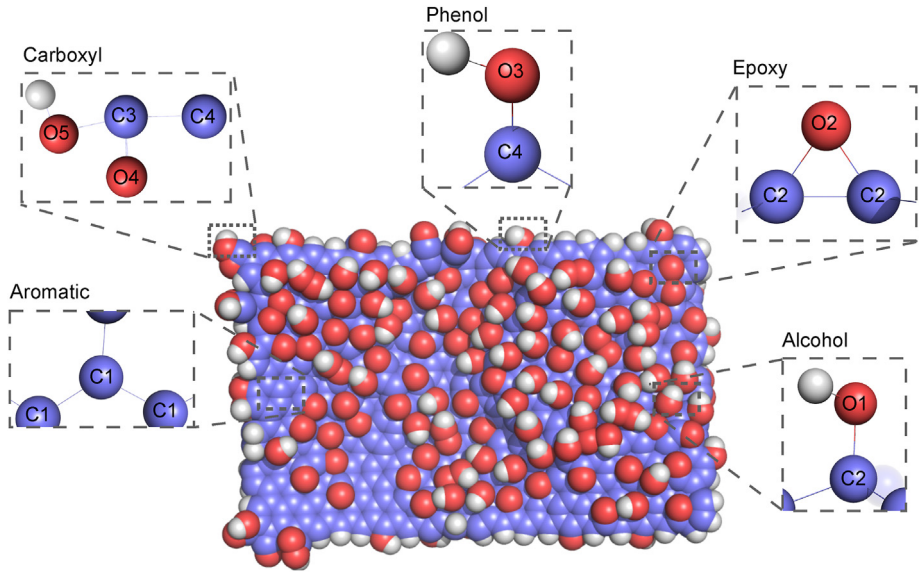


Fig. 1. The semi-ordered 979 atom GO sheet structure, showing regions of oxidised and unoxidised domains. Inset images highlight the structures and naming convention of aromatic carbon and alcohol, epoxy, phenol and carboxyl functional group atoms. (A colour version of this figure can be viewed online.)

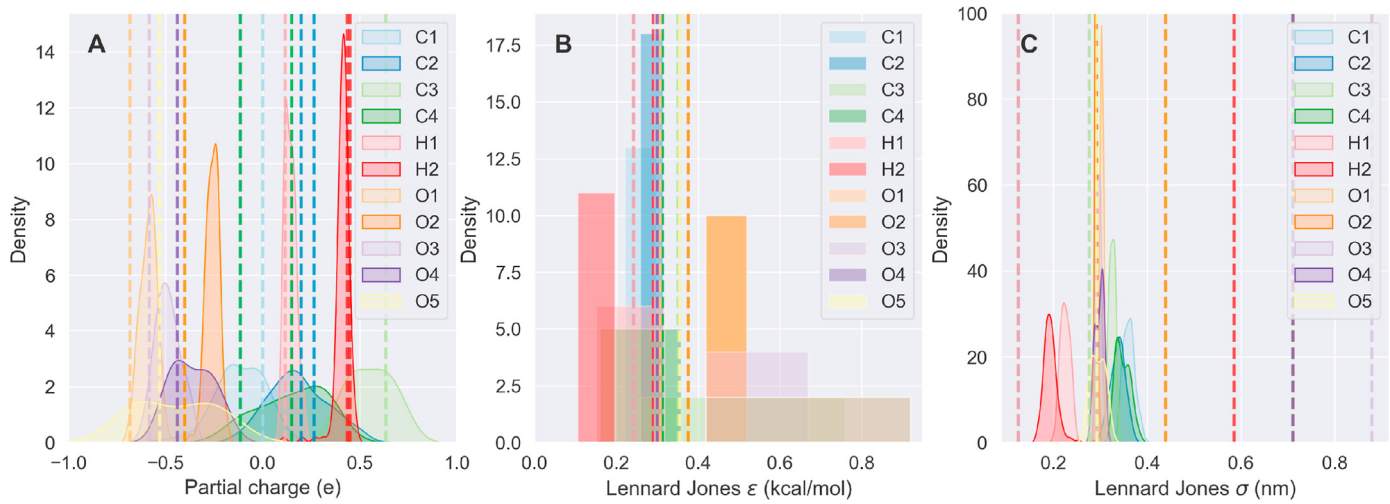


Fig. 2. The distribution of DDEC (A) partial charge, (B) Lennard Jones ϵ and (C) Lennard Jones σ non-bonded forcefield parameters for the component atom types of the GO sheet. OPLS parameters are presented as dashed lines. (A colour version of this figure can be viewed online.)

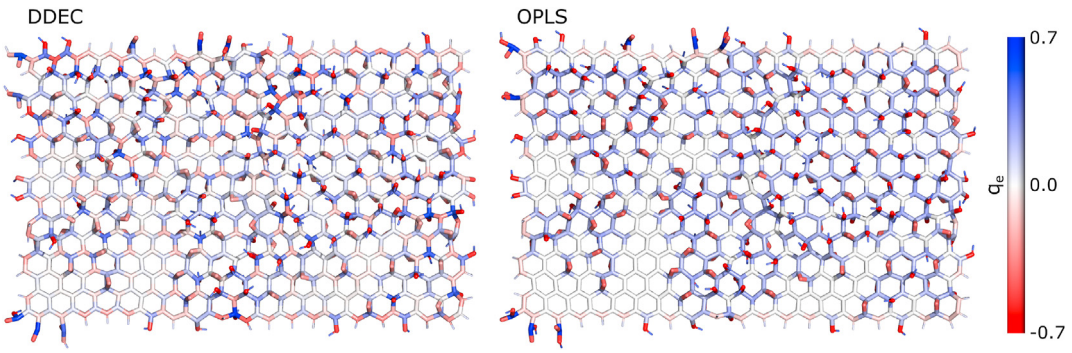


Fig. 3. The lateral illustration of OPLS and DDEC partial charges of the graphene-oxide sheet. (A colour version of this figure can be viewed online.)

coordination by an alcohol group. The water surrounding GO becomes representative of the bulk at 12 Å.

We observe a strong correlation (Pearson coefficient, $R = 0.98$) between the DDEC and OPLS forcefields when considering the

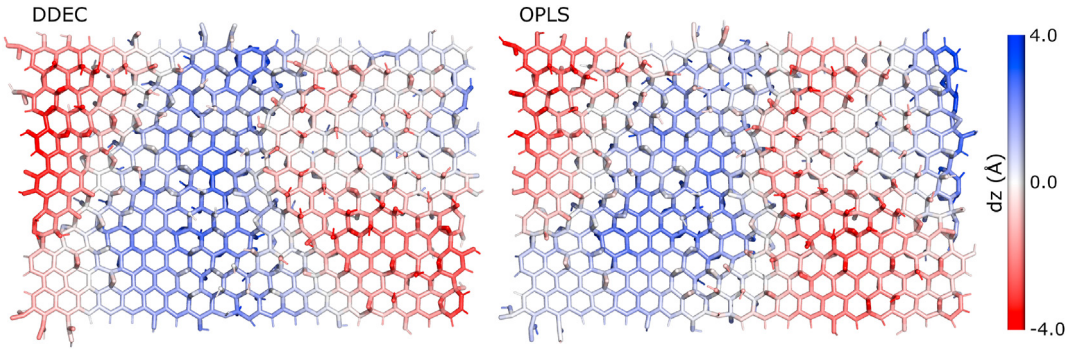


Fig. 4. The structural deformation of the GO sheet in solution, measured by the distance from the mean in the orthogonal plane for the duration of the MD simulation for DDEC (left) and OPLS (right) forcefields. (A colour version of this figure can be viewed online.)

Table 1
Mean number of intramolecular and intermolecular GO hydrogen bonds according to atom type for DDEC and OPLS GO, where the highest numbers of hydrogen bonds are highlighted. Intramolecular hydrogen bonds are normalised by the number of donor atoms. Water-GO hydrogen bonds are normalised by the number of GO atoms. Columns and rows denote accepting and donating species, respectively. Zero hydrogen bonds are denoted as dashes.

DDEC	OW	Alcohol O1	Epoxy O2	Phenol O3	Carboxyl O4	Carboxyl O5
OW		0.32	0.16	0.32	**0.58**	**0.30**
Alcohol O1	**0.16**	**0.04**	**0.02**	–	–	–
Phenol O3	**0.24**	0.01	–	–	–	–
Carboxyl O5	0.12	**0.03**	**0.02**	–	–	–
OPLS						
OW		**0.36**	**0.26**	**0.51**	0.57	0.22
Alcohol O1	0.11	0.02	0.01	–	–	–
Phenol O3	0.14	0.01	–	–	–	–
Carboxyl O5	**0.14**	–	–	–	–	–

mean number of hydrating water molecules around the individual atoms of the GO sheet in solution (Fig. 6A). Collectively, the structure of water as outlined by the atom-to-atom coordination correlation (Fig. 6A) and the RDFs and mean water count (Fig. 5) show a striking similarity between the two forcefields. In particular, this shows that the structure of water around a semi-ordered structure of GO is almost identical independent if it is described by the OPLS generalised forcefield or the bespoke *ab initio* forcefields, despite the variance within the forcefield parameters (Fig. 2).

In ionic solution, however, we find that the number of bound Na^+ and Cl^- ions is significantly different for the systems described by OPLS and by the DDEC forcefield. Further, there is little correlation between the two forcefields for sorption of the sodium ($R = 0.20$) (Fig. 6B) and chlorine ($R = 0.31$) (Fig. 6C) ions to specific atoms in the GO sheet. These disparities are the first signs that a bespoke forcefield parametrisation for exotic materials is necessary to describe the dynamics of GO in complex environments.

2.2.1. Interfacial water

Having established the effect of the local correlated chemical environment of the heterogeneous sheet on structural deformation and the different forcefields on the hydration of GO, we study the structure of water at the GO interface. As the GO sheet is flexible and dynamic, we cannot approximate its surface as a flat 2D sheet. Instead, we construct a discrete intrinsic surface based on the lateral *xy* coordinates of the GO carbon atoms. We linearly interpolate the intrinsic surface to provide a lateral resolution of 1.0×1.0 Å^2. In order to calculate the orientation of water with respect to the surface, we first define surface normals on a per-carbon atom basis then find the angle between the dipole moment of a water molecule and the normal of its nearest carbon atom. The normal to a carbon atom is estimated through a local least squares fitting of the plane formed by the atom and those carbon atoms to which it is covalently bonded. To obtain the surface normal, we use the singular value decomposition (SVD) of the 3D coordinates of the carbon atoms:

$$A = U\Sigma V^T \tag{2}$$

where A is the matrix of 3D coordinates, Σ is a diagonal matrix of the singular values (in descending order), and U and V are orthogonal matrices that contain the left- and right-singular vectors of A, respectively. The normal to the surface is given by the right singular vector with the smallest corresponding singular value. As the SVD can not distinguish between a vector and its additive inverse, we take all normals to be oriented in the positive and negative z direction when considering water above and below the sheet, respectively.

To investigate the structure of water at the interface we report the density of water as a function of the distance in z from the intrinsic surface (Fig. 7A), defined by (Eq. (3))[31]. Unlike the conventional mean density profile, the intrinsic density profile defines the instantaneous shape of a liquid surface as an intrinsic surface $z = \xi(x, y)$ where occupation histograms follow the contour of the instantaneous intrinsic surface for the GO conformation. Where $\xi(x, y)$ is the intrinsic surface for a given GO conformation for a cross-sectional area of the interface A_0. The intrinsic density profile is given by:

$$\tilde{\rho}(z) = \langle \frac{1}{A_0} \sum_{i=1}^{N} \delta(z - z_i + \xi(x, y)) \rangle \tag{3}$$

The results of these measurements show that water above and below the GO sheet is almost symmetric in density — it is excluded at small distances before peaking at around 4 Å (Fig. 7A), roughly the distance of the first minimum in the RDF (Fig. 5A). We see two

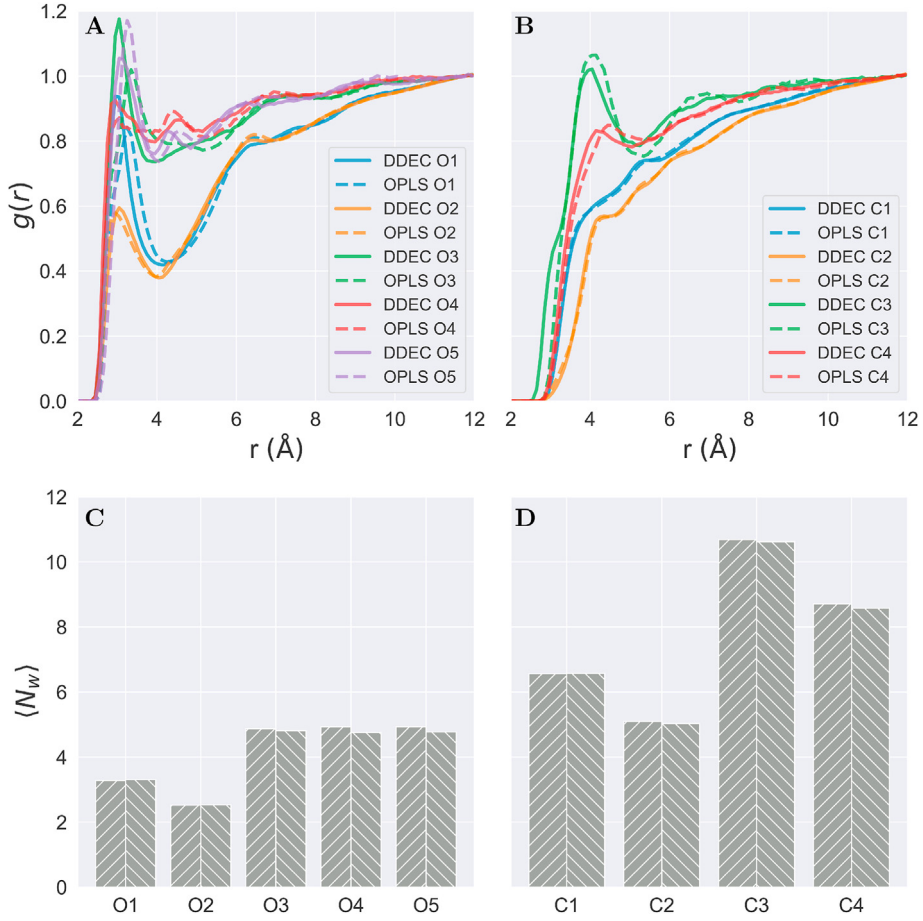


Fig. 5. The radial distribution functions of water oxygen atoms to the component GO (A) oxygen and (B) carbon atom types, according to both DDEC and OPLS forcefields and their respective hydration numbers for (C) oxygen and (D) carbon atom types. (A colour version of this figure can be viewed online.)

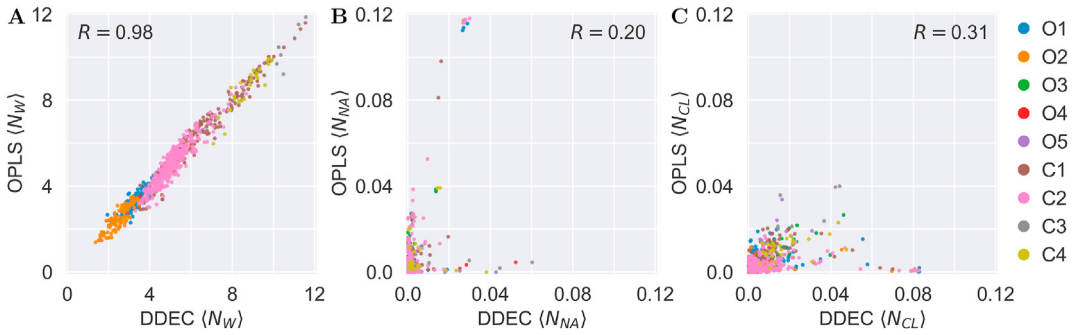


Fig. 6. The correlation of GO atom coordination number by (A) water molecules, (B) Na^+ and (C) Cl^- ions between the OPLS and DDEC forcefields, labeled by atom type. (A colour version of this figure can be viewed online.)

further hydration shells, but at longer distances than the first shell the water quickly becomes more bulk-like.

The mean number of water-water hydrogen bonds as a function of distance z from the intrinsic surface is given by:

$$\langle N_{HB}(z)\rangle = \langle \sum_{i=1} N_{HB,i}\ \delta(z - z_i + \xi(x,y))\rangle \tag{4}$$

Using this as a metric, we see a disruption of the water hydrogen bond network by GO at small distances (Fig. 7B). This may be due to the formation of GO-water hydrogen bonds at the interface (Table 1). GO-water hydrogen bonds directly reduce the number of water-water hydrogen bonds, but may also require a reorientation of water molecules at the interface, which would further disrupt the water hydrogen bond network.

The intrinsic orientational profile is defined by (Eq. 5), where $\hat{\mathbf{p}}_i$ is the unit vector pointing in the direction of the dipole moment of water, $\hat{\mathbf{n}}$ is the normal to the surface of the GO sheet[31].

$$\tilde{P}(z) = \langle \frac{\hat{\mathbf{p}}_i \cdot \hat{\mathbf{n}}}{A_0} \sum_{i=1}^{N} \delta(z - z_i + \xi(x,y))\rangle \tag{5}$$

In Fig. 7C, we see two preferred — and opposite — orientations of water within the first hydration shell. Water nearest the interface ($dz < 3.5$ Å) tends to be oriented with its hydrogen atoms near the

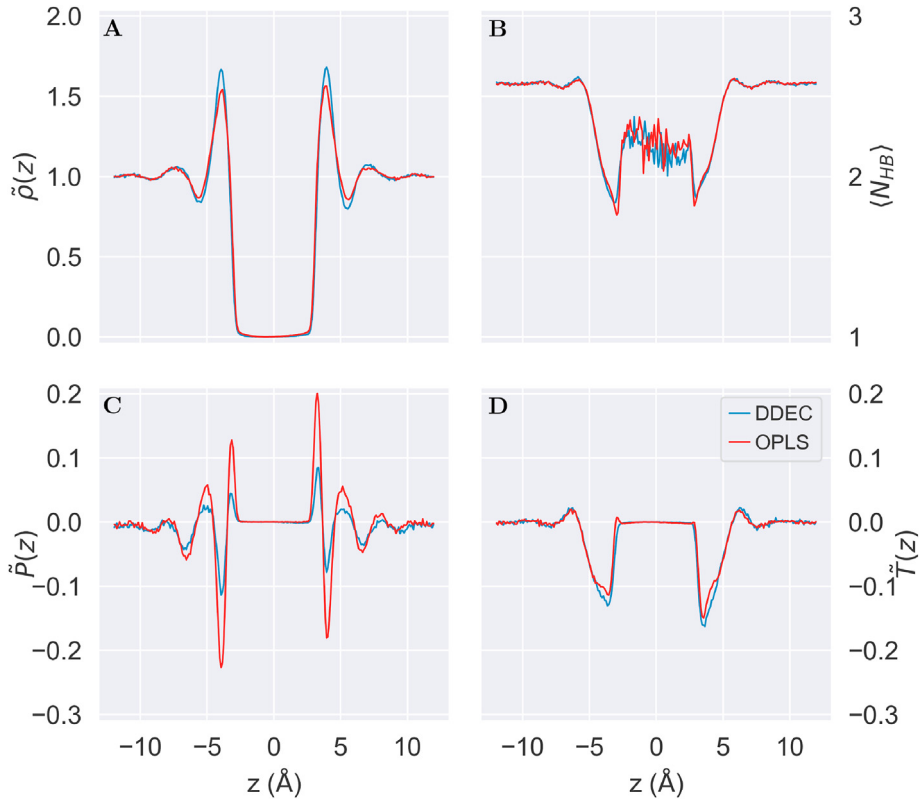


Fig. 7. Intrinsic structure of the GO-water interface for both OPLS and DDEC forcefields, as indicated by (A) the intrinsic density profile, normalised to its bulk value, (B) the number of water-water hydrogen bonds (N_{HB}), (C) the density-weighted profile of dipole orientation ($\tilde{P}$) and (D) the density-weighted intrinsic profile of the second moment of dipole orientation ($\tilde{T}$). (A colour version of this figure can be viewed online.)

GO sheet and oxygen atom facing the bulk ($\theta_{\mu_i} > 0$). The remainder of water within the first hydration shell (3.5 Å < dz < 5 Å) is oriented in the opposite direction, with the oxygen atom nearer the interface ($\theta_{\mu_i} < 0$). The two distinct orientations of water within the first hydration shell of the GO sheet may arise from water either donating or accepting hydrogen bonds with the polar atoms of GO.

The second moment distribution profile is defined by (Eq. (6)), and indicates whether the plane of a water molecule is oriented perpendicular ($\tilde{T} > 0$) or planar ($\tilde{T} < 0$) to the interface[31].

$$\tilde{T}(z) = \langle \frac{(3\hat{\mathbf{p}}_i \cdot \hat{\mathbf{n}})^2 - 1}{2A_0} \sum_{i=1}^{N} \delta(z - z_i + \xi(x,y)) \rangle \quad (6)$$

In Fig. 7D, we see that water molecules tend to be oriented with both hydrogen atoms near the GO interface, rather than with one much nearer the interface than the other.

Fig. 7C suggests that water is more ordered at the GO interface in the OPLS system compared to the DDEC system. We now see this is indeed the case in the joint probability distribution $\rho(z, \theta_\mu)$ of the dipole orientation at a distance z from the GO surface (Fig. 8). Generally, in the semi-ordered model of GO there is favourable orientation of water in the 45°–135° region for either side of the GO sheet, with the preferred orientations creating a mirror image of water orientation across the GO sheet. The orientations in Fig. 8 are calculated with the same reference normal vector (z) and as such, angles above ($z > 0$) and below ($z < 0$) the sheet are offset by 180° from one another.

In order to highlight the differences between the dipole angle orientations between the forcefields, we calculate the joint probability density of the difference (DDEC-OPLS) in dipole angle orientation (Fig. 8). The joint probability density signals the orientational freedom of water at the DDEC GO interface, which is clearly defined for the first interfacial layer from the GO surface. This is illustrated by the difference plot of DDEC – OPLS density, where there is a greater spread of points at a distance 4 Å from the GO surface. This finding is in agreement with previous AIMD results [26], where water was similarly found to display much greater orientational freedom than that observed in OPLS. As such, DDEC GO increases the entropy of the system through interference with non-covalent intermolecular interactions and is therefore considered a chaotropic agent. This property can play a fundamental role and can have vast ramifications on the physics of GO in complex environments, including biomolecules. In proteins, chaotropic agents reduce the stability of the native state formed by water molecules by reducing the hydrophobic effect, leading to destabilisation or complete denaturing[32].

2.2.2. Ion adsorption

Previously, the mean ion count around the constituent species of GO were seen to be grossly disparate between the two forcefields (Fig. 6B and C), where we observed a low atom-to-atom correlation $R = 0.20$ for Na^+ and $R = 0.31$ for Cl^- in ionic solution. The disparities in ion adsorption on GO call for an accurate formulation of forcefields for exotic materials, while establishing their agreement with experiment. The radial distribution functions of ionic atoms highlight the different adsorption patterns around OPLS and DDEC GO (Figs. 9 and 10). The adsorption of Na^+ on the GO sheet oxygen bearing functional groups has a consistent radius for the first solvation shell, differing only in coordination numbers (Fig. 9C). The largest disparity between the OPLS and DDEC GO is that in the latter Na^+ ions preferentially coordinate and are much more tightly

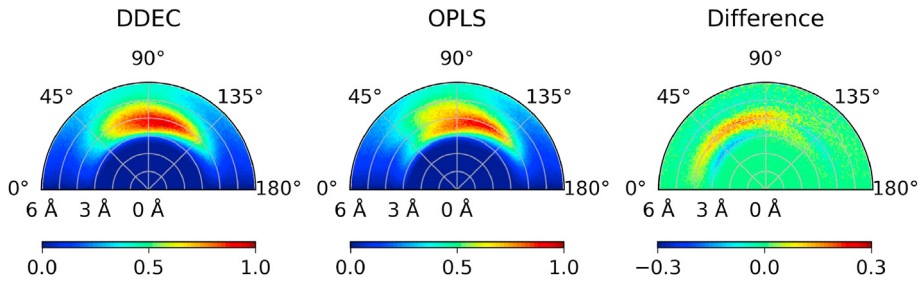


Fig. 8. The joint probability density $\rho(z,\theta_\mu)$ of the water dipole angle θ_μ as a function of z from the intrinsic surface of the GO sheet in solution, for both DDEC and OPLS forcefields. The difference plot shows $\rho^{\mathrm{DDEC}}(z,\theta_\mu) - \rho^{\mathrm{OPLS}}(z,\theta_\mu)$.

bound to edge carboxyl group atoms (carboxyl C3, carboxyl ketone O4 and carboxyl alcohol O5). This preferential adsorption of Na^+ ions to DDEC GO edge functional groups translates to an adsorption half life of 30–40 ps (Tables 2 and 3), highlighting the role of carboxyl groups in trapping Na^+ ions. This result is corroborated by the enhanced desalination performance of carboxyl functionalised GO when compared with pristine GO in experiment[33]. Instead, OPLS GO weakly coordinates Na^+ on edge carboxyl groups at larger distances, with well-ordered peaks in the RDF at 6–8 Å. In both forcefields, pristine graphene regions composed of aromatic carbons (C1) show diffuse and weak coordination of Na^+ , indicating that non-functionalised regions in semi-ordered GO sheets do not play a role in the trapping of Na^+ ions (Fig. 9B). Due to the constant partial charge in all of the OPLS GO atoms, the binding of Na^+ ions to OPLS GO is predictable and wholly dependent on whichever component atoms have a negative partial charge, hence the strong coordination of Na^+ ions to alcohol, epoxy and phenol oxygen atoms as well as carboxyl carbon atoms (Fig. 9C and D).

Cl^- ions consistently adsorb to GO oxygen atoms at larger radii than Na^+ for both forcefields (Fig. 10A). Cl^- adsorption on OPLS and DDEC GO oxygen bearing functional groups have similar coordination numbers; favouring adsorption to alcohol oxygen (O1) and edge phenol group atoms (C4 and O3) in DDEC GO. In OPLS, however, Cl^- preferentially adsorbs to edge carboxyl alcohol oxygen (O5), followed by weaker coordination to edge phenol oxygen (O3) (Fig. 10A,C). Despite weak Cl^- adsorption to GO carbon atoms, it is

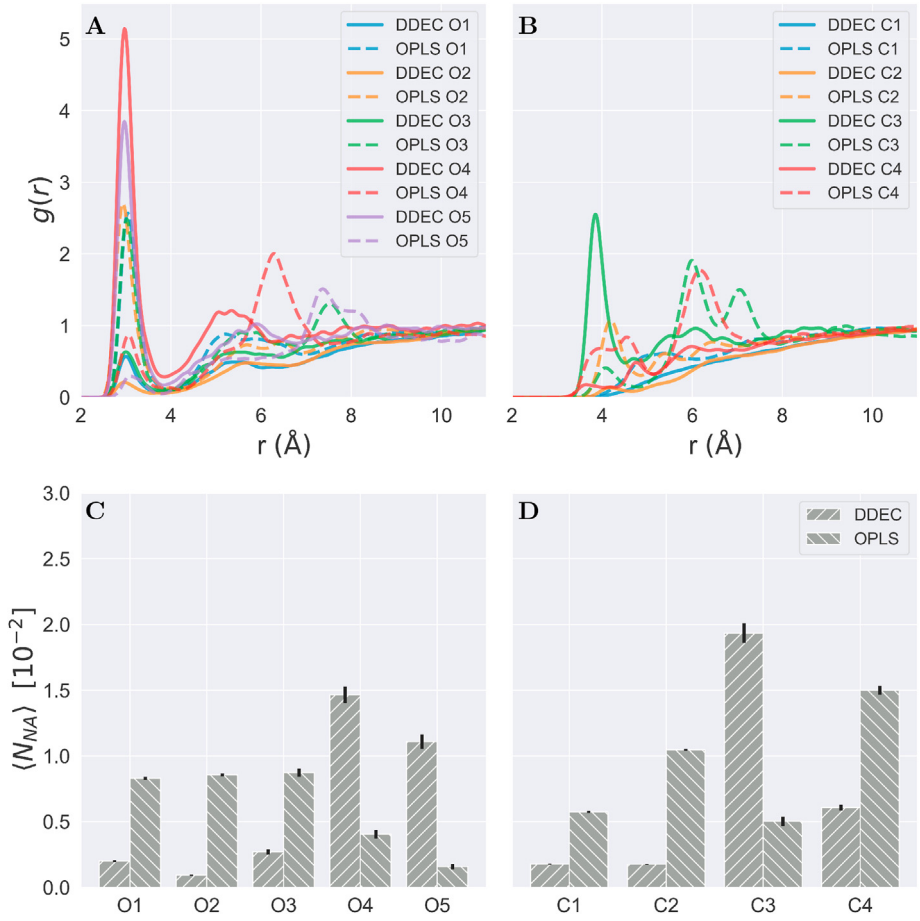


Fig. 9. The radial distribution functions of sodium ions to the component GO (A) oxygen and (B) carbon atom types, according to both DDEC and OPLS forcefields and their respective mean coordination numbers for (C) oxygen and (D) carbon atom types. Error bars indicate the standard error of the mean. (A colour version of this figure can be viewed online.)

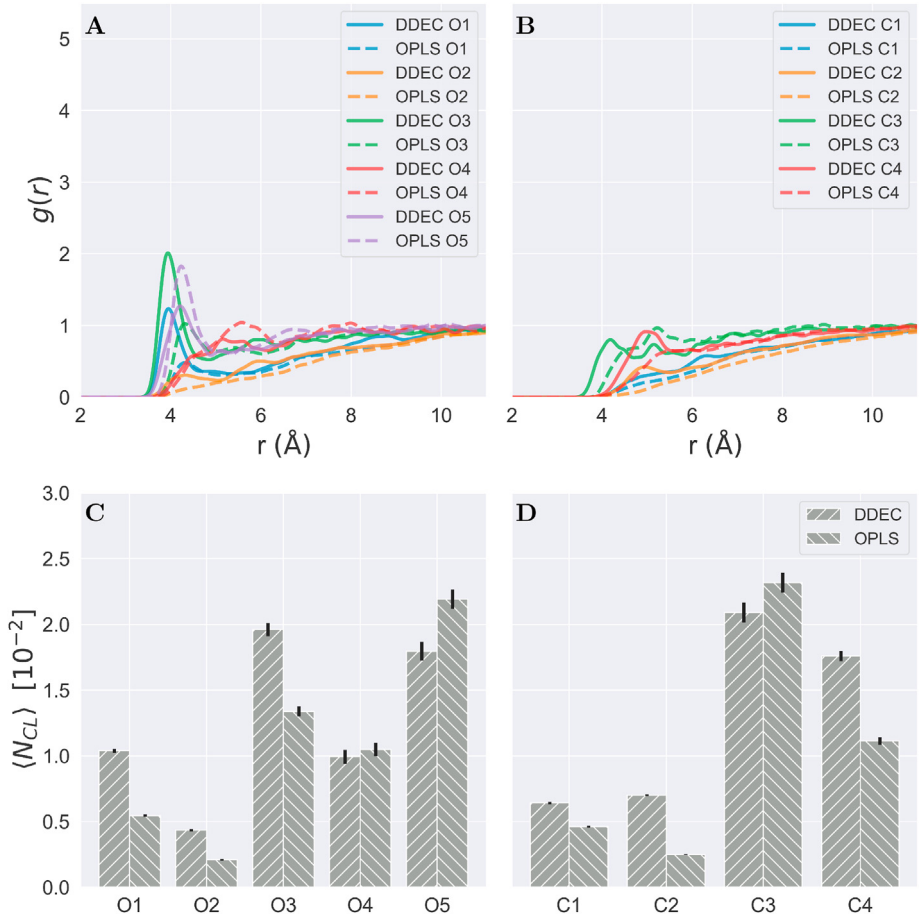


Fig. 10. The radial distribution functions of chlorine ions to the component GO (A) oxygen and (B) carbon atom types, according to both DDEC and OPLS forcefields and their respective mean coordination numbers for (C) oxygen and (D) carbon atom types. Error bars indicate the standard error of the mean. (A colour version of this figure can be viewed online.)

Table 2
Adsorption half-life (ps) of ion atoms around each GO carbon type.

	Aromatic C1	Tertiary alkyl C2	Carboxyl C3	Phenol C4
Na - DDEC	4.3	12.2	37.3	16.4
Na - OPLS	10.3	67.8	11.1	42.5
Cl - DDEC	8.9	15.1	12.0	8.4
Cl - OPLS	4.6	4.1	6.1	4.3

consistently more proximal in DDEC GO (Fig. 10B). Unlike Na^+ ions, the adsorption pattern of Cl^- ions is not significantly disparate to translate to a high variance in the adsorption half-life (Tables 2 and 3).

The survival probability of adsorbed ions at the first RDF minimum for each ion type is calculated using the self-diffusion tensor. Using this, virtual boundary conditions are imposed on the molecular system and survival probabilities and specified time correlation functions of the fluid are computed up to and including the interfacial layer, as proposed by Liu et al. [34] and implemented using MDAnalysis [35,36]. The adsorption half-life is in turn computed by fitting the survival probability to a bi-exponential function in time[37], the results of which are presented in Tables 2 and 3. In order to accurately capture the ultrafast dynamics of ions, the coordinates of the MD trajectory were printed every ps for this analysis.

Using the adsorption half-life of the ions reported in Tables 2 and 3, the OPLS GO mean adsorption half-life over all atoms is 30.7 for Na^+ and 5.9 ps for Cl^- . For DDEC GO the mean adsorption half-life is 18.5 ps for Na^+ and 12.6 ps Cl^- , which is in agreement with experimental desalination studies to investigate ion rejection and permeation rates in GO, which find that Na^+ and Cl^- ions permeate at the same order of magnitude[38].

3. Conclusions

This work provides a framework for simulating graphitic materials, with an accuracy near that obtained with AIMD, while being able to study the dynamics of a system composed of hundreds of

Table 3
Adsorption half-life (ps) of ion atoms around each GO oxygen type.

	Alcohol O1	Epoxy O2	Phenol O3	Carboxyl ketone O4	Carboxyl alcohol O5
Na - DDEC	9.9	9.7	6.8	28.3	41.8
Na - OPLS	41.7	64.7	23.4	8.6	6.3
Cl - DDEC	22.1	16.3	12.5	6.6	11.4
Cl - OPLS	8.7	7.0	6.2	4.4	8.1

thousands of atoms for hundreds of nanoseconds. It extends on pre-existing generalised forcefields (OPLS-AA) and is solely designed to give accessibility to computationally investigate systems that are otherwise not well parameterised for molecular dynamics simulations, as is often the case for exotic materials. It is important to stress that the DDEC non-bonded forcefield parameters derived here are specific to a DFT calculation of the structure in question — a GO sheet with a different distribution of functional groups would require reparameterisation. In the case of no accurate generalised forcefield being available, software is required to apply this framework at scale to streamline the workflow from structure generation, to electronic structure calculation and finally to producing the forcefield files.

By using electronic structure calculations to derive non-bonded forcefield parameters, the MD simulations of a ∼ 1000-atom GO sheet in solution shows that generalised forcefields are incredibly accurate at representing water structure around GO. However, GO in a complex environment — where strong electrostatic interactions with polar or charged moieties are present — ceases to be well described using generalised forcefields. It instead requires *ad hoc* parametrisation, at least until a definitively accurate empirical potential has been designed. Using the *ad hoc* forcefield, our analysis of the structure and dynamics of water near GO revealed a rotational freedom of water molecules within the first hydration shell not present in the generalised forcefield description of GO. The rotational freedom is concomitant with the flexible accepting and donating of hydrogen bonds between GO and water, allowing GO to disrupt the water hydrogen bonding network and acting as a chaotropic agent. Furthermore, when using the *ad hoc* forcefield we observe that the adsorption of Na^+ and Cl^- ions to GO, as described by RDFs, coordination and adsorption half-life, is affected by ion trapping by GO functional groups, which we do not observe to the same degree when employing the OPLS forcefield. This result is consistent with previous experimental findings. The ability of the *ad hoc* forcefield to capture these differences in the interfacial properties of an ionic solution with GO materials will undoubtedly be important in future investigations of the interaction of GO materials with saline solutions in applications like desalination and with biological molecules for its use in drug delivery formulations.

4. Methods

4.1. Geometry

We have modified a Python package [39] to generate rectangular graphene-oxide flakes. The package improves on the commonly applied protocol, discussed above, of randomly placing oxidised functional groups, which we now know is an incomplete model of graphene-oxide structure. Instead, we recreate the two-phase nature of oxidised and unoxidised graphene domains observed in microscopy experiments [6–8,40]. This approach has previously been revealed in a paper by Sinclair et al. [41], where the location of oxygen containing functional groups is determined by the reactivity possible oxidisation sites. Using reactivities calculated by Yang et al. [42], successive oxidised groups can be added to the graphene skeleton in a realistic fashion. In this way, an accurate structure of graphene-oxide, with separate aromatic and highly oxidised domains, can be produced.

The output geometry was minimised using the OPLS molecular dynamics forcefield [43]. OPLS is a versatile forcefield which has been used for similar studies [44]. Minimising the structure in this way prior to performing electronic structure calculations enables the classical forcefield to optimise the approximate geometry generated by the Python package, and inherit the structural characteristics induced by explicit solvation. Hence reducing electrostatic artefacts to an electronic structure calculation performed *in vacuo*, as previously applied and suggested by Lever et al. [45].

4.2. Molecular dynamics

All MD simulations were performed in GROMACS version 2020.1 on the ARCHER Cray XC30 supercomputer on a single 2.7 GHz, 12-core E5-2697 v2 (Ivy Bridge) series processor node. The all-atom OPLS forcefield was used to simulate the classical simulations, and replaced with DDEC non-bonded parameters where indicated. A position restraint algorithm was applied to all non-solvent atoms during equilibration. The ∼80 Å cubic simulation box was fully solvated with TIP3P water molecules. The system was relaxed energetically using steepest-descent energy minimisation for 50,000 steps with an energetic step size of 0.01 kJ/mol. The minimisation was terminated after the maximum energetic contribution was lower than a threshold of 1000.0 kJ/mol/nm. NVT and NPT equilibration was performed for 100 ps using two separate velocity-rescaling thermostat coupling temperature to velocities for graphene and solvent molecules (NVT), where a temperature of 300 K was maintained and 1 bar using the Parrinello-Rahman barostat (NPT). The Verlet cut-off scheme was employed to generate pair lists and the electrostatic interactions were calculated using the Particle-Mesh Ewald algorithm. Both electrostatic and van der Waals interactions were cut off beyond 1.2 nm. All bonds involving hydrogen atoms were constrained using the LINCS algorithm. Production simulations were run for 60ns using a timestep of 1 fs. Analysis was performed using the MDAnalysis package[35,36,46]. All results presented in the text are for GO in pure TIP3P water, unless it is explicitly stated that they are from the simulation we performed with GO in 150 mM (0.15 mol/L) NaCl solution. Note that the results of our analysis of the water dynamics are indistinguishable between the two solutions.

4.3. DFT

The DFT ground-state was obtained using ONETEP[47], which is a linear-scaling DFT code that is formally equivalent to a plane-wave method. Linear-scaling can be achieved by the *in situ* variational optimisation of its atom-centered basis set (spatially-truncated nonorthogonal generalised Wannier functions, or NGWFs)[48]. The total energy is directly minimised with respect to the NGWFs and the single-particle density matrix. The use of a minimal, optimized Wannier function representation of the density-matrix allows for the DFT ground state to be solved with relative ease in large systems. This is particularly useful in molecules, since explicit truncation of the basis functions ensures that the addition of vacuum does not increase the computational cost[49]. The spin-polarised DFT calculations of the 979 atom graphene-oxide system were run with an energy cut-off of 800 eV, using the Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional[50]. Four NGWFs were employed on each of the carbon atoms and oxygen atoms, and one on each hydrogen. NGWFs were truncated using 11 Å cutoff radii. DFT calculations simulated the GO system *in vacuo*, without the explicit treatment of solvent molecules. The pseudopotentials were generated with the OPIUM pseudopotential generation project[51].

Finite-temperature DFT calculations using ensemble density functional theory (EDFT) [52–54] were used to simulate the graphene-oxide sheet containing 979 atoms. Within this formalism, ONETEP minimises the Helmholtz free energy functional to self-consistently find the KS states and their fractional occupancies, which are determined in the Fermi–Dirac [53] scheme using a smearing width k_BT of the Fermi Dirac distribution of $T = 300$ K. The rate of convergence and the search for the electronic ground state using EDFT calculations can be expedited by imposing a kinetic energy

preconditioner of 2.5 Bohr. This works by removing the effect of the kinetic energy operator for high energy states, reducing the width of the eigenspectrum by making the high energy states more degenerate while leaving the low energy states unchanged[55].

4.3.1. DDEC

The DDEC module implemented in ONETEP was used to partition the electron density and assign atom-centered point charges and atomic volumes, it combines two Atoms In Molecule (AIM) approaches – iterative Hirshfeld (IH) and iterated stockholder atoms (ISA) – to assign atomic charges from the electron density. No off-center charges were used in this study. Electron density partitioning was performed using an IH to ISA ratio of 0.02. Lennard-Jones parameters were calculated using the Tkatchenko-Scheffler relations [28] using protocols previously adopted by Cole et al. [49] through the QUBEKit program as an application programming interface [56].

Code availability

A repository containing code and instructions for generating GO sheets is available from https://github.com/maalbadri/Accurate-large-scale-modelling-of-GrapheneOxide.

CRediT authorship contribution statement

Mohamed Ali al-Badri: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Visualization. **Paul Smith:** Methodology, Software, Validation, Formal analysis, Visualization. **Robert C. Sinclair:** Methodology, Software. **Khuloud T. al-Jamal:** Supervision. **Christian D. Lorenz:** Conceptualization, Writing - review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

M.A.B acknowledges Dr. Edward B. Linscott, Dr. Daniel J. Cole and Prof. Arash Mostofi for helpful discussions. This work was supported by BBSRC (grant BB/M009513/1) and EPSRC (grant EP/N509498/1). Via our membership of the UK's HEC Materials Chemistry Consortium, which is funded by EPSRC (EP/L000202, EP/R029431), this work used the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk).

References

[1] D.R. Dreyer, S. Park, C.W. Bielawski, R.S. Ruoff, Chem. Soc. Rev. 39 (2010) 228–240.
[2] A. Lerf, H. He, M. Forster, J. Klinowski, J. Phys. Chem. B 102 (1998) 4477–4482.
[3] C.-J. Shih, S. Lin, R. Sharma, M.S. Strano, D. Blankschtein, Langmuir 28 (2012) 235–241.
[4] H. Dai, Z. Xu, X. Yang, J. Phys. Chem. C 120 (2016) 22585–22596.
[5] M. Trömel, M. Russ, Angew. Chem. 99 (1987) 1037–1038.
[6] D. Pacilé, J. Meyer, A. Fraile Rodríguez, M. Papagno, C. Gómez-Navarro, R. Sundaram, M. Burghard, K. Kern, C. Carbone, U. Kaiser, Carbon 49 (2011) 966–972.
[7] W. Cai, R.D. Piner, F.J. Stadermann, S. Park, M.A. Shaibat, Y. Ishii, D. Yang, A. Velamakanni, S.J. An, M. Stoller, J. An, D. Chen, R.S. Ruoff, Science (New York, N.Y.) 321 (2008) 1815–1817.
[8] S. Saxena, T.A. Tyson, E. Negusse, J. Phys. Chem. Lett. 1 (2010) 3433–3437.
[9] W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, J. Am. Chem. Soc. 118 (1996) 11225–11236.
[10] J.A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser, C. Simmerling, J. Chem. Theor. Comput. 11 (2015) 3696–3713.
[11] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, et al., CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields, J. Comput. Chem. 31 (2010) 671–690.
[12] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case, J. Comput. Chem. 25 (2004) 1157–1174.
[13] J. Wang, W. Wang, P.A. Kollman, D.A. Case, J. Mol. Graph. Model. 25 (2006) 247–260.
[14] H. Dai, Z. Xu, X. Yang, J. Phys. Chem. C 120 (2016) 22585–22596.
[15] N. Wei, C. Lv, Z. Xu, Langmuir 30 (2014) 3572–3578.
[16] C.-J. Shih, S. Lin, R. Sharma, M.S. Strano, D. Blankschtein, Langmuir 28 (2012) 235–241.
[17] J.A.L. Willcox, H.J. Kim, ACS Nano 11 (2017) 2187–2193.
[18] R. Devanathan, D. Chase-Woods, Y. Shin, D.W. Gotthold, Sci. Rep. 6 (2016) 29484.
[19] S. Jiao, Z. Xu, ACS Appl. Mater. Interfaces 7 (2015) 9052–9059.
[20] W. Li, X. Zheng, Z. Dong, C. Li, W. Wang, Y. Yan, J. Zhang, J. Phys. Chem. C 120 (2016) 26061–26066.
[21] X. Sun, Z. Feng, T. Hou, Y. Li, ACS Appl. Mater. Interfaces 6 (2014) 7153–7163.
[22] L. Baweja, K. Balamurugan, V. Subramanian, A. Dhawan, J. Mol. Graph. Model. 61 (2015) 175–185.
[23] N. Willems, A. Urtizberea, A.F. Verre, M. Iliut, M. Lelimousin, M. Hirtz, A. Vijayaraghavan, M.S.P. Sansom, ACS Nano 11 (2017) 1613–1625.
[24] C. Williams, P. Carbone, F. Siperstein, Nanoscale 10 (2018) 1946–1956.
[25] C.D. Williams, P. Carbone, F.R. Siperstein, ACS Nano 13 (2019) 2995–3004.
[26] V. Subasinghege Don, R. David, P. Du, A. Milet, R. Kumar, J. Phys. Chem. B 123 (2019) 1636–1649.
[27] F. Mouhat, F.-X. Coudert, M.-L. Bocquet, Nat. Commun. 11 (2020) 1–9.
[28] A. Tkatchenko, M. Scheffler, Physical Review Letters, vol. 102, Publisher: American Physical Society, 2009, 073005.
[29] P.V. Kumar, N.M. Bardhan, G.-Y. Chen, Z. Li, A.M. Belcher, J.C. Grossman, Carbon 100 (2016) 90–98.
[30] S. Zhou, A. Bongiorno, Sci. Rep. 3 (2013) 2484.
[31] H. Martinez, E. Chacón, P. Tarazona, F. Bresme, Proc. Math. Phys. Eng. Sci. 467 (2011) 1939–1958.
[32] G. Salvi, P.D.L. Rios, M. Vendruscolo, Proteins: Struct. Function Bioinf. 61 (2005) 492–499.
[33] Y. Yuan, X. Gao, Y. Wei, X. Wang, J. Wang, Y. Zhang, C. Gao, Desalination 405 (2017) 29–39.
[34] P. Liu, E. Harder, B. Berne, J. Phys. Chem. B 108 (2004) 6595–6602.
[35] R.J. Gowers, M. Linke, J. Barnoud, T.J.E. Reddy, M.N. Melo, S.L. Seyler, D.L. Dotson, J. Domanski, S. Buchoux, I.M. Kenney, O. Beckstein, MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations, in: S. Benthall, S. Rostrup (Eds.), Proceedings of the 15th Python in Science Conference, SciPy, Austin, TX, 2016, pp. 98–105.
[36] R. Araya-Secchi, T. Perez-Acle, S.-g. Kang, T. Huynh, A. Bernardin, Y. Escalona, J.-A. Garate, A.D. Martınez, I.E. Garcıa, J.C. Saez, et al., Biophys. J. 107 (2014) 599–612.
[37] F. Pedregosa, et al., J. Mach. Learn. Res. 12 (2011) 2825–2830.
[38] Y.H. Cho, H.W. Kim, H.D. Lee, J.E. Shin, B.M. Yoo, H.B. Park, J. Membr. Sci. 544 (2017) 425–435.
[39] Sinclair, R. C. , 2020; https://github.com/velocirobbie/make-graphitics, Accessed: December 15, 2020.
[40] K. Erickson, R. Erni, Z. Lee, N. Alem, W. Gannett, A. Zettl, Adv. Mater. 22 (2010) 4467–4472.
[41] R.C. Sinclair, P.V. Coveney, J. Chem. Inf. Model. 59 (2019) 2741.
[42] J. Yang, G. Shi, Y. Tu, H. Fang, Angew. Chem. Int. Ed. 53 (2014) 10190–10194.
[43] W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, J. Am. Chem. Soc. 118 (1996) 11225–11236.
[44] N. Wei, C. Lv, Z. Xu, Langmuir 30 (2014) 3572–3578.
[45] G. Lever, D.J. Cole, N.D.M. Hine, P.D. Haynes, M.C. Payne, J. Phys. Condens. Matter 25 (2013) 152101. Publisher: IOP Publishing.
[46] P. Smith, R.M. Ziolek, E. Gazzarrini, D.M. Owen, C.D. Lorenz, Phys. Chem. Chem. Phys. 21 (2019) 9845–9857.
[47] C.-K. Skylaris, P.D. Haynes, A.A. Mostofi, M.C. Payne, J. Chem. Phys. 122 (2005), 084119. Publisher: American Institute of Physics.
[48] C.-K. Skylaris, A.A. Mostofi, P.D. Haynes, O. Diéguez, M.C. Payne, Phys. Rev. B 66 (2002), 035119. Publisher: American Physical Society.
[49] D.J. Cole, J.Z. Vilseck, J. Tirado-Rives, M.C. Payne, W.L. Jorgensen, J. Chem. Theor. Comput. 12 (2016) 2312–2323.
[50] J.P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. 77 (1996) 3865–3868. Publisher: American Physical Society.
[51] A.M. Rappe, K.M. Rabe, E. Kaxiras, J.D. Joannopoulos, Phys. Rev. B 41 (1990) 1227–1230. Publisher: American Physical Society.
[52] N.D. Mermin, Phys. Rev. 137 (1965) A1441–A1443. Publisher: American Physical Society.
[53] N. Marzari, D. Vanderbilt, M.C. Payne, Phys. Rev. Lett. 79 (1997) 1337–1340. Publisher: American Physical Society.
[54] C. Freysoldt, S. Boeck, J. Neugebauer, Phys. Rev. B 79 (2009) 241103. Publisher: American Physical Society.
[55] A.A. Mostofi, P.D. Haynes, C.-K. Skylaris, M.C. Payne, J. Chem. Phys. 119 (2003) 8842–8848. Publisher: American Institute of Physics.
[56] J.T. Horton, A.E.A. Allen, L.S. Dodda, D.J. Cole, J. Chem. Inf. Model. 59 (2019) 1366–1381. Publisher: American Chemical Society.

Erratum

Due to an error in the plotting script, Fig. (3.2) has been modified to accurately illustrate the distribution of OPLS and DDEC forcefield parameters. This error has no impact on the discussion, results nor other plots in the chapter. In Fig. (3.11 A,C), the legend colours now correctly reference the respective atom types, unlike Fig. (3.2). The distance between the Lennard-Jones ϵ values is correctly plotted in Fig. (3.11 B), where previously a distribution was wrongly plotted for the DDEC values.

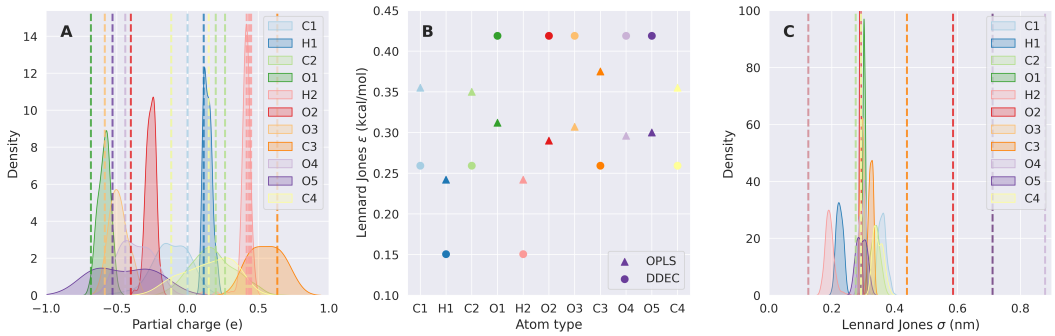


Figure 3.11: The distribution of DDEC (A) partial charge, (B) Lennard Jones ϵ and (C) Lennard Jones σ non-bonded forcefield parameters for the component atom types of the GO sheet. OPLS parameters are presented as dashed lines in (A) and (C), the absolute difference between the OPLS (triangle) and DDEC (circle) Lennard Jones ϵ values (gray: negative, black: positive).

Chapter 4

Nanomaterial functionalisation modulates hard protein corona formation

The work in this chapter extends on the accurate modelling of graphene oxide nanomaterials to investigate changes in protein structure upon adsorption on the material. Using molecular dynamics simulations, we study the evolution of protein structure in response to interfacial interactions on the bio-nano interface. To understand the impact of structural changes over simulation times of hundreds of nanoseconds, we require a comprehensive analysis pipeline that investigates these changes from multiple perspectives. Here, we probe the varying adsorption behaviours of apolipoprotein c-III and the effect of functional groups in modulating protein aggregation on the nanomaterial. We use two functionalisations of graphitic materials — graphene oxide and double-clickable graphene oxide.

Contributions for this work are as follows: **Mohamed Ali al-Badri** and **Christian D. Lorenz** conceived and planned the research. **Mohamed Ali al-Badri** performed the calculations. **Mohamed Ali al-Badri, Paul Smith** and **Christian D. Lorenz** analysed the data and **Mohamed Ali al-Badri** prepared the final manuscript.

The protein corona is an obstacle to exploiting the exotic properties of nanomaterials in clinical and biotechnology settings, with potential applications in DNA sequencing, point of care testing and drug delivery vehicles. The formation of the protein corona is driven by dynamic atomic scale interactions at the bio-nano interface, which are impenetrable using conventional experimental techniques. Here, we use molecular dynamics simulations to study the effect of graphene-oxide (GO) functionalisation on apolipoprotein-c3 (apo-c3) adsorption. We develop an analysis pipeline, encompassing binding energy calculations to protein structure analyses employing Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) and machine learning clustering. We find that apo-c3 is denatured by adsorption on GO, largely driven by the large energetic contributions of electrostatic interactions such as π-π stacking of aromatic amino acids to pristine graphene regions. The enthalpic contribution of such binding event outweighs the intraprotein bond enthalpy required to maintain the protein tertiary structure. Through denaturing and exposing buried hydrophobic residues, the protein backbone is stabilised by forming β-bridges, which serve as binding motifs for protein-protein interactions that drive further protein aggregation on the nanomaterial surface. When adsorbing on double-clickable azide- and alkyne-double functionalised graphene oxide (C2GO), apo-c3 largely retains its tertiary structure. Binding with the nanomaterial surface is dominated by weaker van der Waals interactions that are dispersed over the protein surface, where charged protein residues are sterically hindered by azide functional groups. The apo-c3 N-terminus is the binding motif for C2GO adsorption, leaving the conformation of the C-terminus unchanged, hence conserving the lipid binding function of apo-c3.

4.1 Introduction

Graphene-oxide (GO) is a semi-ordered 2D material that shares many novel mechanical and electronic properties with graphene. It is utilised in ap-

plications including ion trapping, desalination, electronics, chemistry and biomedicine.[102, 103, 104] GO is a promising platform for *ex* and *in vivo* biological applications — such as bio-sensing and therapeutics. Its electrochemical properties makes GO an attractive contender for bio-sensing applications such as single nucleotide polymorphism detection in DNA,[105] next generation nanopore DNA sequencing [106] and point of care (PoC) applications for early virus detection using covalently linked immobilised monoclonal antibodies.[107] Therapeutically, the large surface area to volume ratio of 2D materials results in a large loading capacity for targeting molecules, fluorescent dyes and drug molecules for intravenous administration as well as photothermal therapy for cancer treatment.[108, 109, 110, 111]

Unfortunately the transition from *in vitro* to clinical settings is held back by a chemically driven adsorption of serum proteins — referred to as a protein corona — upon their introduction to a biological medium.[112, 113] A hard corona (HC) is a tightly bound monolayer of proteins at the nanoparticle interface. Subsequent protein layers adherent to the HC are referred to as the soft corona (SC).[114] The character and composition of the HC define the nanomaterial's biological identity and therefore its biological fate, circulation time, cellular uptake and cytotoxicity.[115, 116] HC formation has revealed highly variant protein composition profiles, sensitive not only to inter-species biological media [117] but also disease-specific influences in human GO coronas.[118] Patient-specific or 'personalised' protein corona profiles have since been utilised as a diagnostic tool for high-throughput, inexpensive and highly accurate early cancer detection.[119]

Modulating the HC character to overcome the disadvantages of 2D materials in biological applications through nano-functionalisation remains challenging.[120] Extrapolating the sensitivity of nano-functionalisation to the aforementioned highly variant identity of the corona from experimental

studies is incomplete and requires a better understanding of GO-HC interfacial behaviour to atomistic precision. Such an understanding is paramount to designing the next generation of biosensors and nanomedicines.

Azide- and alkyne-double functionalised graphene oxide (C2GO) has previously been proposed as a cancer targeting nanovector,[121] where click reactions conjugate targeting moieties on azide and trimethylsilyl (TMS)-alkyne functional groups.[122] In vitro, both C2GO and GO show varying HC character when exposed to serum proteins, which subsequently impacts their biological identities.[116] An experimental quality-by-design screening platform was used to unpick the HC character and evaluate its relationship with biological fate, cellular uptake and cytotoxicity. Using this, proteins that reduced material dependent toxicities were identified to play a role in diminishing cytotoxicity, including apolipoprotein C-III (apoc3).[116]

Protein coronas are known to temporally evolve, leaving the total amount of protein constant but varying their composition according to binding affinity, known as the Vroman effect.[123] Upon the introduction of a nanomaterial to a biological medium, highly abundant proteins such as albumin, immunoglobulin G and fibrinogen bind to the nanoparticle surface in the early stages of the Vroman effect, followed by their replacement with high binding affinity proteins such as apolipoproteins and coagulation factors in the late stages of the Vroman effect.[124, 125, 126, 127, 123] The Vroman effect has previously been observed in computational and experimental studies of peptide, cellulose and fatty-acid binding on GO.[128]

Previously, MD has been used to study the interactions of GO with peptides to understand conformational transitions of amyloid-beta during adsorption,[129] the hydration pattern of an adsorbed toy-model alpha-

helix [130] and verification of enzyme active-site deformation following adsorption. [131] To the best of our knowledge, no MD simulations have so far studied protein adsorption on accurate models of GO, and have instead randomly placed oxidised functional groups on the GO surface. Accurate modelling of GO should reflect the semi-ordered structure of GO; composed of inhomogeneous regions of oxidised and unoxidised domains, where amorphous alcohol and epoxy groups make up the oxidised regions.[132] *Ab initio* MD simulations of GO show that semi-ordered models of GO are the most stable structures in vacuum as well as in liquid water.[133] Furthermore, we have recently shown the importance of accurate functionalisation and accounting for steric strain and edge functional groups in large scale MD simulations of GO, using generalised and bespoke electronic structure MD forcefield design.[134] In this work, we use molecular dynamics (MD) simulations to study protein denaturing through adsorption on GO and C2GO, to understand HC conformation-activity relationships through binding free energy, contact map, protein structure and solvent exposure analyses. We apply machine learning dimensionality reduction techniques to decode the spatio-temporal MD data that is hard to interpret into distinct protein secondary structure conformations.

4.2 Results

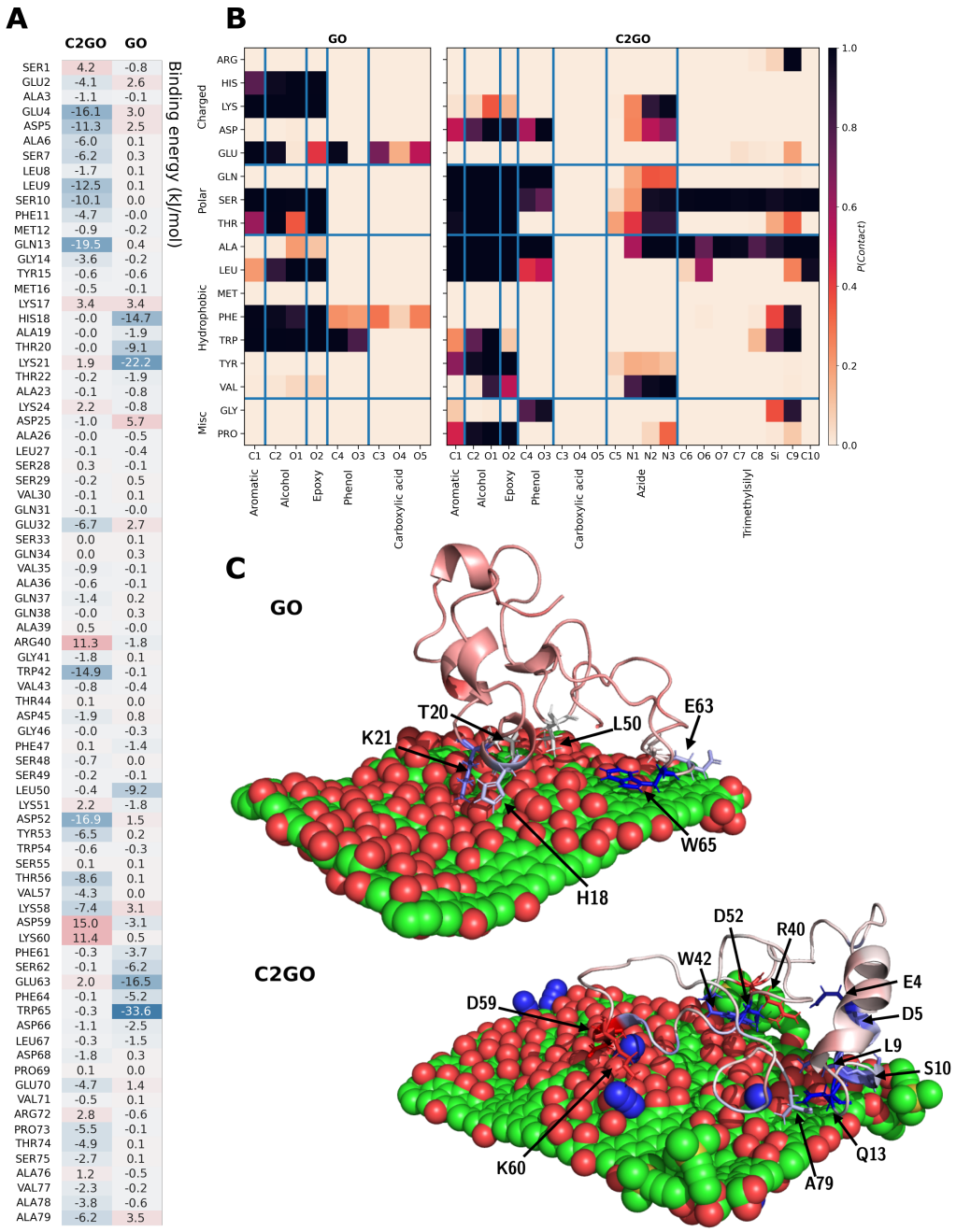
4.2.1 Binding on the bio-nano interface

In both GO and C2GO systems, apo-c3 readily adsorbs to the substrate and reaches a stable conformation over the course of the adsorption trajectory. To investigate the binding of apo-c3 to the GO and C2GO interface, we compute contact maps and free energies of binding that underpin the changes in protein structure following adsorption. These analyses probe interfacial dynamics that may play a large role in the formation of a protein corona around a nanomaterial upon its introduction to a biological medium. To identify adsorption of apo-c3 to the graphitic sheets we calculated the min-

imum distance over time between any heavy (non-hydrogen) atom of the protein to any heavy atom of the sheet (Fig. 4.5). We also compute the minimum distance of each residue to the graphitic sheets over time. These are combined to produce a heatmap of contact probability between apo-c3 and GO/C2GO (Fig. 4.1B). The binding free energy is calculated according to the molecular mechanics with Poisson–Boltzmann and surface area solvation (MM-PBSA) method, implemented using g_mmpbsa.[135]

The average binding free energy components for the GO-apo-c3 and C2GO-apo-c3 systems are given in Table 1. The binding free energy components are decomposed to per-residue contributions (Fig. 4.1A). In this way we can acquire a better understanding of each amino-acid's contribution to the binding free energy of the protein-nanomaterial complex. Per-residue binding free energy contributions are colour-coded by their interaction strength (blue to red) in the protein structure of the final configuration (Fig. 4.1C) for GO and C2GO, where each largely contributing amino acid residue has been illustrated and labelled explicitly. Note that the entropic contribution to the binding free energy is not calculated using high-throughput methods such as g_mmpbsa, due to their computational cost. Therefore, the binding free energies in (Table 1) are computed to evaluate the relative binding free energy instead of the absolute free energy.

The binding of apo-c3 to the graphitic substrate is driven by amino acids with the highest binding affinities from the beginning of the adsorption process. Apo-c3 residues 60-70 initiate binding with GO (Fig. 4.5), most likely due to the high binding affinity of TRP65 to π-π stack with pristine graphene domains on the GO surface, as well as GLU63 binding with positively charged (tertiary alkyl, phenol and carboxylic acid) carbon atoms (Fig.4.1A). From 100 ns apo-c3 binding to GO is driven by positively charged (HIS18, LYS21) and polar (THR20) residues (Fig.4.1A) interacting



	C2GO	GO
SER1	4.2	-0.8
GLU2	-4.1	2.6
ALA3	-1.1	-0.1
GLU4	-16.1	3.0
ASP5	-11.3	2.5
ALA6	-6.0	0.1
SER7	-6.2	0.3
LEU8	-1.7	0.1
LEU9	-12.5	0.1
SER10	-10.1	0.0
PHE11	-4.7	-0.0
MET12	-0.9	-0.2
GLN13	-19.5	0.4
GLY14	-3.6	-0.2
TYR15	-0.6	-0.6
MET16	-0.5	-0.1
LYS17	3.4	3.4
HIS18	-0.0	-14.7
ALA19	-0.0	-1.9
THR20	-0.0	-9.1
LYS21	1.9	-22.2
THR22	-0.2	-1.9
ALA23	-0.1	-0.8
LYS24	2.2	-0.8
ASP25	-1.0	5.7
ALA26	-0.0	-0.5
LEU27	-0.1	-0.4
SER28	0.3	-0.1
SER29	-0.2	0.5
VAL30	-0.1	0.1
GLN31	-0.1	-0.0
GLU32	-6.7	2.7
SER33	0.0	0.1
GLN34	0.0	0.3
VAL35	-0.9	-0.1
ALA36	-0.6	-0.1
GLN37	-1.4	0.2
GLN38	-0.0	0.3
ALA39	0.5	-0.0
ARG40	11.3	-1.8
GLY41	-1.8	0.1
TRP42	-14.9	-0.1
VAL43	-0.8	-0.4
THR44	0.1	0.0
ASP45	-1.9	0.8
GLY46	-0.0	-0.3
PHE47	0.1	-1.4
SER48	-0.7	0.0
SER49	-0.2	-0.1
LEU50	-0.4	-9.2
LYS51	2.2	-1.8
ASP52	-16.9	1.5
TYR53	-6.5	0.2
TRP54	-0.6	-0.3
SER55	0.1	0.1
THR56	-8.6	0.1
VAL57	-4.3	0.0
LYS58	-7.4	3.1
ASP59	15.0	-3.1
LYS60	11.4	0.5
PHE61	-0.3	-3.7
SER62	-0.1	-6.2
GLU63	2.0	-16.5
PHE64	-0.1	-5.2
TRP65	-0.3	-33.6
ASP66	-1.1	-2.5
LEU67	-0.3	-1.5
ASP68	-1.8	0.3
PRO69	0.1	0.0
GLU70	-4.7	1.4
VAL71	-0.5	0.1
ARG72	2.8	-0.6
PRO73	-5.5	-0.1
THR74	-4.9	0.1
SER75	-2.7	0.1
ALA76	1.2	-0.5
VAL77	-2.3	-0.2
ALA78	-3.8	-0.6
ALA79	-6.2	3.5

Figure 4.1: MM-PBSA binding energy contributions per apo-c3 amino acid residue for GO and C2GO sheets, colour coded by magnitude (A), heat map showing contact probability of apo-c3 amino acid type with GO and C2GO functional group atoms (B) and adsorbed apo-c3 structure on GO and C2GO, protein amino acids at the graphitic interface are coloured by MM-PBSA binding energy contribution and hydrogen atoms have been omitted for clarity (C).

Table 4.1: Binding energy components from MM-PBSA calculations performed on the MD trajectories of adsorbed apo-c3 on GO and C2GO sheets. Stronger binding components have been highlighted in bold.

	ΔE_{vdW}(kJ/mol)	ΔE_{elec}(kJ/mol)	ΔG_{polar}(kJ/mol)	ΔG_{apolar}(kJ/mol)	$\Delta G_{binding}$(kJ/mol)
GO	-338.8±0.8	**-208.3±1.5**	**422.2±5.5**	-31.9±0.1	-156.8±5.3
C2GO	**-463.2±0.9**	-135.7±1.0	420.5±3.0	**-47.5±0.1**	**-225.9±2.8**

with oxygen containing surface functional groups. Accordingly, the binding free energy of GO-apo-c3 has a higher net contribution of electrostatic interactions when compared with C2GO-apo-c3, contributing -208 kJ/mol to the total binding free energy (Table 1).

In contrast, binding to C2GO is driven by both of the extreme N- and C-terminal ends of apo-c3, which also have the highest binding affinity to the C2GO substrate according to the MM-PBSA binding free energy analysis. The N-terminus serves as the main binding domain with C2GO, driven by negatively charged (GLU4, ASP5) and polar (SER10, GLN13) residues interacting with a TMS group and positively charged (tertiary alkyl, phenol and carboxylic acid) surface and edge carbon atoms (Fig. 4.7). Meanwhile, charged amino acid residues contribute a net positive change in binding free energy of apo-c3 to C2GO (Fig. 4.1A), stabilising C2GO-apo-c3 binding and a conserved apo-c3 tertiary structure. Charged amino acids are the only residues in apo-c3 that sterically hinder neighbouring amino acids from binding to the C2GO interface, and this is wholly achieved through azide functional groups, with the exception of ARG40 which is sterically hindered by a silanol group. However, the role of azide groups is not exclusively restricted to an agent for steric hindrance of amino acid side chains to the graphitic surface, as they drive the binding of hydrophobic amino acids in the extreme C-terminal end of apo-c3 (Fig. 4.1A,B). Unlike GO-apo-c3, C2GO-apo-c3 binding is dominated by van der Waals interactions, contributing -463.2 kJ/mol to the total binding free energy (Table 1).

In the case of GO-adsorbed apo-c3 binding is dominated by electrostatic interactions (Table 1) with a locus around two binding hotspots — LYS21, TRP65 (Fig. 4.1A) — which stabilises unfolding of the protein. This is due to the enthalpic contribution of this binding outweighing the bond enthalpy of the intraprotein interactions maintaining the tertiary structure. Due to the changes to the apo-c3 native state tertiary structures induced by adsorption, apo-c3 has a significantly higher conformational entropy. The energetic stabilisation of a denatured protein is attributed to the conformational entropy of amino acid side chains, which would be a barrier to recovering the native state tertiary structure.[136] In contrast, the enthalpic contribution in C2GO-adsorbed apo-c3 is more dispersed over the surface (Fig. 4.1A) and is on average dominated by weaker (van der Waals) interactions (Table 1), thus the intraprotein interactions maintaining the tertiary structure are not compromised by any energetic spikes caused by strong binding in any location on the complex interface.

4.2.2 Protein structure

Changes to protein structure during adsorption — leading to protein denaturing or deformation — are evaluated using secondary structure, intramolecular hydrogen bonding and solvent-accessible surface area (SASA) analyses. Protein native contacts indicate changes in secondary structure due to adsorption, their temporal evolution is calculated using the define secondary structure of proteins (DSSP) algorithm[137] and protein intramolecular hydrogen bonds.

4.2.2.1 DSSP and hydrogen bonds

To understand the dynamic progression of denaturing in GO-adsorbed apo-c3, DSSP and hydrogen bond analyses indicate the sequential loss of secondary structure and number of intra-molecular hydrogen bonds respectively. The number of protein intramolecular hydrogen bonds over time (Fig. 4.2B) complement the DSSP results (Fig. 4.2A) at a higher resolu-

tion and are invariant to the DSSP algorithm criteria for defining strands, helices or coils. It shows that over time, unlike C2GO-adsorbed apo-c3, GO-adsorbed apo-c3 has a dynamic character, sporadically gaining or losing hydrogen bonds. These high spatio-temporal hydrogen bond frequencies stabilise intrachain contacts by recovering the loss of hydrogen bonds, indicating structural reorganisation is taking place during adsorption.

Loss of secondary structure in C2GO-adsorbed apo-c3 is transient, with temporary loss of α-helix character spanning residues 30-40 (Fig. 4.2A). In contrast, GO-adsorbed apo-c3 α-helix character is lost irreversibly for residues 30-35 and 55-60 (Fig. 4.2B), concomitantly with the formation of β-turns whose lifetime range from tens to hundreds of ns and persist throughout adsorption (Fig. 4.2C). A decrease of α-helices accompanied by an increase of β-strands elsewhere in the protein is a characteristic observed in experimental secondary-structure analysis of denatured protein libraries using vacuum-ultraviolet circular dichroism.[138] Within the initial 150 ns of adsorption, α-helices spanning residues 1-10, 30-33 and 40-50 transition to coils (Fig. 4.2A). Subsequently at 200 ns, the C-terminus loses it's α-helix character in residues 55-65, the lost hydrogen bonds (Fig. 4.2B) are recovered through forming β-turn motifs (Fig. 4.2A) and stabilising the apo-c3 protein backbone (Fig. 4.2C) for the remainder of adsorption until 400 ns. β-turns remain stable in a solvated state, potentially serving as recognition motifs of protein-protein interaction,[139] until binding takes place with other proteins. Both L4-S7 and W54-T56 transitioned from an alpha-helix to type-I and type-II β-turns, respectively, whereas S48-K51 transitioned from a coil to a type II β-turn.

4.2.3 Clustering

UMAP dimensionality reduction is a useful approach to map high dimensional data of MD protein backbone coordinates trajectories to a lower dimensional protein configuration space.[140] This is due to the better per-

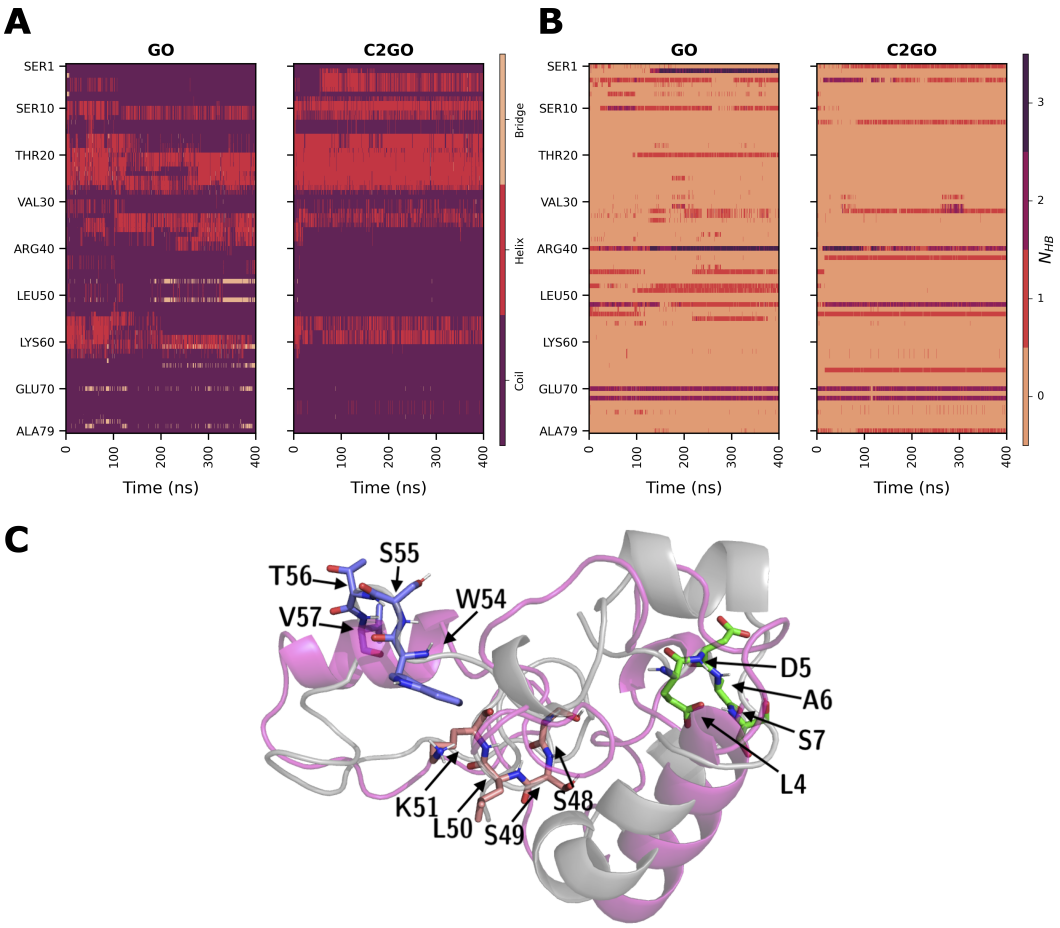


Figure 4.2: Define Secondary Structure of Proteins (DSSP) algorithm applied to the MD trajectory of apo-c3 adsorption with GO and C2GO (A), the number of intramolecular hydrogen bonds per amino acid residue throughout adsorption (B) and illustration of β-turns induced in GO-adsorbed apo-c3 following denaturing (residues L4-S7, S48-K51 and W54-V57), pink and grey structures respectively correspond to the initial and final adsorption conformations (C).

formance of UMAP in preserving local and global structure relationships compared with conventional dimensionality reduction techniques such as principle component analysis (PCA) or t-distributed stochastic neighbour embedding (t-SNE). A projection of the high dimensional spatio-temporal MD data into a low-dimensional space is used to identify distinct protein structures, bridging pathways and global relationships between distinct configurations. It is also a practical tool for visualising what is otherwise a vast dataset.

The number of clusters in the protein configuration space reflects the number of ensembles the protein backbone adopts, namely the distinct structures of apo-c3 during the entire adsorption process. Using this and the visualisation of protein structures corresponding to each cluster, we can see that apo-c3 has higher structural variance when adsorbing on GO than on C2GO, which have 13 and 10 clusters, respectively (Fig. 4.3). GO-adsorbed apo-c3 clusters show how the protein sequentially undergoes a process of reorganisation (Fig. 4.3) driven by interchain hydrogen bonding interactions (Fig. 4.2B) due to the binding interactions with the GO surface (Fig. 4.1A).

The temporal evolution of apo-c3 adsorption follows the cluster numbers in the protein configuration space (Fig. 4.3). The pathways between clusters are synonymous with the DSSP analysis, corresponding to transitions between configuration states where apo-c3 secondary structure either gains or loses α-helix, β-bridge or coil character. In both cases of GO and C2GO adsorption, the reorganisation process requires transitions to discrete intermediary states (clusters) before converging to a stable complex (Fig. 4.3). The stable complex may or may not recover its secondary structure following these transitions, corresponding to whether it has or has not been denatured via the adsorption process. This is indeed the case in GO-adsorbed apo-c3, which has undergone drastic structural changes and retains some of its N-terminus (Fig. 4.3) but loses the C-terminus secondary structure. In contrast C2GO-adsorbed apo-c3 does recover most of its secondary structure, where different cluster pathways transition between the highly conserved native backbone (Fig. 4.3), which is indicative of binding deformations instead of protein denaturing. Note that the images of the protein backbone annotating the clusters only approximate secondary structure character and are not as accurate as the state-of-the-art characterisation

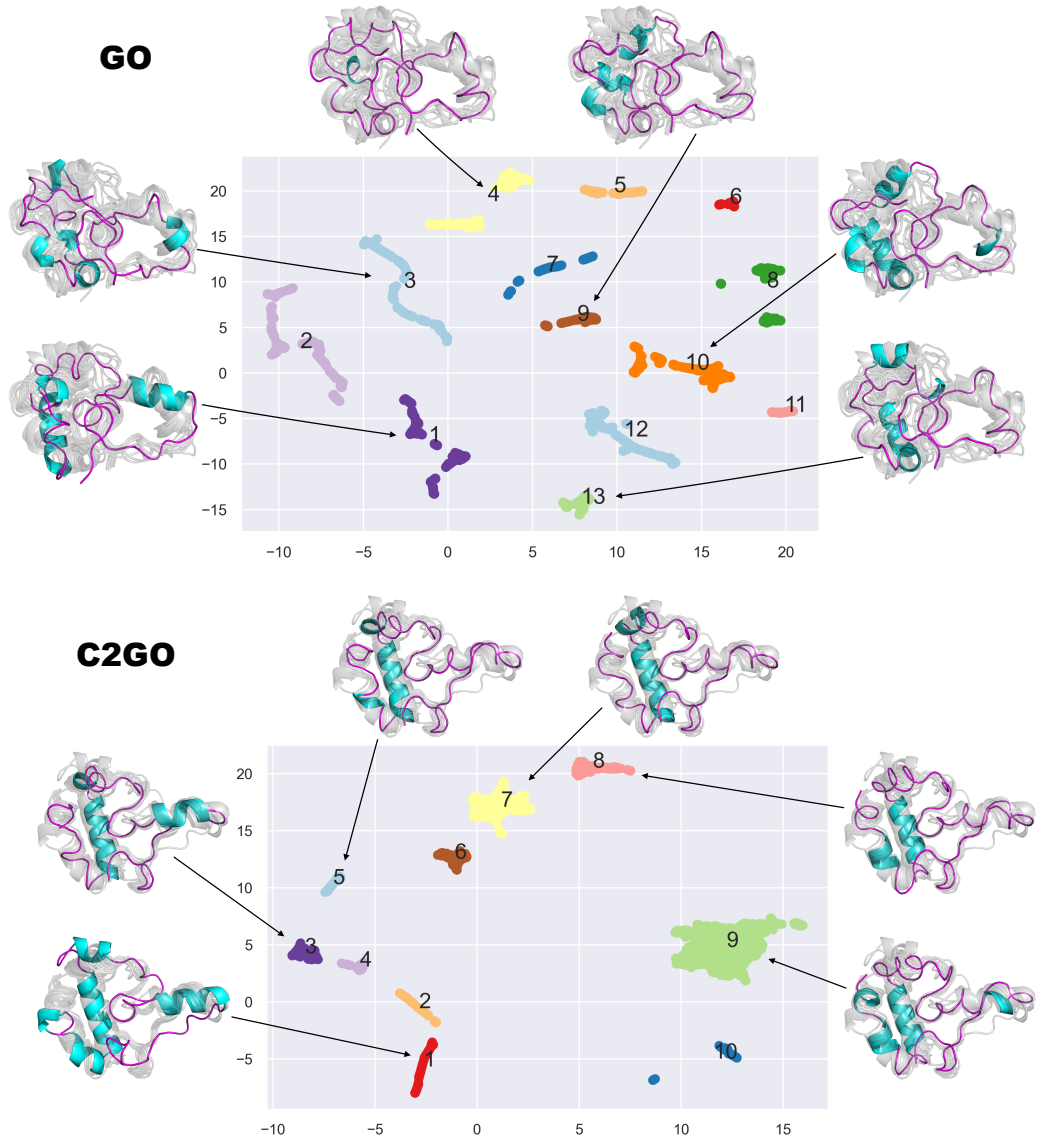


Figure 4.3: Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction of protein backbone denaturing during adsorption on GO (top) and C2GO (bottom) nanosheets. Separate clusters show clear separation of distinct protein backbone secondary structures. Protein structures corresponding to each cluster are coloured by secondary structure (helices in blue, loops in pink) on top of an overlay of all cluster conformations (grey).

of the secondary structure using DSSP analysis (Fig. 4.2).

4.2.4 Solvent exposure

As a result of structural changes induced by adsorption, the hydration of the apo-c3 amino acids change and contribute to potential protein aggregation *in vivo*. As the structure of apo-c3 changes during adsorption on GO, solvent exposure is increased significantly over multiple regions of the amino acid sequence (Fig. 4.4A). Some residues will have a larger solvent-accessible surface area (SASA) after adsorption due to change in conformation. Some residues will have a smaller SASA after adsorption either due to a change in conformation leading to new protein-protein contacts or due to interaction of the residue with the nanomaterial in question. There is a correlation with increased solvent exposure in regions where apo-c3 forms β-turns, as analysed by DSSP and illustrated in (Fig. 4.2). The β-turns remain in this solvated state, potentially serving as recognition motifs for protein-protein interactions until binding to other proteins *in vivo* or *in vitro*.[139] As well as increased solvent exposure of the GO-adsorbed apo-c3 C-terminal helix (residues 47-65), the EVRPTSAVAA minimotif (residues 70-79) has a consistently higher solvent exposure than the native state of apo-c3 in solution (Fig. 4.4A,B). These exposed hydrophobic residues keep the adsorbed complex in a disordered state until it can coalesce with its protein/lipid environment. In contrast, the solvent exposure of apo-c3 adsorbed to C2GO is limited to isolated short sequences leaving the conformation of most of the C-terminal helix region (residues 47-65) unchanged (Fig. 4.4A).

Previous work has found that the C-terminal region of apo-c3 is essential for mediating lipid binding, with the N-terminal region (residues 1-40) playing a limited secondary role.[141, 142, 143] Solvent exposure analysis of the C-terminal helix expresses the conservation of its lipid-binding function from a physio-chemical perspective. These results may in part delineate the experimentally observed preferential uptake of corona-coated C2GO to corona-coated GO by J774 cells.[116]

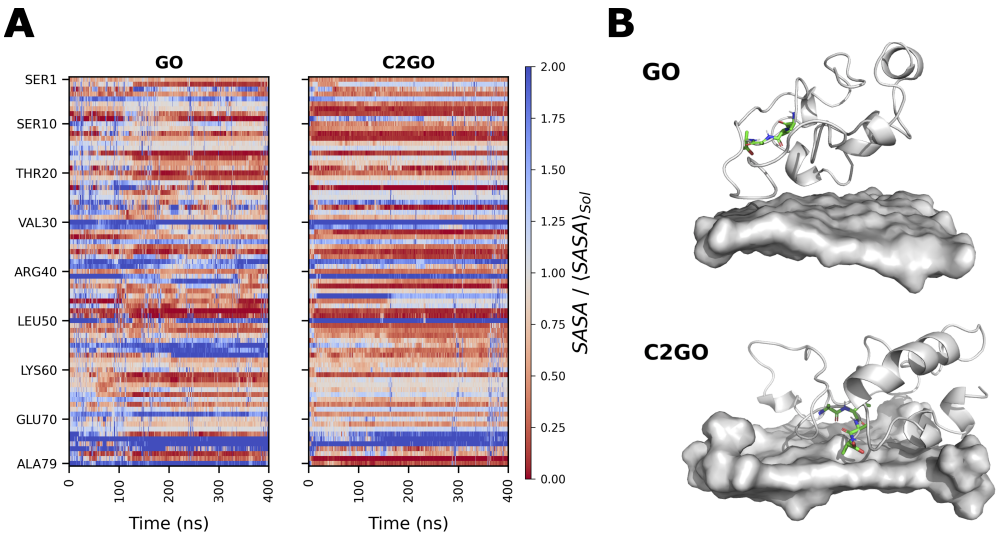


Figure 4.4: The surface accessible surface area (SASA) of apo-c3 amino acid residues during adsorption to GO and C2GO sheets, normalised by average SASA of apo-c3 in solution (A) and illustration of exposure of the AVAA minimotif in the C-terminal region of apo-c3 to solvent in GO adsorption and contrasting structure in C2GO adsorption, the nanomaterials are represented as a surface for clarity (B).

4.3 Methods

4.3.1 Nanomaterial structure

We have modified a Python package [144] to generate azide and trimethylsilyl (TMS)-alkyne functional group conjugated rectangular graphene-oxide flakes.[59] The package improves on the commonly applied protocol, of randomly placing oxidised functional groups, which we now know is an incomplete model of graphene-oxide structure.[134, 133] Instead, we recreate the two-phase nature of oxidised and unoxidised graphene domains observed in microscopy experiments [145, 146, 147, 148] in accordance with our recent study of accurate large scale modelling of graphene oxide.[134]

4.3.2 Molecular Dynamics

MD simulations were performed in GROMACS version 2020.1 on the ARCHER2 AMD EPYC Zen2 (Rome) 64 core CPUs at 2.2 GHz or Nvidia V100 GPUs. The all-atom OPLS forcefield was used to simulate the classical

simulations. A position restraint algorithm was applied to all non-solvent atoms during equilibration. The ∼130 Å cubic simulation box was fully solvated with TIP3P water molecules. The system net charge was neutralised by adding four sodium ions into the system. The system was relaxed energetically using steepest-descent energy minimisation for 50000 steps with an energetic step size of 0.01 kJ/mol. The minimisation was terminated after the maximum energetic contribution was lower than a threshold of 1000.0 kJ/mol/nm. NVT and NPT equilibration was performed for 100 ps using two separate modified velocity-rescaling thermostat — with a stochastic term to ensure generating the canonical ensemble — coupling temperature to velocities for graphene and solvent molecules (NVT),[65] where a temperature of 300K was maintained and 1 bar using the Parrinello-Rahman barostat (NPT). The Verlet cut-off scheme was employed to generate pair lists and the electrostatic interactions were calculated using the Particle-Mesh Ewald algorithm. Both electrostatic and van der Waals interactions were cut off beyond 1.2 nm. All bonds involving hydrogen atoms were constrained using the LINCS algorithm. Production simulations were run for 400 ns using a timestep of 1 fs. Analysis was performed using the MDAnalysis package [149, 150, 151] and its analysis modules.[152, 153] To identify hydrogen bonds we used the hydrogen bond analysis tool [153] implemented in MDAnalysis. We used the MDTraj [154] implementation of DSSP.

4.3.3 Contact maps

Contact maps between apo-c3 and the graphitic sheets were calculated using a hard cutoff of 5.0 Å between any two heavy (non-hydrogen) atoms. Data over the final 50 ns were used for the contact maps, as during this period apo-c3 was stably adsorbed to both GO and C2GO. We constructed contact maps between between GO/C2GO and each specific residue of apo-c3 (Fig. 4.6 and 4.7), as well as contact maps between GO/C2GO and each amino acid regardless of its position in the sequence (Fig. 4.1B).

4.3.4 PBSA binding energies

We post process the last 10 ns of MD trajectories using molecular mechanics with Poisson–Boltzmann and surface area solvation (MM-PBSA) analysis implemented using g_mmpbsa,[135] to obtain a relative order of binding of apo-c3 to the different graphitic nanosheets. The energetic components of the binding free energy are the changes in the system potential energy *in vacuo*, the polar and non-polar solvation energies. The potential energy accounts for bonded (bond, angle and torsion) and non-bonded (van der Waals and electrostatic) energetic terms. Polar solvation energy is the electrostatic contribution to the solvation free energy and is estimated using the Poisson-Boltzmann equation. The non-electrostatic contribution to the solvation free energy accounts for forces between solute and solvent, which are calculated using the solvent accessible surface area (SASA).[135]

4.3.5 UMAP dimensionality reduction

We used the atomic coordinates of apo-c3 to obtain an ensemble of distinct structural conformations when adsorbed to the graphitic sheets. We first centre and align the structures along their backbone atoms, using every tenth frame (50 ps) from the trajectory. We then use the uniform manifold approximation and projection for dimension reduction (UMAP[140]) algorithm to embed the atomic coordinates of heavy atoms of apo-c3 into a two-dimensional space. UMAP constructs a graph of the points in the high-dimensional space, then optimises a low-dimensional representation of the graph such that the topological distance is preserved to a degree in the embedding.[140] Therefore, similar conformations that are close in the high-dimensional space will also be close in the reduced space. We set the `n_neighbours` and `min_dist` hyperparameters to 15 and 0.0, respectively, with the latter being a requirement if the points in the reduced space are to be clustered. We use HDBSCAN[155], implemented in scikit-learn[156], to cluster the apo-c3 conformations in the embedded space. We identify representative structures of each conformation by calculating the mean structure

of a conformation, then finding the structure with the smallest RSMD from this mean structure.

4.3.6 SASA solvent exposure

To quantify the solvent exposure of amino acids during adsorption, we calculated the solvent accessible surface area of each residue of apo-c3 over time. To understand how the conformational changes lead to exposure of residues that are buried in the native state, we normalised the SASA time-series by the mean SASA of each residue of the protein in solution. We used the final 50 ns of apo-c3 solution for determining the mean SASA of each residue. Values greater than and less than 1 indicate, respectively, increased and decreased exposure to the solvent (Fig. 4.4A). The SASA was calculated using the “rolling-ball" Shake-Rupley algorithm,[157] implemented in MDTraj.[154] This approximates the SASA by effectively rolling a ball over each atom of the protein to define a surface, then examining how much of the each atom’s surface is exposed to solvent as opposed to overlapping with surfaces of neighbouring atoms. The radius of the probe in typically set to 1.4 Å — approximately the radius of a water molecule.

4.4 Conclusions

This work has dissected the protein structural changes induced by interactions of functional groups and protein residues on the bio-nano interface. Through adsorption on GO, apo-c3 shows denaturing of the secondary structure over large swathes of the protein sequence. Whereas C2GO-adsorbed apo-c3 has limited deformation, with most variance displayed in the N-terminus. The C-terminus of apo-c3 has previously been found to be responsible for lipid binding, with the N-terminus playing a limited secondary role.[141, 142, 143] These findings correlate with experimental evidence of the increased cellular uptake of C2GO, as compared to GO, hard protein corona.[116]

Apo-c3 is found to denature upon adsorption to GO, according to solvent exposure, DSSP and conformational clustering analyses. Following adsorption on GO, apo-c3 forms 3 separate β-turns that serve as binding motifs for protein-protein interactions. These can aid in subsequent aggregation of serum proteins to the corona, contributing to the corona identity in vivo, hence complicating the targetability of the nanomaterial-corona complex.

The functional groups at the binding interface between the nanomaterials and apo-c3 set off a series of dynamic events that result in large-scale secondary structure changes. Contact distances, heatmaps and binding free energy calculations collectively describe the driving forces of the changes in protein structure following adsorption. The introduction of azide and TMS functional groups was sufficient to stop the denaturing of apo-c3 and the formation of β-turns that serve as protein-protein interaction binding motifs. We find that contact with azide functional groups correlate strongly with contact to other surface functional groups and therefore work cooperatively to maximise the binding surface at the bio-nano interface. Consequently, the C2GO-adsorbed protein is stabilised and is not pushed to structural deformation as is the case in GO-adsorbed protein. The results from this analysis pipeline suggest the causes of protein aggregation and cellular uptake are mediated by the aforementioned changes in protein structure.

This work indicates that the dynamic and functional role of adsorbed proteins may be a more important probe to understanding the protein corona, rather than quantifying a static image of its constituent proteins. MD simulations are a valuable tool for such an investigation and our analysis pipeline serves as a transferable method for understanding the structure/function relationship of dynamic protein-nanomaterial adsorption.

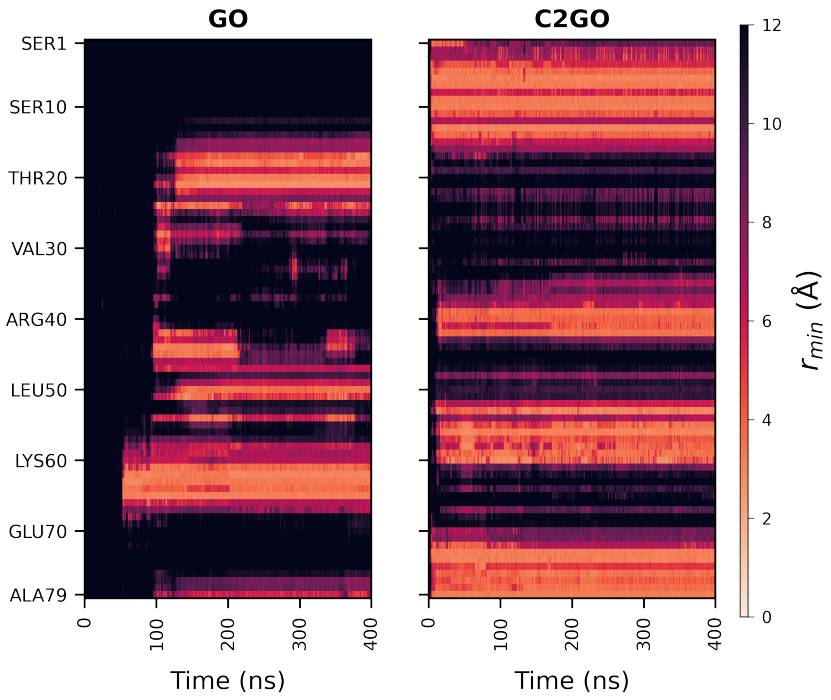


Figure 4.5: Minimum distance to any GO/C2GO heavy atom for each residue in apo-C3.

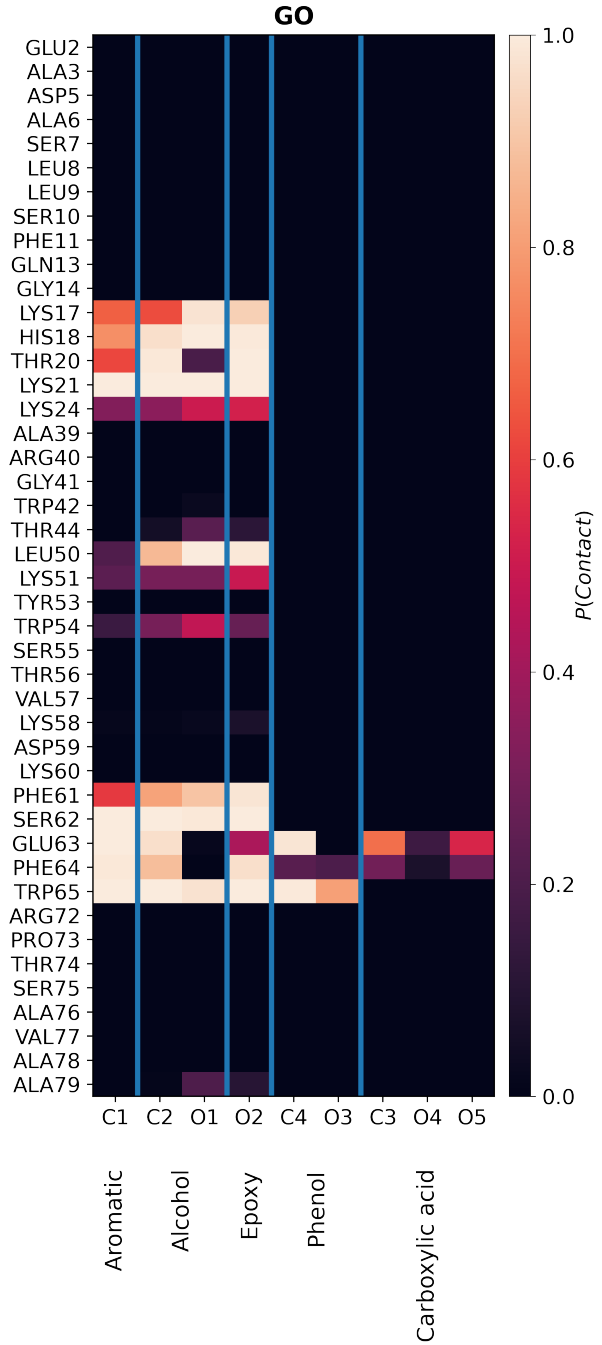


Figure 4.6: Contact probability between each apo-C3 residue and each atom type of GO. Data from the final 50 ns of the trajectory. For clarity, only those residues with $P(\text{Contact}) \geq 0.1$ for either GO or C2GO are shown.

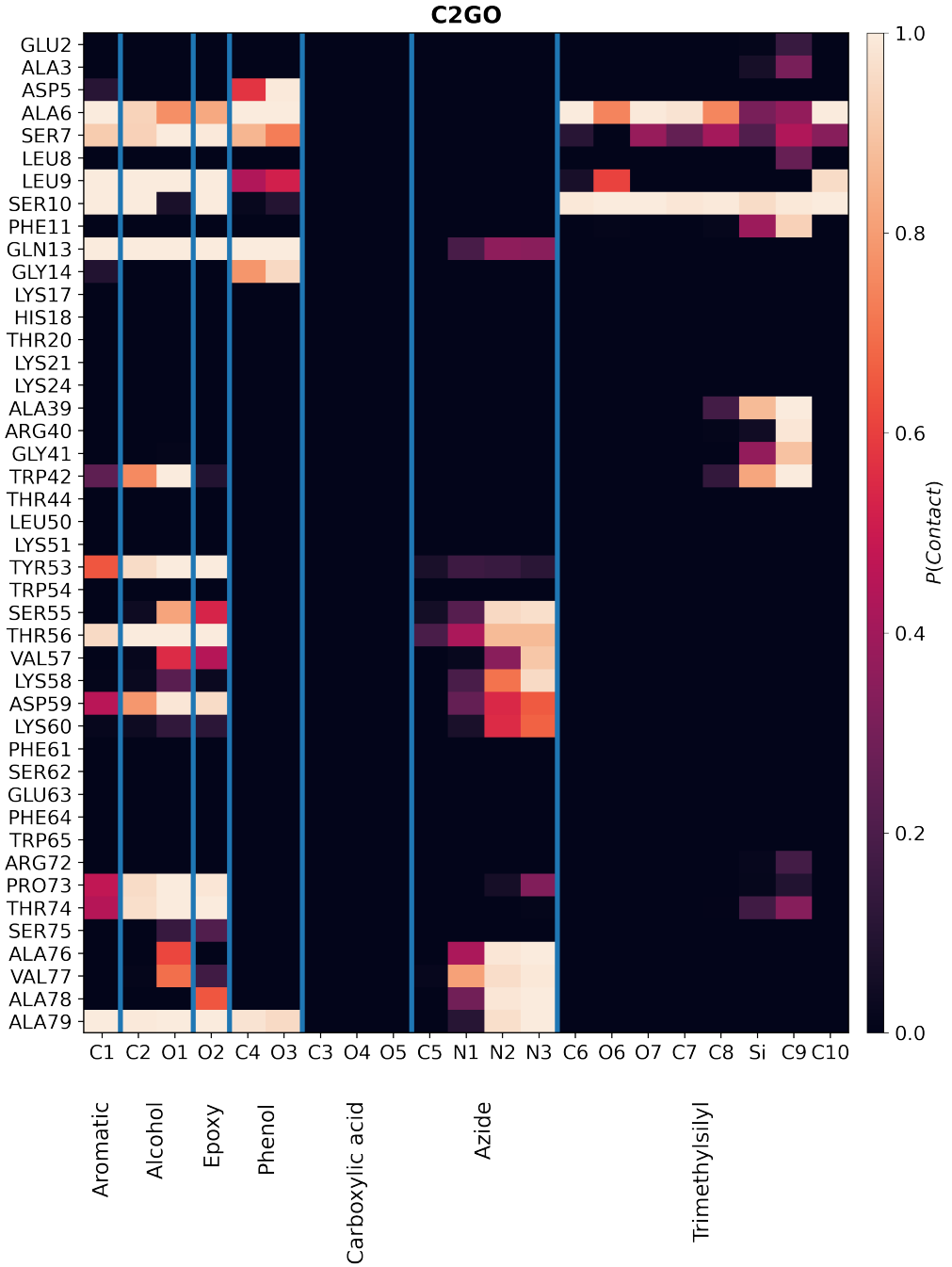


Figure 4.7: Contact probability between each apo-C3 residue and each atom type of C2GO. Data from the final 50 ns of the trajectory. For clarity, only those residues with $P(\text{Contact}) \geq 0.1$ for either GO or C2GO are shown.

Chapter 5

Allosteric regulation of the SARS-CoV-2 main protease

During the coronavirus disease 2019 (COVID-19) pandemic, the scientific community widely attempted to assist in the characterisation of potential therapeutics for the treatment of the disease. With the help of efforts to crystallise SARS-CoV-2 proteins, the molecular dynamics community was able to study structures of SARS-CoV-2 viral proteome. In this work, we were able to identify a method to computationally accelerate the disruption of the catalytic dyad in the SARS-CoV-2 main protease. The mechanism of dyad disruption is an important factor to identifying an inhibitor for the SARS-CoV-2 main protease, the implementation of the metadynamics bias applied in this work can greatly reduce the simulation times of continuing drug discovery efforts.

Contributions for this work are as follows: **Mohamed Ali al-Badri** and **Khaled Abdel-Maksoud** conceived and planned the research. **Mohamed Ali al-Badri** performed the molecular dynamics calculations. **Khaled Abdel-Maksoud** performed the metadynamics calculations. **Mohamed Ali al-Badri, Khaled Abdel-Maksoud, Christian D. Lorenz** and **Jonathan W. Essex** analysed the data and **Mohamed Ali al-Badri** and **Khaled Abdel-Maksoud** prepared the final manuscript.

The Coronavirus Disease of 2019 (COVID-19) is caused by a novel coronavirus known as the Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2). The main protease (Mpro) of SARS-CoV-2 mediates viral replication through proteolytic activity and the subsequent generation of infectious virus particles. Current computational efforts towards SARS-CoV-2 inhibitor design generally neglect an allosteric mechanism linked to His41-Cys145 catalytic dyad disruption and thus do not target the open conformational state. We identify that the orientation of the His41 imidazole side chain away from Cys145 entails an allosteric mechanism for mediating activity. In this work, we show that molecular dynamics and metadynamics simulations are fundamental for performing computer-aided inhibitor design where the sampling of this allosteric mechanism within a computationally feasible timescale is essential. We calculate a 4.2 ± 1.9 kJ/mol free energy difference between the open and closed states of the SARS-CoV-2 active site, indicating that favourable ligand interactions with His41 over the Cys145-His41 dyad interaction can stabilise the open state.

5.1 Introduction

Coronaviruses have proven to be a challenge for drug discovery since Severe Acute Respiratory Syndrome Coronavirus 1 (SARS-CoV-1), better known as SARS (2002-2004) and the Middle East Respiratory Syndrome (MERS-CoV) (2012-2013). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the cause of the coronavirus disease 2019 (COVID-19). It is a positive-sense single-stranded RNA virus that is contagious in humans, with zoonotic origins and genetic similarities to bat coronaviruses.[158] The virus mainly enters human cells through the receptor angiotensin converting enzyme 2 (ACE2).[159]

To date, more than 21 million cases of COVID-19 have been confirmed, with more than 760,000 deaths worldwide.[160] SARS-CoV-2 has a combi-

nation of high transmissibility, longer incubation period and a much shorter interval between symptom onset and maximum infectivity when compared with SARS and MERS-CoV.[161] As a result, even with a relatively low mortality rate, COVID-19 is proving much harder to eradicate and will therefore remain an epidemiological problem until a therapeutic agent is developed.

Broad spectrum antiviral medication such as remdesivir have been proposed to decrease SARS-CoV-2 RNA production, however remdesivir was not associated with statistically significant clinical benefits.[162] Recent clinical research has shown that administration of dexamethasone[163] or systemic corticosteroids,[164] were found to be associated with lower 28-day all-cause mortality when compared with usual care or placebo.

The most promising tools for the cessation of the epidemic spread of COVID-19 are vaccines, with many in latter stages of clinical trials, that are expected to be available in late 2020 or early 2021.[165] The majority of these vaccines are based on platforms such as inactivated viral vectors or RNA sequences encoding the spike glycoprotein of SARS-CoV-2 that trigger an immunogenic response. The SARS spike protein has been identified as the major target of selective pressure in the adaptive evolution of SARS coronaviruses.[166] Recently, molecular dynamics (MD) simulations have been used to study the druggability of the SARS-CoV-2 spike.[167] They have identified vulnerabilities in the spike glycan shield — utilised to frustrate an immune response — that can be harnessed for vaccine development.

Protease inhibitors, however, do not depend on an immunogenic response to elicit immunity. Unlike immunogenic approaches, any inhibitor identified for the SARS-CoV-2 would very likely also serve as an inhibitor of further evolution of this virus as the sequence and structure of the

are closely related to those from other betacoronaviruses.[168] Elsewhere, protease inhibitors have been extensively used for the treatment of HIV-AIDS[169, 170] and hepatitis-C[171]. Molecular dynamics (MD) simulations were paramount in identifying the dynamic bound and free states of the HIV-1 protease flaps (two glycine-rich β-hairpins) that cover a large substrate-binding pocket used as a target for antiviral drugs.[172] Also, MD simulations were integral in identifying a cryptic trench within the HIV integrase, which became the target for the first FDA approved HIV integrase inhibitor (raltegravir).[13, 169]

The SARS-CoV-2 is largely responsible for the proteolytic processing of the polyproteins transcribed by the SARS-CoV-2 genome, which are responsible for viral transcription and replication. cleaves the polyproteins at 11 conserved sites using the catalytic dyad.[173, 174] A catalytic dyad is a set of two coordinated amino acids, common to some enzyme active sites. Within a histidine-cysteine (His-Cys) dyad, the His amino acid will act as a base and activate the Cys mercaptan as a nucleophile for polypeptide cleavage.

Owing to the high sequence similarity, the conservation of primary structure about the catalytic sites and the high degree of tertiary structure similarity between both proteases, it has been surmised that the His41-Cys145 catalytic dyad plays the same role of regulating protease activity within both SARS-CoV proteins.[174, 168] Disrupting this dyad, then, disables activity and subsequent virus replication. In order to be able to cleave the SARS-CoV-2 polyprotein, hydrolysis must be facilitated by priming the Cys145 mercaptan group for nucleophilic attack via deprotonation by the His41 imidazole group. The proteolytic mechanism in the SARS-CoV active site regulates protease activity through activation of the Cys145 mercaptan. The rotation of the His41 imidazole towards Cys145 serves as an allosteric

trigger to inducing proteolytic activity in SARS-CoV . One way to disrupt the proteolytic mechanism is to induce a conformational change in the more flexible His41 imidazole side chain of the dyad and prevent it from abstracting a proton from the Cys145 mercaptan group. An inhibitor would serve to stabilise this disrupted conformation and therefore inhibit activity.

Recently, MD simulations have been used to study the druggability of the SARS-CoV-2.[175, 176] However, conventional MD fails to disrupt the catalytic dyad and thus deactivate activity within computationally achievable timescales. As such, screening potential ligands from vast drug libraries to inhibit is inaccurate and could fail to identify potent inhibitors. Instead, metadynamics (MetaD) enhances the sampling of rare events to reconstruct the free energy landscape by discouraging revisiting of sampled states. It is a useful tool for studying mechanisms of drugs binding to flexible targets where conventional MD may otherwise fail to ergodically sample the free energy landscape. Defining the free energy landscape of a complex simulation is non-trivial and depends on a choice of a few collective variables (CV).[177] Other enhanced sampling techniques (Gaussian accelerated MD[178]) have been applied in elucidating cryptic pockets not detectable from the crystal structure, identifying additional pockets for studying inhibition beside the active site.[179]

In this work, we apply MetaD to identify the allosteric mechanism of SARS-CoV-2 using the inhibition of SARS-CoV-1 with a previously identified potent inhibitor (D3F)[180] (Fig. 5.1A). In order to comprehensively study whether a contender ligand successfully inhibits the SARS-CoV-2, the allosteric mechanism must be sampled within feasible timescales. This sampling is difficult to achieve using conventional computational methods as there is a high free energy cost to dyad disruption that these approaches cannot account for. We show that the allosteric mechanism can be sampled

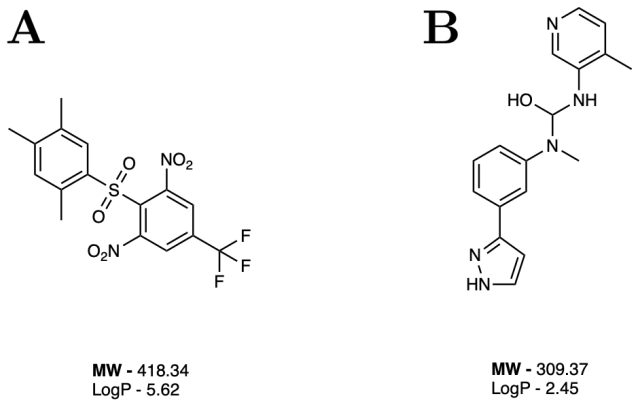


Figure 5.1: Inhibitor candidate ligand structures with molecular weight (MW) and calculated lipophilicity (LogP). (A) D3F, a strong binder and inhibitor of SARS-CoV-1. (B) The drug candidate (code named LIG herein) considered for binding and inhibition of SARS-CoV-2. LogP values were calculated using the ChemDraw LogP estimation tool.

using MetaD simulations through biasing rotations about a single dihedral within the dyad, and illustrate this by incorporating a contender ligand (LIG) (Fig. 5.1B) to SARS-CoV-2.

5.2 Results

5.2.1 Allosteric regulation of SARS-CoV-1 and SARS-CoV-2 activity is linked to the His41-Cys145 interaction

Co-solvent MD simulations [181] are a useful tool for identifying binding hotspots on protein surfaces by simulating proteins in a solution of water and co-solvent molecules. This dynamic approach to identifying binding sites incorporates proteins' inherent flexibility, allowing the cosolvent molecules to compete with water to bind to the protein surface.[182] Our co-solvent MD simulations were initiated by placing five LIG molecules randomly surrounding SARS-CoV-2 . Of the five LIG molecules, only one entered the binding site of SARS-CoV-2 in order to stabilise a binding mode (Fig. 5.2).

A short simulation of SARS-CoV-1 with the inhibitor D3F bound in the

active site was also performed and was used to evaluate the success of the observed LIG binding in inhibiting SARS-CoV-2, as the initial D3F binding mode has been proven to successfully inhibit SARS-CoV-1 in previous studies.[180] When observing the bound state of D3F to SARS-CoV-1, the nitrate group most proximal to His41 forms a strong electrostatic interaction between a D3F nitro O atom and the N1-H of the His41 imidazole (D3F N-O · · · H-N1 His41) (Fig. 5.2A). This interaction stabilises the orientation of His41 away from Cys145 which comprises the open (“holo") conformation of. Despite the ligand appearing as bound within the SARS-CoV-2 active site, His41 is not seen to interact with the bound ligand at all (Fig. 5.2B). The primary difference between the binding modes of D3F to SARS-CoV-1 and LIG to SARS-CoV-2 is the orientation of the His41 imidazole with respect to Cys145. In the latter case, the His41-Cys145 catalytic dyad is maintained and thus the is within a closed (“apo") conformation. The disruption of the catalytic dyad via reorientation of the His41 imidazole or interaction with the Cys145 mercaptan side chain[174] therefore serves as a prospective diagnostic tool for successful inhibition.

5.2.2 Generating the SARS-CoV-2 active state via His41 side chain reorientation

The disruption of the catalytic dyad to promote ligand binding was further explored through the use of enhanced sampling simulations. The binding mode of D3F to SARS-CoV-1 indicates an ability to achieve inhibition of SARS-CoV-2 through promoting interactions with the His41 imidazole in the holo form. Metadynamics (MetaD) is herein employed with a bias potential over the χ_1 dihedral of His41 in order to induce rotation of the His41 side chain imidazole group about the χ_2^{backbone} dihedral. The χ_2^{backbone} dihedral cannot be used as a CV due to an intrinsic conformational restriction, since the dihedral is part of the protein backbone. Instead, we have applied a bias to the χ_1 dihedral torsion, which allows us to investigate the free rotational orientation of the His41 side chain imidazole (Fig. 5.3).

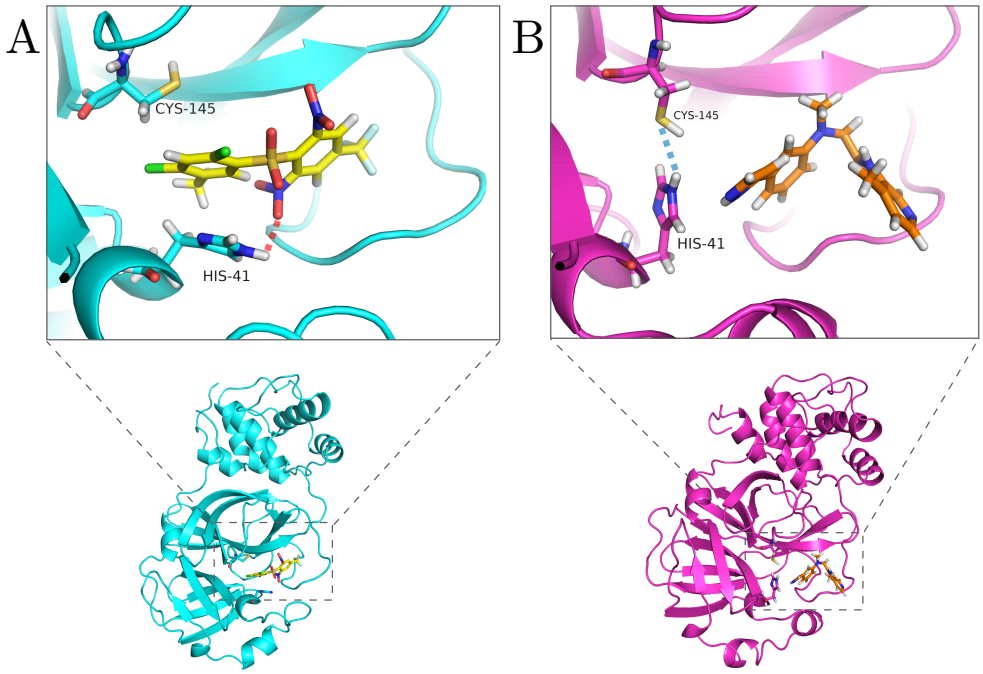


Figure 5.2: The observed binding modes of (A) D3F (yellow) within the SARS-CoV-1 catalytic binding site (cyan), showing the disrupted catalytic dyad (“holo") and the strong D3F interaction with His41 (dashed red line) and (B) LIG (orange) within the SARS-CoV-2 catalytic binding site (pink) with a maintained catalytic dyad (“apo"). The dyad residues and their interactions (dashed blue line) are labeled His41 and Cys145.

Ergodic sampling along the χ_1 torsional CV space was achieved using MetaD simulations. The free energy surface with respect to the χ_1 dihedral was calculated and is shown in (Fig. 5.4, top). This surface was used to verify metadynamics convergence by evaluating the relative free energy differences between pairs of free energy minima, which stabilised after 70 ns (Fig. 5.4, bottom).

5.2.3 catalytic dyad conformational changes

In order to verify that our choice of CV of χ_1 torsion couples to and changes the degree of χ_2^{backbone} torsion, we compare the probability densities found from our MetaD simulation to the probability distributions observed in our co-solvent MD simulations of LIG-SARS-CoV-2 and the D3F-SARS-

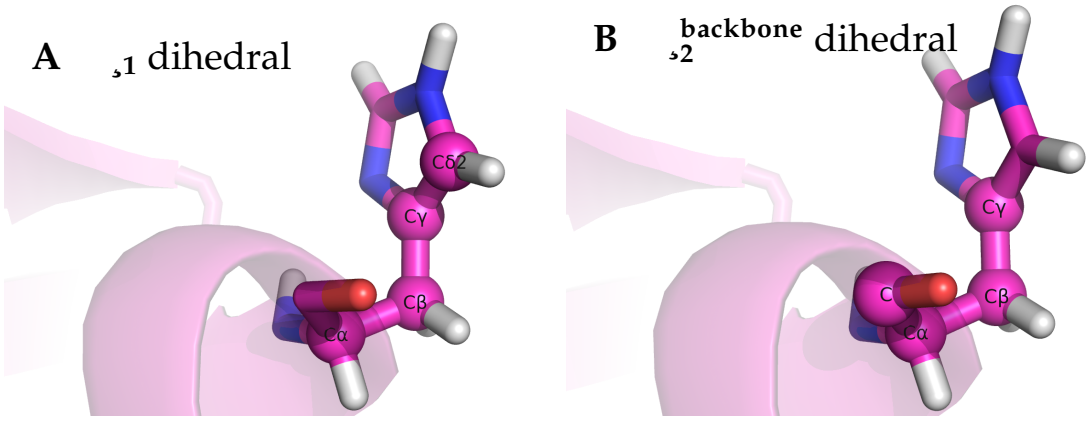


Figure 5.3: His41 imidazole side chain dihedrals. (A) χ_1 dihedral used for construction of the MetaD bias. (B) χ_2^{backbone} dihedral considered within analysis of the performance of the bias and characterising the disassociation of His41 from Cys145.

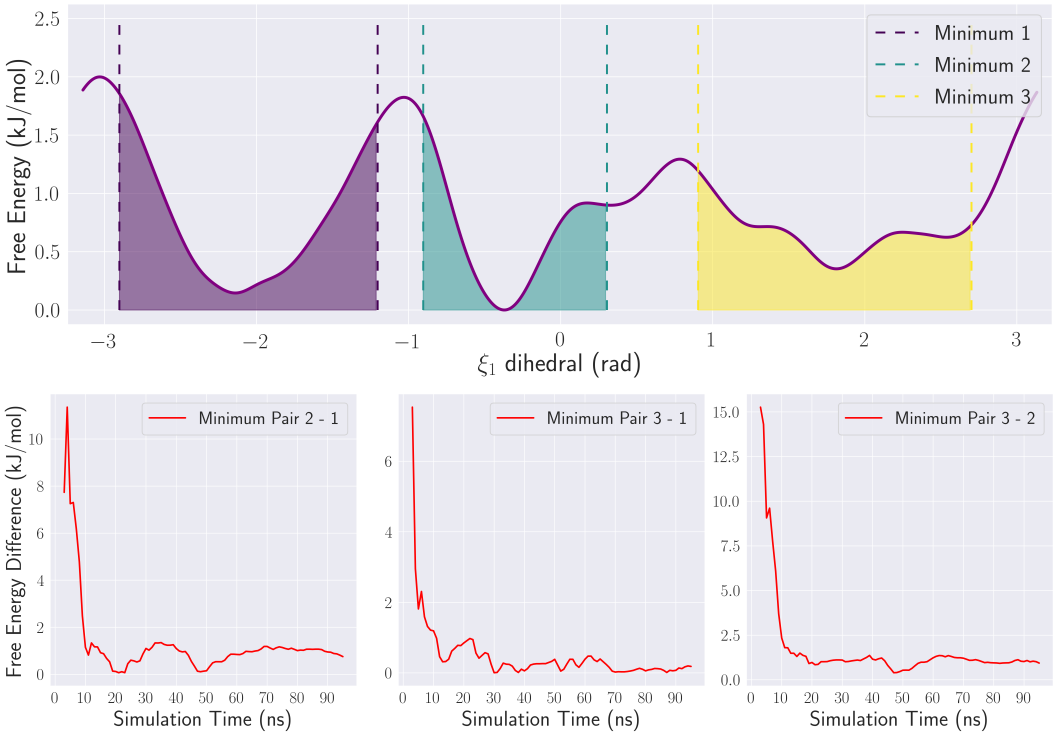


Figure 5.4: Analysis of convergence of His41 imidazole dihedral MetaD bias potential. Potential of mean force (PMF) projected on the χ_1 dihedral space of the bias potential. Relative free energy differences (in kJ/mol) were calculated between minima indicated by the dashed lines (top). Relative free energy difference between each pair of minima defined within the above PMF over the simulation time (bottom).

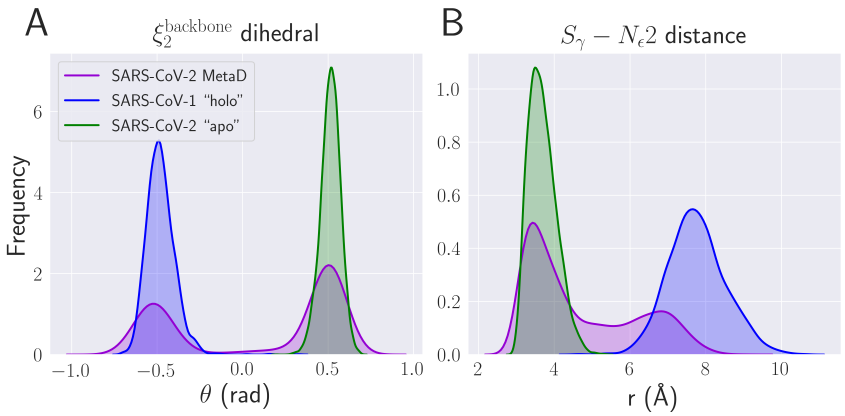


Figure 5.5: 1D probability density functions obtained from the co-solvent MD simulations of SARS-CoV-1 "holo" (blue), SARS-CoV-2 "apo" (green) and SARS-CoV-2 MetaD simulation (purple) over the (A) ξ_2^{backbone} dihedral space and (B) (Cys145-S$_\gamma$)-(His41-Nϵ) distance.

CoV-1, which we use as the "apo" (closed conformation) and "holo" (open conformation) states, respectively. In (Fig. 5.5A), the MetaD probability distribution clearly samples both "apo" and "holo" states. In (Fig. 5.5B), this choice of ξ_1 torsion CV also biases the the catalytic dyad into an open state by orienting the His41 imidazole away from Cys145, measured through the (His41-N$_{\epsilon 2}$)-(Cys145-S$_\gamma$) distance.

To demonstrate the interconnection of the ξ_2^{backbone} torsion and (Cys145-S$_\gamma$)-(His41-Nϵ) distance, we can represent the sampling of co-solvent MD simulations ("holo" - (Fig. 5.6A) and "apo" - (Fig. 5.6C)) with the interchangeable MetaD sampling of both "apo" and "holo" states in a 2D space (Fig. 5.6B). Comparing against the three density functions, it is noted that the region about ξ_2^{backbone} = -0.5 rad overlaps with the "holo" state, while the region about ξ_2^{backbone} = 0.5 rad overlaps with the sampled density of the "apo" in which the catalytic dyad remains intact throughout. This comparison allows for the characterisation of each of the respective regions in the metadynamics density function above to be considered as "apo" and "holo" conformational states (Fig. 5.6B).

In (Fig. 5.6B), the bimodal probability density in the 2D CV space includes points at shorter (Cys145-S_γ)–(His41-Nϵ) distances at $\zeta_2^{\text{backbone}}$ = -0.5 rad. This region indicates that, in the absence of a potent inhibitor, the "holo" state cannot be stabilised by biased ζ_1 sampling alone. As such, the dyad will reorient to maintain the Cys145-His41 side chain interaction.

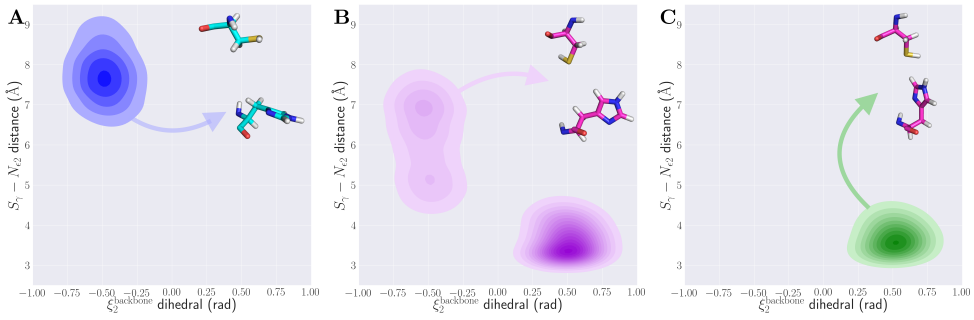


Figure 5.6: Probability density functions of the (A) SARS-CoV-1 "holo" , (B) SARS-CoV-2 MetaD (inset showing MetaD "holo") and (C) SARS-CoV-2 "apo" MD simulation defined within a 2D CV space of the His41 $\zeta_2^{\text{backbone}}$ dihedral and the (Cys145-S_γ)-(His41-Nϵ) atomic distance.

To confirm that the choice of a ζ_1 bias samples the "apo"–"holo" transition, the time-dependent behaviour of both the $\zeta_2^{\text{backbone}}$ dihedral and the (Cys145-S_γ)-(His41-Nϵ) distance was evaluated using K-means clustering.[156] Three distinct clusters were found (Fig. 5.7) corresponding to the "apo" conformation ($\zeta_2^{\text{backbone}}$ = 0.5 rad) , "holo" conformation ($\zeta_2^{\text{backbone}}$ = -0.5 rad) and transition region. The time dependent behaviour of $\zeta_2^{\text{backbone}}$ dihedral and (Cys145-S_γ)-(His41-Nϵ) distance show the choice of CV freely samples the reversible "apo"–"holo" transition throughout the 100 ns trajectory.

5.2.4 free energy surfaces

In order to verify the reproducibility of the constructed bias, four replicas of the MetaD simulation were performed until convergence was observed in each. The free energy profiles obtained in each of the replicas show agree-

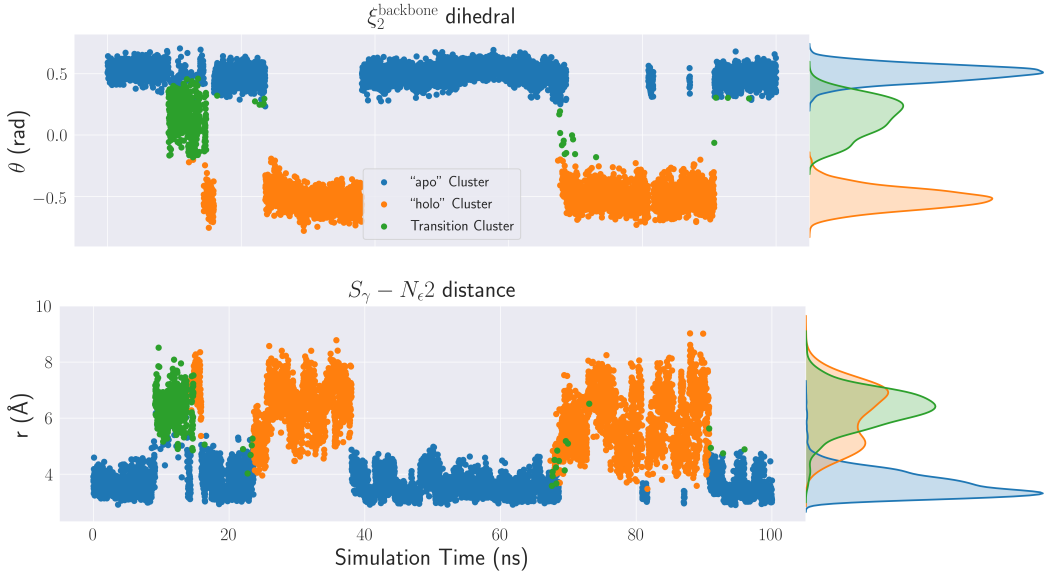


Figure 5.7: Time evolution of the MetaD simulation trajectory and the associated cluster identity of points with respect to the His41 χ_2^{backbone} dihedral (top) and the (Cys145-S_γ)-(His41-Nϵ) distance (bottom).

ment, with deviation from the average free energy values observed chiefly at the transition states and at the minimum at χ_1 = 1.9 rad. The deviation at this minimum is due to the corresponding state being sparsely sampled in all replicas compared to the other system states as a result of the high free energy barrier restricting sampling from other observed minima (Fig. 5.8).

As convergence of the bias was observed for all four replicas, it can be assumed that any sampling beyond the point of convergence is within the desired thermodynamic ensemble. Thus, the mean free energy surface with respect to the His41 χ_2^{backbone} dihedral over the 4 replicas was computed (Fig 5.9). The mean surface is bimodal, with positions of the minima at 0.5 and -0.5 rad corresponding the the sampled “apo" and “holo" states of the SARS-CoV-2 catalytic dyad, respectively. The “apo"-“holo" relative free energy difference of 4.2 ± 1.9 kJ/mol (1.0 ± 0.5 kcal/mol) indicates that the “holo" state is less energetically stable than the “apo" state in the absence of any ligands interacting with His41. Considering this free energy difference between states and the high free energy barrier (15.0 ± 1.3 kJ/mol /

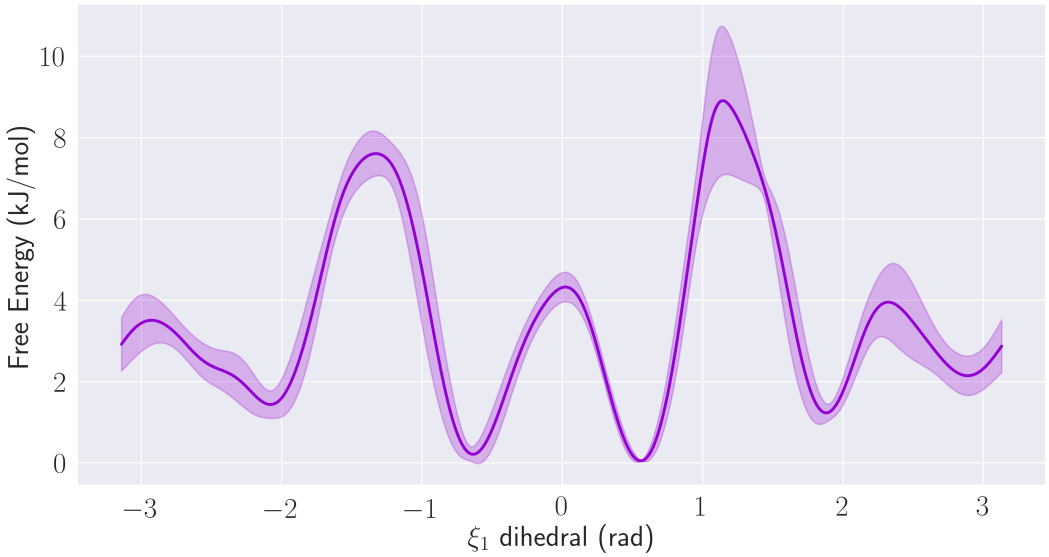


Figure 5.8: Mean free energy surface obtained over 4 replicas of His41 torsional metadynamics defined within the ¸1 dihedral space. The shaded region corresponds to the standard deviation about each free energy point calculated over the set of replicas.

3.6 ± 0.3 kcal/mol) between them, it can be assumed that the dyad “apo"–“holo" transition is unlikely to be sampled using unbiased conventional MD simulations alone.

The free energy surface of SARS-CoV-1 “holo" (dashed blue line) was computed from the reference simulation of the D3F-SARS-CoV-1 complex using population analysis (Fig. 5.9). This surface is used as a reference state to compare against the mean MetaD free energy surface as the “holo" state. The SARS-CoV-1 “holo" conformation is more energetically stable than the SARS-CoV-2 MetaD surface “holo" minimum by approximately 4.2 ± 1.9 kJ/mol (1.0 ± 0.5 kcal/mol). This energy difference highlights how the energetic penalty associated with dyad disruption in the “apo"–“holo" transition may be recovered by ligand interactions in the active site. All MetaD replicas yielded sampling of the two rotamers of His41 about the backbone dihedral. The larger error about the “dyad disrupted" conformation at -0.5 rad (Fig. 5.9) is due to sampling with a ligand-free active site for, meaning that less stable conformation isn’t stabilised by ligand interactions unlike

the case of D3F binding.

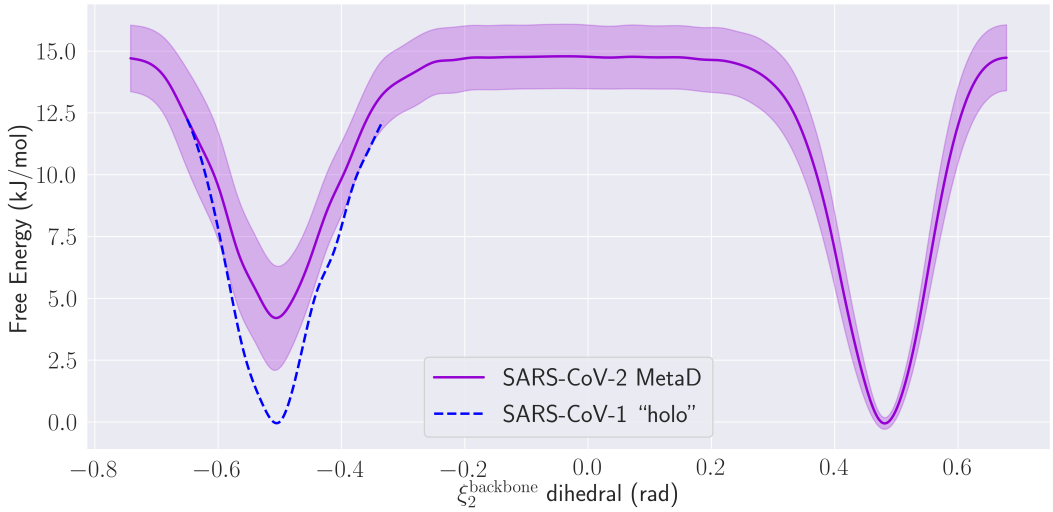


Figure 5.9: Mean free energy surface obtained from the MD simulation of D3F-SARS-CoV-1 "holo" (blue) and four replicas of His41 torsional MetaD simulations (purple), defined within the χ_2^{backbone} dihedral space. The shaded region corresponds to the standard deviation about each free energy point calculated over the set of MetaD replicas.

5.3 Methods

5.3.1 Molecular Dynamics

All MD simulations were performed in GROMACS version 2019.4 on the ARCHER Cray XC30 supercomputer on a single 2.7 GHz, 12-core E5-2697 v2 (Ivy Bridge) series processor node, NVIDIA GeForce RTX 2060 or GeForce GTX1080Ti GPUs with the CUDA 10.2 toolkit. The AMBER14SB forcefield [183] was used to model the system, where ligand molecule forcefield parameters were generated using the General Amber ForceField v2 (GAFF2) with charges calculated using the AM1-BCC semi-empirical method.[184].

The structure of the candidate ligand LIG was derived from a virtual fragment expansion, docking, and screening exercise provided by Gabriel

Grand, Elana Simon, Michael Bower, Bruce Clapham and Jonah Kallenbach of Reverie Labs.[185] Protein X-Ray diffraction (XRD) structures for SARS-CoV-1 holo structure (PDB code: 2GZ7) [186] and SARS-CoV-2 apo structure (PDB code: 5RE4) were used for these simulations. Missing residues from the SARS-CoV-2 XRD structure were modelled using MODELLER [187] and subsequently solvated in a 90 cubic box of TIP3P water molecules.

For co-solvent MD simulations, five of the candidate ligand (LIG) were placed in random coordinates within the box and the system net charge was neutralised by adding 4 sodium ions into the system. The co-solvent MD simulations was performed for 100 ns. The simulation of D3F bound in SARS-CoV-1 was performed for 10 ns. The net charge in the system was neutralised by adding 3 sodium ions. In all structures prepared for simulation, the His41 side chain was maintained at the N1-H tautomeric state, as this state is primed for nucleophillic attack from the Cys145 mercaptan in the active catalytic dyad.

The system was relaxed energetically using steepest-descent energy minimisation for 50000 steps with an energetic step size of 0.01 kJ/mol. The minimisation was terminated after the maximum energetic contribution was lower than a threshold of 10.0 kJ/mol. NVT and NPT equilibration was performed for 1 ns using two separate velocity-rescaling thermostat coupling temperature to velocities for protein, drug and solvent molecules (NVT), where a temperature of 300 K was maintained and 1 bar using the Parrinello-Rahman barostat (NPT).[66] The Verlet cut-off scheme was employed to generate pair lists and the electrostatic interactions were calculated using the Particle-Mesh Ewald algorithm.[188] Both electrostatic and van der Waals interactions were cut off beyond 1.2 nm. All bonds involving hydrogen atoms were constrained using the LINCS algorithm.[189] Production simulations ran with an integration stepsize of 2 fs. MDAnalysis was

used to postprocess the MD trajectories for analysis.[190, 191]

5.3.2 Metadynamics

Four independent replicas of the non-tempered MetaD simulation were run to convergence.[192] The bias potential was setup to sample energetically hindered rotations about the χ_2^{backbone} dihedral of His41 by biasing the sampling along the χ_1 torsional profile directly. The bias was accumulated with a Gaussian deposition rate $\tau = 1$ ps. The deposited Gaussians had a fixed height of 0.1 kJ/mol and employed an adaptive width scheme in which the correlation between the biased CV space and the microscopic configuration space is utilised to recalculate the covariance matrix. A correlation length of 0.5 was used.[193] MetaD simulations were performed using Plumed 2.5.4.[194]. All the data and Plumed input files required to reproduce the results reported in this paper are available on PLUMED-NEST (www.plumed-nest.org), the public repository of the PLUMED consortium [194], as plumID:20.027.

5.4 Conclusions

We have presented the application of metadynamics to molecular dynamics simulations of the SARS-CoV-2 to better inform the drug discovery efforts including virtual screening, molecular docking and unbiased molecular dynamics simulations. We show that the proteolytic mechanism of is contingent on the integrity of the His41-Cys145 catalytic dyad and that the rotation of the His41 imidazole side chain to Cys145 acts as an allosteric trigger to regulating this proteolytic activity.

Using metadynamics, we find that promoting ligand binding and sampling of the active site of the protease is achieved through disrupting the catalytic dyad by biasing over the His41 χ_1 dihedral to subsequently sample over the χ_2^{backbone} dihedral. Using the (Cys145-S_γ)-(His41-Nϵ) distance and

the χ_2^{backbone} dihedral to define the collective variable (CV) space, we identify clusters of unbound, intermediary and bound conformations of the flexible active site throughout the simulation. We show that repeated replicas of the His41 torsional metadynamics reproduce the free energy surface along the 1D χ_1 and χ_2^{backbone} spaces. Our results detail the allosteric regulation of the SARS-CoV-2, irrespective of the choice of ligand. The candidate ligand LIG acts as a toy model and its inadequacy in disrupting the catalytic dyad was established by drawing comparisons of its interaction in the active site against the ligand D3F with SARS-CoV-1.

The application of our analysis uncovers the changes on the receptor structure as a result of an allosteric mechanism resolved using enhanced sampling. This observation can assist in selecting an optimal strategy for screening ligands from drug libraries. The free energy comparison between the D3F unbound and bound “holo" state minima suggests that the energy expended in disrupting the dyad can be readily recovered using ligand interactions in the active site, and so this open dyad state should be considered alongside the closed state as a target for virtual screening. We provide coordinates for structures corresponding to the open catalytic dyad state of SARS-CoV-2 as supplementary data for use in further studies.

Chapter 6

Conclusions

I am leaving the regions of fact, which are difficult to penetrate, but which bring in their train rich rewards, and entering the regions of speculation, where many roads lie open, but where a few lead to a definite goal.

William Ramsay

Using the methods described throughout this thesis, I have attempted to employ accurate multiscale modelling to address problems within the domain of biology where conventional investigation techniques may be less suitable. This work collectively shows the importance of a multipronged approach to accurate modelling, namely the interdependence of the correct characterisation of structure with the accuracy of the theoretical treatment of a model.

I have used density functional theory as a basis to derive a quantum mechanical forcefield for accurate classical molecular dynamics simulations of a graphitic nanomaterial composed of around 1000 atoms. Using this, we investigated the impact of the correlated semi-ordered functionalisation as well as the bespoke forcefield parameters on the nanomaterial's interactions

with an ionic solution and its chaotropic potential. This study illustrated the capacity for extrapolating bespoke forcefield parametrisation methods to very large molecular systems, without the need to sacrifice accurate characterisation, which we show is principal to reproducing both experimental and state of the art *ab initio* measurements. Having developed an accurate forcefield model for the system, it is extendable to MD simulations of hundreds of thousands of atoms.

Using the accurate structural model of graphitic materials, we used our model to address an open question relating the atomistic interactions at the bio-nano interface. The formation of the protein corona is an obstacle to effectively translating the technological advancements of nanomaterials to biotechnology. The dynamic character of the protein corona has made it a challenging problem to tackle using conventional computational and experimental methods alike. The investigation reported in this thesis utilised numerous analysis techniques to form a collective analysis pipeline to unpick the impact of nano-functionalisation on adsorbed protein structure. Using our results, we were able to make sense of experimentally observed behaviours of the much larger protein corona from a single adsorbed common serum protein. These findings included the way in which some nanomaterial functional groups induce the adsorbed protein to denature its tertiary structure and form binding motifs for protein aggregation; this instigates the formation of a protein corona as observed in experiment. Furthermore, using our analyses we explained the experimentally observed contrast in cellular uptake between different nanomaterial-corona complexes, differing only by functionalisation type. Our analyses showed the impact of conserving functionally important sequences in the protein on the cellular uptake of the nanomaterial-protein complex. Much of the observed behaviour was sensitive to the nanomaterial characterisation, again reinforcing the importance of accurate modelling to associate molecular modelling with

experiment.

The accurate atomistic modelling of a protein active site, aided by data analysis techniques to unscramble and interpret the dynamics, formed the basis for understanding the role of a catalytic dyad in the function of the SARS-CoV-2 main protease. The protease plays the role of cleaving the viral polyprotein, inhibiting its function therefore disables subsequent viral replication. Unlike the now prevalent preventative immunogenic approaches, protease inhibitors do not require an immunogenic response and can be used to treat both severely afflicted or immunocompromised patients. Unlike many computational approaches to drug discovery, MD simulations clarify the change in dynamics of the protein due to the influencing presence of an inhibitor. Without the previous identification of a potent inhibitor for the similar active site of the SARS-CoV-1 main protease, we would not have been able to unpick the mechanism of the catalytic dyad disruption and use it to calculate the free energy of binding. Since that work was written, the continued efforts to identify protease inhibitors through X-ray crystallography screening have observed numerous ligands that disrupt the His41-Cys145 catalytic dyad. Accelerating the rare-event sampling of the dyad disruption is achieved through the use of Metadynamics simulations, which is completely transferable and open to use by other MD-based projects. This can be used to sample the catalytic dyad disruption within computationally feasible timescales, in order to perform high-throughput screening for as many molecules as possible while retaining the advantages of modelling a dynamic target.

The use of electronic structure calculations that account for strong electronic correlations have mostly been reserved for small molecules or periodic systems. We have translated the hybrid DFT+DMFT method to a molecular system to study the active site of the hemocyanin oxygen-

transporting protein found in the hemolymph of some invertebrates. It is the first application of cluster DMFT to a biological system, where it identifies the electronic structure of an uncommon open-shell singlet ground state. The singlet ground state is a quantum-entangled superposition of two localised magnetic moments on two distant copper atoms in the hemocyanin active site. This investigation details the function of a superexchange mechanism where electron hopping between the copper d-shells is mediated by the bridging dioxygen ligand p-orbitals. Its role in reversible oxygen binding is inevitable but is yet to be described using this new information.

This work stresses that it is biology that defines the warranted accuracy with which it is to be treated, in line with experimental results to which we return for verifiability. In the pursuit of accuracy, no single theoretical approach can be used to tackle an array of questions aimed at the same class of biomolecular systems. The progress of interdisciplinary collaboration, the advancement of open-source scientific software and increasing experimental data all coalesce into a flourishing environment for collaborative scientific research to advance the multiscale modelling of biomolecular systems and their interface with state-of-the-art nanomaterials.

6.1 Future work

Having demonstrated the implementation of theoretical approaches to the biological domain in different applications, an extension of this work would build on the most significant divergence from experiment in preexisting methods. The parametrisation of forcefield parameters for graphitic materials is well suited to describing the interfacial properties of a complex chemical environment where strong electrostatic interactions drive the interfacial phenomena. As such, the model could be extended to study the interactions with lipid bilayer systems for which there is abundant demand for multiscale modelling. Furthermore, the availability of linear-scaling electronic

structure calculations and forcefield parametrisation software can be used to extend the accuracy to much larger systems in the biological domain. Unlike the predominant application of the interface between biomolecular systems and bespoke forcefield parametrisation of small molecules in the drug discovery domain, the work in this thesis describes its extension to the rapidly growing field of biotechnology.

The accurate characterisation and function of the SARS-CoV-2 main protease active site as well as the acceleration of rare-event sampling has been described. This particular work can be extended into an automated pipeline for high-throughput screening of inhibitor ligands. Given the time and computational limitations, this extension remained outside our capacity but nonetheless forms a qualitative proof of principle of how accurate modelling can inform drug delivery efforts for a subset of inhibitor ligands that require the rare-event disruption of the catalytic dyad.

The strongly correlated ground state of the hemocyanin active site was resolved using DFT+DMFT and can be extended to other metalloproteins with quantum-mechanically driven biological function. Interestingly, the accurate characterisation of the electronic structure of the hemocyanin active site following this work was extended to bio-mimetic applications in the selective catalytic reduction of nitrogen-oxide pollutants.[195]

The field of biotechnology will necessitate the accessible coherent research from both computational and experimental perspectives. Accurate tools are therefore an invaluable requisite for the advancement of the field where the bio-nano interface is concerned. The concomitant advancement of computational architectures and *good* scientific software will improve accessibility to flawlessly research and inform the design of biotechnological solutions.

Bibliography

[1] Alan M. Turing. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1):230–265, 1937.

[2] Nicholas Metropolis. The beginning. *Los Alamos Science*, 15:125–130, 1987.

[3] Enrico Fermi, John Pasta, Stanislaw Ulam, and Mary Tsingou. Studies of the nonlinear problems. Technical report, Los Alamos Scientific Lab., N. Mex., 1955.

[4] Berni J Alder and Thomas Everett Wainwright. Studies in molecular dynamics. I. general method. *The Journal of Chemical Physics*, 31(2):459–466, August 1959.

[5] John E Lennard-Jones. Processes of adsorption and diffusion on solid surfaces. *Transactions of the Faraday Society*, 28:333–359, 1932.

[6] Aneesur Rahman. Correlations in the motion of atoms in liquid argon. *Physical review*, 136(2A):A405, 1964.

[7] Arieh Warshel and Michael Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of molecular biology*, 103(2):227–249, 1976.

[8] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253(5494):694–698, 1975.

[9] Juergen M Schmidt, Rafael Brueschweiler, Richard R Ernst, Roland L Dunbrack Jr, Diane Joseph, and Martin Karplus. Molecular dynamics simulation of the proline conformational equilibrium and dynamics in antamanide using the charmm force field. *Journal of the American Chemical Society*, 115(19):8747–8756, 1993.

[10] ISY Wang and Martin Karplus. Dynamics of organic reactions. *Journal of the American Chemical Society*, 95(24):8160–8164, 1973.

[11] Peter L Freddolino, Anton S Arkhipov, Steven B Larson, Alexander McPherson, and Klaus Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14(3):437–449, 2006.

[12] Steven B Larson, John S Day, and Alexander McPherson. Satellite tobacco mosaic virus refined to 1.4 Å resolution. *Acta Crystallographica Section D: Biological Crystallography*, 70(9):2316–2330, 2014.

[13] Julie R Schames, Richard H Henchman, Jay S Siegel, Christoph A Sotriffer, Haihong Ni, and J Andrew McCammon. Discovery of a novel binding trench in HIV integrase. *Journal of Medicinal Chemistry*, 47(8):1879–1881, 2004.

[14] Maxwell I. Zimmerman, Justin R. Porter, Michael D. Ward, Sukrit Singh, Neha Vithani, Artur Meller, Upasana L. Mallimadugula, Catherine E. Kuhn, Jonathan H. Borowsky, Rafal P. Wiewiora, Matthew F. D. Hurley, Aoife M Harbison, Carl A Fogarty, Joseph E. Coffland, Elisa Fadda, Vincent A. Voelz, John D. Chodera, and Gregory R. Bowman. SARS-CoV-2 simulations go exascale to capture spike opening and reveal cryptic pockets across the proteome. *ArXiv*, June 2020.

[15] Roberto Car and Michele Parrinello. Unified approach for molecular dynamics and density-functional theory. *Physical Review Letters*, 55(22):2471–2474, November 1985.

[16] Shirley W. I. Siu, Kristyna Pluhackova, and Rainer A. Böckmann. Optimization of the OPLS-AA force field for long hydrocarbons. *Journal of Chemical Theory and Computation*, 8(4):1459–1470, March 2012.

[17] Edward Harder, Wolfgang Damm, Jon Maple, Chuanjie Wu, Mark Reboul, Jin Yu Xiang, Lingle Wang, Dmitry Lupyan, Markus K Dahlgren, Jennifer L Knight, et al. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *Journal of chemical theory and computation*, 12(1):281–296, 2016.

[18] Kevin J Bowers, David E Chow, Huafeng Xu, Ron O Dror, Michael P Eastwood, Brent A Gregersen, John L Klepeis, Istvan Kolossvary, Mark A Moraes, Federico D Sacerdoti, et al. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *SC'06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, pages 43–43. IEEE, 2006.

[19] Yudong Qiu, Daniel Smith, Simon Boothroyd, Hyesu Jang, Jeffrey Wagner, Caitlin C. Bannan, Trevor Gokey, Victoria T. Lim, Chaya Stern, Andrea Rizzi, and et al. Development and benchmarking of Open Force Field v1.0.0, the Parsley small molecule force field. *ChemRxiv*, 2020. 10.26434/chemrxiv.13082561.v2.

[20] Lee-Ping Wang, Jiahao Chen, and Troy Van Voorhis. Systematic parametrization of polarizable force fields from quantum chemistry data. *Journal of chemical theory and computation*, 9(1):452–460, 2013.

[21] Inc. Daylight Chemical Information Systems. `https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html`, 2019. Accessed: 14.04.2021.

[22] Claudio Zeni, Kevin Rossi, Aldo Glielmo, Ádám Fekete, Nicola Gaston, Francesca Baletto, and Alessandro De Vita. Building machine learning force fields for nanoclusters. *The Journal of chemical physics*, 148(24):241739, 2018.

[23] Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters*, 104(13):136403, 2010.

[24] Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Computational Materials*, 6(1), March 2020.

[25] Aldo Glielmo, Claudio Zeni, and Alessandro De Vita. Efficient nonparametric n-body force fields from machine learning. *Physical Review B*, 97(18):184307, 2018.

[26] Yu Xie, Jonathan Vandermause, Lixin Sun, Andrea Cepellotti, and Boris Kozinsky. Bayesian force fields from active learning for simulation of inter-dimensional transformation of stanene. *npj Computational Materials*, 7(1):1–10, 2021.

[27] Elena Uteva, Richard S. Graham, Richard D. Wilkinson, and Richard J. Wheatley. Active learning in gaussian process interpolation of potential energy surfaces. *The Journal of Chemical Physics*, 149(17):174114, November 2018.

[28] Jie Cui and Roman V Krems. Efficient non-parametric fitting of potential energy surfaces for polyatomic molecules with gaussian processes. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 49(22):224001, 2016.

[29] Aldo Glielmo, Claudio Zeni, Ádám Fekete, and Alessandro De Vita. Building nonparametric n-body force fields using gaussian process

regression. In *Machine Learning Meets Quantum Physics*, pages 67–98. Springer, 2020.

[30] Stefan Chmiela, Huziel E Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature communications*, 9(1):1–10, 2018.

[31] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical science*, 8(4):3192–3203, 2017.

[32] Pilsun Yoo, Michael Sakano, Saaketh Desai, Md Mahbubul Islam, Peilin Liao, and Alejandro Strachan. Neural network reactive force field for C, H, N, and O systems. *npj Computational Materials*, 7(1):1–10, 2021.

[33] Joshua T Horton, Alice EA Allen, Leela S Dodda, and Daniel J Cole. QUBEKit: automating the derivation of force field parameters from quantum mechanics. *Journal of chemical information and modeling*, 59(4):1366–1381, 2019.

[34] Alice EA Allen, Michael C Payne, and Daniel J Cole. Harmonic force constants for molecular mechanics force fields via Hessian matrix projection. *Journal of chemical theory and computation*, 14(1):274–281, 2018.

[35] Joshua T Horton, Alice EA Allen, and Daniel J Cole. Modelling flexible protein–ligand binding in p38α map kinase using the QUBE force field. *Chemical Communications*, 56(6):932–935, 2020.

[36] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning*, pages 8459–8468. PMLR, 2020.

[37] Samuel S. Schoenholz and Ekin D. Cubuk. JAX M.D. a framework for differentiable physics. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

[38] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xiii. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1011–1020, 2019.

[39] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, January 2020.

[40] David E. Shaw, Martin M. Deneroff, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Ierardi, István Kolossváry, John L. Klepeis, Timothy Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, and Stanley C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the ACM*, 51(7):91–97, July 2008.

[41] David E Shaw, Ron O Dror, John K Salmon, JP Grossman, Kenneth M Mackenzie, Joseph A Bank, Cliff Young, Martin M Deneroff, Brannon Batson, Kevin J Bowers, et al. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the conference on high performance computing networking, storage and analysis*, pages 1–11, 2009.

[42] Ian Kuon and Jonathan Rose. Measuring the gap between FPGAs and ASICs. *IEEE Transactions on computer-aided design of integrated circuits and systems*, 26(2):203–215, 2007.

[43] Chen Yang, Tong Geng, Tianqi Wang, Rushi Patel, Qingqing Xiong, Ahmed Sanaullah, Chunshu Wu, Jiayi Sheng, Charles Lin, Vipin Sachdeva, et al. Fully integrated FPGA molecular dynamics simulations. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–31, 2019.

[44] Donald A Mcquarrie. Statistical mechanics, 1965.

[45] Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Role of molecular dynamics and related methods in drug discovery. *Journal of medicinal chemistry*, 59(9):4035–4061, 2016.

[46] William L Jorgensen and Julian Tirado-Rives. The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6):1657–1666, 1988.

[47] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, 1996.

[48] Jay W Ponder, Chuanjie Wu, Pengyu Ren, Vijay S Pande, John D Chodera, Michael J Schnieders, Imran Haque, David L Mobley, Daniel S Lambrecht, Robert A DiStasio Jr, et al. Current status of the amoeba polarizable force field. *The journal of physical chemistry B*, 114(8):2549–2564, 2010.

[49] Adri CT Van Duin, Siddharth Dasgupta, Francois Lorant, and William A Goddard. Reaxff: a reactive force field for hydrocarbons. *The Journal of Physical Chemistry A*, 105(41):9396–9409, 2001.

[50] Kenno Vanommeslaeghe, Elizabeth Hatcher, Chayan Acharya, Sibsankar Kundu, Shijun Zhong, Jihyun Shim, Eva Darian, Olgun Guvench, P Lopes, Igor Vorobyov, et al. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of computational chemistry*, 31(4):671–690, 2010.

[51] Alexander D MacKerell Jr, Michael Feig, and Charles L Brooks. Improved treatment of the protein backbone in empirical force fields. *Journal of the American Chemical Society*, 126(3):698–699, 2004.

[52] Stephen K Burley, Helen M Berman, Charmi Bhikadiya, Chunxiao Bi, Li Chen, Luigi Di Costanzo, Cole Christie, Jose M Duarte, Shuchismita Dutta, Zukang Feng, Sutapa Ghosh, David S Goodsell, Rachel Kramer Green, Vladimir Guranovic, Dmytro Guzenko, Brian P Hudson, Yuhe Liang, Robert Lowe, Ezra Peisach, Irina Periskova, Chris Randle, Alexander Rose, Monica Sekharan, Chenghua Shao, Yi-Ping Tao, Yana Valasatava, Maria Voigt, John Westbrook, Jasmine Young, Christine Zardecki, Marina Zhuravleva, Genji Kurisu, Haruki Nakamura, Yumiko Kengaku, Hasumi Cho, Junko Sato, Ju Yaen Kim, Yasuyo Ikegawa, Atsushi Nakagawa, Reiko Yamashita, Takahiro Kudou, Gert-Jan Bekker, Hirofumi Suzuki, Takeshi Iwata, Masashi Yokochi, Naohiro Kobayashi, Toshimichi Fujiwara, Sameer Velankar, Gerard J Kleywegt, Stephen Anyango, David R Armstrong, John M Berrisford, Matthew J Conroy, Jose M Dana, Mandar Deshpande, Paul Gane, Romana Gáborová, Deepti Gupta, Aleksandras Gutmanas, Jaroslav Koča, Lora Mak, Saqib Mir, Abhik Mukhopadhyay, Nurul Nadzirin, Sreenath Nair, Ardan Patwardhan, Typhaine Paysan-Lafosse, Lukas Pravda, Osman Salih, David Sehnal, Mihaly Varadi, Radka Vařeková, John L Markley, Jeffrey C Hoch, Pedro R Romero, Kumaran Baskaran, Dimitri Maziuk, Eldon L Ulrich, Jonathan R Wedell, Hongyang Yao, Miron Livny, and Yannis E Ioannidis. Protein data bank: the single

global archive for 3d macromolecular structure data. *Nucleic Acids Research*, 47(D1):D520–D528, October 2018.

[53] Marc A Martí-Renom, Ashley C Stuart, András Fiser, Roberto Sánchez, Francisco Melo, and Andrej Šali. Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure*, 29(1):291–325, 2000.

[54] Benjamin Webb and Andrej Sali. Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, 54(1):5–6, 2016.

[55] Marcus D Hanwell, Donald E Curtis, David C Lonie, Tim Vandermeersch, Eva Zurek, and Geoffrey R Hutchison. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of cheminformatics*, 4(1):1–17, 2012.

[56] Sunhwan Jo, Taehoon Kim, Vidyashankara G Iyer, and Wonpil Im. Charmm-gui: a web-based graphical user interface for charmm. *Journal of computational chemistry*, 29(11):1859–1865, 2008.

[57] Jumin Lee, Manuel Hitzenberger, Manuel Rieger, Nathan R. Kern, Martin Zacharias, and Wonpil Im. CHARMM-GUI supports the amber force fields. *The Journal of Chemical Physics*, 153(3):035103, July 2020.

[58] Leandro Martínez, Ricardo Andrade, Ernesto G Birgin, and José Mario Martínez. Packmol: a package for building initial configurations for molecular dynamics simulations. *Journal of computational chemistry*, 30(13):2157–2164, 2009.

[59] Mohamed Ali al Badri and Robert C. Sinclair. `https://github.com/maalbadri/Accurate-large-scale-modelling-of-GrapheneOxide`, 2020. Accessed: 14.04.2021.

[60] Osmar N. de Souza and Rick L. Ornstein. Effect of periodic box size on aqueous molecular dynamics simulation of a DNA dodecamer with particle-mesh ewald method. *Biophysical Journal*, 72(6):2395–2397, June 1997.

[61] Haskell B. Curry. The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics*, 2(3):258–261, October 1944.

[62] Ian Ford. *Statistical Physics: an entropic approach*. John Wiley & Sons, 2013.

[63] Herman JC Berendsen, JPM van Postma, Wilfred F van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.

[64] Stephen C Harvey, Robert K-Z Tan, and Thomas E Cheatham III. The flying ice cube: velocity rescaling in molecular dynamics leads to violation of energy equipartition. *Journal of computational chemistry*, 19(7):726–740, 1998.

[65] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126(1):014101, 2007.

[66] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

[67] Paul P Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der physik*, 369(3):253–287, 1921.

[68] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: AnN·log(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, June 1993.

[69] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593, November 1995.

[70] Naoki Karasawa and William A Goddard III. Acceleration of convergence for lattice sums. *The Journal of Physical Chemistry*, 93(21):7320–7327, 1989.

[71] Erwin Schrödinger. An undulatory theory of the mechanics of atoms and molecules. *Physical review*, 28(6):1049, 1926.

[72] Llewellyn H. Thomas. The calculation of atomic fields. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23(5):542–548, January 1927.

[73] Enrico Fermi. A statistical method for the determination of some priority of the atom. *Rend. Happened. Nat. Lincei*, 6(602-607):32, 1927.

[74] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[75] Walter Kohn and Lu Jeu Sham. Quantum density oscillations in an inhomogeneous electron gas. *Phys. Rev.*, 137:A1697–A1705, Mar 1965.

[76] Max Born and Robert Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389:457–484, 1927.

[77] David M Ceperley and Berni J Alder. Ground state of the electron gas by a stochastic method. *Phys. Rev. Lett.*, 45:566–569, Aug 1980.

[78] John P. Perdew and Alex Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, 23:5048–5079, May 1981.

[79] David C. Langreth and M. J. Mehl. Beyond the local-density approximation in calculations of ground-state electronic properties. *Phys. Rev. B*, 28:1809–1834, Aug 1983.

[80] Chris-Kriton Skylaris, Peter D. Haynes, Arash A. Mostofi, and Mike C. Payne. Introducing onetep: Linear-scaling density functional simulations on parallel computers. *The Journal of Chemical Physics*, 122(8):084119, 2005.

[81] Nicholas DM Hine, Peter D Haynes, Arash A Mostofi, C-K Skylaris, and Mike C Payne. Linear-scaling density-functional theory with tens of thousands of atoms: expanding the scope and scale of calculations with ONETEP. *Comput. Phys. Commun.*, 180:1041–1053, 2009.

[82] Peter D. Haynes, Chris-Kriton Skylaris, Arash A. Mostofi, and Mike C. Payne. Onetep: linear-scaling density-functional theory with local orbitals and plane waves. *physica status solidi (b)*, 243(11):2489–2499, 2006.

[83] Alexandre Tkatchenko and Matthias Scheffler. Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data. *Physical review letters*, 102(7):073005, 2009.

[84] Thomas A Manz and David S Sholl. Improved atoms-in-molecule charge partitioning functional for simultaneously reproducing the electrostatic potential and chemical states in periodic and nonperiodic materials. *Journal of chemical theory and computation*, 8(8):2844–2867, 2012.

[85] Patrick Bultinck, Paul W Ayers, Stijn Fias, Koen Tiels, and Christian Van Alsenoy. Uniqueness and basis set dependence of iterative hirshfeld charges. *Chemical physics letters*, 444(1-3):205–208, 2007.

[86] Timothy C Lillestolen and Richard J Wheatley. Redefining the atom: atomic charge densities produced by an iterative stockholder approach. *Chemical communications*, pages 5909–5911, 2008.

[87] Patrick Bultinck, David L Cooper, and Dimitri Van Neck. Comparison of the hirshfeld-i and iterated stockholder atoms in molecules schemes. *Physical Chemistry Chemical Physics*, 11(18):3424–3429, 2009.

[88] Xiang Chu and Alexander Dalgarno. Linear response time-dependent density functional theory for van der waals coefficients. *The Journal of chemical physics*, 121(9):4083–4088, 2004.

[89] Vladimir Ilich Anisimov. Electronic structure of strongly correlated materials. In *AIP Conference Proceedings*, volume 1297, pages 3–134. American Institute of Physics, 2010.

[90] Mohamed Ali al Badri, Edward Linscott, Antoine Georges, Daniel J Cole, and Cédric Weber. Superexchange mechanism and quantum many body excitations in the archetypal di-cu oxo-bridge. *Communications Physics*, 3(1):1–8, 2020.

[91] Cédric Weber, David D O'Regan, Nicholas DM Hine, Peter B Littlewood, Gabriel Kotliar, and Mike C Payne. Importance of many-body effects in the kernel of hemoglobin for ligand binding. *Phys. Rev. Lett.*, 110:106402, 2013.

[92] Cédric Weber, Daniel J Cole, David D O'Regan, and Mike C Payne. Renormalization of myoglobin–ligand binding energetics by quantum many-body effects. *Proc. Natl. Acad. Sci.*, 111(16):5790–5795, 2014.

[93] Junjiro Kanamori. Electron correlation and ferromagnetism of transition metals. *Progress of Theoretical Physics*, 30(3):275–289, 1963.

[94] John Hubbard. Electron correlations in narrow energy bands. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 276(1365):238–257, 1963.

[95] Martin C Gutzwiller. Effect of correlation on the ferromagnetism of transition metals. *Physical Review Letters*, 10(5):159, 1963.

[96] Vladimir I Anisimov, Jan Zaanen, and Ole K Andersen. Band theory and Mott insulators: Hubbard U instead of Stoner I. *Phys. Rev. B*, 44(3):943–954, jul 1991.

[97] Walter Metzner and Dieter Vollhardt. Correlated lattice fermions in $d = \infty$ dimensions. *Physical review letters*, 62(3):324, 1989.

[98] Philip Warren Anderson. Localized magnetic states in metals. *Physical Review*, 124(1):41, 1961.

[99] Antoine Georges, Gabriel Kotliar, Werner Krauth, and Marcelo J Rozenberg. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Rev. Mod. Phys.*, 68(1):13–125, jan 1996.

[100] Edward B Linscott, Daniel J Cole, Nicholas DM Hine, Michael C Payne, and Cédric Weber. onetep+ toscam: Uniting dynamical mean field theory and linear-scaling density functional theory. *Journal of Chemical Theory and Computation*, 16(8):4899–4911, 2020.

[101] Thomas A Maier, Thomas Pruschke, and Mark Jarrell. Angle-resolved photoemission spectra of the Hubbard model. *Phys. Rev. B*, 66:075102, 2002.

[102] Yiqing Yuan, Xueli Gao, Yi Wei, Xinyan Wang, Jian Wang, Yushan Zhang, and Congjie Gao. Enhanced desalination performance of carboxyl functionalized graphene oxide nanofiltration membranes. *Desalination*, 405:29–39, 2017.

[103] Yanwu Zhu, Shanthi Murali, Weiwei Cai, Xuesong Li, Ji Won Suk, Jeffrey R Potts, and Rodney S Ruoff. Graphene and graphene oxide: synthesis, properties, and applications. *Adv. Mater.*, 22(35):3906–3924, 2010.

[104] Chul Chung, Young-Kwan Kim, Dolly Shin, Soo-Ryoon Ryoo, Byung Hee Hong, and Dal-Hee Min. Biomedical applications of graphene and graphene oxide. *Acc. Chem. Res.*, 46(10):2211–2224, 2013.

[105] Alessandra Bonanni, Chun Kiang Chua, Guanjia Zhao, Zdenek Sofer, and Martin Pumera. Inherently electroactive graphene oxide nanoplatelets as labels for single nucleotide polymorphism detection. *ACS Nano*, 6(10):8546–8551, 2012.

[106] Stephanie J Heerema and Cees Dekker. Graphene nanodevices for dna sequencing. *Nat. Nanotechnol.*, 11(2):127–136, 2016.

[107] Savannah Afsahi, Mitchell B Lerner, Jason M Goldstein, Joo Lee, Xiaoling Tang, Dennis A Bagarozzi Jr, Deng Pan, Lauren Locascio, Amy Walker, Francie Barron, et al. Novel graphene-based biosensor for early detection of zika virus infection. *Biosens. Bioelectron.*, 100:85–88, 2018.

[108] Joshua T Robinson, Scott M Tabakman, Yongye Liang, Hailiang Wang, Hernan Sanchez Casalongue, Daniel Vinh, and Hongjie Dai. Ultrasmall reduced graphene oxide with high near-infrared absorbance for photothermal therapy. *J. Am. Chem. Soc.*, 133(17):6825–6831, 2011.

[109] Jingquan Liu, Liang Cui, and Dusan Losic. Graphene and graphene oxide as new nanocarriers for drug delivery applications. *Acta Biomater.*, 9(12):9243–9257, 2013.

[110] Xiaoming Sun, Zhuang Liu, Kevin Welsher, Joshua Tucker Robinson, Andrew Goodwin, Sasa Zaric, and Hongjie Dai. Nano-graphene oxide for cellular imaging and drug delivery. *Nano Res.*, 1(3):203–212, 2008.

[111] Liming Zhang, Jingguang Xia, Qinghuan Zhao, Liwei Liu, and Zhijun Zhang. Functional graphene oxide as a nanocarrier for controlled

loading and targeted delivery of mixed anticancer drugs. *Small*, 6(4):537–544, 2010.

[112] Eudald Casals, Tobias Pfaller, Albert Duschl, Gertie Janneke Oostingh, and Victor Puntes. Time evolution of the nanoparticle protein corona. *ACS Nano*, 4(7):3623–3632, 2010.

[113] Pu Chun Ke, Sijie Lin, Wolfgang J Parak, Thomas P Davis, and Frank Caruso. A decade of the protein corona. *ACS Nano*, 11(12):11773–11776, 2017.

[114] Carlheinz Röcker, Matthias Pötzl, Feng Zhang, Wolfgang J Parak, and G Ulrich Nienhaus. A quantitative fluorescence study of protein monolayer formation on colloidal nanoparticles. *Nat. Nanotechnol.*, 4(9):577–580, 2009.

[115] Daniel Nierenberg, Annette R Khaled, and Orielyz Flores. Formation of a protein corona influences the biological identity of nanomaterials. *Rep. Pract. Oncol. Radiother.*, 23(4):300–308, 2018.

[116] Kuo-Ching Mei, Artur Ghazaryan, Er Zhen Teoh, Huw D Summers, Yueting Li, Belén Ballesteros, Justyna Piasecka, Adam Walters, Robert C Hider, Volker Mailänder, et al. Protein-corona-by-design in 2d: A reliable platform to decode bio–nano interactions for the next-generation quality-by-design nanomedicines. *Adv. Mater.*, 30(40):1802732, 2018.

[117] Aidee Solorio-Rodríguez, Vicente Escamilla-Rivera, Marisela Uribe-Ramírez, Alicia Chagolla, Robert Winkler, Claudia Maria García-Cuellar, and Andrea De Vizcaya-Ruiz. A comparison of the human and mouse protein corona profiles of functionalized sio 2 nanocarriers. *Nanoscale*, 9(36):13651–13660, 2017.

[118] Mohammad Javad Hajipour, Jamshid Raheb, Omid Akhavan, Sareh Arjmand, Omid Mashinchian, Masoud Rahman, Mohammad Ab-

dolahad, Vahid Serpooshan, Sophie Laurent, and Morteza Mahmoudi. Personalized disease-specific protein corona influences the therapeutic impact of graphene oxide. *Nanoscale*, 7(19):8978–8994, 2015.

[119] Massimiliano Papi, Valentina Palmieri, Luca Digiacomo, Francesca Giulimondi, Sara Palchetti, Gabriele Ciasca, Giordano Perini, Damiano Caputo, Maria Cristina Cartillone, Chiara Cascone, et al. Converting the personalized biomolecular corona of graphene oxide nanoflakes into a high-throughput diagnostic test for early cancer detection. *Nanoscale*, 11(32):15339–15346, 2019.

[120] Riccardo Rampado, Sara Crotti, Paolo Caliceti, Salvatore Pucciarelli, and Marco Agostini. Recent advances in understanding the protein corona of nanoparticles and in the formulation of "stealthy" nanomaterials. *Front. Bioeng. Biotechnol.*, 8, 2020.

[121] Kuo-Ching Mei, Noelia Rubio, Pedro M Costa, Houmam Kafa, Vincenzo Abbate, Frederic Festy, Sukhvinder S Bansal, Robert C Hider, and Khuloud T Al-Jamal. Synthesis of double-clickable functionalised graphene oxide for biological applications. *Chem. Commun.*, 51(81):14981–14984, 2015.

[122] Noelia Rubio, Kuo-Ching Mei, Rebecca Klippstein, Pedro M. Costa, Naomi Hodgins, Julie Tzu-Wen Wang, Frederic Festy, Vincenzo Abbate, Robert C. Hider, Ka Lung Andrew Chan, and Khuloud T. Al-Jamal. Solvent-free click-mechanochemistry for the preparation of cancer cell targeting graphene oxide. *ACS Applied Materials & Interfaces*, 7(34):18920–18923, August 2015.

[123] Leo Vroman, Ann L Adams, Gena C Fischer, and Priscilla C Munoz. Interaction of high molecular weight kininogen, factor XII, and fibrinogen in plasma at interfaces. *Blood*, 55(1):156–159, January 1980.

[124] Parisa Foroozandeh and Azlan Abdul Aziz. Merging worlds of nanomaterials and biological environment: factors governing protein corona formation on nanoparticles and its biological consequences. *Nanoscale Res. Lett.*, 10(1):1–12, 2015.

[125] Morton S Ehrenberg, Alan E Friedman, Jacob N Finkelstein, Günter Oberdörster, and James L McGrath. The influence of protein adsorption on nanoparticle association with cultured endothelial cells. *Biomaterials*, 30(4):603–610, 2009.

[126] Stephan Harnisch and Rainer H Müller. Adsorption kinetics of plasma proteins on oil-in-water emulsions for parenteral nutrition. *Eur. J. Pharm. Biopharm.*, 49(1):41–46, 2000.

[127] Torsten M Göppert and Rainer H Müller. Polysorbate-stabilized solid lipid nanoparticles as colloidal carriers for intravenous targeting of drugs to the brain: comparison of plasma protein adsorption patterns. *J. Drug Target.*, 13(3):179–187, 2005.

[128] Slaven Radic, Nicholas K Geitner, Ramakrishna Podila, Aleksandr Käkinen, Pengyu Chen, Pu Chun Ke, and Feng Ding. Competitive binding of natural amphiphiles with graphene derivatives. *Sci. Rep.*, 3(1):1–8, 2013.

[129] Lokesh Baweja, Kanagasabai Balamurugan, Venkatesan Subramanian, and Alok Dhawan. Effect of graphene oxide on the conformational transitions of amyloid beta peptide: A molecular dynamics simulation study. *J. Mol. Graph. Model.*, 61:175–185, September 2015.

[130] Lokesh Baweja, Kanagasabai Balamurugan, Venkatesan Subramanian, and Alok Dhawan. Hydration patterns of graphene-based nanomaterials (GBNMs) play a major role in the stability of a helical protein: a molecular dynamics simulation study. *Langmuir*, 29(46):14230–14238, 2013.

[131] Xiaotian Sun, Zhiwei Feng, Tingjun Hou, and Youyong Li. Mechanism of Graphene Oxide as an enzyme inhibitor from molecular dynamics simulations. *ACS Appl. Mater. Interfaces*, 6(10):7153–7163, May 2014.

[132] Robert C. Sinclair and Peter V. Coveney. Modelling nanostructure in graphene oxide: inhomogeneity and the percolation threshold. *Jo. Chem. Inf. Model.*, 59(6):2741, 2019.

[133] Félix Mouhat, François-Xavier Coudert, and Marie-Laure Bocquet. Structure and chemistry of graphene oxide in liquid water from first principles. *Nat. Commun.*, 11(1):1–9, 2020.

[134] Mohamed Ali al Badri, Paul Smith, Robert C Sinclair, Khuloud T al Jamal, and Christian D Lorenz. Accurate large scale modelling of graphene oxide: ion trapping and chaotropic potential at the interface. *Carbon*, 2020.

[135] Rashmi Kumari, Rajendra Kumar, Open Source Drug Discovery Consortium, and Andrew Lynn. g_mmpbsa: A GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model.*, 54(7):1951–1962, 2014.

[136] Andrew R Leach and Andrew P Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the a* algorithm. *Proteins: Struct., Funct., Bioinf.*, 33(2):227–239, 1998.

[137] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[138] Koichi Matsuo, Yoshie Sakurada, Ryuta Yonehara, Mikio Kataoka, and Kunihiko Gekko. Secondary-structure analysis of denatured proteins by vacuum-ultraviolet circular dichroism spectroscopy. *Biophys. J.*, 92(11):4088–4096, 2007.

[139] Joel DA Tyndall, Bernhard Pfeiffer, Giovanni Abbenante, and David P Fairlie. Over one hundred peptide-activated g protein-coupled receptors recognize ligands with turn structure. *Chem. Rev.*, 105(3):793–826, 2005.

[140] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.*, 3(29):861, 2018.

[141] James T Sparrow, Henry J Pownall, Fu-Juan Hsu, Lee D Blumenthal, Alan R Culwell, and Antonio M Gotto. Lipid binding by fragments of apolipoprotein C-III-1 obtained by thrombin cleavage. *Biochemistry*, 16(25):5427–5431, 1977.

[142] Nathan L Meyers, Mikael Larsson, Evelina Vorrsjö, Gunilla Olivecrona, and Donald M Small. Aromatic residues in the C terminus of apolipoprotein C-III mediate lipid binding and LPL inhibition. *J. Lipid Res.*, 58(5):840–852, 2017.

[143] Daniel A Lambert, Alberico L Catapano, Louis C Smith, John T Sparrow, and Antonio M Gotto Jr. Effect of the apolipoprotein C-IIC-III1 ratio on the capacity of purified milk lipoprotein lipase to hydrolyse triglycerides in monolayer vesicles. *Atherosclerosis*, 127(2):205–212, 1996.

[144] Robert C. Sinclair. `https://github.com/velocirobbie/make-graphitics`, 2020. Accessed: 14.04.2021.

[145] Daniela. Pacilé, Jannik C. Meyer, Arantxa Fraile Rodríguez, Marco Papagno, Cristina. Gómez-Navarro, Ravi S. Sundaram, Marko Burghard, Klaus Kern, Carlo Carbone, and Ute Kaiser. Electronic properties and atomic structure of graphene oxide membranes. *Carbon*, 49(3):966–972, 2011.

[146] Weiwei Cai, Richard D Piner, Frank J Stadermann, Sungjin Park, Medhat A Shaibat, Yoshitaka Ishii, Dongxing Yang, Aruna Velamakanni, Sung Jin An, Meryl Stoller, Jinho An, Dongmin Chen, and Rodney S Ruoff. Synthesis and solid-state NMR structural characterization of 13C-labeled graphite oxide. *Science*, 321(5897):1815–7, 2008.

[147] Sumit Saxena, Trevor A. Tyson, and Ezana Negusse. Investigation of the local structure of graphene oxide. *J. Phys. Chem. Lett.*, 1(24):3433–3437, 2010.

[148] Kris Erickson, Rolf Erni, Zonghoon Lee, Nasim Alem, Will Gannett, and Alex Zettl. Determination of the local chemical structure of graphene oxide and reduced graphene oxide. *Adv. Mater.*, 22(40):4467–4472, 2010.

[149] Richard J Gowers, Max Linke, Jonathan Barnoud, Tyler John Edward Reddy, Manuel N Melo, Sean L Seyler, Jan Domanski, David L Dotson, Sébastien Buchoux, Ian M Kenney, et al. Mdanalysis: a python package for the rapid analysis of molecular dynamics simulations. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2019.

[150] Richard Gowers, Max Linke, Jonathan Barnoud, Tyler Reddy, Manuel Melo, Sean Seyler, Jan Domański, David Dotson, Sébastien Buchoux, Ian Kenney, and Oliver Beckstein. MDAnalysis: A python package for the rapid analysis of molecular dynamics simulations. In *Proceedings of the 15th Python in Science Conference*. SciPy, 2016.

[151] Naveen Michaud-Agrawal, Elizabeth J. Denning, Thomas B. Woolf, and Oliver Beckstein. Mdanalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.*, 32(10):2319–2327, 2011.

[152] Raul Araya-Secchi, Tomas Perez-Acle, Seung-gu Kang, Tien Huynh, Alejandro Bernardin, Yerko Escalona, Jose-Antonio Garate, Agustin D

Martínez, Isaac E García, Juan C Sáez, et al. Characterization of a novel water pocket inside the human cx26 hemichannel structure. *Biophys. J.*, 107(3):599–612, 2014.

[153] Paul Smith, Robert M Ziolek, Elena Gazzarrini, Dylan M Owen, and Christian D Lorenz. On the interaction of hyaluronic acid with synovial fluid lipid membranes. *Phys. Chem. Chem. Phys.*, 21(19):9845–9857, 2019.

[154] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.*, 109(8):1528 – 1532, 2015.

[155] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[156] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[157] Andrew Shrake and John A Rupley. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.*, 79(2):351–371, 1973.

[158] Kristian G Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, and Robert F Garry. The proximal origin of SARS-CoV-2. *Nature Medicine*, 26(4):450–452, 2020.

[159] Jian Shang, Gang Ye, Ke Shi, Yushun Wan, Chuming Luo, Hideki Aihara, Qibin Geng, Ashley Auerbach, and Fang Li. Structural basis of receptor recognition by SARS-CoV-2. *Nature*, 581(7807):221–224, 2020.

[160] World Health Organization. Coronavirus disease (COVID-19): situation report, 209, 2020. Accessed: 02.09.2020.

[161] Eskild Petersen, Marion Koopmans, Unyeong Go, Davidson H Hamer, Nicola Petrosillo, Francesco Castelli, Merete Storgaard, Sulien Al Khalili, and Lone Simonsen. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. *The Lancet Infectious Diseases*, 20(9):e238–e244, September 2020.

[162] Yeming Wang, Dingyu Zhang, Guanhua Du, Ronghui Du, Jianping Zhao, Yang Jin, Shouzhi Fu, Ling Gao, Zhenshun Cheng, Qiaofa Lu, Yi Hu, Guangwei Luo, Ke Wang, Yang Lu, Huadong Li, Shuzhen Wang, Shunan Ruan, Chengqing Yang, Chunlin Mei, Yi Wang, Dan Ding, Feng Wu, Xin Tang, Xianzhi Ye, Yingchun Ye, Bing Liu, Jie Yang, Wen Yin, Aili Wang, Guohui Fan, Fei Zhou, Zhibo Liu, Xiaoying Gu, Jiuyang Xu, Lianhan Shang, Yi Zhang, Lianjun Cao, Tingting Guo, Yan Wan, Hong Qin, Yushen Jiang, Thomas Jaki, Frederick G Hayden, Peter W Horby, Bin Cao, and Chen Wang. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. *The Lancet*, 395(10236):1569–1578, May 2020.

[163] The RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with covid-19 — preliminary report. *New England Journal of Medicine*, July 2020.

[164] Jonathan A. C. Sterne, Srinivas Murthy, Janet V. Diaz, Arthur S. Slutsky, Jesús Villar, Derek C. Angus, Djillali Annane, Luciano Cesar Pontes Azevedo, Otavio Berwanger, Alexandre B. Cavalcanti, Pierre-Francois Dequin, Bin Du, Jonathan Emberson, David Fisher, Bruno Giraudeau, Anthony C. Gordon, Anders Granholm, Cameron Green, Richard Haynes, Nicholas Heming, Julian P. T. Higgins, Peter Horby, Peter Jüni, Martin J. Landray, Amelie Le Gouge, Marie Leclerc, Wei Shen Lim, Flávia R. Machado, Colin McArthur, Ferhat

Meziani, Morten Hylander Møller, Anders Perner, Marie Warrer Petersen, Jelena Savovic, Bruno Tomazini, Viviane C. Veiga, Steve Webb, and John C. Marshall. Association between administration of systemic corticosteroids and mortality among critically ill patients with COVID-19. *JAMA*, September 2020.

[165] World Health Organization. DRAFT landscape of COVID-19 candidate vaccines, 2020. Accessed: 02.09.2020.

[166] Chi-Yu Zhang, Ji-Fu Wei, and Shao-Heng He. Adaptive evolution of the spike gene of SARS coronavirus: changes in positively selected sites in different epidemic groups. *BMC microbiology*, 6(1):88, 2006.

[167] Lorenzo Casalino, Zied Gaieb, Jory A. Goldsmith, Christy K. Hjorth, Abigail C. Dommer, Aoife M. Harbison, Carl A. Fogarty, Emilia P. Barros, Bryn C. Taylor, Jason S. McLellan, Elisa Fadda, and Rommie E. Amaro. Beyond shielding: The roles of glycans in the SARS-CoV-2 spike protein. *ACS Central Science*, 6(10):1722–1734, September 2020.

[168] Sven Ullrich and Christoph Nitsche. The SARS-CoV-2 main protease as drug target. *Bioorganic & Medicinal Chemistry Letters*, 30(17):127377, September 2020.

[169] Jacob D Durrant and J Andrew McCammon. Molecular dynamics simulations and drug discovery. *BMC biology*, 9(1):1–9, 2011.

[170] Steven G Deeks, Frederick M Hecht, Melinda Swanson, Tarek Elbeik, Richard Loftus, PT Cohen, and Robert M Grant. HIV RNA and CD4 cell count response to protease inhibitor therapy in an urban AIDS clinic: response to both initial and salvage therapy. *Aids*, 13(6):F35–F43, 1999.

[171] Daniel Lamarre, Paul C Anderson, Murray Bailey, Pierre Beaulieu, Gordon Bolger, Pierre Bonneau, Michael Bös, Dale R Cameron,

Mireille Cartier, Michael G Cordingley, et al. An NS3 protease inhibitor with antiviral effects in humans infected with hepatitis C virus. *Nature*, 426(6963):186–189, 2003.

[172] Darrin M. York, Tom A. Darden, Lee G. Pedersen, and MW Anderson. Molecular dynamics simulation of HIV-1 protease in a crystalline environment and in solution. *Biochemistry*, 32(6):1443–1453, 1993.

[173] Annette Hegyi and John Ziebuhr. Conservation of substrate specificities among coronavirus main proteases. *Journal of General Virology*, 83(3):595–599, 2002.

[174] Zhenming Jin, Xiaoyu Du, Yechun Xu, Yongqiang Deng, Meiqin Liu, Yao Zhao, Bing Zhang, Xiaofeng Li, Leike Zhang, Chao Peng, Yinkai Duan, Jing Yu, Lin Wang, Kailin Yang, Fengjiang Liu, Rendi Jiang, Xinglou Yang, Tian You, Xiaoce Liu, Xiuna Yang, Fang Bai, Hong Liu, Xiang Liu, Luke W. Guddat, Wenqing Xu, Gengfu Xiao, Chengfeng Qin, Zhengli Shi, Hualiang Jiang, Zihe Rao, and Haitao Yang. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature*, 582(7811):289–293, April 2020.

[175] Marina Macchiagodena, Marco Pagliai, and Piero Procacci. Identification of potential binders of the main protease 3CLpro of the COVID-19 via structure-based ligand design and molecular modeling. *Chemical Physics Letters*, 750:137489, July 2020.

[176] Mohammad M. Ghahremanpour, Julian Tirado-Rives, Maya Deshmukh, Joseph A. Ippolito, Chun-Hui Zhang, Israel Cabeza de Vaca, Maria-Elena Liosi, Karen S. Anderson, and William L. Jorgensen. Identification of 14 known drugs as inhibitors of the main protease of SARS-CoV-2. *ArXiv*, August 2020.

[177] Alessandro Laio and Francesco L Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in bio-

physics, chemistry and material science. *Reports on Progress in Physics*, 71(12):126601, November 2008.

[178] Yinglong Miao, Victoria A. Feher, and J. Andrew McCammon. Gaussian accelerated molecular dynamics: Unconstrained enhanced sampling and free energy calculation. *Journal of Chemical Theory and Computation*, 11(8):3584–3595, 2015.

[179] Terra Sztain, Rommie Amaro, and J. Andrew McCammon. Elucidation of cryptic and allosteric pockets within the SARS-CoV-2 protease. *ArXiv*, July 2020.

[180] I-Lin Lu, Neeraj Mahindroo, Po-Huang Liang, Yi-Hui Peng, Chih-Jung Kuo, Keng-Chang Tsai, Hsing-Pang Hsieh, Yu-Sheng Chao, and Su-Ying Wu. Structure-based drug design and structural biology study of novel nonpeptide inhibitors of severe acute respiratory syndrome coronavirus main protease. *Journal of Medicinal Chemistry*, 49(17):5154–5161, 2006.

[181] Jesus Seco, F Javier Luque, and Xavier Barril. Binding site detection and druggability index from first principles. *Journal of Medicinal Chemistry*, 52(8):2363–2371, 2009.

[182] Phani Ghanakota and Heather A Carlson. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics: Miniperspective. *Journal of Medicinal Chemistry*, 59(23):10383–10399, 2016.

[183] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.

[184] DA Case, Josh Berryman, RM Betz, DS Cerutti, TE Cheatham Iii, TA Darden, RE Duke, TJ Giese, H Gohlke, AW Goetz, et al. AMBER 2015. *University of California: San Francisco, CA, USA*, 2015.

[185] Gabriel Grand, Elana Simon, Michael Bower, Bruce Clapham, and Jonah Kallenbach. Reverie Labs PostEra COVID Moonshot submission GAB-REV-df6, 2020. Accessed: 10.09.2020.

[186] I-Lin Lu, Neeraj Mahindroo, Po-Huang Liang, Yi-Hui Peng, Chih-Jung Kuo, Keng-Chang Tsai, Hsing-Pang Hsieh, Yu-Sheng Chao, and Su-Ying Wu. Structure-based drug design and structural biology study of novel nonpeptide inhibitors of severe acute respiratory syndrome coronavirus main protease. *Journal of Medicinal Chemistry*, 49(17):5154–5161, 2006.

[187] Andrej Šali and Tom L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, 1993.

[188] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An $Nlog(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[189] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. LINCS: a linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[190] Richard J. Gowers, Max Linke, Jonathan Barnoud, Tyler J. E. Reddy, Manuel N. Melo, Sean L. Seyler, Jan Domański, David L. Dotson, Sébastien Buchoux, Ian M. Kenney, and Oliver Beckstein. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In Sebastian Benthall and Scott Rostrup, editors, *Proceedings of the 15th Python in Science Conference*, pages 98 – 105, 2016.

[191] Naveen Michaud-Agrawal, Elizabeth J. Denning, Thomas B. Woolf, and Oliver Beckstein. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry*, 32(10):2319–2327, April 2011.

[192] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[193] Davide Branduardi, Giovanni Bussi, and Michele Parrinello. Metadynamics with Adaptive Gaussians. *Journal of Chemical Theory and Computation*, 8(7):2247–2254, 2012. PMID: 26588957.

[194] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. "PLUMED 2: New feathers for an old bird". *Computer Physics Communications*, 185(2):604 – 613, 2014.

[195] Lin Chen, Ton VW Janssens, and Henrik Grönbeck. A comparative test of different density functionals for calculations of NH 3-SCR over Cu-Chabazite. *Physical Chemistry Chemical Physics*, 21(21):10923–10930, 2019.