High Performance Computing-based QM/MM Simulations for Drug Design: Application to the Non-Invasive Diagnosis of IDH1-Associated Glioma

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Bharath Raghavan, M.S.

aus

Kochi, Indien

Berichter: Univ.-Prof. Paolo Carloni, Ph. D.
Univ.-Prof. Dr. rer. nat. Marc Spehr

Tag der mündlichen Prüfung: 09.07.2024

Kurzfassung

Strukturbasiertes Wirkstoffdesign und Molekulardynamik (MD) werden routinemäßig eingesetzt, um die präklinischen Phasen der Arzneimittelentdeckung zu beschleunigen, indem Ligand-Target-Positionen vorhergesagt werden. Die derzeitigen MD-Methoden (die auf der klassischen Physik beruhen) können jedoch bei der Beschreibung einer großen Anzahl von Zielen, einschließlich vieler metallhaltiger Proteine und Enzymübergangszustände, auf Schwierigkeiten stoßen. Dies schränkt die Anwendbarkeit von MD bei der Arzneimittelentdeckung stark ein. Eine quantenmechanische (QM) Behandlung kann erforderlich sein, um diese Ziele genau zu untersuchen. Dies ist jedoch sehr kostspielig und in den derzeitigen Pipelines für die Arzneimittelentwicklung nicht realisierbar. Hier zeige ich, dass die auf der Dichtefunktionaltheorie basierende Multiskalen-Quantenmechanik/Molekülmechanik (QM/MM), wie sie im kürzlich eingeführten MiMiC-Framework mit High Performance Computing (HPC) implementiert ist, ein großes Potenzial hat, zu solchen Bemühungen beizutragen. Ein quantenbasiertes virtuelles HPC-Screening-Protokoll (QHPC-VS), das MiMiC-QM/MM-MD integriert, wird hier formuliert und verfeinert. Dazu gehört die Entwicklung des MiMiCPy-Softwarepakets für die nahtlose Konvertierung von MM-Ausgaben in QM/MM-Dateien großer Biomoleküle. Mein Protokoll wird zur Untersuchung des therapeutisch wichtigen Enzyms Isocitrat-Dehydrogenase 1 verwendet. Für diese Anwendung wurde eine sehr hohe Leistung und Skalierung von MiMiC erreicht, was sein Potenzial für QM/MM-Simulationen in der Biologie unter Beweis stellt. Die mutierte (R132H) Isoform von IDH1 hat sich als potenzieller prädiktiver Biomarker für Gliome erwiesen, was die Entwicklung nicht-invasiver Bildgebungsverfahren für diesen Marker wie die Positronen-Emissions-Tomographie (PET) nahelegt. Dabei wird das aktive Zentrum der IDH1-Mutation mit selektiven radioaktiven PET-Liganden oder Radiotracern ins Visier genommen. Aufgrund der Beschaffenheit des Proteins wurden solche selektiven Liganden in der Literatur bisher jedoch nicht vorgeschlagen. Außerdem führen klassische MD-Simulationen des komplizierten aktiven Zentrums zu einer extremen Verzerrung der Struktur, was die Anwendung von Berechnungsmethoden erschwert. Mit der QHPC-VS-Pipeline konnte das aktive Zentrum von mut-IDH1 auf Quantenebene genau simuliert werden, und es wurden Liganden vorgeschlagen, die als radioaktive Tracer für PET fungieren könnten. Unsere Simulationen sagen 15 solcher kleinen Moleküle voraus, darunter auch Fluorothymidin, ein bereits bekannter PET-Radiotracer-Vorläufer für verschiedene andere Krebsarten. Diese Entdeckung hat das Potenzial, die Entwicklung einer nicht-invasiven Diagnose von Gliomen erheblich zu beschleunigen.

Abstract

Structure-based drug design and molecular dynamics (MD), are routinely used to speed up the pre-clinical stages of the drug discovery process by predicting ligand-target poses. Nevertheless, current MD methods (based on classical physics), may encounter difficulties in describing a large range of targets, including many metal-containing proteins, and enzyme transition states. This greatly limits the applicability of MD to drug discovery. A quantum mechanical (QM) treatment may be required to accurately investigate these targets. However, this is very expensive and infeasible to implement in current drug design pipelines. Here I show that the density functional theory-based multiscale Quantum Mechanical/Molecule Mechanical (QM/MM) MD, as implemented in the recently introduced MiMiC framework with high performance computing (HPC), has a great potential to contribute towards such efforts. A quantum-based HPC virtual screening (QHPC–VS) protocol, integrating MiMiC-QM/MM MD, is formulated and refined here. This includes the development of the MiMiCPy software package for seamless conversion of MM outputs to QM/MM files of large biomolecules. My protocol is used to study the therapeutically important Isocitrate Dehydrogenase 1 enzyme. Very high performance and scaling of MiMiC was achieved for this application, demonstrating its potential for QM/MM simulations of biology. The mutant (R132H) isoform of IDH1 has been noted as a potential predictive biomarker for glioma, introducing a strong rationale to develop non-invasive imaging methods of this marker like positron emission tomogrpahy (PET). This involves targeting the mutant IDH1 active site with selective PET radioactive ligands or radiotracers. However, due to the nature of the protein, such selective ligands have not been proposed in the literature so far. Furthermore, classical MD simulations of the complicated active site lead to extreme distortion of the structure, making application of computational techniques dificult. The QHPC–VS pipeline allowed for accurately simulating the mut-IDH1 active site at the quantum level, and suggest ligands that would function as radioactive tracers for PET. Our simulations predict 15 such small molecules, out of which is fluorothymidine, an already well-known PET radiotracer precursor for various other cancers. This discovery has the potential to greatly accelerate the development of non-invasive diagnosis of glioma.

Contents

List of Figures

List of Tables

1. Introduction

Drug discovery (Figure 1.1) is a highly capital and time-intensive process. The complete pipeline takes an average of 10.5 years with a failure rate higher than 90%.[1, 2] A recent report estimated the median R&D costs to develop 63 therapeutic agents by 47 pharmaceutical companies for the period of 2009 and 2018 to be $985 million, with the mean being $1.3 billion.[3] These costs had increased to $2.3 billion for 2022, reflecting a return to the pre-pandemic levels of 2019-2020.[4] At the same time, reduction in average peaks sales due to increased competition has lead to a consistent decline (barring the pandemic year of 2021) in the R&D return rate for the top pharmaceutical companies from 8% in 2014 to 1.5% in 2022. To mitigate this problem, the report stressed the need for increasing digitization, including computer-aided drug design (CADD) and artificial intelligence.

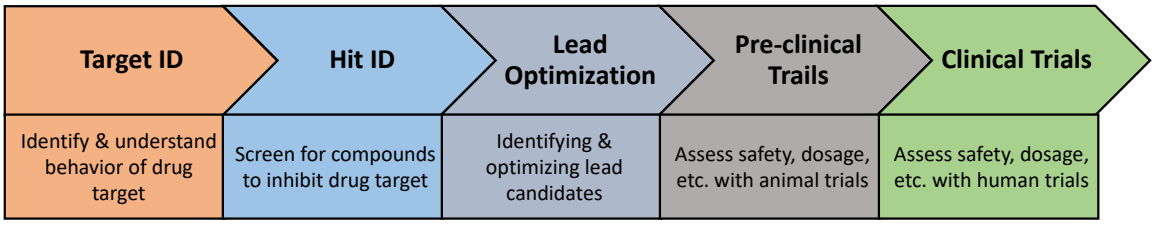


Figure 1.1. Schematic of the different stage of the drug discovery process.

CADD dramatically speeds up the pre-clinical stages of drug discovery[5], which currently involves 43% of the total R&D expenditure.[1] The identification, validation and study of drug targets[1] forms the first and one of the most important stages of CADD.[6] Classical molecular dynamics (cMD) simulations has proven extremely useful towards target identification and ligand pose prediction, especially for protein targets (which form the vast majority of drug targets).[6, 7] MD forms a core component of structure-based virtual screening protocols, one of the best methods of CADD (discussed in detail in Chapter 2).

However, cMD is not always the most efficient method for CADD.[8] Many targets are difficult to study with current classical MD approaches, including the three broad classes of metal-containing targets (metalloproteins and RNA/DNA-metal complexes), enzyme/ribozyme transition states for design of superior transition state

[1]Drug targets are biological molecules, pathways or physiological responses in the body that when inhibited or activated would change the course of the related disease in a positive way.

analogs, and covalent inhibitors.[9, 10] Hybrid quantum mechanical/molecular mechanical simulations (QM/MM)[11, 12] can alleviate this problem (see Section 1.1.1). In this approach, the region of interest (often the active site of an enzyme) is treated at the QM level, while the rest is described at the classical MM level (as shown in Figure 1.2).[13] Non-empirical density functional theory (DFT) is well-suited for the QM description due to the good compromise between accuracy and efficiency.[14] However, long-timescale DFT-QM/MM MD of drug targets has also largely remained difficult to achieve so far. This, among other reasons, leads to DFT to be severely underutilised within computational drug design and reduces the chemical space of druggable targets.[8] Recent developments in deep learning and high performance computing techniques (HPC) have pointed to potential solutions. The exascale revolution, emerging out of the field of HPC, is of particular interest here.[15] The Frontier supercomputer at the Oak Ridge National Laboratory and the upcoming JUPITER system at Forschungzsentrum Jülich could allow for routine long-timescale dynamics simulations of biological systems at the DFT level. This would expand the chemical space of druggable targets, and greatly reduce the cost associated with and time-to-market associated with the drug discovery process.

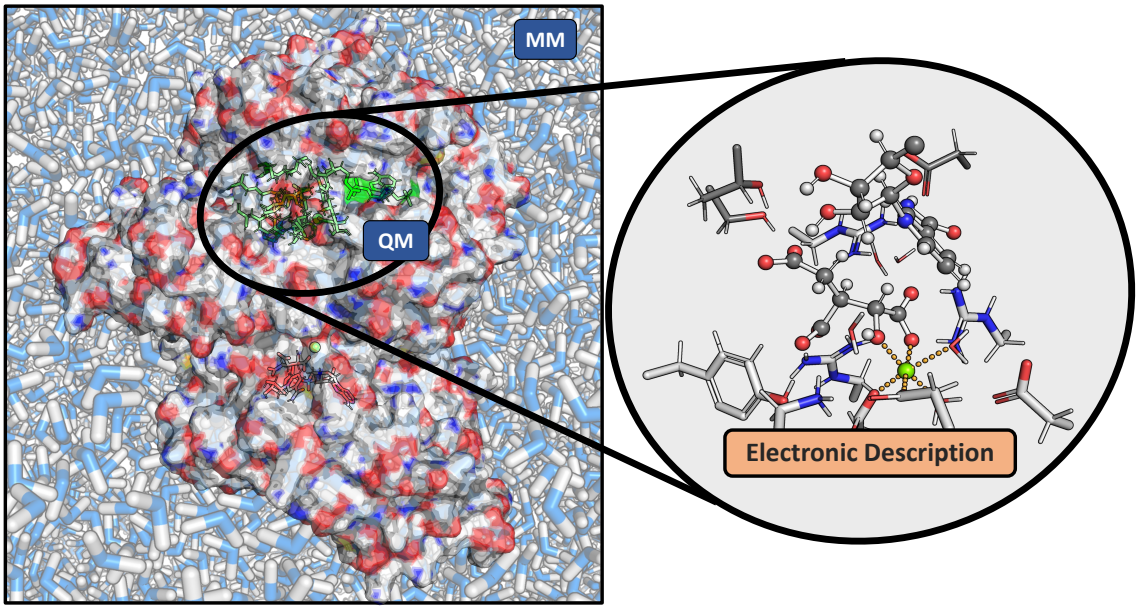


Figure 1.2. A representation of the QM and MM partition for the IDH1 protein under the QM/MM scheme.

A vital prerequisite to transcend the current limitations of DFT-QM/MM MD is the development of highly scalable software that take advantage of the huge resources provided by exascale supercomputers to perform efficient quantum MD simulations. The recently developed Multiscale Modeling in Chemistry (MiMiC) interface, coupling the CPMD[16] and GROMACS[17] codes, is an excellent candidate to break the limits of currently achievable time scales relevant to pharmacology.[18, 19, 20] The development and maintenance of this complex code is carried out within a collabora-

tion among different European research groups (including the current one), involving Dr. Davide Mandelli (Forschungszentrum Jülich), Prof. Jògvan Magnus Haugaard Olsen (Denmark Technical University), Prof. Simone Meloni (University of Ferrara), Dr. Viacheslav Bolynkh (formerly Forschungszentrum Jülich), Prof. Paolo Carloni (Forschungszentrum Jülich), and Prof. Ursula Rothlisberger (EPFL). The software has been utilized previously to study various systems of biological interest at the DFT-QM/MM level.[21, 22, 23, 24] However, so far, it has not been tested with therapeutically relevant proteins, and thus its applicability to drug design is still to be established.

Here, I have illustrated the suitability of MiMiC-QM/MM MD, and highly scalable QM/MM MD codes in general, to efficiently contribute towards drug design. Specifically, MiMiC was utilized to study a specific target important in glioma diagnosis. This is the Isocitrate dehydrogenases 1 or IDH1 enzyme.

IDH1 is a metabolic, homodimeric enzyme that plays a key role in the TCA cycle in human cells.[25] Point mutations (like R132H, R132C, R100A) on IDH1 (mut-IDH1) may be cancer-causing[26], and found in the majority of WHO grades II and III gliomas and secondary glioblastomas.[27, 26] This makes mut-IDH1 a promising therapeutic target for glioma. It has also emerged as a potential predictive biomarker with great prognostic value. This is because mut-IDH1 glioma patients have better survival rates compared to patients with wild-type IDH1 (wt-IDH1) gliomas, and may further be more sensitive to certain targeted therapies.[27, 28] Hence, there is a strong rationale for the development of non-invasive imaging methods for the detection of gliomas associated with mut-IDH1.

A promising candidate for this is the positron emission tomography (PET) technique.[29, 30] Here a precursor molecule is used to develop a radiotracer that selectively binds to and illuminates mut-IDH1 within glial cells in the brain during a PET scan. The major requirement is that the radiotracer (and hence the precursor) should bind selectively to the cancer-causing mut-IDH1 enzyme, a condition that all known mutant IDH1 inhibitors do not satisfies.[31, 32] In practice, this renders the method not feasible. By developing a QM/MM-based drug design protocol with MiMiC (Figure 9.1), I performed DFT-QM/MM MD simulations of the mut-IDH1 enzyme for the first time. This quantum-level understanding of the enzyme active site allowed for the design of substrate analogs which satisfy the selectivity requirement of a PET radiotracer. This opens up the possibility of making the non-invasive detection of mut-IDH1 associated glioma with PET more feasible, allowing for early diagnosis and greater survivability of patients. This work was supported within the Helmholtz European Partnership program ("Innovative high-performance computing approaches for molecular neuromedicine") between Forschungszentrum Jülich and the Italian Institute of Technology (IIT). This included Dr. Marco De Vivo from IIT, and our experimental collaborators Prof. Bernd Neumaier and Dr. Roberta Cologni from Forschungszentrum Jülich, who were involved in developing the PET radiotracers in

the wet lab.

1.1. Motivation for the Thesis

1.1.1. Why QM/MM MD in Drug Design?

The main goal of this thesis is to reliably include MiMiC QM/MM MD within a drug design pipeline. This would greatly expand drug design for these broad classes of drug targets:

1. Metalloproteins: Generation of an ensemble of drug target structures and/or target-inhibitor complexes (see Section 2.1.2) close to the ground state or Michaelis complex is vital information for accurate drug design. In metalloproteins, the effect of the metal in the active site is purely a quantum phenomenon.[9, 33] Generating an ensemble of metalloprotein structures would then require QM/MM MD. Metalloproteins form an integral part of the human proteome, and despite being an important therapeutic target, has been difficult to drug.[34, 35, 36, 37]

2. Ribonucleic acids: RNAs have also recently emerged as an exciting new therapeutic target.[38, 39] Analogous to metalloproteins, many important functions in RNAs are mediated by metal centers, necessitating inclusion of a quantum description.[40, 41] In fact, it is not clear if classical force fields can even describe many basic properties of RNA. For e.g., during H-bonding between base pairs, the X–H covalent bonds in the H-bond donor elongates by up to ~0.04 Å, clearly demonstrating a change in the electronic structure.[42] This is not describable by current force fields.[40] RNA-protein interactions might also benefit from a QM/MM treatment.[10]

3. Covalent Inhibitors: Traditional drug design aims to design small molecules that interact with the biological target through non-covalent means (H-bonding, van der Waals interactions) in a fast and reversible fashion. However, focus has recently shifted to irreversible, covalent inhibitors with a more prolonged duration of drug-target interactions.[43, 44] A rational approach to covalent drug design will benefit from the explicit inclusion of quantum effects. This also includes metal-containing drugs like cisplatin, which is an active area of study within chemotherapy.[45, 46, 47, 48] Covalent inhibitors have also emerged as an attractive solution for targeting the notoriously 'undruggable' KRAS, involved in cancer.[49] QM/MM MD simulations predicting the binding poses of these inhibitors would be extremely valuable for structure optimization.

4. Transition State Analogs: QM/MM MD simulations can be used to obtain in-

formation on the transition state of enzymes and ribozymes.[9, 50, 51] Complex chemical reaction can only be reliably described with a full quantum description, preferably with dynamics. As mentioned in Section 2.1.2, knowledge of the transition state can allow for the design of transition state analogs.[52, 53] These, according to transition state theory, are predicted to have unusually high binding affinities and selectivity for the enzyme/ribozyme.[54] This leads to reduced dosage requirements and toxicity, potentially a huge boost to drug design efforts. Routine QM/MM MD of biological targets would allow for more rational design of transition state analogs as inhibitors.

QM/MM MD within a drug design pipeline allows us to expand the druggable chemical space, and design drugs with improved binding affinity and lower off-target effects. This could form a major part of the solution to the 90% failure rate in drug design, greatly reducing the time and cost investment for pharmaceutical companies.

1.1.2. Problems Integrating MiMiC in Drug Design

Usually an initial step of cMD equilibiration is performed before performing QM/MM MD.[21, 22] Starting a QM/MM simulation from a raw experimental structure will lead to extremely unstable simulations. Thus incorportating MiMiC QM/MM MD in a drug design protocol would lead to the tentative Quantum HPC-based virtual screening (QHPC–VS) protocol proposed in Figure 1.3. Despite the potential, this pipeline faces the following problems:

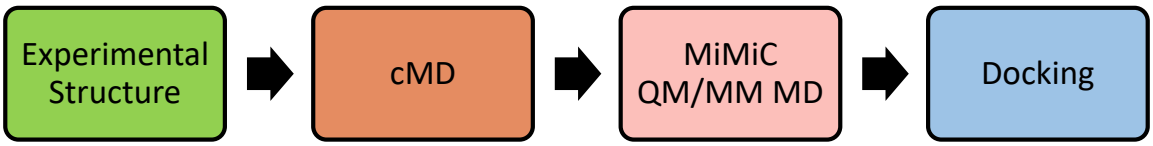


Figure 1.3. A tentative structure based-virtual screening protocol involving MiMiC QM/MM MD, which we refer to as the QHPC–VS (Quantum HPC-based virtual screening) protocol.

1. **MiMiC Input Preparation**. This pipeline involves 3 transitions, (i) between the experimental structure and cMD, (ii) between cMD (GROMACS) and MiMiC, (iii) MiMiC and docking (Schrödinger). GROMACS and Schrödinger used in this work are highly established codes, meaning that transitions (i) and (iii) would be very easy. However, since MiMiC is a relatively new software, a seamless way to move from GROMACS MM structures to the corresponding MiMiC QM/MM structures is not available. This is absolutely necessary to include MiMiC QM/MM within any (automated) drug design pipeline. The loose-coupling nature of the software makes it is often tedious and difficult to

generate CPMD and GROMACS in a consistent framework. This reduces the utility of MiMiC as a CADD tool.

2. **Efficiency for Drug Targets**. MiMiC is of one best candidates to perform efficient simulations of large biological systems available in the literature.[19, 21, 23, 24] However, its performance for larger biomolecules relevant for drug design has not been so thoroughly studied. Specifically, the scaling of MiMiC (although theoretically predicted to be excellent) has not been studied for a therapeutically-relevant drug target. This leaves the question open on whether MiMiC can reach the time and length scales necessary for drug design.

3. **Not simulatable with this Pipeline.** QM/MM MD is often specifically useful in cases where classical force fields fail to adequately reproduce the behavior of certain drug targets (see Section 1.1.1). However in such situations, the need to perform an initial cMD step for a sufficient period of time might push the system so far away from the quantum energy surface that it would be irrecoverable even within the QM/MM regime. In such cases, the protocol in Figure 1.3, with a primary cMD step feeding into a subsequent QM/MM MD step, would fail. A more complex pipeline will be needed for such complex targets.

4. **Applicability in CADD.** As discussed in Section 1.1.1, QM/MM could potentially greatly improve CADD. Furthermore, the MiMiC framework presents compelling QM/MM performance for biomolecules on supercomputers. However, even if such QM/MM simulations are achieved, it remains to be seen if it can distinct results in drug design over current classical methods. Specifically, it remains to be seen if QM/MM MD can provide significant advantages in drugging targets not amenable to current methods.

These challenges have, so far, prevented the community from reliably incorporating MiMiC—or QM/MM MD in general—into drug design or virtual screening protocols. Starting from the QHPC–VS protocol of Figure 1.3, I remedy the problem explained above one-by-one in Chapters 5–7. The protocol will be refined further, eventually allowing us to use it to perform drug design against the hard-to-drug mut-ID1 active site for glioma diagnosis.

1.2. Thesis Content

The theoretical background required to understand the results are explained in Chapters 2–4.

CADD and MD form the primary background of this work. The theory behind these are presented in Chapter 2, including a summary of classical, and multiscale

QM/MM MD for biological drug targets.

This work uses QM/MM MD as implemented in the MiMiC package. An introduction to the software, and the HPC-based parallelism techniques used to achieve performance on supercomputers, are presented in Chapter 3.

The IDH1 enzyme is the therapeutic of focus in this work. The biological and therapeutic role of this are discussed in Chapter 4. Difficulties in drugging this target are also mentioned.

The results involve solving each of the four problems of the proposed QHPC–VS protocol (Figure 1.3) discussed in Section 1.1.2.

The first of the contributions of this thesis is presented in Chapter 5 and encompasses the MiMiCPy software tool. It expedites the conversion of the GROMACS MM inputs to MiMiC-QM/MM inputs. This makes MiMiC, hitherto a software with a steep learning curve, much easier to use for the novice user. This is envisaged as the primary front-end for the MiMiC framework, and is crucial to make MiMiC-QM/MM MD more palatable for the drug design community.

QM/MM MD simulations for wt-IDH1 are presented in Chapter 6, which constitute the first use of MiMiC for a large, therapeutically-relevant enzyme. Importantly, the scalability of MiMiC was pushed beyond any previous work, up to an unprecidented number of cores (85,000) on the JUWELS Jülich Supercomputing System.

Chapter 7 is devoted to QM/MM simulations of mut-IDH1 enzyme. This protein, for various reasons, is not treatable with the QHPC–VS protocol (Figure 1.3). So innovative modifications to this protocol were proposed to simulate this enzyme at the *ab-initio* level.

Finally, a molecular docking procedure to suggest candidates for the mut-IDH1 active site is discussed in Chapter 8. No drug candidates have been found for this binding site before. This results in small-molecule precursors of PET radiotracers for the detection glioma. This is hoped to advance the field of glioma therapy by providing a non-invasive method of diagnosis.

1.3. List of Publications

In accordance with §5(3) of the doctoral regulations extract of this thesis have been published or submitted for publication in the following papers:

[A] B. Raghavan, F. K. Schackert, A. Levy, S. K. Johnson E. Ippoliti, D. Mandelli, J. M. H. Olsen, U. Röthlisberger, and P. Carloni. MiMiCPy: An Efficient

Toolkit for MiMiC-based QM/MM Simulations. *Journal of Chemical Information and Modeling*, 63(5), 1406–1412, **2023**.

[B] B. Raghavan, M. Paulikat, K. Ahmad, L. Callea, A. Rizzi, E. Ippoliti, D. Mandelli, L. Bonati, M. D. Vivo, and P. Carloni. Drug Design in the Exascale Era: A Perspective from Massively Parallel QM/MM Simulations. *Journal of Chemical Information and Modeling*, 63(12), 3647–3658, **2023**.

Chapter 5 is based on [A], and Chapter 6 corresponds to the IDH1 simulations of [B]. A paper corresponding to the results of Chapter 7–8 is in preparation. According to §5(6) of the doctoral regulations, I declare my contributions to the papers above:

[A] The code was developed by F. K. Schackert and me. Writing of the manuscript was spearheaded by me, with all authors commenting on the manuscript. Figures were prepared by F. K. Schackert.

[B] I carried out all simulations and analyzed results pertaining to the IDH1 enzyme. Writing of the paper was a collaborative effort.

2. Theory: Molecular Dynamics and Drug Design

This chapter will briefly introduce the theoretical background to the methods used in computer-aided drug design (CADD) and molecular dynamics (MD). It is divided into two sections. The first one will introduce the various terminologies used in CADD, and then focus on structure-based drug design. The second will present the background of MD and focus on the different intramolecular potential energy functions such as: molecular mechanics (MM), quantum mechanics (QM), and the mixed QM/MM methods.

2.1. Structure-based Drug Design in CADD

The high costs and failure rate associated with the traditional path of drug discovery and development has prompted the need for CADD.[5] Over the past decades, it has proven to provide a better hit rate of novel drug compounds because with a more targeted search than traditional high throughput screening and combinatorial chemistry.[55] Referring back to the complete drug design pipeline (Figure 1.1) from the Chapter 1, CADD fits in perfectly into the pre-clinical stages[6]:

1. Target ID: Simulate the drug target biomolecule in physiological conditions to gain a deeper understanding of its behavior.

2. Hit ID: Filter large compound libraries of small molecules into smaller sets of potential active compounds against the drug target. Then, only this set needs to be tested experimentally.

3. Lead Optimization: Optimize lead compounds for affinity, drug metabolism and pharmacokinetics (DMPK), and properties including absorption, distribution, metabolism, excretion, and toxicity (ADMET). This may involve adding/subtracting functional groups and structural optimization of the lead candidate.

CADD broadly encompasses structure-based and ligand-based methods.[55] The former requires accurate structural information of the target protein structure to

rank compounds on the interaction with the protein, whereas the latter exploits the knowledge of known active and inactive molecules of the target and attempts to filter the compound database using chemical similarity searches or construction of predictive, quantitative structure-activity relation (QSAR) models. Structure-based CADD is generally preferred where high-resolution structural data of the target protein are available. A lack of such data necessitates the usage of less accurate ligand-based techniques. The most common of these ligand-based techniques is pharmacophore analysis.[56] The IUPAC definition of a pharmacophore is 'the ensemble of steric and electronic features that is necessary to ensure the optimal supra-molecular interactions with a specific biological target structure and to trigger (or to block) its biological response'.[57, 56] In pharmacophore analysis, the (3D or 2D) spatial arrangement of chemical features of known active compounds are determined. These are then aligned against unknown molecules, and the degree of matching is used to predict if the molecule would be active or not. This presents an extremely fast method to filter a large database of small molecules, and is often used as a first step before performing more expensive molecular docking (see below). The pharmacophore of active ligands can be generated with just the free ligand structures, or alternatively using protein-ligand complex.[58] Using the latter method would push the pharmacophore analysis closer to structure-based drug design.

The main goal of structure-based CADD is to design lead compounds that bind tightly to the target (high binding affinity), improved DMPK/ADMET properties, and are target specific (have reduced off-target effects).[59] A successful application of this method will result in a compound that has been validated *in vitro*/*in vivo* with its binding location confirmed by experiments. The process begins with the identification of the biological target (usually a protein) involved in the disease using various wet-lab techniques. During the process, the drug target is crystallised using various techniques (NMR, X.Ray, etc.) to obtain the 3D determinants of the protein. Two approaches can then be taken: virtual screening (VS) and *de novo* drug design.[60] In VS, databases of millions of commercially available drug-like molecules are computationally screened against targets of known structure, and are ranked based on their predicted bind affinity (workflow shown in Figure 2.1). The exact screening and ranking of compound libraries is accomplished by molecular docking, where ligands are filtered based on a calculated docking score. The top hits of the VS are then further sieved for desirable properties and tested *in vitro*. However, using existing compound databases to initiate VS does not result in molecules that are structurally "novel". In the *de novo* drug design approach, the 3D structure of the receptor is used to design structurally novel molecules that have never been synthesized before using various computational techniques.

The central problem of this work is the design of mut-IDH1 specific small molecules, which can be utilized as radiotracers in the PET imaging of glioma. For this, we will largely utilize the structure-based VS method. The exact protocol used in described in Chapter 8. From the previous discussion it is clear that for accurate structure-

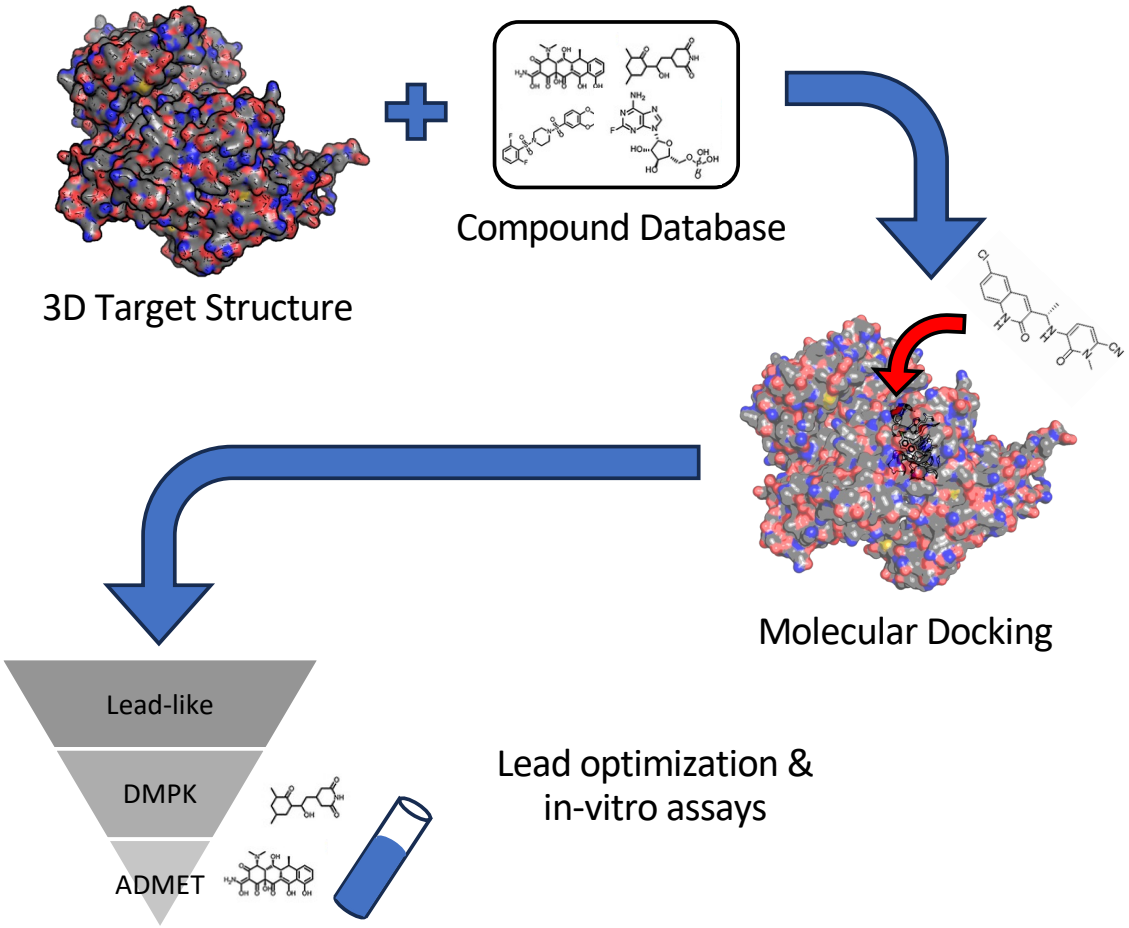


Figure 2.1. An overview of steps involved in structure-based virtual screening.

based drug design, careful considerations must be given to the (i) molecular docking procedure, and (ii) quality of the 3D target structure(s). We will explore these in some depth in Sections 2.1.1 and 2.1.2.

2.1.1. Molecular Docking

The aim of molecular docking is to predict the ligand-receptor complex structure in a relatively fast manner, so a large library of compounds can be tested. Docking involves two steps: first prediction of the various poses of the ligand within the binding site using sampling methods; then ranking these poses via a scoring function.[61, 62] Ideally, this process should reproduce the experimental binding mode of the ligand as the pose with the highest scoring function among all generated conformations.

Sampling methods are search algorithms that generate ligand pose at a target's

binding site, taking into consideration the rotational, translational and internal degrees of freedom of the ligand.[63, 64] The first type of sampling is where none is carried out, i.e., both the ligand and the target receptor are fixed.[65] This rigid docking approximation is analogous to the Emil Fischer's famous "lock-key" binding model[66] and is largely restricted to cases where the number of conformational degrees of freedom is too high to be sampled (for e.g., protein-protein docking). Algorithms were the conformations are actually searched can be systematic, where each ligand's degree of freedom is searched incrementally. As the number of free rotatable bonds increases, the computer time required to explore all evaluations can become prohibitively expensive. On the other hand, stochastic search algorithms perform random changes in the ligand's degrees of freedom. This can deal with more degree of freedom, but does not guarantee convergence to the best solution.

After generating the various ligand poses, an associated docking score is used to rank the conformations. The functions used to calculate the docking are usually loosely based on the calculations of the binding energy, free energy, or approximate interaction energies. These scoring functions can be divided into into three major types: force field, empirical and knowledge-based.[61, 67] Force field-based functions consist of a sum of energy terms, similar to the classical force fields used in MD (see Section 2.2.2.1). This include potential energy terms like bonded (covalent bond, angle, dihedrals, torsional) and nonbonded (van der Waals, electrostatic) terms. The DockThor[68] employs a force-field based scoring function based on the MMFF94S force field[69], utilizing just the torsonial bonded interactions, the electrostatic interactions in the buffered coulombic form and the Buffered-14-7 form of the van der Waals potential. Force fields can be extended to include emperical terms capturing the hydrogen bonds, solvation and entropy contributions. For example, AutoDock[70] uses the following semi-empirical scoring function:

$$\begin{aligned} V = \Delta G_{vdm} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \Delta G_{hbond} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ + \Delta G_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + \Delta G_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{\left(\frac{-r_{ij}^2}{2\sigma^2}\right)} \end{aligned} \tag{2.1}$$

The four terms on the right-hand side represents the pair-wise van der Waals interaction, hydrogen bond, coulomb energy and desolvation energies between two atoms i and j. Purely empirical scoring functions consist of a similar sum of terms such as hydrogen bond, ionic interaction, hydrophobic effect and binding entropy, each with an associated coefficient.[64] These coefficients are obtained from empirical means like fitting to a test set of ligand-protein complexes with known binding affinities. SP GlideScore is an example of this consisting of the following sum of terms:

$$\begin{aligned}\Delta G_{bind} = & C_{lipo-lipo} \sum f(r_h) + C_{hbond-neut-neut} \sum g(\Delta r)h(\Delta\alpha) \\ & + C_{hbond-neut-charged} \sum g(\Delta r)h(\Delta\alpha) \\ & + C_{hbond-charged-charged} \sum g(\Delta r)h(\Delta\alpha) \\ & + C_{max-metal-ion} \sum f(r_{lm}) \\ & + C_{rotb}H_{rotb} + C_{polar-phob}H_{polar-phob} + C_{coul}H_{coul} \\ & + C_{vdW}H_{vdW} + solvation\ terms\end{aligned} \tag{2.2}$$

The fifth term, the metal-ligand interaction term, is to be noted here. It has special considerations for modelling the coordination bonds between anionic acceptor ligand atoms (such as either of the two oxygens of a carboxylate group) and net-positive metal centers.[71, 72] This allows for the greater performance of GlideScore in recognizing the strong preference for coordination of anionic ligand functionality to metal centers in metalloproteins. The metal term is primarily fit to zinc and magnesium ions, with lesser contributions from other metals relevant in biology.[73] These effects are difficult to include in a purely classical force field approach, without resorting to some kind of quantum calculation. This scoring function (albeit through the modified Glide XP procedure[74]) is utilized in Chapter 7 to perform docking studies against mut-IDH1.

Empirical scoring functions, as opposed to force-field functions, are simpler to evaluate.[61] However, it is unclear as to how well they are suited for ligand-protein complexes beyond the training set. Recently developed machine learning and deep learning-based scoring functions, which can be classified as more advanced empirical scoring functions, might help in this regard.[75, 76] Lastly, knowledge-based functions involve statistical analysis of the frequency of atom pair interactions observed in experimentally determined 3D structures of ligand-target complexes.[77, 62] They are based on the assumption that the more favorable an interaction is, the greater the frequency of its occurrence will be. The score is calculated by favoring the more frequent contacts and penalizing repulsive interactions between the ligand and protein. An example of this is the PMF scoring function.[78] These functions are even more simpler than empirical-based, and can furthermore capture uncommon behavior like sulphur-aromatic or cation-π interactions.[61, 79] But, they also highly suffer from biases in the training set.

2.1.2. Target Flexibility and MD Simulations

In Section 2.1.1, sampling techniques were touched upon including rigid docking (where the both the ligand and the target are kept fixed) and semi-rigid docking (where the ligand conformations are sampled). In the last case, the target is kept fixed. This approximation is often not accurate, for example, protein drug targets

possess an inherent flexibility that can greatly affect drug-receptor binding. Thus, there is need to introduce flexible docking schemes.[80, 81, 82] Various methods attempt to account for the protein flexibility induced by the approach of the ligand on the fly. These include predicting side-chain conformations either from discrete rotamer libraries[83], on a continuum[84], or by local optimization[85]. For example, Glide/Prime include and induced-fit docking procedure to account for the full flexibility of a limited number of receptor residues.[86]

These methods may not go far enough in accounting for the protein flexibility, as they capture only localized phenomena. A solution is to use a pre-existing or pre-generated set of multiple receptor conformations of the binding pocket, and sequentially dock to these structures. This is called ensemble docking.[81] These structure would ideally be obtained from X-Ray and/or NMR analysis of the drug target. However, the conformations space explored by crystallography is often limited and are biased towards a few configurations. Under physiological conditions, the drug target is in thermodynamics equilibrium at around 300 K with the surrounding solvent and ions. From a physical perspective, this is what imparts the intrinsic flexibility to the biomolecule. However, the crystal structure of protein is obtained by cooling the protein to low temperatures (around 100 K) to freeze it into a single (often biased) stationary structure. Furthermore, the structure of enzymes are often modified to prevent the protein from undergoing catalysis or other activities. Thus, there is a need to correct the crystal structure of the protein, by bringing it back into physiological conditions and sampling as many configurations as possible.[65]

The method of molecular dynamics (MD) is an excellent way to achieve this.[7, 87, 88, 89] The use of MD with docking is summarized in Figure 2.2. MD starts from a single initial crystal structure of the target, and generates successive configurations of the apo protein evolving in time. This allows for larger exploration of the configurational landscape. One of the earliest attempts at utilizing the superior exploration capabilities of MD was by McCammon and co-workers[90, 91], and since been successfully used to design inhibitors against the cancer-relevant MDM2/MDMx-p53 interaction[92], among others[93, 94]. MD can also be used to postprocess docked structures to reveal unstable binding modes, filter out physically unreliable docking solutions, identify new ones, and allow for rescoring of the docking score.[89] This is achieved because MD can naturally account for the flexibility and induced fit effects, albeit as a post-docking effect.

The sampling problem takes on an expanded meaning in the specific case of enzymes as drug target (which form the vast majority of cases). Enzyme catalysis can be described by the Michaelis-Menten model:

$$\mathrm{E} + \mathrm{S} \rightleftharpoons \mathrm{ES} \xrightarrow{\mathrm{TS}} \mathrm{EP} \longrightarrow \mathrm{E} + \mathrm{P} \tag{2.3}$$

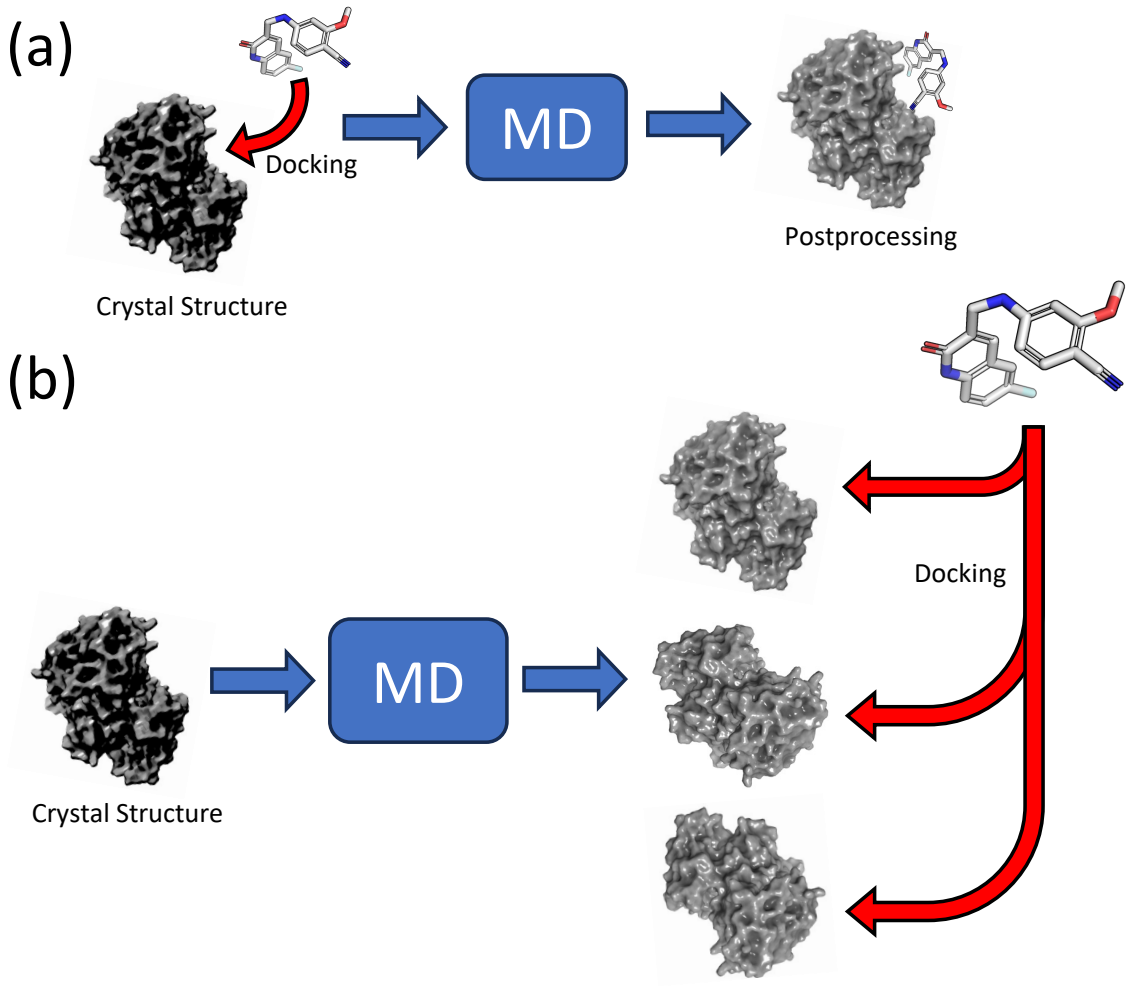
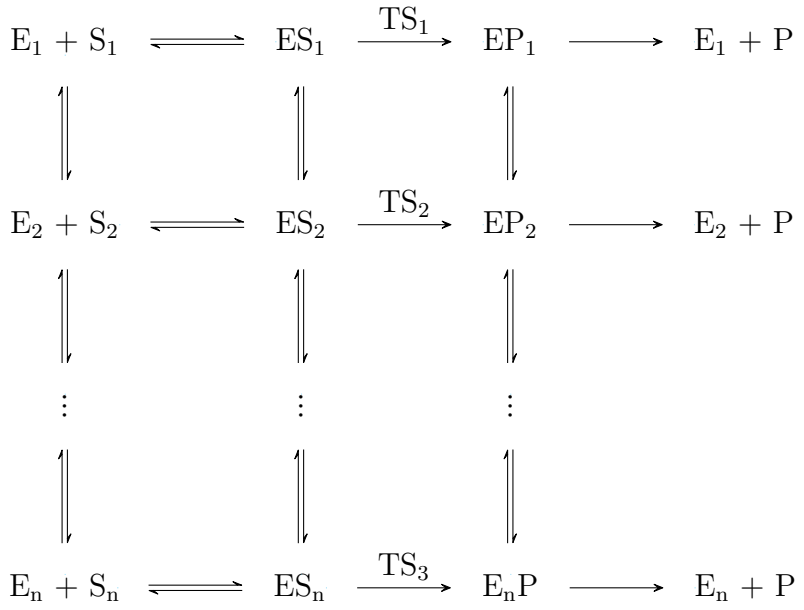


Figure 2.2. MD can be employed, within a structure-based virtual screening protocol, as a (a) postprocessing tooling for rescoring or refining docking poses; (b) conformational ensemble generator for ensemble docking.

This involves the enzyme E reversibly binding to a substrate S to form a complex ES (the Michaelis complex). The Michaelis complex ES reacts to form the Enyzme-Product complex EP, finally leading to the release of the product P and regeneration of the enzyme E. The conversion of ES to EP occurs through the transition state TS. The state E corresponds to what is generally referred to as the enzyme in the 'open' configuration, and the binding of S would transition the enzyme from the open to the 'closed' configuration.[95] Enzymes are evolutionary designed to tightly bind substrates at the transition state and the intermediates, due to their high-energy nature. Any ligand that mimics these states would also bind tightly to the enzyme and should function as excellent inhibitors (the so-called 'transition state analog').[52, 53] Furthermore, they should theoretically be highly selective to the that specific enzyme, reducing off-target effects and toxicity. Thus, performing molecular docking

with enzyme conformations sampled along the ES to TS conversion spectrum would theoretically lead to excellent drug candidates. The relevance of conformer sampling up to TS takes further significance when we consider the arguments of Ma and Nussinov.[96] As opposed to the static picture of the classical Michaelis-Menten equation (Equation 2.3), fluctuations of the enzyme lead to an ensemble of ES, TS, and EP complexes. Thus, a more holistic description of the overall catalytic process should consider the interconverting conformers at each step:



(2.4)

Performing ensemble docking with various conformations of ES_1, ES_2, TS_1, TS_2, etc. should greatly aid in the design of superior transition state analogs. Unfortunately, configurations sampled by crystallography would usually correspond to an 'early' Michaelis complex, or configuration located somewhere between the binding of S to E and formation of the tightly bound Michaelis Complex ES. This is the ground state of the enzyme, and anything beyond this (the spectrum along the ES to TS) would be too short-lived and transient to capture experimentally. MD simulations can help in pushing the system from the early Michaelis complex to a later structure, closer to the transition states, and sampling many possible configurations at this stage of the catalysis. This is especially possible if we augment MD with various enhanced sampling techniques and a quantum description of the active site to completely describe bond formation/breaking. Thus, MD is an integral part of the modern CADD pipeline. A theoretical background on MD is described in Section 2.2.

2.2. Molecular Dynamics

MD is a simulation method to study a system of atoms and molecules by propagating it in time. Molecular systems relevant to biology and chemistry are fundamentally non-relativistic quantum mechanical systems, and we start with the time-dependent Schrödinger's equation of the system to describe such systems:[1]

$$i\hbar\frac{\partial}{\partial t}\Phi(\{\boldsymbol{r_i}\},\{\boldsymbol{R_I}\};t) = \mathcal{H}\Phi(\{\boldsymbol{r_i}\},\{\boldsymbol{R_I}\};t) \tag{2.5}$$

Here $\mathcal{H}$ is the quantum Hamiltonian of the system:

$$\begin{aligned}\mathcal{H} &= \sum_I \frac{\hbar^2}{2M_I}\nabla_I^2 - \frac{\hbar^2}{2m_e}\nabla_I^2 + \frac{1}{4\pi\varepsilon_0}\sum_{i<j}\frac{e^2}{|\boldsymbol{r_i}-\boldsymbol{r_j}|} - \frac{1}{4\pi\varepsilon_0}\sum_{i<j}\frac{e^2 Z_I}{|\boldsymbol{R_I}-\boldsymbol{r_i}|} \\ &+ \frac{1}{4\pi\varepsilon_0}\sum_{i<j}\frac{e^2 Z_I Z_J}{|\boldsymbol{R_I}-\boldsymbol{R_J}|} \\ &= \sum_I \frac{\hbar^2}{2M_I}\nabla_I^2 + \mathcal{H}_e(\{\boldsymbol{r_i}\},\{\boldsymbol{R_I}\})\end{aligned} \tag{2.6}$$

The Hamiltonian depends on the electronic $\{\boldsymbol{r_i}\}$ and nuclear $\{\boldsymbol{R_i}\}$ degrees of freedom, and the atomic mass M_I, the electronic mass m_e, and charge Z_I. To understand the behaviour of the system, we need to solve for $\Phi(\{\boldsymbol{r_i}\},\{\boldsymbol{R_i}\};t)$, which is the total wavefunction of both the electron and nuclei. This is usually done by employing two approximations:

(1) **Born-Oppenheimer approximation.** First the ansatz, introduced by Born, is applied to the total wavefunction for the *time-independent* Schrödinger's equation to separate the light electrons from the heavy nuclei.[98] On top of this, we introduce the adiabatic approximation, which implies that the motion of the nuclei proceeds without changing the quantum state of the electronic subsystem during time evolution. This leads to the splitting of the total wavefunction into two components:

$$\Phi(\{\boldsymbol{r_i}\},\{\boldsymbol{R_I}\};t) \approx \Psi(\{\boldsymbol{r_i}\};\{\boldsymbol{R_I}\})\chi_k(\{\boldsymbol{R_I}\};t). \tag{2.7}$$

Here k is the quantum state of the electrons. Inserting 2.7 into Equation 2.5, the equation can be split into two coupled equations. One is the time-dependent equation describing the nuclear motion:

[1] This derivation is adapted from Ref [97].

$$\left[-\sum_I \frac{\hbar^2}{2M_I} \nabla_I^2 + E_k(\{\boldsymbol{R_I}\})\right] \chi_k = i\hbar \frac{\partial}{\partial t} \chi_k \tag{2.8}$$

The second is the time-independent equation describing the electronic state:

$$\mathcal{H}_e(\{\boldsymbol{r_i}\};\{\boldsymbol{R_I}\})\Psi_k = E_k(\{\boldsymbol{R_I}\})\Psi_k(\{\boldsymbol{r_i}\};\{\boldsymbol{R_I}\}) \tag{2.9}$$

We can rewrite the nuclear equation as a set of coupled equations (Equation 2.11 and 2.12) by choosing to express the nuclear wavefunction in terms of an amplitude factor A_k and a phase S_k (both considered to be real) as in Equations 2.10.

We can now choose to express the nuclear wavefunction in terms of an amplitude factor A_k and a phase S_k (both considered to be real):

$$\chi_k(\{\boldsymbol{R_I}\};t) = A_k(\{\boldsymbol{R_I}\};t) exp\left[iS_k(\{\boldsymbol{R_I}\};t)/\hbar\right] \tag{2.10}$$

This allows as to rewrite the nuclear equation as a set of coupled equations in the phase S_k:

$$\frac{\partial S_k}{\partial t} + \sum_I \frac{1}{2M_I} (\nabla_I S_k)^2 + E_k(\{\boldsymbol{R_I}\}) = \hbar^2 \sum_I \frac{1}{2M_I} \frac{\nabla_I A_k}{A_k} \tag{2.11}$$

and amplitude factor A_k:

$$\frac{\partial A_k}{\partial t} + \sum_I \frac{1}{M_I} (\nabla_I A_k)(\nabla_I S_k) + \sum_I \frac{1}{2M_I} A_k \left(\nabla_I^2 S_k\right) \tag{2.12}$$

Notice that Equation 2.11 has one term that depends explicitly on $\hbar$. Here we apply the second approximation.

(2) **Classical limit.** We assume that the nuclei can be treated as classical particles, and $\hbar \to 0$ within this limit. This gives us:

$$\frac{\partial S_k}{\partial t} + \sum_I \frac{1}{2M_I} (\nabla_I S_k)^2 + E_k(\{\boldsymbol{R_I}\}) = 0 \tag{2.13}$$

This is isomorphic to the Hamiltonian-Jacobi equation of classical mechanics, with the nuclear phase S_k being equivalent to the classical action functional of the nuclei. The classical momentum is related to the derivative of the action functional, which

then allows us to write the classical momentum of the nuclei $\boldsymbol{P_I}$ is terms of the nuclear phase S_k:

$$\boldsymbol{P_I} = \nabla_I S_k \tag{2.14}$$

This definition of momentum allows us to write the electronic energy at the kth state in terms of the classical nuclear momenta:

$$\frac{d\boldsymbol{P_I}}{dt} = M_I \ddot{\boldsymbol{R}}_{\boldsymbol{I}}(t) = -\nabla_I E_k(\{\boldsymbol{R_I}; t\}) \tag{2.15}$$

With Equation 2.15, we have recovered Netwon's equation of motion. Thus, the nuclei can be approximated as classical particles moving in an effective potential E_k, which is the Born-Oppenheimer potential energy surface obtained by solving the electronic time-independent Schrödinger's equation (Equation 2.9) at the given configuration. Solving Equations 2.9 and 2.15 simultaneously is the essential formalism of MD, or more specifically Born-Oppenheimer MD (BO-MD). Depending on the exact approximation used, other flavors of MD (like Ehrenfest and Car-Parrinello), can be derived. But these will not be discussed here as this work only utilises BO-MD.

Practically implementing MD on a computer requires consideration towards finding efficient algorithms for solving: (i) the classical equation of motion of the nuclei, and (ii) time-independent Schr

2.2.1. Nuclear Equation of Motion

Equation 2.15 is a many-body problem of n nuclei, with $n \sim 10^6$ for the most common biological drug targets (in a solvated environment). This is not analytically solvable. Hence, we utilize various numerical integration techniques, by discretizing the time into Δt timesteps, to simulate its time evolution.[2] The most simplest of these techniques is the Euler method, where we consider the Taylor series expansion of $\boldsymbol{R_I}(t + \Delta t)$ around t:

$$\boldsymbol{R_I}(t + \Delta t) = \boldsymbol{R_I}(t) + \boldsymbol{V_I}(t)\,\Delta t - \frac{\nabla_I E_k(t)}{2M_I}\Delta t^2 + \mathcal{O}\left(\Delta t^3\right) \tag{2.16}$$

And the same for velocity $\boldsymbol{V_I}(t)$:

[2]The discussion of numerical intergration in MD is adapted from Ref [99].

$$\boldsymbol{V_I}(t+\Delta t) = \boldsymbol{V}(t) - \frac{\nabla_I E_k(t)}{M_I}\Delta t + \mathcal{O}\left(\Delta t^2\right) \tag{2.17}$$

This scheme is not reliable as it does not converge the energy. Furthermore, the accuracy of the position and velocities are too low. This means that Δt should be chosen to be small enough to maintain the validity of neglecting higher order terms. Simultaneously, a very small term will only allow for limited exploration of the time evolution of the system. An improvement to the Euler method is the Verlet method, where the Taylor series expansion of $\boldsymbol{R_I}(t \pm \Delta t)$ around t is considered:

$$\boldsymbol{R_I}(t+\Delta t) = \boldsymbol{R_I}(t) + \boldsymbol{V_I}(t)\Delta t - \frac{\nabla_I E_k(t)}{2M_I}\Delta t^2 + \frac{\Delta t^3}{6}\frac{d^3\boldsymbol{R_I}}{dt^3} + \mathcal{O}\left(\Delta t^4\right) \tag{2.18}$$

$$\boldsymbol{R_I}(t-\Delta t) = \boldsymbol{R}(t) - \boldsymbol{V_I}(t)\Delta t - \frac{\nabla_I E_k(t)}{2M_I}\Delta t^2 - \frac{\Delta t^3}{6}\frac{d^3\boldsymbol{R_I}}{dt^3} + \mathcal{O}\left(\Delta t^4\right) \tag{2.19}$$

Summing these two equations gives us the Verlet algorithm:

$$\boldsymbol{R_I}(t+\Delta t) = 2\boldsymbol{R_I}(t) - \boldsymbol{R_I}(t-\Delta t) - \frac{\nabla_I E_k(t)}{m}\Delta t^2 + \mathcal{O}\left(\Delta t^4\right) \tag{2.20}$$

This has improved the accuracy of the position, and also ensures energy conservation. This allows for a stable integration algorithm at a relatively larger Δt. In practice, the time can range from 0.1 to 2 fs, depending on the solution to E_k (discussed in Section 2.2.2). Notice here that this scheme does not use the velocity to compute the new position. The velocity is however important to calculate the kinetic energy and observables like the temperature of the system. We can recover the velocity by using Equation 2.21.

$$\boldsymbol{V_I}(t) = \frac{\boldsymbol{R_I}(t+\Delta t) - \boldsymbol{R_I}(t-\Delta t)}{\Delta t} + \mathcal{O}\left(t^2\right) \tag{2.21}$$

The velocity derived this way is of lower accuracy that the positions, and is a step behind the position term (since this is for the velocity at time t). This might lead to inaccuracies of crucial quantities calculated from the MD trajectory. To remedy this, the leapfrog formulation was introduced. Here the Taylor series expansion of $\boldsymbol{R_I}(t)$ and $\boldsymbol{R_I}(t+\Delta t)$ around $\Delta t/2$ is taken:

$$\boldsymbol{R_I}(t) = \boldsymbol{R_I}\left(t+\frac{\Delta t}{2}\right) - \frac{1}{2}\boldsymbol{V_I}\left(t+\frac{\Delta t}{2}\right)\Delta t + \frac{\nabla_I E_k\left(t+\frac{\Delta t}{2}\right)}{8m}\Delta t^2 + \mathcal{O}\left(\Delta t^3\right) \tag{2.22}$$

$$\boldsymbol{R_I}(t+\Delta t) = \boldsymbol{R_I}\left(t+\frac{\Delta t}{2}\right) + \frac{1}{2}\boldsymbol{V_I}\left(t+\frac{\Delta t}{2}\right)\Delta t + \frac{\nabla_I E_k\left(t+\frac{\Delta t}{2}\right)}{8m}\Delta t^2 + \mathcal{O}\left(\Delta t^3\right) \quad (2.23)$$

Subtracting these two equations gives us the update for the positions within the leapfrog scheme:

$$\boldsymbol{R_I}(t+\Delta t) = \boldsymbol{R_I}(t) + \Delta t \cdot V_I\left(t+\frac{\Delta t}{2}\right) + \mathcal{O}\left(\Delta t^3\right) \quad (2.24)$$

Putting this back in Equation 2.20, gives the update for the velocities:

$$\boldsymbol{V_I}\left(t+\frac{\Delta t}{2}\right) = \boldsymbol{V_I}\left(t-\frac{\Delta t}{2}\right) - \frac{\nabla_I E_k(t)}{M_I}\Delta t + \mathcal{O}\left(\Delta t^3\right) \quad (2.25)$$

The accuracy of the velocities has been improved, but it is still staggered with respect to the positions by $\Delta t/2$. To increase the accuracy of the positions and velocities, and also maintain them in-sync, the velocity Verlet scheme was introduced. The Taylor series expansion of $\boldsymbol{R_I}(t+\Delta t)$ around Δt is taken:

$$\begin{aligned}\boldsymbol{R_I}(t+2\Delta t) = \boldsymbol{R_I}(t+\Delta t) + \boldsymbol{V_I}(t+\Delta t)\Delta t - \frac{\nabla_I E_k(t+\Delta t)}{2M_I}\Delta t^2 \\ + \frac{\Delta t^3}{6}\frac{d^3 r}{dt^3} + \mathcal{O}\left(\Delta t^4\right)\end{aligned} \quad (2.26)$$

This and Equation 2.18 are subtracted to give Equation 2.27:

$$\begin{aligned}\boldsymbol{V_I}(t+\Delta t) = \boldsymbol{V_I}(t) + \frac{\nabla_I E_k(t) - \nabla_I E_k(t+\Delta t)}{2M_I}\Delta t \\ + \frac{\left[\boldsymbol{R_I}(t+2\Delta t) + \boldsymbol{R_I}(t) - 2\boldsymbol{R_I}(t+\Delta t)\right]}{\Delta t} + \mathcal{O}\left(\Delta t^4\right)\end{aligned} \quad (2.27)$$

Expressing the Verlet scheme (Equation 2.20) around $2\Delta t$, we can substitute for the position dependent term in 2.27 to give the update for the velocities within the velocity Verlet scheme:

$$\boldsymbol{V_I}(t+\Delta t) = \boldsymbol{V_I}(t) + \frac{\nabla_I E_k(t+\Delta t) + \nabla_I E_k(t)}{2M_I}\Delta t + \mathcal{O}\left(\Delta t^4\right) \quad (2.28)$$

This can be coupled with the kinematic equation of motion, to give the position update in velocity Verlet:

$$\boldsymbol{R_I}(t+\Delta t) = \boldsymbol{R_I}(t) + \boldsymbol{V_I}(t)\,\Delta t - \frac{\nabla_I E_k(t)}{2M_I}\Delta t^2 \tag{2.29}$$

Equations 2.28 and 2.29 are very similar to the Euler scheme, except for the improvement in velocity.

2.2.1.1. Sampling Configurations with MD

Before we move further, it is useful to consider the context in which MD was introduced in Section 2.1.2. The key use of MD within drug design is as a configuration generator of the therapeutic target, e.g., an enzyme-substrate complex. In the language of statistical mechanics, this amounts to an exhaustive exploration of the phase space of the system under the specified thermodynamic conditions.[3] The phase space represents all configurations of the generalized nuclear positions $\boldsymbol{R}$ and conjugated momenta $\boldsymbol{P}$ accessible under the considered conditions. Each microscopic state (microstate) can therefore be described by a vector $\boldsymbol{X} = (\boldsymbol{Q}, \boldsymbol{P})$ in the phase space. A sufficient large exploration of the phase space allows us to calculate the 'ensemble average' of an observable A of the system using the corresponding microscopic phase space function $a(\boldsymbol{X})$.

$$A = \langle a \rangle = \int \mathrm{d}\boldsymbol{X}\, a(\boldsymbol{X}) f(\boldsymbol{X}) \tag{2.30}$$

Any measurable (structural) property of the drug target must be an average over the statistical ensemble, i.e., a theoretically infinite large set of identical systems that exhibit the same macroscopic properties (e.g. volume, temperature, or pressure)and different microscopic properties. A set of independent macroscopic quantities defines a thermodynamic system. The probability density function $f(\boldsymbol{X})$ in Equation 2.30 indicates the likelihood to find a system in a given microstate. It is defined via an ensemble-specific function $\mathcal{F}(E(\boldsymbol{x}))$ of the total energy $E(\boldsymbol{X})$.

$$f(\boldsymbol{x}) = \frac{\mathcal{F}(E(\boldsymbol{X}))}{\int \mathrm{d}\boldsymbol{X}\, \mathcal{F}(E(\boldsymbol{X}))} \tag{2.31}$$

MD is a method to evolve the system through time t, not through $\boldsymbol{X}$ of the phase space. This means that we can only measure the time dependent microscopic function $a(t)$ as opposed to the phase space function $a(\boldsymbol{X})$. To then use MD to sufficiently explore the phase space, we introduce the ergodic hypothesis. If we assumes that the phase space is not divided into disjoint regions, then the system can reach all microstates at the same energy, with the same probability given a sufficiently long simulation time T. If this is true, then at thermodynamic equilibrium the ensemble

[3] The following discussion of statistical mechanism is mostly taken from Ref [100] and Ref [101].

average (Equation 2.30) can then be replaced by the time average $\overline{a}$ of the system:

$$\langle a \rangle = \overline{a} = \lim_{T \to \infty} \frac{1}{T} \int_0^t \mathrm{d}t \, a(t) \tag{2.32}$$

Dynamical systems where this is valid are termed as ergodic and have the property of ergodicity. It is difficult to prove explicitly that a system is ergodic for all but the most simplest of systems. However, so far most dynamical systems appear ergodic, and it is often assumed that the dynamical system user study is also ergodic. This assumption is called the ergodic hypothesis. It can only be valid given that T of the MD simulation is sufficiently long. In practice, the time is discretized to $T = n\Delta t$. The timestep Δt is chosen to describe the fastest movement possible, which depends on the level of theory used to calculate $E_k(\{\boldsymbol{R_I}; t\})$ (see Section 2.2.2). This usually lies within a range of 0.5 fs (for a quantum description) to 2 fs (for a classical description), implying $n > 100,000$ to reach the pico- and nanosecond scales. Solving the coupled set of Equations 2.15 and 2.9 for so many iterations is a non-trivial task, and requires efficient software (see Chapter 3). Special considerations must be given for the ability of MD to sufficiently sample the configurations of the drug target before ensemble docking.

2.2.1.2. Maintaining Temperature and Pressure

Any system that evolves according to Newton's equations of motion (Equations 2.15) will naturally function as a isolated system that conserves the total energy (E), volume (V) and number of paticles (N). In the jargon of statistical mechanics (Section 2.2.1.1) this is referred to as an NVE ensemble. But, therapeutic drug targets in physiological conditions are not isolated system and are in contact with the cell environment (e.g., solvent molecules from the cytoplasm). These interaction mean that the total pressure (P) and temperature (T) of the system in conserved, instead of E and V. Thus to mimic physiological conditions, we need to simulate biomolecules in the NPT ensemble, which requires us to modify the equation of motion.

The first step is to conserve the temperature, by introducing a thermostat algorithm. Here we introduce the popular Nose-Hoover thermostat[102, 103, 104], where the system is in contact with an external heat bath via a dynamic friction coefficient η/Q, where Q determines the inertia of the friction. The equations of motion for the extended system are:

$$\ddot{\boldsymbol{R}}_{\boldsymbol{I}}(t) = -\frac{\nabla_I E_k(\{\boldsymbol{R_I}; t\})}{M_I} - \frac{\eta}{Q} \boldsymbol{R_I}(t) \tag{2.33}$$

$$\dot{\eta} = \sum_{N} (\dot{\boldsymbol{R}_I}^2 - 3Nk_bT) \tag{2.34}$$

This allows use to constraint the system into an NVT ensemble.

To fix the pressure, we further employ a barostat. The most popular choice for this is the Parrinello-Rahman barostat[105], which allows the simulation box vectors $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\} \equiv \boldsymbol{h}$ to evolve dynamically in order to maintain a constant reference pressure tensor $\boldsymbol{P}$. With scaled coordinates $\boldsymbol{R}'_I \equiv \boldsymbol{h}^{-1}\boldsymbol{R}_I$ and the metric tensor $\boldsymbol{G} \equiv \boldsymbol{h}^T\boldsymbol{h}$, the equations of motion are:

$$M_I \boldsymbol{R}''_I = -\frac{\nabla E_k(\{\boldsymbol{R_I}; t\})}{\boldsymbol{h}} - M_I \frac{\dot{\boldsymbol{G}}}{\boldsymbol{G}} \dot{\boldsymbol{R}}'_I \tag{2.35}$$

$$\ddot{\boldsymbol{h}} = \frac{\boldsymbol{P}(t) - \boldsymbol{P}}{\boldsymbol{W} \cdot \boldsymbol{h^T}} \det \boldsymbol{h} \tag{2.36}$$

Here the mass parameter $\boldsymbol{W}$ determines the coupling strength. [106] Applying both the thermostat and barostat algorithms simultaneously allows us to sample configurations of the biomolecule under the NPT ensemble, mimicking physiological conditions. Notice that $\boldsymbol{R_I}$ is present on both sides in Equations 2.33 and 2.35. This makes numerical integration with algorithms like Velocity Verlet more complex.[99] Solutions often include an iterative approach, or utilising the predictor-corrector scheme.

2.2.2. Time-Independent Electronic Equation

The most computationally intensive part of integrating Equation 2.15 is the energy term $E_k(\{\boldsymbol{R_I}; t\})$ of the electrons in the k^{th} state. This would ideally involve solving for Equation 2.9 at every Δt timestep. This is computationally infeasible, especially given that the number of iterations should be large to sufficiently explore the phase space of the drug target (Section 2.2.1.1). Hence we introduce approximations to 2.9 to make the computation more tractable.

2.2.2.1. Molecular Mechanics

Instead of explicitly considering the electronic wavefunction in Equation 2.9, we can completely disregard the dependence of $E_k(\{\boldsymbol{R_I}; t\})$ on the quantum state k of the electrons. This allows us to completely decouple the nuclear motion from the calculation of the quantum energy surface. This energy surface is then approximated

as an analytical function depending only on the nuclear positions. For systems with large number of atoms, this still leaves us with the 'dimensionality bottleneck' as the nuclear degrees of freedom increases.[97] We further simplify the potential to a truncated expansion of many-body contribution to the energy surface. This approximation leads us into the field of molecular mechanics (MM), where classical approximations is used to model molecular systems. Here atoms are treated as point charges, and electronic/quantum effects are neglected. The truncated, analytical approximation of $E_k(\{\boldsymbol{R_I}; t\})$ are referred to as force fields or $E^{FF}(\{\boldsymbol{R_I}; t\}$ (where the dependence on the electronic quantum state k has been dropped). A general form is given in Equation 2.37.

$$E^{FF} = \underbrace{U_{\text{bond}} + U_{\text{angle}} + U_{\text{imp.}} + U_{\text{dihedral}}}_{bonded} + \underbrace{U_{\text{es}} + U_{\text{vdw}}}_{non\text{-}bonded} \tag{2.37}$$

The nuclear equations of motion can then be propagated as:

$$M_I \ddot{\boldsymbol{R}}_{\boldsymbol{I}}(t) = - \bigtriangledown_I E^{FF}(\{\boldsymbol{R_I}; t\}) \tag{2.38}$$

This reduces the mixed quantum electronic and classical nuclear problem of BO-MD to a fully classical problem. MD under this scheme is referred to as classical MD (cMD), and is routinely used in modelling large biomolecular system for drug design. E^{FF} is represented as a sum of bonded as non-bonded interactions. The bonded interactions between two atoms are approximated as springs (Figure 2.3), and hence are modelled with the harmonic potential as:

$$U_{\text{bond}} = \sum_{\text{bond}\, ij} k^b_{ij} \left(b_{ij} - b^0_{ij}\right)^2 \tag{2.39}$$

$$U_{\text{angle}} = \sum_{\text{angle}\, ijk} k^\theta_{ijk} \left(\theta_{ijk} - \theta^0_{ijk}\right)^2 \tag{2.40}$$

These are a functions of bond lengths b_{ij} , and angles θ_{ijk}. The equilibrium values of spring systems are b^o_{ij}, and θ^o_{ijk} with corresponding force constants k^x_{ij}, and k^θ_{ijk}. A term for the proper dihedrals ϕ_{ijkl} with force constants k^ϕ_{ijkl} is also incorporated as a periodic function with a maxima at δ and periodicity pf n:

$$U_{\text{dihedral}} = \sum_{\text{dihedral}\, ijkl} k^\phi_{ijkl} \left(1 + \cos\left(n\phi_{ijkl} - \delta\right)\right) \tag{2.41}$$

Improper dihedrals are also often included to force atoms to remain in a plane or to prevent transition to a configuration of opposite chirality:

$$U_{\text{improper}} = \sum_{\text{improper}\, ijk} k^{\xi}_{ijkl} \left(\xi_{ijkl} - \xi^{0}_{ijkl}\right)^2 \tag{2.42}$$

This modelled a spring similar to Equations 2.39 and 2.40 as a functions of improper dihedrals ξ_{ijkl}. The equilibrium value is ξ^{o}_{ijkl}, and the corresponding force constant is k^{ξ}_{ijkl}.

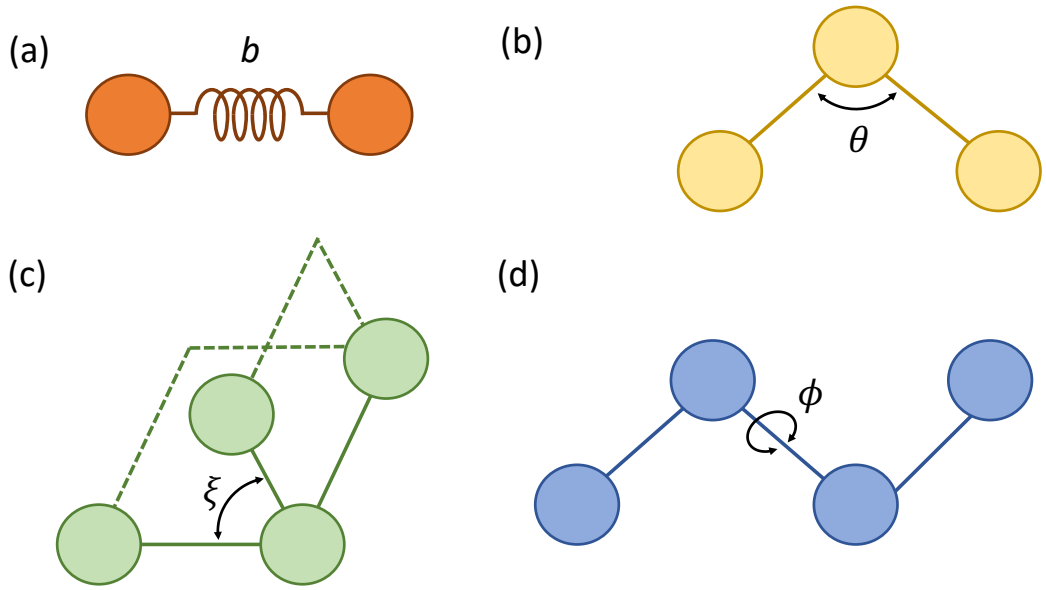


Figure 2.3. Definition of the variables used in the MM force field for the (a) bonded terms, (b) angular terms, (c) improper and (d) proper dihedrals.

2.2.2.2. Quantum Mechanics

The non-bonded terms are summed over all non-bonded pairs ij, and is calculated as a sum of the Coulomb electrostatic interactions described between point charges q_i and q_j (Equation 2.43) and Lennard-Jones Van der Waals interactions well depth ϵ_{ij} at equilibrium distance r^{0}_{ij} (Equation 2.44).

$$U_{\text{es}} = \sum_{\text{n.b. pair}\ ij} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \tag{2.43}$$

$$U_{\text{vdw}} = \sum_{\text{n.b. pair}\ ij} \epsilon_{ij} \left[\left(\frac{r^{0}_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r^{0}_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.44}$$

The exact form and parameters used for MM force fields vary with the exact application and implementation. The most commonly used force fields for biological systems are CHARMM[107] and AMBER[108]; a version of the latter is used in this work. The main advantage of MM force fields is that they are extremely fast to compute, especially in comparison to QM methods (see Section 2.2.2.2). This allows for exploration of drug targets in the microsecond timescale, resulting in a wealth of sampling data for accurate ensemble docking as discussed in Section 2.1.2. However, not explicitly treating electronic degrees of freedom means that cMD cannot reproduce quantum effects. These include polarization, metal binding, bond breakage and formation. These effects are important in biological drug targets, especially as we wish to sample closer to an enzyme transition state (see Section 2.1.2). Although polarizable and reactive force fields have been developed to include these effects in classical force fields, these have not yielded the desired level of quantum accuracy yet. Only by solving Equation 2.9 explicitly would we be able to currently reproduce quantum phenomenon in biomolecular simulations.

If a complete description of the quantum effects in the system is desired, Equation 2.9 must be tackled directly. This contains the many-electron wavefunction $\Psi_k(\boldsymbol{r_i})$, which makes it impossible to analytically solve for more than one electron. Consequently, the most common approach to this problem is to split the many-electron wavefunction to multiple single-electron functions. This is exemplified by the Hartree-Fock method, where the many-electron wavefunction is approximated as a Slater determinant of N single-electron states:

$$\Phi_{i_1 \ldots i_N}(r_1\sigma_1 \ldots r_n\sigma_N) = \frac{1}{\sqrt{N}} \begin{vmatrix} \phi_{i_1}(r_1\sigma_1) & \ldots & \phi_{i_N}(r_1\sigma_1) \\ \vdots & & \vdots \\ \phi_{i_1}(r_N\sigma_N) & \ldots & \phi_{i_N}(r_N\sigma_N) \end{vmatrix} \tag{2.45}$$

Here σ denotes the electronic spin variable. With Equation 2.45, the set of Hartree-Fock equations can be derived:

$$\frac{-\hbar^2 \bigtriangledown^2}{2m}\phi_i(\boldsymbol{r}\sigma) + \int d(\boldsymbol{r}'\sigma' v_{eff}^{HF}(\boldsymbol{r}\sigma, \boldsymbol{r}'\sigma')))\phi_i(\boldsymbol{r}'\sigma') = \epsilon_i \phi_i(\boldsymbol{r}\sigma) \tag{2.46}$$

$$v_{eff}^{HF}(\boldsymbol{r}\sigma, \boldsymbol{r}'\sigma') = \delta_{\sigma\sigma'}\delta(r - r')\left[v_{ext}\boldsymbol{r} + v_H\boldsymbol{r}\right] + v_x^{HF}(\boldsymbol{r}\sigma, \boldsymbol{r}'\sigma') \tag{2.47}$$

$$v_H(\boldsymbol{r}) = \int d\boldsymbol{r}' \frac{e^2}{|\boldsymbol{r} - \boldsymbol{r}'|} \sum_{\sigma'=\uparrow,\downarrow} \sum_{j=1}^{N} |\phi_j(\boldsymbol{r}'\sigma')|^2 \tag{2.48}$$

$$v_x^{HF}(\boldsymbol{r}\sigma, \boldsymbol{r}'\sigma') = -\frac{e^2}{|\boldsymbol{r} - \boldsymbol{r}'|} \sum_{j=1}^{N} \phi_j(\boldsymbol{r}\sigma)\phi_j^*(\boldsymbol{r}'\sigma') \tag{2.49}$$

Here equation 2.48 is the Hartree potential and Equation 2.49 is the exchange potential. The applicability of Hartree-Fock method to biochemical systems is rather limited due to intrinsic limitation of the method, the most critical of which is the lack of the electronic correlation. For this reason, a plethora of post-HF approaches have been developed to mitigate this issue. These methods include: Coupled Cluster (CC), Configuration Interaction (CI), perturbation theory approaches such as Møller-Plesset (MP2, MP3, ...), etc. However, in most cases the computational complexity of these formulations prevents their usage for the large systems of interest in drug design (even within the QM/MM scheme as discussed in Section 2.2.2.3). The method that best balances computation cost with accuracy, especially in context of biology, is density functional theory (DFT).[14] DFT implicitly takes into account electronic correlation effects, and is built around the Hohenberg-Kohn theorems:

1. Given an external potential v_{ex}, the ground state energy of a non-degenerate n electron system is a unique functional of the electron density $\rho(\boldsymbol{r})$, denoted as $E[\rho]$.

2. $E[\rho]$ is minimal for the exact ground state density [109].

This allowed Kohn and Sham to propose the following approach to approximating the functional of the density. They introduced a fictious system of n non-interacting electrons with one-particle orbitals $\phi_i(\boldsymbol{r})$:

$$\mathcal{H}_e^{KS}\phi_i(\boldsymbol{r}) = \epsilon_i\phi_i(\boldsymbol{r}) \tag{2.50}$$

In this system, the Hamiltonian can be written exactly as:

$$\mathcal{H}_e^{KS} = -\frac{1}{2}\nabla^2 + v^{\text{KS}}(\boldsymbol{r}) \tag{2.51}$$

$$v^{\text{KS}}(\boldsymbol{r}) = v_{\text{ex}}(\boldsymbol{r}) + \int \mathrm{d}\boldsymbol{r}' \, \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} + \frac{\delta E_{\text{xc}}[\rho]}{\delta\rho(\boldsymbol{r})} \tag{2.52}$$

Here v_{xc} is the external potential, the second term is the Hartree term describing the electron–electron Coulomb repulsion, and the last term depends on the exchange–correlation potential E_{xc}. This is simply the sum of the error made in using a non-interacting kinetic energy and in treating the electron-electron interaction classically. The correct choice of E_{xc}, allows the Kohn-Sham approach to achieve an exact

correspondence between the density and ground state energies of a system consisting of non-interacting electron and the 'real' interacting, many-electron system. The exchange-correlation potential, despite its name, also contains an element of the kinetic energy and is not just the sum of the exchange and correlation energies as they are understood in Hartree-Fock.

Equations 2.50 and 2.52 have to be solved self-consistently: Starting with an initial guess for ρ, one computes v^{KS}. Then, one can solve 2.50 to obtain a new set of ϕ_i. These, in turn, are used to calculate the new electron density:

$$\rho(\boldsymbol{r}) = \sum_{i=1}^{n} |\phi_i(\boldsymbol{r})|^2 \tag{2.53}$$

These steps are repeated until the electron density has converged [110]. With the converged (ground state) electron density and orbitals, one can integrate the equation of motion and advance the system in time. Overall, DFT drastically reduces the complexity of the problem, since one has to solve only n one-particle equations (Equation 2.50) instead of the original n-particle Schrödinger equation (Equation 2.9) [97].

2.2.2.2.1. Exchange-Correlation Functional

In order to utilize DFT for interacting, multi-electron system, the exchange-correlation function in Equation 2.52 should be known. An exact representation for this is not known, so several approximations have been developed. Local Density Approximation (LDA) is the most simplest, where the spatial exchange-correlation energy density $\varepsilon_{\mathrm{xc}}$ of the original system is replaced by that of an homogeneous electron gas with constant electron density ρ_0.The exchange-correlation energy thus becomes: [111]

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[\rho] = \int \mathrm{d}\boldsymbol{r}\, \varepsilon_{\mathrm{xc}}^{\mathrm{HEG}}(\rho_0 = \rho(\boldsymbol{r})) \tag{2.54}$$

LDA is limited to various applications in solid state systems.[112, 113] For (bio)chemical systems, the accuracy of DFT can be improved by taking into account not only the local electron density $\rho(\boldsymbol{r})$, but also its gradient $\nabla\rho(\boldsymbol{r})$, leading to the class of Generalized Gradient Approximation (GGA) functionals.[97, 113] This includes the BLYP functional, utilized thought this work.[114, 115, 116] An additional step towards higher accuracy are hybrid functionals, which include a fraction of exact Hartree-Fock (HF) exchange [117]:

$$E_{\mathrm{x}}^{\mathrm{HF}}[\rho] = -\frac{1}{2}\sum_{ij}\int \mathrm{d}\boldsymbol{r}\int \mathrm{d}\boldsymbol{r}'\, \phi_i^*(\boldsymbol{r})\phi_j^*(\boldsymbol{r}')|\boldsymbol{r}-\boldsymbol{r}'|^{-1}\phi_j(\boldsymbol{r})\phi_i(\boldsymbol{r}') \tag{2.55}$$

B3LYP [118, 119] is one of the most popular hybrid functional:

$$E_{\mathrm{xc}}^{\mathrm{B3LYP}}[\rho] = (1-a)E_{\mathrm{x}}^{\mathrm{LDA}} + aE_{\mathrm{x}}^{\mathrm{HF}} + b\Delta E_{\mathrm{x}}^{\mathrm{B88}} + (1-c)E_{\mathrm{c}}^{\mathrm{LDA}} + cE_{\mathrm{c}}^{\mathrm{LYP}} \tag{2.56}$$

It combines exact exchange, local contributions $E_{\mathrm{x}}^{\mathrm{LDA}}$ [120] and $E_{\mathrm{c}}^{\mathrm{LDA}}$ [121], the Becke gradient correction $\Delta E_{\mathrm{x}}^{\mathrm{B88}}$ [114], and the correlation functional of Lee, Yang, and Parr $E_{\mathrm{c}}^{\mathrm{LYP}}$ [115] The parameters were fitted to experimental data, yielding $a = 0.20$, $b = 0.72$, $c = 0.81$. [122] B3LYP is one the best functionals for studying biochemical systems related to drug design. However, they are often prohibitively expensive for codes that utilize the plane-wave basis set (employed in CPMD/MiMiC as described in Section 2.2.2.2.2), and so most simulations in this work use the reasonable accurate BLYP functional.

2.2.2.2.2. Plane Wave Basis Set

To numerically evaluate the continuous integrals over $\rho(\boldsymbol{r})$ in Equations 2.52, the Kohn-Sham orbitals in Equation 2.53 must be discretized into a finite basis set. An often used form (most common in solid-state physics) is the plane wave basis set.[4] For a periodic system with Bravais vectors $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\} \equiv \boldsymbol{h}$ and volume $\Omega \equiv \det \boldsymbol{h}$, the Kohn-Sham potential (Equation 2.52) is periodic as well. This allows us to write the Kohn-Sham orbitals (Equation 2.50) in Bloch form as:

$$\phi_i(\boldsymbol{r}, \boldsymbol{k}) = \exp(i\,\boldsymbol{k}\cdot\boldsymbol{r})\; u_i(\boldsymbol{k}, \boldsymbol{r}) \tag{2.57}$$

Here $\boldsymbol{k}$ denotes the orbital momentum within the first Brillouin zone and the functions u_i exhibit the same periodicity as the potential. This can be expanded with coefficients c_i as:

$$u_i(\boldsymbol{r}, \boldsymbol{k}) = \frac{1}{\sqrt{\Omega}} \sum_{\boldsymbol{G}} c_i(\boldsymbol{G}, \boldsymbol{k}) \exp(i\,\boldsymbol{G}\cdot\boldsymbol{r}) \tag{2.58}$$

Substituting for u_i, Equation 2.57 can be rewritten into Equation 2.59 as:

$$\phi_i(\boldsymbol{r}, \boldsymbol{k}) = \frac{1}{\sqrt{\Omega}} \sum_{\boldsymbol{G}} c_i(\boldsymbol{G}, \boldsymbol{k}) \exp(i(\boldsymbol{k}+\boldsymbol{G})\cdot\boldsymbol{r}) \tag{2.59}$$

This represents the complete expansion of the wavefunction in the plane-wave basis set. The reciprocal space vector $\boldsymbol{G}$ is defined as $2\pi\left(\boldsymbol{h}^T\right)^{-1}\boldsymbol{g}$ with $\boldsymbol{g} = (\alpha, \beta, \gamma)$ and

[4] This discussion is adapted from Ref [97].

$\alpha, \beta, \gamma \in \mathbb{N}$. The summation over $\boldsymbol{G}$ needs to be truncated at some point, usually when a cutoff energy E_{cut} is exceeded as:

$$|\boldsymbol{k} + \boldsymbol{G}|^2 < E_{\text{cut}} \tag{2.60}$$

For biological systems, only the center of the first Brillouin zone, i.e., the Γ-point with $\boldsymbol{k} = \boldsymbol{0}$, is considered.

Using the plane-wave basis set allows for the design of DFT algorithms with greater parallel performance and efficiency of (Section). However, a complete plane-wave approach is not suitable for non-crystalline, (bio)organic systems. Wavefunctions of 'core' electronic orbitals close to the nucleus have several nodes, necessitating the use of large number of plane-waves to describe. This will make the calculation very expensive. However, it is well known that chemical reactivity and bonding is mainly controlled by the 'valence' electronic orbitals, i.e. the orbitals farther away from the nuclei, and not the core orbitals. This implies that for our purposes, an explicit treatment of core orbitals is unnecessary. Therefore, to reduce the computational cost of a calculation using plane wave basis set, the core obitals can be described by so-called pseudopotentials that are specifically constructed to reproduce the effective potential created by core electrons. This then, not only reduces the number of explicit orbitals needed, but also allows for a smaller plane wave energy cutoff E_{cut} as the remaining valence orbitals are smoother than the ones closer to the core. Thus, in the QM simulations in this work, the core electrons are described using norm-conserving pseudopotentials [123] and only the valence electrons were treated explicitly with DFT.

2.2.2.3. Quantum Mechanics/Molecular Mechanics

Solvated biomolecular system routinely consists of >100,000 atoms, a full QM treatment (as discussed in Section 2.2.2.2) of which is not possible. On the other hand, as mentioned in Section 2.2.2.1, a complete MM approach would neglect important effects like chemical reactivity, polarization and metal coordination. To solve this problem, the following attributes of quantum effects within biomolecules are to be noted:[12]

1. Electronic effects are highly localized phenomenon within the 'active site' and efficiently screened by surrounding biomolecular atoms.

2. These active sites are often embedded in a complex hydrogen-bond network with steric contributions from the rest of the biomolecule. A complete neglect of the environment can result in spurious QM behavior. Although a continuum

representation and/or arbitrary position constraints will not suffice, MM force fields may provide enough accuracy to describe these environment effects.

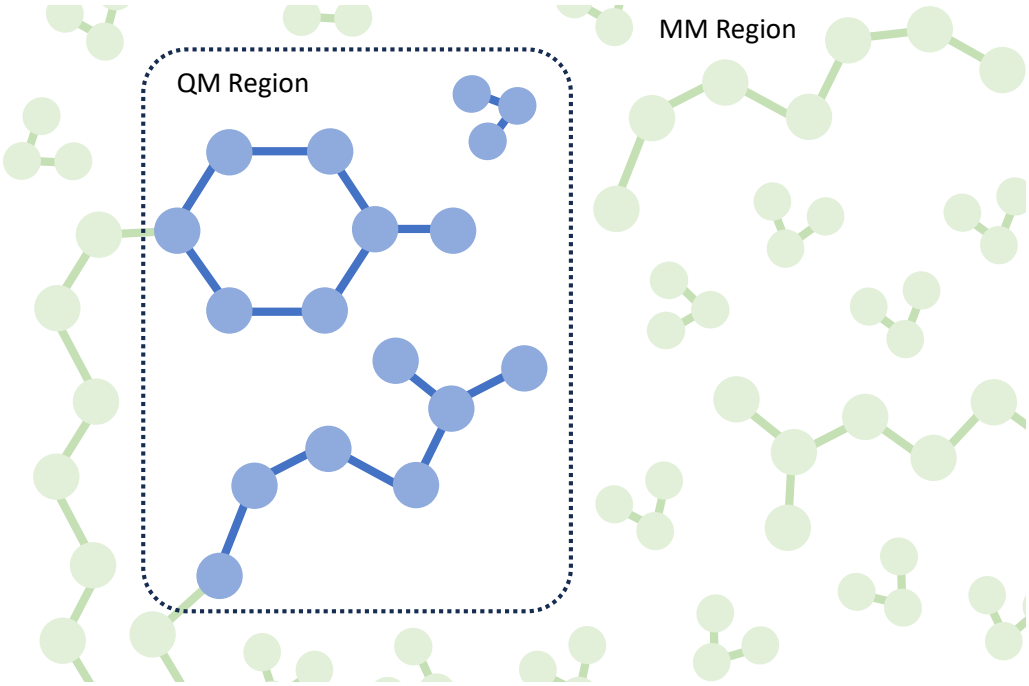


Figure 2.4. Illustration of the QM/MM scheme, where atoms are divided into the QM region (blue) and MM region (green).

The combination of the above two points leads us to consider a multiscale or hybrid approach, in which only a portion of the system is treated at the QM level while the rest of the system is represented with the computationally simpler MM force field, leading to a mixed Quantum Mechanical/Molecular Mechanical (QM/MM) partitioning of the system (Figure 1.2).[124] Within this strategy a system is partitioned into two parts: an active site (where the quantum effects in the biomolecule are localized) that is described at the QM level and the rest of the system, which is treated at the MM level.[13] This naturally leads to a third subsystem, the coupling across the QM-MM region. This QM/MM coupling must be treated accurately, leading to the two broad methods of QM/MM: additive and subtractive schemes.[12, 11] Sections 2.2.2.3.1 and 2.2.2.3.2 discussing these techniques are based on Ref [125].

2.2.2.3.1. Additive Scheme

In this scheme, we take the very straightforward approach of splitting the total energy into separate QM, MM and mixed QM-MM interactions as in Equations 2.61.

$$E = E_{\mathrm{QM}}(\boldsymbol{R_{QM}}) + E_{\mathrm{MM}}(\boldsymbol{R_{MM}}) + E_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) \tag{2.61}$$

Despite its apparent simplicity, Equation 2.61 can be very complicated to implemented (especially compared to the subtractive scheme described in Section 2.2.2.3.2) owing to the mixed term $E_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}})$. This can be separated further into bonded and non-bonded interactions (Equations 2.62).

$$\begin{aligned} E_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) &= V^{b}_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) + V^{nb}_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) \\ &= V^{b}_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) + V^{vdw}_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) \\ &+ V^{es}_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) \end{aligned} \tag{2.62}$$

Non-bonded var der Waals interactions $V^{b}_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}})$ are treated using the same function as in the MM force field. This often amounts to using the Lennard-Jones potential.

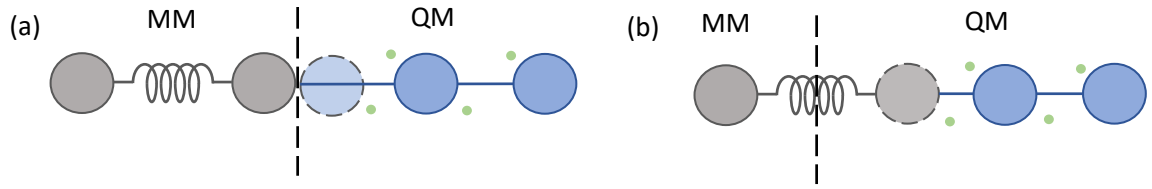


Figure 2.5. (a) Monovalent atom where the monovalent capping atom is present only in QM calculations, and (b) boundary pseudopotential where the MM atom is replaced with a QM pseudopotential that participates in the MD.

Bonded terms arise when there is the QM/MM boundary cuts across the QM-MM boundary. In biological systems, the QM is often the active site of a protein consisting of the side chains of certain chemically relevant residues. This means the the QM-MM boundary will naturally cut across covalent bonds, and lead to dangling bonds and open valences of the QM atoms on the boundary. Therefore, the energy will involves 'capping term' that saturate the boundary QM atoms. There are many ways of doing this, of which the two most common are (depicted in Figure 2.5):

1. **Monovalent atom**: Here a monovalent atom, typically hydrogen, is put in the QM region (added to $E_{\mathrm{QM}}(\boldsymbol{R_{QM}})$ in Equation 2.61) on the line between the QM and MM atom. This atom is not included in MD and force computations, and the position is constrained on the QM-MM bond.

2. **Boundary pseudopotential**: The MM atom at the boundary is replaced with a specifically designed capping pseudopotential (added to $E_{\mathrm{QM}}(\boldsymbol{R_{QM}})$ in Equation 2.61). Here the capping atom is included in the MD (MM terms for the capping atom is added to $V^{b}_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}})$ in Equation 2.62), providing a more consistent description of the interface. However, these potentials can be rather tricky to build. In this work, this approach is used with monovalent pseudopotentials[126] that mimics the C-C single bond.

The electrostatic interactions between QM electrostatic potential and MM partial charges can be incorporated in different ways:

1. **Mechanical embedding**: In the simplest embedding technique, the wavefunction of the electronic subsystem is computed for the isolated QM part. MM charges of QM atoms may either be fixed or may be recomputed each time, when the wavefunction changes using some of the charge fitting procedure (i.e. Hirshfield or RESP charges [104, 105]). Lennard-Jones parameters are kept fixed throughout the simulation. The mechanical embedding is simple to implement. However, it suffers from the obvious shortcoming that the polarization of the electronic charge in the QM region due to MM atoms are completely disregarded. Often these effects are necessary to reproduce many phenomena seen in biomolecules.

2. **Electrostatic embedding**: Polarization of the QM charge distribution due to MM charges can be achieved by adding a specific QM/MM electrostatic potential to the external potential of the time-independent electronic equation. The QM electrons are made to feel MM atoms as a special type of nuclei with non-integer charge. Many choices for the electrostatic potential exists, with this term often being expensive to compute. The exact form and implementation used in the MiMiC framework, employed in this work, is described in Chapter 3. This method accounts for polarization and QM/MM interactions for a fully periodic systems, greatly improving its usability in biology.

3. **Polarizable embedding**: In the final type of embedding, the polarization of the MM atoms due to QM region is also considered. This is significantly more complicated than electrostatic embedding and is not employed in this work. Discussion of the details of this are beyond the scope of this work.

2.2.2.3.2. Subtractive Scheme

Computation of $E_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}})$ in Equation 2.61 can become quite complex. A different approach, is to perform the QM calculation of the QM region as an isolated system. The influence of the environment is then estimated as the difference of energies between the entire system (QM + MM) and the QM subsystem, both evaluated at the MM level. This is the subtractive approach, and can be written as:

$$E = E_{\mathrm{QM}}(\boldsymbol{R_{QM}}) + E_{\mathrm{MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) - E_{\mathrm{MM}}(\boldsymbol{R_{QM}}) \tag{2.63}$$

Notice that we avoid introducing the mixed term $E_{\mathrm{QM/MM}}(\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}})$. This makes the implementation of this method very simple. One of the most popular examples of this coupling scheme is ONIOM.[127] However, such a coupling scheme

is prone to a number of problems, including the lack of interactions (polarization, unsaturated covalent bonds, etc.) between QM and MM subsystems. Furthermore, a major disadvantage of this method is that it requires accurate parameters for all atoms, including the complete QM region.[128] This is a big disadvantage in drug design applications where exotic targets, where only a quantum description can suffice (like those described in Section 1.1.1), are being studied. Although dummy parameters could be used for the QM region, the additive QM/MM scheme is far superior in this regard. Only the additive scheme is discussed in this work.

3. Methods: MiMiC for Multiscale Modeling in Chemistry

DFT-based QM/MM MD have been previously used to study various biophysical phenomena like enzymatic reactions [124, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138], metalloproteins [139, 48, 140, 141], proton transfers [22, 21, 24, 142] and photophysical processes [143, 144, 145, 146]. However, using this within a drug design pipeline has proved difficult.[147] This is mainly because large QM regions (~100 atoms) are required with timescales of ~100 ps for adequate conformational sampling. These are computationally very expensive, and require software that can perform these computations efficiently. A few candidate codes (like the TeraChem protocol buffers[148] and the QUICK-Amber interface[149]) have been introduced that run efficiently on a handful of GPUs. Another promising direction to increase QM/MM performance is to efficiently scale across an extremely large number of CPU nodes in a supercomputer. The MiMiC framework is one of these candidates, showing excellent QM/MM performance by efficient strong scaling.[18, 19] In this work, we seek to show how MiMiC can significantly benefit drug design efforts. The software heavily profits from high performance computing (HPC) concepts to achieve efficient performance. A summary of the methodology behind MiMiC, and the HPC concepts used to achieve efficient performance are detailed in this chapter.

3.1. Introduction

MiMiC implements DFT-QM/MM MD within the additive scheme with electrostatic embedding (see Section 2.2.2.3.1), which implies that the energy computation involves performing (i) a DFT computation of the QM region, (ii) an MM computation of the MM region, and (iii) computation of the mixed QM-MM interactions across the boundary.[18, 19, 20] Each of these computations are implemented as different algorithms, with their own implementation and parallelization strategies. Integrating them into a monolithic piece of software would be tedious, and could easily result in an overall inefficient implementation. Furthermore, standalone QM and MM codes have already been developed and highly optimized for many decades. With this in mind, MiMiC has been designed to perform QM/MM MD in a loose-coupling, multiple data multiple program (MPMD) fashion. In practice, separate QM and MM

external programs are used for the computation in the respective subsystems, while MiMiC provides libraries for the computation of the mixed QM/MM interactions, for which information from both the external programs are needed. Specifically, MiMiC is constituted by:

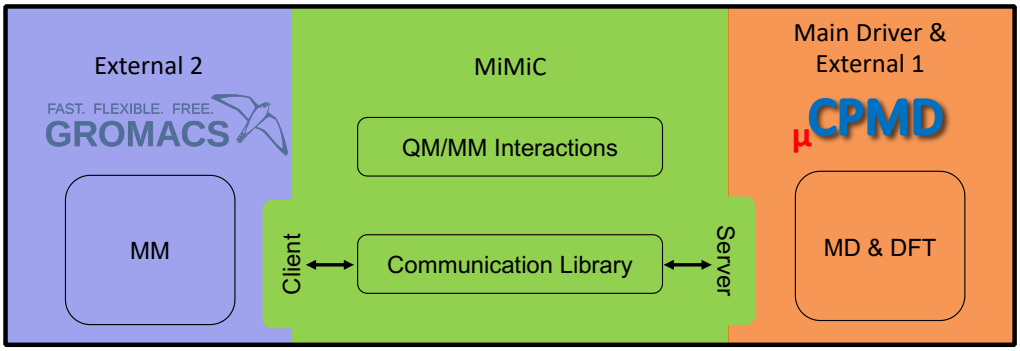


Figure 3.1. Illustration of the strategy used in MiMiC to run QM/MM MD.

1. The MiMiC Communication Library (CommLib): An MPI-based communication library used to exchange information between the external programs[150]

2. The main MiMiC Library: A collection of subroutines that implement the computation of the electrostatic QM/MM interactions [151]

MiMiC was designed to be program agnostic, allowing one (in principle) to couple any number of external programs to treat different part of the system at different levels of theory. The current implementation links the CPMD QM code[16] with the GROMACS MM code[152] (Figure 3.1).[19] Here, CPMD functions as the server, where all information is collated and the equations of motion are integrated to propagate the dynamics. The GROMACS acts as a client, receiving and providing information on the MM subsystem. The workflow is shown in Figure 3.2. All communication is performed through the CommLib. This seamlessly allows shuttling of information from different resource areas of a supercomputer through high-speed interconnects, such as InfiniBand and Omni-Path Architecture, introducing only a very minimal overhead. This allows CPMD and GROMACS to run concurrently, using separately allocated resources where they can apply their own parallelization strategies, obtaining the highest possible efficiency for the overall QM/MM simulations.[18] The main MiMiC Library currently share resources with CPMD while it implements its own parallelization strategies. This implementation of the QM/MM MD algorithm is what potentially allows for the simulation of large biological targets for drug design.[23, 21]

The key to the success of the above mentioned approach is the specialized parallelization techniques used by each program. Running complex MD code on a single processor can easily require thousands of years to complete and just increasing the

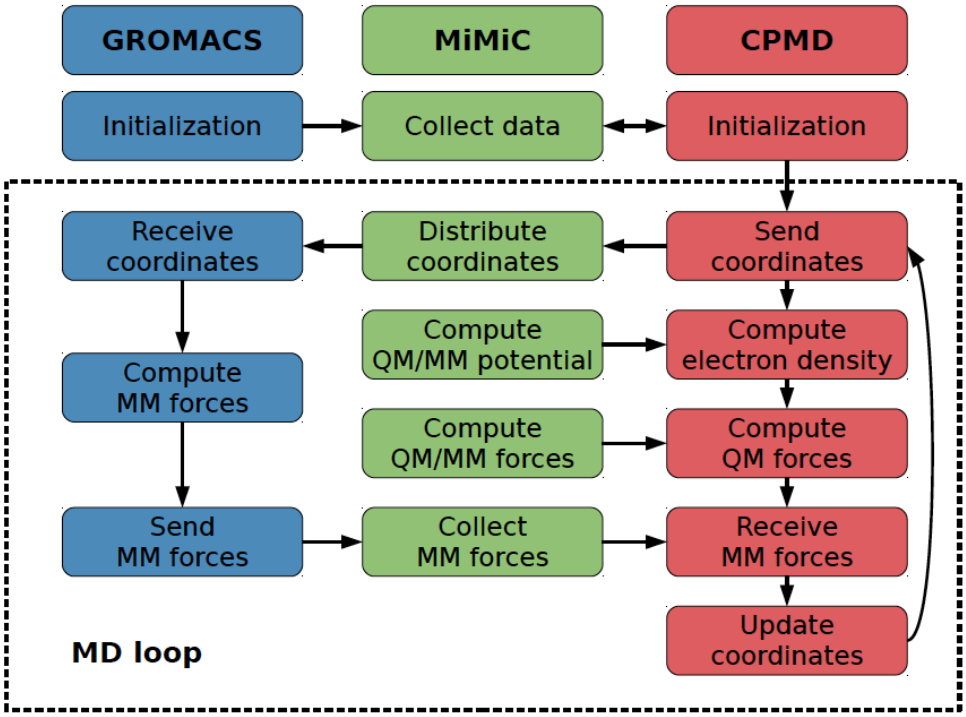


Figure 3.2. Illustration of the MiMiC-based QM/MM-MD workflow using BO-MD. Adapted from Ref. [18].

speed of single-core hardware cannot physically deliver the required performance levels to obtain reasonable time to solution. This can be overcome by increasing the amount of cores, and moving from a serial to parallel execution of computation.[153] Computer hardware are nowadays explicitly designed for this purpose, and scientific software should be optimized for such hardware.[154] These tasks are classed into the field of HPC.[155] Here, software is designed to split each computation into multiple sub-components and parallely run them across the various cores available. This provides the best possible performance, provided the adopted algorithms are well designed for this purpose. A quick overview of the basic terms used in HPC is given in Section 3.2. These are then used to explain how CPMD, GROMACS and MiMiC utilize them for efficient QM/MM simulations in Section 3.3. Both the sections are based on ref [156].

3.2. Basics of HPC

The basic computational unit in a HPC system is a **compute node**. (see scheme in Figure 3.3). Nowadays, a node typically consists of multiple **central processing units (CPUs)**, where each CPU itself consists of a few dozen **cores**. Each core

usually supports **hardware threading**, which enables simultaneous execution of independent sequence of instructions at the level of a single CPU. Nodes can also be fitted with **graphical processing units (GPUs)**, which can function as accelerators for certain numerically-intensive operations. The increase of computational power of one node is limited by the CPU area and by overheating. This presents a physical barrier to the size of scientific problems we can treat on a single node. To mitigate this, a second level of parallelization can be achieved by joining (or networking) various compute nodes together through a fast network. This is really what lends the immense power associated with HPC systems, where (currently) up to several thousands of compute nodes are connected to work on the same scientific problem. As opposed to improvement of single-node performance, the number of connected compute nodes can, in principle, increase arbitrarily. This results in theoretically no limit on the size of scientific problems that can be treated with HPC. In practice, building larger and more powerful supercomputing systems is an engineering challenge that needs to take into account many factors to optimize performance while reducing costs and avoiding waste of resources. This makes HPC an active area of research, prioritized by many governmental organizations like USA and the EU as next-generation technologies of national interest.[15]

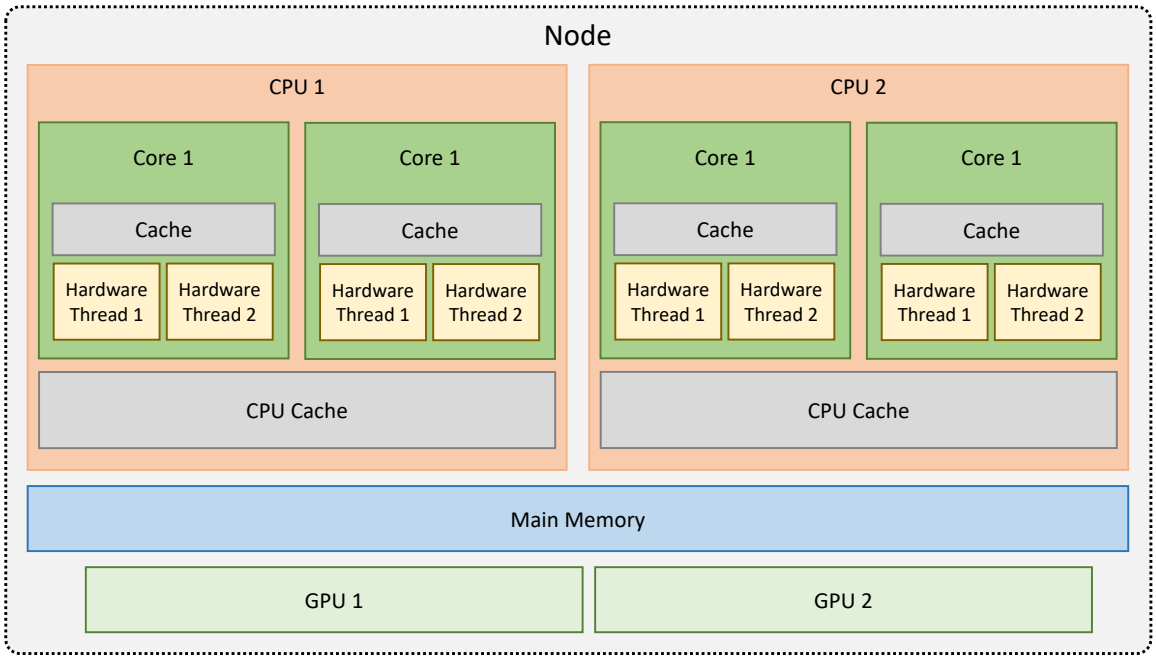


Figure 3.3. Schematic of a multicore node with Graphics Processing Units (GPUs).

To employ multiple networked computational nodes to work on the same scientific problem, computations are split into multiple **tasks** or **processes**. These are in practice multiple instances of the same program operating on independent sets of data. Each task may be then executed using one thread or more. Typically, several tasks are launched on a certain number of nodes of the HPC system, and may share computer

resources. The software developer (and, to a certain extent, the user) decides the way compute nodes are split into different tasks. This is the so called distributed memory parallelization strategy, in which each task has no possibility of directly accessing data belonging to other tasks. Communication, when and if needed, occurs via a message passing protocol, such as Message Passing Interface (MPI).[157] Running software on multiple nodes potentially allows for indefinite scaling. **Scalability** refers to the parallelization efficiency and is measured as the ratio between the actual speedup of a computation and the ideal speedup obtained when using a certain number of processors. In this thesis, speedup is calculated as the ratio of CPU time to execute a program on one core $t(1)$ to that for N cores $t(N)$ (Equation 3.1).

$$Speedup = \frac{t(1)}{t(N)} \tag{3.1}$$

A specific way to quantify the scalability of an application is by using the concept of **strong scaling**. This is the speedup of the algorithm measured with a fixed problem size and increasing number of computational units. Strong scaling indicates how well the algorithm is parallelized. The ideal case is represented by a linear behavior of the speedup with respect to the number of processors. In reality, only very few algorithms can scale linearly indefinitely. The point of saturation is commonly called the scaling limit. Usually, the speedup of a parallel algorithm saturates at a certain number of processes. The strong scaling efficiency is given by Equation 3.2.

$$Efficiency = \frac{t(1)}{t(N)N} \times 100 \tag{3.2}$$

The saturation in speedup is imposed by either inefficient communication patterns or by the point where the serial part of the code takes a comparable amount of time as the parallel one. The communication overhead associated with each message passing operation will add some "idle" (i.e., not computing) time to the process. Therefore, the overall parallel performance is compromised. Typically, this can be alleviated by "overlapping" computation and communication via asynchronous communications. To design an efficient QM/MM MD software, one of important goals is to ensure that it scales strongly for the maximum number of processor units. This requires splitting the computation into roughly equal computational loads, where the communication and computation can be neatly overlapped.

3.2.1. Domain Decomposition

The most used method of parallelizing the energy computation (the methods discussed in Section 2.2.2) is using the concept of domain decomposing (DD). Here the whole

domain over which the computation is carried out (this could be the collection of atom or DFT plane waves) is split into smaller "chunks" or sub-domains which are distributed across computational units. This enables both to reduce the memory footprint of the software, and to reduce the amount of data communicated each time. The type of the DD is then defined by the shape of sub-domains (Figure 3.4):

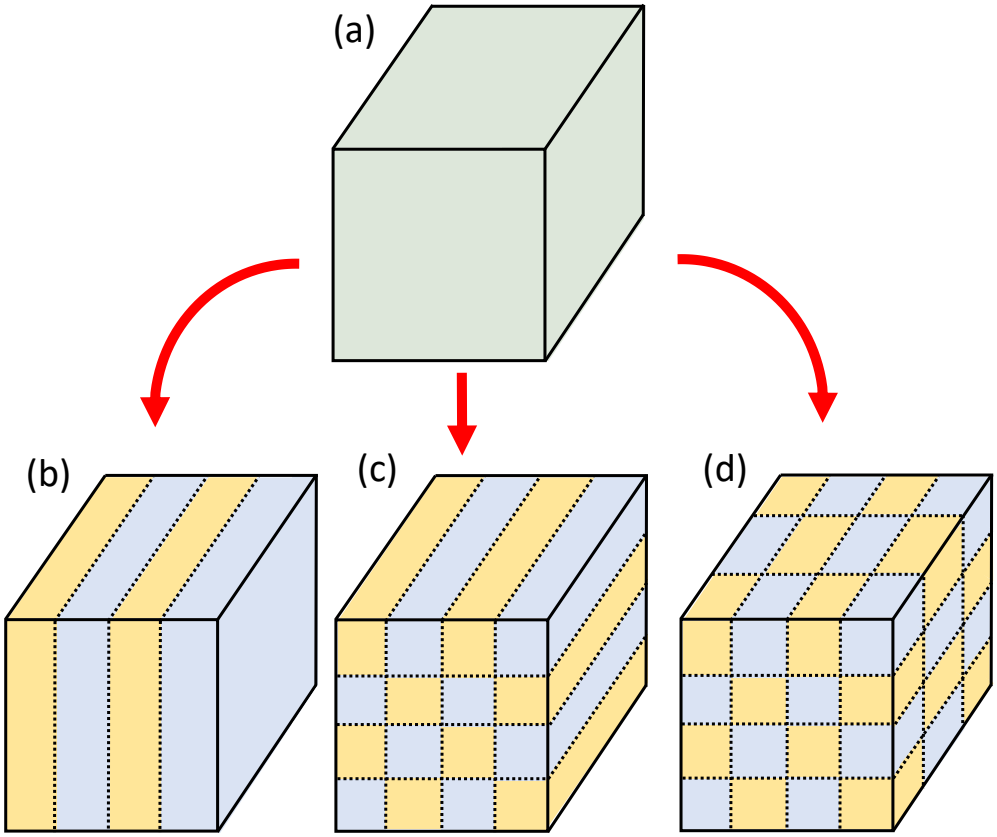


Figure 3.4. Domain decomposition involves splitting (a) a full 3D domain into (b) slabs, (b) pencils, or (c) volumetric DDs.

Slabs. Here the 3D domain is split into 2D sub-domains along a chosen dimension. This is the most straightforward DD scheme. However, it imposes a strict scaling limit, which is equal to the number of planes available in the grid along the decomposition dimension.

Pencils. Here the 3D domain is split into a set of 1D "pencils". The latter are then distributed across computational units.

Volumetric. Here the domain is split into a set of 3D "bricks". Potentially this DD type yields the highest scalability.

The crucial point to keep in mind is that the computational load of each of the sub-domains has to be uniform. When it is, this is called a balanced computation.

Whenever there is an imbalance in computational load between computational units this typically leads to a drop in the scaling efficiency. This problem is especially prominent in the distributed and accelerated codes as the added communication overhead further degrades the strong scaling of the algorithm. Load balancing can be achieved either via static or dynamic algorithms.

3.3. Parallelizing QM/MM MD with MiMiC

As mentioned in 3.1, a MiMiC QM/MM run involves evaluating forces in GROMACS, CPMD and MiMiC itself, in parallel. The forces are calculated as a gradient of the energy calculated by Equation 2.61. We can recast the general additive QM/MM Equation 2.61 to the specific case of CPMD/GROMACS/MiMiC QM/MM as Equation 3.3.

$$\begin{aligned} E[\rho_{QM}](\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) = E_{\text{CPMD}}[\rho_{QM}](\boldsymbol{R_{QM}}) + E_{\text{GROMACS}}(\boldsymbol{R_{MM}}) \\ + E_{\text{MiMiC}}[\rho_{QM}](\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) \end{aligned} \tag{3.3}$$

Each of these involves parallelisms tailored to the level of theory being dealt with. The parallelism strategies used by each code is described in Sections 3.3.1—3.3.3.

3.3.1. Parallelism in CPMD

Evaluation of $E_{\text{CPMD}}[\rho_{QM}](\boldsymbol{R_{QM}})$ within the DFT formalism involves solving Equation 2.50 and 2.52 self-consistently. This is the most computationally heavy component of a QM/MM calculation.[158] In CPMD, the wavefunctions are expanded on a plane-wave basis set according to Equation 2.59. By arranging the plane waves on a 3D grid in reciprocal space, domain decomposition can be utilized for efficient parallelism (Figure 3.5).[159] The Kohn-Sham equations are calculated with a mixed real space/reciprocal space formalism in which the analytical expression of Equation 2.52 turn out to be rather simple.[160] The transition between the two spaces is achieved via 3D Fast Fourier Transforms (FFTs).[161] 3D FFT in CPMD is implemented in such a manner so as to minimize the communication between the different parallel processes. This is achieved by a mixture of slab and pencil decomposition, where slabs are distributed across processes and pencils across the threads associated to each process (for shared-memory optimization).[159] The scalability of the 3D FFT depends on the longest dimension of the real space grid points of the QM box.[162] In practice, CPMD parallelizes slabs along the arbitrarily chosen X axis, and it is the job of the user to rotate the system in such a way that the X-axis is the longest QM box dimension. If there are M_x real-space grid points along the X-axis, then the load

on each process P is:

$$Load = \frac{M_x}{P} \tag{3.4}$$

To achieve good efficiency, the load on each process should be equal. Thus for proper load balancing in CPMD, the total number of processors used must an integral divisor of M_x. Furthermore, the scalability is limited by M_x, as going beyond $P = M_x$ processors would mean that some processes will have no work to do. With the plane wave basis set, M_x depends only on E_{cut} from Equation 2.60 and the QM box size, not on the number of QM atoms (unlike with a localized basis set). Nevertheless, the number of planes available is typically about 50 for small systems and 200 to 300 for large systems.[162] This restricts the maximum number of processors that can be efficiently used. A second level of parallelism was introduced in CPMD to solve this problem, via the task-grouping approach.[159, 163] The processes are grouped into G groups or 'CP groups', with $p_G = P/G$ processes associated with each group (Figure 3.5).[163] The real-space grid points data M_x are replicated across all groups. However, each process in a groups holds only M_x/p_G points instead of M_x/P points. The load on each process then becomes:

$$Load = \frac{M_x}{p_G} = \frac{M_x \times G}{P} \tag{3.5}$$

All-to-all communication is limited to within each group.[162] Although this scheme introduces roughly double the amount of communication as compared to the original scheme, it improves the FFT load balancing and reduces the latency in the needed all-to-all communication. Increasing the load also allows us to apply more processors (and nodes) to the same problem, resulting in stronger scaling, and reduced time to solution.

We have now effectively increased the scaling limit from M_x to $(M_x \times G)$. Therefore, when using task-grouping in CPMD, in order to achieve load balance, the total number of processors used must be an integral divisor of $(M_x \times G)$. Up to the limit where communication becomes the bottleneck, G is a parameter that can be freely chosen, affording incredible flexibility to extend the strong scaling behavior of the simulation. In realistic HPC systems, we usually specify the number of nodes N and the number of processors per node p_N, giving us the total number of processors as $P = N \times p_N$. We can then rewrite Equation 3.5 to give us:

$$Load = \frac{M_x \times G}{N \times p_N} \tag{3.6}$$

The task grouping approach is what really allows CPMD to scale to an incredibly large number of cores on supercomputers. This makes expensive hybrid functionals, like B3LYP, feasible for large systems. CPMD has been shown to scale up to several million cores for large systems on massively parallel platforms such as BlueGene/Q.[163] This makes CPMD an ideal candidate for a highly scalable QM/MM application, and the primary reason it is utilized in MiMiC. Equation 3.6 is the main tool used to optimize the scaling of MiMiC-QM/MM simulations in Chapter 6.

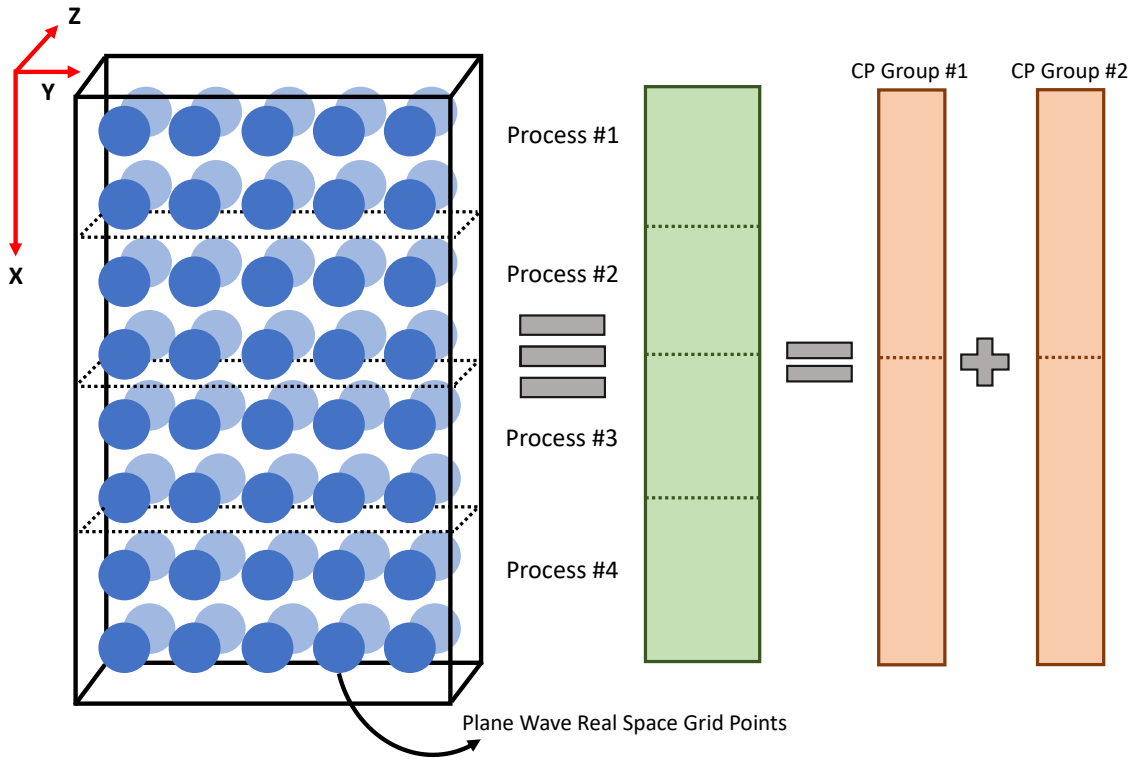


Figure 3.5. Domain decomposition of the QM region in CPMD. Here the planes waves are distributed into 4 processes, and further groups into 2 CP groups using the task group approach.

3.3.2. Parallelism in GROMACS

Evaluation of $E_{\text{GROMACS}}[\rho_{MM}](\boldsymbol{R_{MM}})$ in Equation 3.3 is the least expensive subroutine of the overall QM/MM simulation. It is sufficient to assign just 1 node to GROMACS in a MiMiC-QM/MM simulation. Hence, only a brief summary of the parallelization techniques used in GROMACS are discussed here. The most computationally part of Equation 2.37 is the non-bonded interactions. These are split into: (i) Lennard-Jones potential (Equation 2.44) and short range electrostatics (Equation 2.43) and (ii) long-range electrostatics evaluated using Particle Mesh Ewald

(PME)[164] summation.

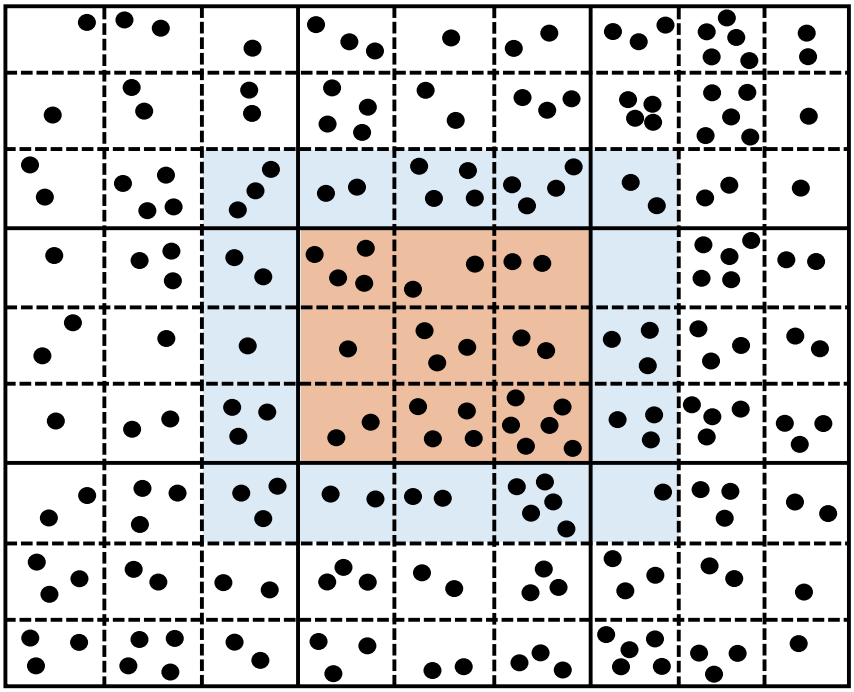


Figure 3.6. Illustration of a linked-cell approach to DD for a 2D domain. Domains are separated by thick lines, while cells with thin lines. Atoms are represented with black circles. Orange represent the reference cells for force computation. Additional cells needed for computations are in blue. GROMACS implements this algorithm in a 3D domain.

For the short-range interaction, the MM atoms are decomposed into volumetric cells using the linked-list approach.[17, 165, 166] Here the volumetric domains are distributed across processes. The domains consist of multiple orthorhombic cells, with sizes slightly larger than the non-bonded interaction cutoff (Figure 3.6). Evaluating the non-bonded interaction among atoms within each domain is straightforward, and made even faster with multithreading. Inter-domain interactions, on the other hand, require communication across processes. The parallelism is designed to minimize this communication overhead. Finally, fixing the cell sizes to a constant value can lead to atoms accumulating in a few cells. This can lead to load imbalances as the simulation progresses. GROMACS implements dynamic load balancing by allowing changes in the cell sizes during the simulation.[167]

The long-range PME summation is evaluated in reciprocal space with 3D FFT. The pencil decomposition scheme is used with multithreading to increase the scaling and performance of this subprogram.

3.3.3. Parallelism in MiMiC

Although the DFT subroutine as performed by CPMD is the main determinant to the overall QM/MM performance, the mixed QM/MM interaction as calculated by MiMiC should also be optimized to introduce minimal additional latency. Evaluation of $E_{\text{MiMiC}}[\rho_{QM}](\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}})$ in Equation 3.3 requires MiMiC to evaluate Equation 2.62. This is done using the electrostatic embedding scheme, where the computationally-expensive electrostatic embedding between the QM and MM subsystems can be expressed generally as:

$$V^{es}_{\text{QM/MM}}[\rho_{QM}](\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) = \sum_{i=1}^{N_{MM}} \frac{Q_i}{4\pi\epsilon_0} \left(\int dr \rho(\boldsymbol{r}) \frac{1}{|\boldsymbol{r} - \boldsymbol{R_i}|} + \sum_{j=1}^{N_{QM}} \frac{Z_i^{QM}}{|\boldsymbol{R_j} - \boldsymbol{R_i}|} \right) \quad (3.7)$$

Equation 3.7 involves the Coulomb interaction of the MM atoms (consisting of MM partial charges Q_i) with both the electron density $\rho(\boldsymbol{r})$ and the nuclear charges of the QM atoms Z_i^{QM}. This implementation leads to the following problems:

a) At short range, the use of standard MM point charges Q_i to represent the MM atoms in Equation 3.7 may lead to over-polarization near the QM-MM boundary originating by the lack of Pauli repulsion between the QM and MM subsystems. This means that the point charges on the MM side may attract/repel the electrons too strongly, causing spurious results. Such artifacts can become serious if large flexible basis sets like plane waves are used in the QM calculations. This problem is known as electron spill-out.

b) At long range, the computational cost to evaluate the integral in Equation 3.7 could become prohibitively expensive. To compute the integral, the electronic density of the QM subsystem is mapped onto a 3D grid (in a similar way to how the planes waves are distributed in CPMD as discussed Section 3.3.1), with the number of points typically in the order 10^6. If not handled properly, this could add a significant cost and degrade the parallel efficiency of the overall QM/MM simulation.

To overcome these two problem, MiMiC reformulates Equation 3.7 in the generalized version of the electrostatic coupling scheme by Laio et al.[168]. The MM atoms are grouped into short-range and long-range domains depending on their distance from the QM subsystem (Figure 6.2.3.1(a)). This allows us to rewrite Equation 3.7 as Equation 3.3.3.

$$\begin{aligned} V^{es}_{\text{QM/MM}}[\rho_{QM}](\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) = V^{es,sr}_{\text{QM/MM}}[\rho_{QM}](\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) \\ + V^{es,lr}_{\text{QM/MM}}[\rho_{QM}](\boldsymbol{R_{MM}}) \end{aligned} \quad (3.8)$$

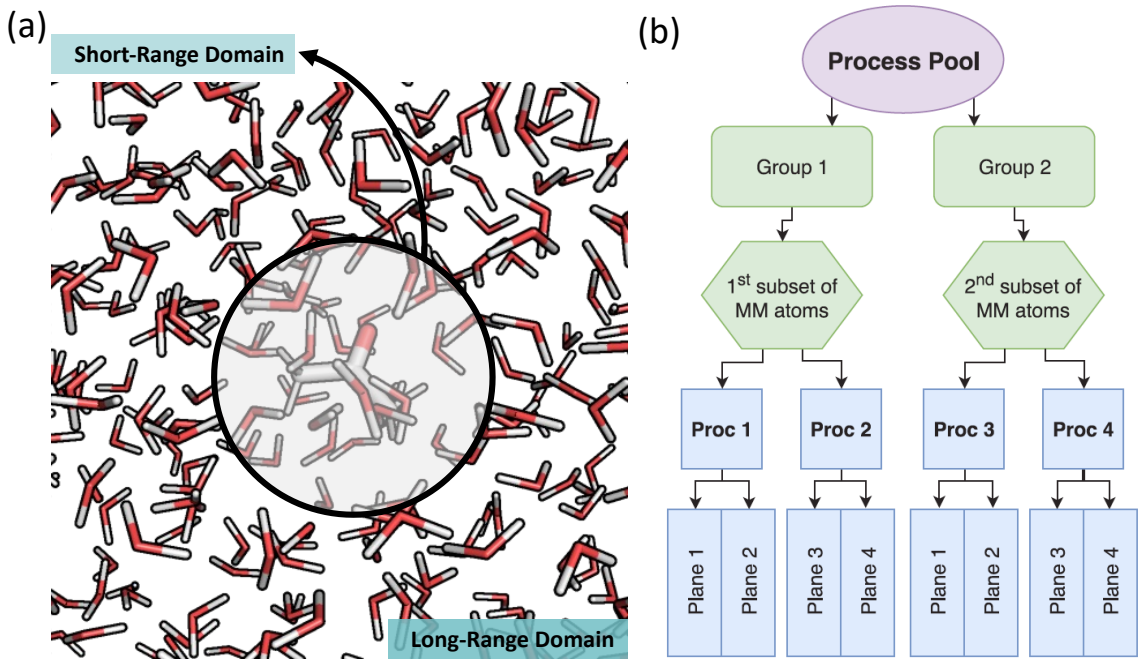


Figure 3.7. (a) Splitting of the system into short and long-range domains in MiMiC. Here acetone is the QM region, surrounded by MM water. (b) Parallelization scheme of mixed QM/MM computations in MiMiC using the process pool allocated to CPMD. Adapted from Ref. [19].

The short-range MM atoms interact directly with the QM electron density and nuclei. $V^{es,sr}_{\mathrm{QM/MM}}[\rho_{QM}](\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}})$ is calculated exactly, except that electron spill-out is avoided by 'smeared' out the partial charges of the MM atoms Q_i with a smearing function $\nu_i^{smear}(r)$:

$$V^{es}_{\mathrm{QM/MM}}[\rho_{QM}](\boldsymbol{R_{QM}}, \boldsymbol{R_{MM}}) = \sum_{i=1}^{N_{MM}} \frac{Q_i}{4\pi\epsilon_0} \int d\boldsymbol{r}\rho(\boldsymbol{r}) \frac{\nu_i^{smear}(|\boldsymbol{r} - \boldsymbol{R_i}|)}{|\boldsymbol{r} - \boldsymbol{R_i}|} \tag{3.9}$$

In CPMD and MiMiC, $\nu_i^{smear}(r)$ of atom i is expanded around the covalent radius $r_{c,i}$ as:

$$\nu_i^{smear}(r) = \frac{r_{c,i}^4 - \boldsymbol{r^4}}{r_{c,i}^5 - \boldsymbol{r^5}} \tag{3.10}$$

In contrast to the point charge model, the Coulomb interactions between the MM electrons and the smeared charge distributions does not diverge if the electrons approach the MM atoms.

The remaining long-range MM atoms, interact with the QM region through a mul-

tipole expansion of the electrostatic potential created by the QM electrons and nuclei. $V^{es,lr}_{\mathrm{QM/MM}}[\rho_{QM}](\boldsymbol{R_{MM}})$ is calculated using:

$$
\begin{aligned}
V^{es,lr}_{\mathrm{QM/MM}}[\rho_{QM}](\boldsymbol{R_{MM}}) &= \sum_{i=1}^{N^{lr}_{MM}} \frac{Q_i}{4\pi\epsilon_0}\left[\frac{C_{QM}}{|\boldsymbol{R_i}-\bar{r}|}+\frac{D_\alpha}{\left|\boldsymbol{R_i}-\bar{R}_{QM}\right|^3}+\frac{E_{\alpha\beta}}{\left|\boldsymbol{R_i}-\bar{R}_{QM}\right|^5}+\cdots\right] \\
D_\alpha &= \sum_\alpha \mu_\alpha\left(\boldsymbol{R_i}^\alpha-\bar{R}^\alpha_{QM}\right) \\
E_{\alpha\beta} &= \sum_{\alpha\beta} O_{\alpha\beta}\left(\boldsymbol{R_i}^\alpha-\bar{R}^\alpha_{QM}\right)\left(\boldsymbol{R_i}^\beta-\bar{R}^\beta_{QM}\right)
\end{aligned}
\tag{3.11}
$$

The N^{lr}_{MM} long-range MM atoms interact with the total charge C_{QM}, dipole μ_α, quadrupole $O_{\alpha\beta}$, etc., of the QM charge distribution. $\bar{R}_{QM}$ is the centroid of the QM region. MiMiC implements Equation 3.9 up to an arbitrary number of multipoles (definable by the user) in the full generalized formulation as discussed in ref [19]. Equations – have been validated to provide the same accuracy at a reduced cost as compared to Equation 3.7.[18, 168] Nevertheless, the exact evaluation of the MM interactions with the QM density in the short-range region remains, resulting in N^{sr}_{MM} integrals over the dense electronic density grid. This forms the most computationally intensive part of the mixed QM/MM interaction in MiMiC. The parallelization strategy used to speedup the computation consists of distributing subsets of the (real-space) grid planes among processes, very similar to how the plane waves are distributed in CPMD.[156, 19] The processes are divided among task groups that receive a certain amount of MM atoms each. Then, real-space grid points of the electronic density are decomposed into slabs and distributed among the process within each group. Finally, a pencil decomposition is performed within each slab, where the threads associated with each process receives a subset of pencils. This is depicted in Figure 6.2.3.1(b). The computation of the relatively cheaper electronic part of the multipole expansion in Equation 3.3.3 is done exclusively by slab decomposition of the real-space grid across processes. After calculating the electrostatic interaction, forces on the QM nuclei due to the MM atoms, and vice versa, for both the short and long-range domains can be calculated, and the corresponding equations of motion integrated.

4. Biology: Isocitrate Dehydrogenase 1 and Glioma

To demonstrate that MiMiC is suitable to solve problems 1–4 of Section 1.1.2, a specific biomolecule as an ideal test candidate must be chosen. The test candidate should have the following conditions:

- **Therapeutically relevant**: Since we wish to establish the utility of MiMiC for drug design, we choose a therapeutically relevant biomolecule. This should preferably be an enzyme, as they are easier to handle.
- **Sufficiently complex**: The system would require a quantum description for full treatment. This would not only entail studying the catalytic mechanics of the enzyme, but also preferably involve a system with metals (metalloproteins).
- **Previous literature**: I attempt to confirm if MiMiC can correctly reproduce the behavior of the enzyme both at the group state and the transition state. Sufficient previous literature on the enzyme (both experimental and computational) on the catalysis is required for comparison.

An enzyme that satisfies all of these conditions is the human Isocitrate Dehydrogenase 1 or IDH1 enzyme.

4.1. Biological Role of IDH1

IDH1 is an enzyme that catalyzes the oxidative decarboxylation of isocitrate (ICT) to α-ketoglutarate (α-KG).[169, 170] The enzyme requires a cofactor, as an electron acceptor, and a Mg(II) ion in the catalytically active form. Eukaryotic cells express two distinct classes of IDH1s, one that utilize NAD^+ as the electron-acceptor cofactor, and another that uses $NADP^+$.[25] NAD–dependent IDH1 is localized in the mitochondria. It plays and important role in the Krebs or TCA cycle for cellular energy production, catalysing the following reaction:

$$\mathrm{ICT} + \mathrm{NAD}^{+} \xrightarrow[\mathrm{IDH1}]{\mathrm{Mg}^{2+}} \alpha\mathrm{KG} + \mathrm{NADH} \tag{4.1}$$

NADP–dependent IDH1 are located either in the mitochondria or the cytoplasm, catalyzing the following reaction:

$$\mathrm{ICT} + \mathrm{NADP}^{+} \xrightarrow[\text{wt-IDH1}]{\mathrm{Mg}^{2+}} \alpha\mathrm{KG} + \mathrm{NADPH} \tag{4.2}$$

As shown in Equation (4.2), the primary role of NADP–dependent IDH1 is as a source of cytoplasmic NADPH in human cells. This contributes towards lipid synthesis and cellular defense against damage caused by reactive oxygen species (ROS). Mice overexpressing IDH1 in the liver and adipose tissues experienced increase in triglyceride and cholesterol content, resulting in obesity and hyperlipidemia.[171] Conversely, interference of IDH1 activity resulted in loss of body weight and the reduction of triglyceride level in diet-induced obese mice.[172] IDH1 has also been shown to regulate phospholipid metabolism in developing astrocytes.[173] The role of IDH1 (through the production NADPH) as protectors against ROS and other stresses has been confirmed in mice studies. It has been shown that IDH1 deficiency in mouse embryonic fibroblasts leads to increased oxidative DNA damage, and decreased survival after oxidative stress, while overexpression prevents these effects. [174, 175] Apart from the primary role in the TCA cycle, αKG is also a major cofactor in many dioxygenases enzymes. This includes regulating prolyl-hydroxylase activity and stabilization of the hypoxia-inducible factor-1α (HIF-1α).[176] The regulation of multiple demethylases by αKG (and by consequence IDH1), like ten-eleven translocation (TET) DNA hydroxylases and the Jumonji histone demethylases, has a direct impact on cell stemness, and differentiation.[177]

Mutation at the Arg132 position to histidine (among other minor mutations) in the active site of the NADP–dependent IDH1 results in the loss of ability to catalyse Equation 4.2, but at the same time imparts a new ability to convert αKG to 2-hydroxyglutarate (2-HG) [178]:

$$\alpha\mathrm{KG} + \mathrm{NADPH} \xrightarrow[\text{mut-IDH1}]{\mathrm{Mg}^{2+}} 2\text{–HG} + \mathrm{NADP}^{+} \tag{4.3}$$

From here on, the NADP–dependent IDH1 is referred to as the wild type IDH1 (wt-IDH1) while the mutant form is indicated as mut-IDH1. In addition, Equation 4.2 will be referred to as the 'normal reaction', whereas Equation is the 'neomorphic reaction'. A complete summary of the complex interplay of various aspects of IDH1 functioning in human cells is provided in Figure 4.1.

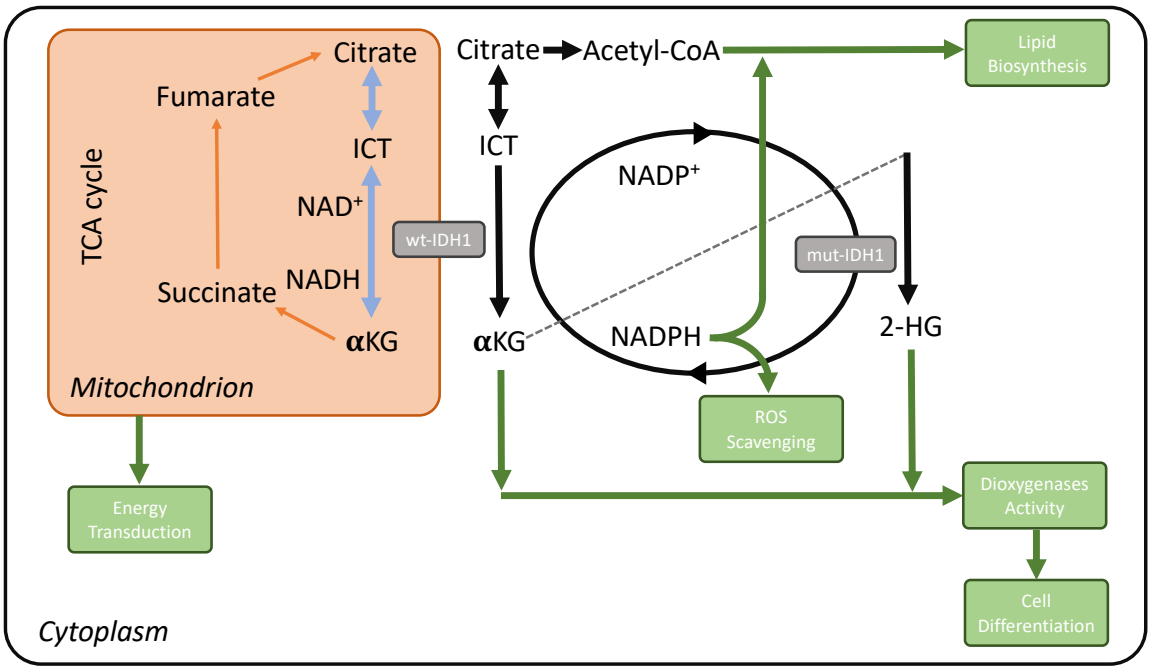


Figure 4.1. Various cellular processes involving the various isoforms of IDH1, both in mitochondria and cytoplasm.

4.1.1. Mut-IDH1 as a Predictive Biomarker for Glioma

Apart from the biological significance of IDH1, it has also become increasingly clear that it plays a crucial role in the progression of glioma. Glioma is an extremely deadly form of cancer, where the tumor starts in the glial cells of the brain of spine.[179] Multiple types of glioma exist in which IDH1 is involved: (i) Astrocytoma, which affects glial cells called astrocytes. (ii) Glioblastoma, which is the most aggressive form of astrocytoma. (iii) Oligodendroglioma, which affects glial cells called oligodendrocytes. wt-IDH1 is know to account for almost 95% of all glioblastomas.[180] Specifically, it plays a major role in WHO grade 4 glioblastoma.[181] Although the mechanism of this action is not completely clear, the promotion of cell stemness by αKG is a candidate for tumor progression.[180] mut-IDH1, on the other hand, is involved in the progress of WHO grade 2–4 astrocytoma (where grade 4 astrocytoma was previously classified under glioblastoma) and grade 2–3 oligodendroglioma. It is even implicated in acute myeloid leukaemia (AML).[182, 181, 183] 2-HG, produced by mut-IDH1, is a known oncometabolite that promotes stemness in human cells and inhibits DNA demethylases.[184, 185] At a molecular level, this is due to the structural similarity of 2-HG to αKG, whereby it can inhibit αKG-dependent dioxygenases. From this discussion, it should be evident that both wt- and mut-IDH1 enzymes are incredibly important for glioblastoma therapeutics and are interesting actionable therapeutic targets.

In fact, multiple mut-IDH1 inhibitors have already been developed.[186] However, these cannot be directly used for wt-IDH1 glioma. In fact, mut-IDH1 glioma patients generally have better survival rates compared to patients with the wt-IDH1 counterparts.[27] While mut-IDH1 display one unique and specific neomorphic activity (Equation 4.1), wt-IDH enzymes catalyze several metabolic reactions involved in different cellular processes depending on various biological conditions (Section 4.1 discussed only some of these processes).[180] One of these is the normal reaction (Equation 4.2), which is not an ideal target as it is displayed by both normal and tumoral cells. There is thus a clinical need to be able to differentiate between glioma involving the mutant and wild type isoforms of IDH1. In this regards, the ability to non-invasively image mut-IDH1 expression in gliomas can serve as a valuable tool in the clinical setting to predict prognosis and evaluate treatment response based on IDH1 mutation status.[30]

Positron emission tomography (PET) is a widely-applied method in oncology, and neuro-oncology in particular, for such diagnosis.[187, 188] PET is a molecular imaging technique which uses specific probes that are labeled with positron-emitting radioisotopes to visualize and measure changes in biological processes *in vivo*.[29] These probes are referred to as radiotracers, and are often labelled with radioactive fluorine isotopes ^{18}F. The intricacies involved in designing a PET radiotracer for mut-IDH1 are discussed in Section 4.2.

4.2. Designing PET Radiotracers Targeting Mut-IDH1

The key to utilizing the PET technique as a diagnostic tool for mut-IDH1 glioma is the successful design of the radiotracer. This is a scientifically challenging problem, and requires that various aspects of the mut-IDH1 be taken into account. A precursor molecule, which will be subsequently labelled with ^{18}F, should be developed with the following properties:

1. Since mut-IDH1 is an intracellular target, this excludes precursors that are macromolecules and unable to cross the cell membrane.[29] Small molecules obeying Lipinski's rules are prefered.

2. To image mut-IDH1 in glial cells of the brain, the precursor must cross the blood-brain barrier (BBB)

3. The precursor must either (i) already posses an F group, or (ii) an F group can be safely substitued without disrupting its biological activity.

4. It must selectively bind only to mut-IDH1, and not wt-IDH1. If not, there would be no diagnose of the specific isoform.

All current mut-IDH1 inhibitors satisfy points 1-2,[186, 189, 190, 191, 192, 193] and further more, at least one of these inhibitors (AGI-5198[194]) contains a fluorine that can be easily substituted with ^{18}F (satisfying point 3). However, it was shown recently that all current inhibitors fail point 4 (discussed in Section 4.2.1). Thus, they not suitable as radiotracers, and this renders the PET method of mut-IDH1 diagnosis infeasible.[32]

4.2.1. Binding of Mut-IDH1 Inhibitors

Current inhibitors of mut-IDH1 show great inhibition selectivity for the mutant isoform, over the wild type, as measured by the IC_{50} values.[186]. This would naively imply that these inhibitors are also binding selective for only mut-IDH1. However, the results of Liu et. al. delinked the direct relationship between the binding affinity and inhibitory potency of the mut-IDH inhibitors to wt-IDH.[31] In fact, it is found that the inhibitors bind to both isoform, but do not, or only weakly, inhibit wt-IDH1. Such a lack of a direct relationship reflects the special allosteric nature of the inhibitor binding to IDH1, with this allosteric binding site close to Mg^{2+} of the active site.[195]

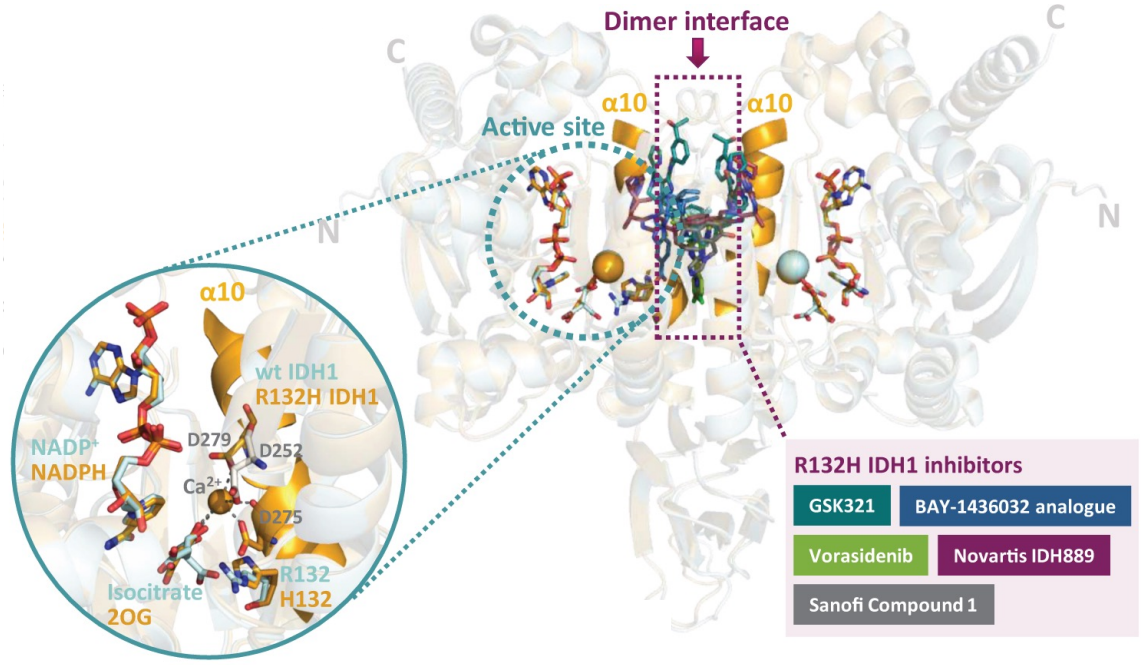


Figure 4.2. Crystal structure views of mut-IDH1–inhibitor complexes reveal the allosteric binding of the inhibitors at the dimer interface. Adapted from Ref. [31].

Mut-IDH1 inhibitors bind not competitively to the substrate (αKG) in the active site, but at the dimer interface (Figure 4.2).[31] This region is structurally similar in

both mut- and wt-IDH1, explaining the lack of binding selectivity of the inhibitors. Inhibition selectivity, on the other hand, can be explained by noting that the binding at the dimer interface disrupts binding of Mg^{2+} to the substrate.[31] Due to the weaker αKG–Mg^{2+}, as opposed to the stronger ICT–Mg^{2+} binding, the mut-IDH1 is much more susceptible to allosteric inhibition as opposed to the wild type.[195]

To develop a PET radiotracer for mut-IDH1 (i.e., to specifically satisfy point 4 from Section 4.2), active-site inhibitors, as opposed to allosteric inhibitors, are required. The primary structural difference in the mut- and wt- isoforms is the mutation at the 132 position of the active site. This allows binding of αKG in mut-IDH1, as opposed to ICT in wt-IDH1. Only by developing radiotracers that act as an analog for the αKG substrate in the active site, can binding selectivity be induced.

Drugging the deep pocket of the mut-IDH1 active site would require a negatively charged, hydrophobic molecule that would not necessarily cross the BBB. This is the most likely reason why high throughput screening studies, that discovered the previously mentioned mut-IDH1 inhibitors, produced molecules that bind outside the hydrophobic active site. Given this difficulty so far in producing substrate analogs for mut-IDH1, we can label the active site of mut-IDH1 as an 'undruggable' target. These are targets that are difficult to target with current pharmacological methods, whether it be due to the complexity of the active site or other reasons.[196, 197] The case of the mut-IDH1 active site is similar to protein tyrosine phosphatases, which is considered to be undruggable due to selectivity concerns among closely related family members.[198] Based on discussions in Sections 2.1.2 and 1.1.1, MD simulations of mut-IDH1 combined with docking at the active site, could be a possible route to drug this target. Multiple snapshots of both isoforms of IDH1, from MD, are required for a specific design of molecules that dock to the complex active site. These simulations are carried out in Chapters 6 and 7.

4.3. Active Site of Wt-IDH1

The catalytically active wt-IDH1 in complex with ICT, $NADP^+$ and CA(II), has been crystallized using X-ray crystallography of the protein (Figure 4.3a).[25] It occurs as a homodimer, with each of the two monomers consisting of: (i) a large domain, (ii) a small domain and (iii) a clasp domain. The active site (where ICT, $NADP^+$ and CA(II) bind) is formed by the large and small domains of one subunit and a small domain of the other subunit, held together in the dimer by the clasp domain.[25] Since the two identical active sites include residues from both monomers, in this text, residues from the second subunit are labelled by the superscript B and those from the first subunit are left unmarked. ICT is anchored in each of the active sites through the following interactions: (i) the α-carboxylate group of ICT forms a direct H-bond with Arg100 and Arg109, (ii) the β-carboxylate group interacts with $Lys212^B$, Arg132

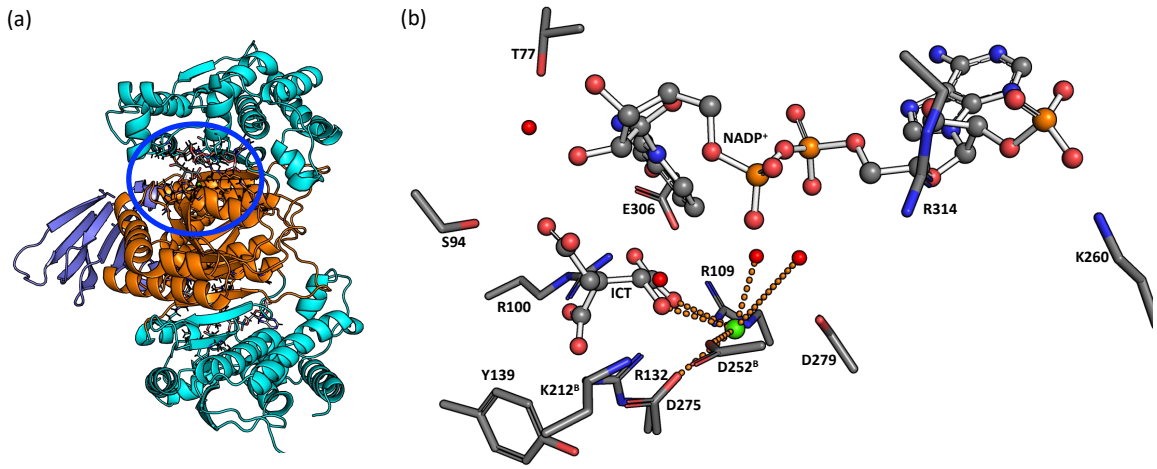


Figure 4.3. (a) Cartoon representation of the wt-IDH1 enzyme with ICT and $NADP^+$. (b) Representation of the IDH1 active site from the X-ray structure. ICT and $NADP^+$ are shown in ball-and-sticks representation, while the protein residues are shown as sticks. The Ca^{2+} ion (shown in green) coordination interactions are shown as orange dotted lines. Adapted from Ref. [B].

and Tyr139 through H-bonds and (iii) the γ-carboxylate is held in place by Thr77 (through a warer molecule), Ser94, $Thr214^B$ and the $NADP^+$ ribose. Along with the latter interaction, the nicotinamide ring of $NADP^+$ is held close to ICT by virtue of interactions with Glu306 and Asn96. Finally, the phosphate group of the ribose ring carrying the adenine moiety is held in the active site by interactions with Arg314 and Lys260. A Ca^{2+} ion, present in the active site, is vital for charge dispersal during the catalysis. Its coordination polyhedron consists of the α-carboxylate group of ICT, the α-alcohol of ICT, Asp275, $Asp252^B$, and two water molecules. This Ca^{2+} is replaced with Mg^{2+} in the simulations of Chapter 6 to study the catalytically active configuration.

4.3.1. Catalytic Mechanism of the Normal Reaction

The reaction of wt-IDH1 is known to occur in a multi-step way. [199] The first step comprises two further sub-steps (Figure 4.4): a base-initiated deprotonation of the C_α hydroxyl of ICT to Oxalosuccinate (OXS), and the subsequent reduction of $NADP^+$ to NADPH with simultaneous oxidation of OXS to produce a ketone at the C_α position. The second step involves loss of C_β carboxylate of OXS to give enolate and release of CO_2. The final step is the protonation of this enolate by an acid results in α-ketoglutarate as the product.

The exact pathway is predominantly determined by the base in step one, as it

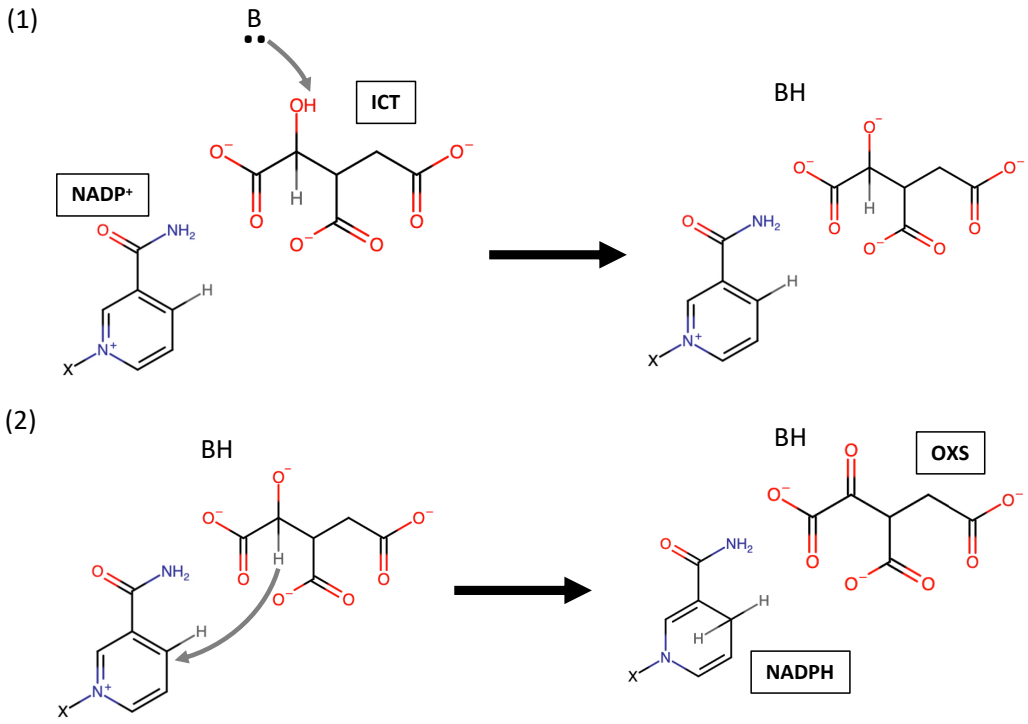


Figure 4.4. A schematic of the mechanism of the first of the normal reaction, split into sub-steps.

plays the role of triggering the catalysis. Significant efforts have been devoted in proposing the most likely base candidate. Hurley et. al. [200] initially suggested that Asp252^{B} could play a significant role as the base in the catalysis. However, Grodsky et al. found that mutating this Asp252^{B} to Asn did not affect the activity of the enzyme.[201] This lead to the proposal that Asp279 is the most likely base. Later, Kim et. al. demonstrated Lys212^{B} mutated to Arg, Gln, or Tyr lowers the activity of IDH1, suggesting that this residue could somehow play a major role. [202] Lys usually possess a protonated N in the side chain, where it cannot act as a base. However within the IDH1 environment, Lys212^{B} might exist in the deprotonated state where it could be well positioned to abstract the proton from the C_{α} hydroxyl of ICT. Classical MD of the protein in which Lys212^{B} was either protonated or deprotonated, along with static QM/MM calculations were performed by Neves and coworker.[203] The activation free energy of the first step was found to be significantly higher with Asp279 as the initiator base than when deprotonated Lys212^{B} as the base (see Table 6.2 for the values).[1] The free energy of the latter pathway (deprotonated Lys212^{B}) agreed fairly well with the experimentally observed $k_{cat} \approx 16$ kcal/mol [204].

Despite the latest evidence pointing towards deprotonated Lys212^{B} as the base,

[1] These calculations used a two-layered ONIOM model at the B3LYP/6-31G(d) level of theory with entropic effects included via harmonic approximation

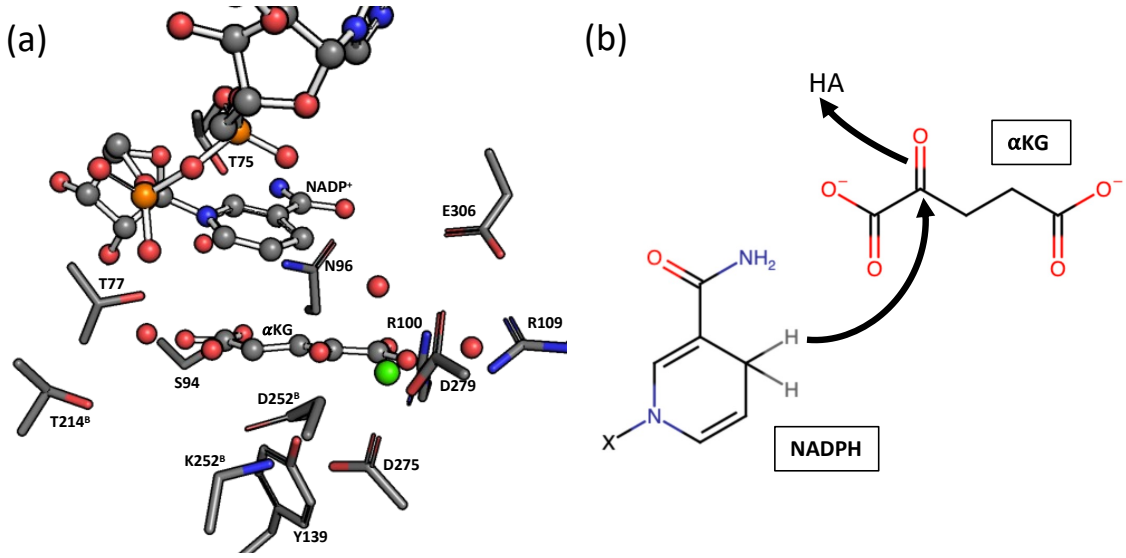


Figure 4.5. (a) Representation of the mut-IDH1 active site from the X-ray structure. αKG and NADPH are shown in ball-and-sticks representation, while the protein residues are shown as sticks. The Ca^{2+} ion (shown in green) coordination interactions are shown as orange dotted lines. (b) A schematic of the mechanism of the neomorphic reaction.

I decided to investigate if there is a possibility of the catalysis taking place with a protonated Lys in Chapter 6. For this, I utilize the QHPC-VS protocol of Figure 5.3 to calculate the free energy of reaction at the quantum level with MiMiC in Section 6.1.

4.4. Active Site of Mut-IDH1

The structure of the catalytically active mut-IDH1 in complex with αKG, NADPH and CA(II) has been crystallized using X-ray crystallography (Figure 4.5a).[178] It occurs as a homodimer, with each of the two monomers consisting of the large, small and clasp domains like wt-IDH1 (see Section 4.3). Two identical active sites of mut-IDH1 (where αKG, NADPH and CA(II) bind) include residues from both monomers, in this text, residues from the second subunit are labelled by the superscript B and those from the first subunit are left unmarked. αKG establishes the following interactions within the active site: (i) the α-carboxylate group of αKG forms a direct H-bond with Arg100 and Arg109, (ii) the β-ketone group interacts with $Lys212^{B}$ and (iii) the γ-carboxylate is held in place by Thr77, Thr75 (through a water molecule), Ser94, $Thr214^{B}$, Asn96 and the NADPH ribose (again through a water molecule). Along with the latter interaction, the nicotinamide ring of NADPH is held close to αKG by virtue of interactions with Glu306 and Asn96. Finally, the phosphate group

of the ribose ring carrying the adenine moiety is held in the active site by interactions with Arg314 and Lys260.

Most of the interactions are very similar to those in the wt-IDH1 active site. except for: (i) The wt-active site had an interaction between Arg132 and the β-carboxylate of ICT, which is not present in the mut-isoform. The Arg132 is mutated to a His, and αKG has a ketone in the β position. The mutation is what likely promotes binding of αKG over ICT, but His132 does not interact closely with αKG itself. (ii) Tyr139 is rotated, establishing interactions with Asp275. Asp275 itself is also in the vicinity of $Lys212^{B}$ and the β-ketone of αKG.

The Ca^{2+} ion anchors the substrate to the active site. The coordination sphere is heptacootdination consists of the α-carboxylate of αKG, the α-ketone of ICT, Asp275, $Asp252^{B}$, Asp279 and two water molecules. This Ca^{2+} is replaced with Mg^{2+} to study the catalytically active configuration. However, this introduces difficulties as described in Section 7.1.

4.4.1. Catalytic Mechanism of the Neomorphic Reaction

Unlike the normal reaction catalyzed by wt-IDH1, exhaustive studies on the neomorphic reaction have not been carried out so far. However, the conversion of αKG to 2-HG by mut-IDH1 can be compared to the that of pyruvate to lactate by lactacte dehydrogenase (LDH).[205] Both involve reduction of a ketone to alcohol, and a nicotinamide ring (NADPH in mut-IDH1 and NADH in LDH). The conversion of pyruvate to lactate involves donation of hydride by NADH, and a proton from an acidic residue of LDH.[206] A similar mechanism, with NADPH and an acidic residue from mut-IDH1, can be expected (Figure 4.5b). To fully understand the catalysis, two questions need to be answered:

- Which acidic residue in the protein is the proton donor?
- Are the hydride transfer (from NADPH) and the proton transfer (from the acidic residue) concerted or subsequent? If subsequent, in what order do they occur?

We can begin to formulate an answer to the first question from inspection of the crystal structure itself. A lysine residue in the protonated state is very acidic, and can act as a potent proton donor. One such residue in the active site, $Lys212^{B}$ is close enough to interact with the C_{α} ketone of αKG. At the same time, there is the Asp275 residue, which is also well positioned to act as an acid if protonated. For a thorough investigation, both the deproronated $Lys212^{B}$/protonated Asp275 pair and proronated $Lys212^{B}$/deprotonated Asp275 pair must be investigated. This protonated Asp275, or even Tyr139 in the vicinity, may be able to take up the role

of the acid. In fact, the latter is the preferred explanation by some experimental groups.[207] However, it must be noted that, Rendina et. al. showed that IDH1 with Tyr139 mutated to Asp could still catalyse the neomorphic reaction, albeit at a reduced rate.[204] This indicates that this Tyr may not be the primary proton donor in the catalysis. The investigation of both configurations is done in Chapter 7.

5. MiMiCPy for MiMiC Input Preparation

The discussions in this chapter are based on publication [A] (see Section 1.3).

The proposed QHPC–VS protocol of Figure 1.3 involves first performing cMD simulations with GROMACS, and then subsequently moving to MiMiC-QM/MM MD. This brings us back to the discussion of problem 1 in Section 1.1.2. Preparing and performing classical MD simulations of biological targets in well-developed programs like GROMACS is mostly—although tedious—a well standardized and documented procedure; moving from GROMACS to MiMiC-QM/MM, on the other hand, is not. A MiMiC-QM/MM simulation involves running both CPMD and GROMACS simultaneously. Due to the loose-coupling nature, the QM atoms involved have to be fed to the QM and MM engine separately. This information has to be tailored to the input formats of the specific software package, with it still being consistent. In the specific case of CPMD and GROMACS, this can be particularly tedious to keep track of. GROMACS internally groups the atoms by molecules as specified in the input coordinates, but CPMD organizes the atoms in blocks of atomic elements. As one can imagine, performing the conversion from GROMACS to CPMD atom formats by hand is an error prone process and neither CPMD nor GROMACS will raise easy-to-understand messages to warn the user about any mismatches. This can lead to segmentation faults or unphysical dynamics during the QM/MM simulation, making these errors difficult to debug for the inexperienced user. This is especially true in the case of QM regions with large number of atoms, an expected situation if MiMiC-QM/MM were to be used within the proposed QHPC–VS pipeline.

Preceding the start of my thesis, MiMiC had most frequently been used to study systems with QM regions of ~50 atoms.[18, 19, 21, 22] In these use cases, it was possible to perform the conversion of MM to QM/MM inputs by hand or with "quick-and-dirty" scripts. In this study, and for future drug design applications, the QM region would consists of more than 100 atoms. A standardized, self-contained code to reliably perform this conversion is vital to the efficient implementation of the QHPC–VS protocol. Furthermore, if the wide adoption of MiMiC among the drug design community is to be achieved, a toolset for the automated conversion of MM inputs to QM inputs is required. For this purpose, the MiMiCPy package was developed as part of this work (in collaboration with F. K. Schackert from [A]). This chapter

provides an overview of the package, including a description of the software design as well as its usage.

5.1. General Overview

MiMiCPy is a python package that aids in the preparation of MiMiC QM/MM simulations. In line with MiMiC, MiMiCPy currently supports building CPMD and GROMACS input files. Furthermore it is designed with a modular object-oriented programming structure in mind, and can be expanded relatively smoothly to add support for new QM and MM engines as required by MiMiC. The code is available in the MiMiC-projects repository on GitLab, and the current version (v0.1.0) has been uploaded to the python repository PyPI. MiMiCPy exposes it functionalities in three ways:

- **Command-line** subcommands. This is the easiest and most user friendly. The following subcommands are currently available:
  - PrepQM outputs the MiMiC-compliant CPMD and GROMACS input files for a QM/MM run from the GROMACS MM data.
  - CPMD2Coords writes the QM atoms selected in a MiMiC-compliant CPMD input file to a gro or pdb file.
  - FixTop fixes missing information in GROMACS topology files that are required by CPMD in a MiMiC run.
  - CPMDid provides the indices that CPMD assigns to each atom; this is especially useful for the MM atoms because in general such indices are reshuffled in a non-obvious way with respect to the GROMACS ordering.
  - Geom2Coords converts a CPMD GEOMETRY file to a gro or pdb file for easy visualization.
- A **PrepQM plugin** for molecular visualization software. This is ideal for selecting complex QM regions where visual inspection is preferred. Currently, interfaces with VMD[208] and PyMOL[209] are supported.
- A bare **python library**. This is useful when developing an automated python workflow for running MiMiC simulations.

In particular, PrepQM is the chief tool used to generate the GROMACS tpr and the CPMD input files. A workflow diagram depicting the general scheme of how

MiMiC-compliant input scripts (i.e., the individual CPMD and GROMACS inputs) are generated with MiMiCPy is shown in Figure 5.1. The central role of the `mimicpy prepqm` command is to be noted here. With the right topology and coordinate information, the tool generates the CPMD input script for a MiMiC run. It also generates a GROMACS index file of the QM atoms selected, and can optionally generate the GROMACS run tpr file automatically by calling `gmx grompp`. This is a complicated tool, and the main features are described in Section 5.2. A further discussion on generating MiMiC-compatible QM/MM input files with PrepQM is given in Appendix A.

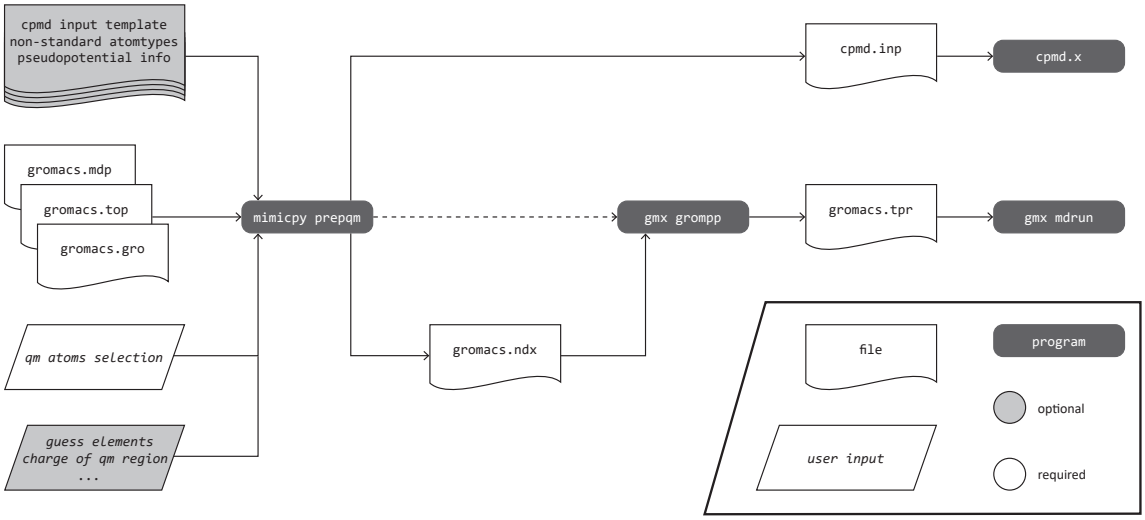


Figure 5.1. Flowchart of the generation of the CPMD and GROMACS input files for a MiMiC-based QM/MM simulation. Adapted from Ref. [A].

5.2. The PrepQM Subcommand

The most basic command to create MiMiC input files is:

```
$ mimicpy prepqm -top topol.top -coords coords.gro
```

Listing 5.1 Executing a basic PrepQM command in the command line.

Executing Listing 5.1 passes the GROMACS topology file topol.top and the initial coordinate file for the QM/MM run coords.gro to the `mimicpy prepqm` subprogram. Before running the command, GROMACS must be sourced, i.e. available on the command line. This is because MiMiCPy requires the `GMXDATA` environment variable

to be set, so that it can access the force field directory in the GROMACS installation. Another option is to copy the complete force field data from the GROMACS installation directory to the current directory. Still another option is to pass the path of the force field with the `-ff` option.

5.2.1. Selection of QM Atoms

Running Listing 5.1 starts an interactive session, where instructions can be given to add and/or deleted atoms to the QM region. An example of such an instruction is:

```
Please enter selection below. For more information type 'help'
> add resname is ARG
> q
```

Listing 5.2 Selecting QM atoms within the PrepQM interactive session.

The instructions after `add` in Listing 5.2 correspond to the selection query that identifies the atoms to be added to the QM region. After selecting the desired atoms, `q` is used to exit the interactive session. MiMiCPy provides a selection language similar to PyMOL and VMD for selecting QM atoms. The syntax for the selection query follows the following general structure:

```
<atom selection keyword> <logical operator> <value>
```

The `atom selection keywords` can include `add` for protein residue names, `resid` for protein residue ID name for the atom name,`type` for the atom type, `id` for the atom ID, and `mol` for the molecular/chain. All the IDs and names are as per the conventions of MM engine, in this case from the GROMACS topology. Logical operators can be `is`, `not`, `>`, `<`, `>=`, or `<=`. Many selection queries can be strung together by using `and` or `or` operators, and grouped with brackets. An example of a more complex selection is shown below:

```
Please enter selection below. For more information type 'help'
> add (resname is ARG and name not C3) or (resid < 15)
```

Listing 5.3 An example of a complex selection query within PrepQM.

Here, all atoms with residue name `ACT` are selected, except for those with atom name `CA`, along with atoms with residue ID less than 15. Atoms can also be deleted from the QM region with the same selection syntax using the `delete` command. To

achieve the same result as Listing 5.3, a series of `add` and `delete` commands can also be used as shown below:

```
Please enter selection below. For more information type 'help'
> add resname is ACT
> add resid < 15
> delete name is C3
> q
```

Listing 5.4 Adding and deleting QM atoms in PrepQM.

5.2.1.1. Selecting Boundary Atoms

As described in Section 2.2.2.3.1, dealing with covalent bounds across the QM-MM boundary is very important. MiMiC treats this using the boundary pseudopotential method, were atoms at the QM-MM border are supplied to CPMD with a special pseudopotential. Writing a CPMD input file with these details can get complicated for large systems. Hence, PrepQM provides options to achieve this seamlessly. Consider the IDH1 enzyme containing multiple residues. If we wish to include Arg100 and Arg109 residues in the QM region, we launch PrepQM as in Listing 5.1 and enter the following selection:

```
Please enter selection below. For more information type 'help'
> add resid is 100 or resid is 109
> delete name is C or name is O
> delete name is HA or name is N or name is H
> delete name is CA
> add-bound name is CA
```

Listing 5.5 Selecting boundary atoms in the QM region in the PrepQM interactive session.

In Listing 5.5, we add the atoms of residues 100 and 109 while also deleting those not part of the side chain (`C`, `O`, `HA`, `N`, and `H`). The alpha carbons of protein residues (`CA`) often form a convenient point to place the cut between the QM and MM regions. To mark these as QM atoms, these are deleted and re-added using the `add-bound` command rather than the generic `add`.

The detection and assignment of QM boundary atoms can also be done automatically by launching PrepQM with the `-bound` option set to `True`:

```
$ mimicpy prepqm -top topol.top -coords coords.gro -bound True
```

Listing 5.6 PrepQM command to automatically detect boundary atoms.

The following selection must be entered:

```
Please enter selection below. For more information type 'help'
> add resid is 100 or resid is 109
> delete name is C or name is O
> delete name is HA or name is N or name is H
```

Listing 5.7 Adding QM atoms for automatic boundary detection in PrepQM.

With the selection as in Listing 5.7, we make sure that the boundary atoms (`CA`) are included in the selection as 'normal' atoms. PrepQM will transverse through the bond connectivity network and automatically detect that the `CA` atoms lie on the QM-MM boundary. This will result in a separate atoms block in the CPMD input file, containing only boundary atoms, where the right boundary pseudopotential can be specified.

5.2.2. Handling Non-standard Atomtypes

The most important information to be passed from GROMACS to CPMD is the QM atom coordinates and atomic elements. The coordinates can be easily read from the appropriate file, but reading the atomic element information may be more difficult in certain cases. For standard atomtypes (those that make up the protein residue), the atomic element information is stored in the GROMACS forcefield data. However, for non-standard atom types generated using programs like ACPYPE[210], the information is not found in the forcefield. MiMiCPy will automatically guess the atomic elements information, based on a combination of atomic mass, name and type. However, the user needs to verify that the guessed atomic elements for each atom type are meaningful, as it is essential for CPMD to run correctly. When atomic elements are guessed, they are reported as follows:

```
Some atom types had no atom number information.
They were guessed as follows:

+----------------------+
| Atom Type | Element |
+----------------------+
|     c     |    C    |
```

```
+---------------------+
|      c3     |    C    |
+---------------------+
|      o      |    O    |
+---------------------+
|      hc     |    H    |
+---------------------+
```

The automatic behavior of guessing atomic elements can be toggled on or off using the `-guess` option. If `False` is passed and non-standard atoms are present, MiMiCPy will exit with an error message instead of attempting to guess the elements. If you are not satisfied with the element information guessed, a file containing the list of all non-standard atom types with the correct element information can be also be passed to `mimicpy prepqm` with the `-nsa` option:

```
$ mimicpy prepqm -top topol.top -coords coords.gro -nsa atomtypes.
    dat
```

Listing 5.8 Passing custom atomtypes information to PrepQM.

5.2.3. Plugins

The console version of the PrepQM tool allows for selecting the QM region with only the MiMiCPy selection language. This is sufficient for simple selection commands. But, if more complex QM regions are required, PrepQM plugins for the popular molecular visualization software VMD and PyMOL are provided. For example, to specify the protein residues that are 3.5 Å around Zinc (atom name `ZN`) in the GB1 protein from ref [18], we first load the cMD equilibriated structure into VMD, then, we run the following command within the Tkconsole:

```
$ set sel [atomselect top "(same residue as within 3.5 of name ZN)
    and not (protein and (name O N C H HA CA))"]
$ set sel_bound [atomselect top "(same residue as within 3.5 of name
    ZN) and name CA"]
$ prepqm -top topol.top -sele $sele -pad 3.5 -sel_bound $sel_bound
```

Listing 5.9 Invoking the PrepQM plugin within the VMD Tkconsole enviornment.

The selection of the QM atoms and the boundary C-α atoms are made using `$sel` and `$sel_bound` according to the usual syntax of VMD. After that, the `prepqm` command is called, where `$sel` and `$sel_bound` are passed separately using the `-sel` and

`-sel_bound` options respectively. Similar to Listing 5.6, the boundary atoms can be detected automatically:

```
$ set sel [atomselect top "(same residue as within 3.5 of name ZN)
   and not (protein and (name O N C H HA))"]
$ prepqm -top topol.top -sele $sele -pad 3.5 -bound True
```

Listing 5.10 Automatic detection of boundary atoms by the PrepQM VMD plugin

As in the case of Listing 5.6, the selection passed to `PrepQM` (`$sel`) in Listing 5.10 must include the atoms to be detected as boundary (in this case, the C-α atoms).

A similar tool for PyMOL is also provided:

```
    prepqm topol.top, sele0, pad=3.5, sele_bound=sele_bound0
    prepqm topol.top, sele0, pad=3.5, bound=True
```

Listing 5.11 Invoking the PrepQM plugin within the PyMOL console enviornment.

The first line involves passing the PyMOL selection of the QM atoms (`sele0`) and the boundary selection (`sele_bound0`) explicitly to `prepqm`. The second command triggers automatic detection of the boundary atoms.

5.3. Other Subcommands

5.3.1. FixTop

In Section 5.2.2 we described how to include the atomic element information of all QM atoms in CPMD. In certain cases, there may exist non-standard atom types in the MM region for which GROMACS does not pass on the atomic element information to CPMD. Although in such cases MiMiC would still run, it could lead to garbage values for these atoms in CPMD and may cause hard to diagnose errors further down the simulation. It is recommend to fix the GROMACS topology itself, so that the correct information is passed on to CPMD. This can be either done using options the specialized subprogram FixTop.

FixTop guesses missing atomic species information in the topology file (this includes atoms in the MM region, which PrepQM will not fix) and prints a consolidated `[ atomtypes ]` section into a GROMACS .itp file. Within the GROMACS topology format, the `[ atomtypes ]` section specifies the definition of atom types with the

atomic element information. The easiest way to incorporate this information into an existing GROMACS force field is to write it to the ffnonbonded.itp file containing the `[ atomtypes ]` definition of the whole system for all default GROMACS force fields. For e.g., a copy of the AMBER force field directory from GROMACS is created locally under amberff/. Now run the following command:

```
$ mimicpy fixtop -top topol.top -cls amberff/ffnonbonded.itp -out
    amberff/ffnonbonded.itp
```

Listing 5.12 An example of running the FixTop subcommand.

FixTop replaces the `[ atomtypes ]` section in amberff/ffnonbonded.itp with the updated one containing all species information and clears other `[ atomtypes ]` sections from the topology (as `-cls` was specified). This is done to avoid conflicting atomtypes definition errors which would be raised by the GROMACS preprocessor while generating the tpr. Running `mimicpy prepqm` using this topology should result in all the atomic element information being present for both MM and QM atoms.

5.3.2. CPMDid

Constraining the motions of selected atoms with the use of external forces is often used in MD to limit the exploration of the phase space to a region of interest. This technique is utilized, for example, within the thermodynamics integration scheme of Chapter 6). Applying constraints, among other commands, requires the user to specify the indices of the atoms involved. Within a MiMiC-QM/MM run with CPMD and GROMACS, the atom IDs should be specified according to the order internally stored by CPMD. QM atoms are numbered according to the order in which they occur within the `&ATOMS` section, whereas the MM atoms are shuffled according to the atomtypes in order of appearance in the topology. This leads to an overall non-obvious ordering of atoms, which makes it extremely difficult to parse out the atom IDs. The MiMiCPy tool CPMDid has been provided to solve this problem.

Suppose we want to add a distance constraint between the `C1` atom of `AKG` and the `OW` atom of residue `1`. The former is in the QM region and the latter in in the MM region. In this case we launch the following command:

```
$ mimicpy cpmdid -top topol.top -inp cpmd.inp
```

Listing 5.13 Example of running the CPMDid command.

This will then launch an interactive session similar to PrepQM, where the selection for the atoms can be entered:

```
Please enter selection below. For more information type 'help'
> name is C1 and resname is AKG
> name is OW and resid is 1
```

Listing 5.14 An example of an interactive selection session with CPMDid.

This will output the atom IDs of the selected atoms, which can then be directly used in the CPMD input file. The indices can be printed in a table format (for debugging), list format (for quickly copying into the input), or as a range (for certain tasks like multiple thermostats). The printing format can be set with the `-print` option in 5.13.

5.3.3. Other Debugging Tools

MiMiC-QM/MM runs may often crash due to a number of reasons. Visualizing the systems is often the first and best approach to diagnosing the problem. Most discrepancies with the structure will often manifest themselves within the GEOMETRY file (or GEOMETRY.xyz file) outputted be CPMD. However, due to the atom shuffling described in Section X, visualzing this file within VMD or PyMOL becomes difficult. For this reason, the Geom2Coords tool is provided:

```
$ mimicpy geom2coords -geom GEOMETRY -top topol.top -inp cpmd.inp
   -coords GEOMETRY.gro
```

Listing 5.15 Example of running the Geom2Coords command.

In 5.15, the GEOMETRY files, topology and MiMiC-compliant CPMD input file (with the right QM atoms) are given. This will then output the file GEOMETRY.gro, which can be visualized easily.

A lighter version of this tool is CPMD2Coords, which converts just the QM atoms in the CPMD input file into a standalone GRO/PDB file:

```
$ mimicpy cpmd2coords -top topol.top -inp cpmd.inp -coords GEOMETRY.
   gro
```

Listing 5.16 Example of running the CPMD2Coords command.

5.4. Importing MiMiCPy within Python

MiMiCPy can also be used directly as a Python library. If available in `PYTHONPATH`, it can be imported as per the usual procedure. A simple snippet of Python code to generate the a CPMD input file and GROMACS index file for a MiMiC run is shown below:

```
1 import mimicpy as mp
2
3 prep = mp.Preparation(mp.DefaultSelector("topol.top", "coords.gro"))
4 prep.add("resname is ACT")
5 ndx, cpmd = prep.get_mimic_input(box_padding=0.35)
```

Listing 5.17 A basic example of using MiMiCPy as a Python library.

The code in Listing 5.17 initializes an object of the `Preparation` class with right topology and coordinate files passed (line 3). The right QM atoms are selected using the `Preparation` class instance (in line 4). Here `prep` can be used to further add/delete atoms to the QM region using the `add()` and `delete()` methods, respectively:

```
1 prep.add("resname is ACT")
2 prep.add("resid is 10")
3 prep.delete("name is C3")
4 prep.add("name is C2", is_bound=True)
```

Listing 5.18 An example of selecting QM atoms within the MiMiCPy Python library.

The QM atoms are stored internally as a Pandas `DataFrame`, which can be viewed (but not edited directly) using the `prep.qm_atoms` property. Atoms can be manually added with boundary pseudopotentials by passing `is_bound=True` to `add()` (line 5 in Listing 5.18). The QM atoms can also be automatically detected:

```
1 prep.clear()
2 prep.add("resname is ACT")
3 prep.add("resid is 10")
4 prep.delete("name is C3")
5 prep.find_bound_atoms()
```

Listing 5.19 Automatic detection of QM boundary atoms within the MiMiCPy Python library.

The QM atoms in the `DataFrame` are cleared in Line 1 of Listing 5.19, to remove the selection made in Listing 5.18. Line 4 then triggers detection of boundary atoms automatically.

After making the required QM atoms' selection with the boundary atoms, calling the `get_mimic_input()` method (line 5 in Listing 5.17) returns the required CPMD and GROMACS run files as instances of type `CpmdScript` and `Ndx`, respectively (line 5 in Listing 5.19). Many options that can be passed to the PrepQM command-line tool can also be passed as parameters to this function (e.g., the `pad` parameter in this case).

5.4.1. Software Design

A further explanation of the code in Listing 5.17, and a deeper dive into the intricacies of using the code, necessitates a discussion on the software design and philosophy behind MiMiCPy (Figure 5.2). It was designed with a fully object-oriented and modular architecture in mind. This includes the four main modules: `core` (including the `selector` module), `topology`, `coords` and `scripts`.

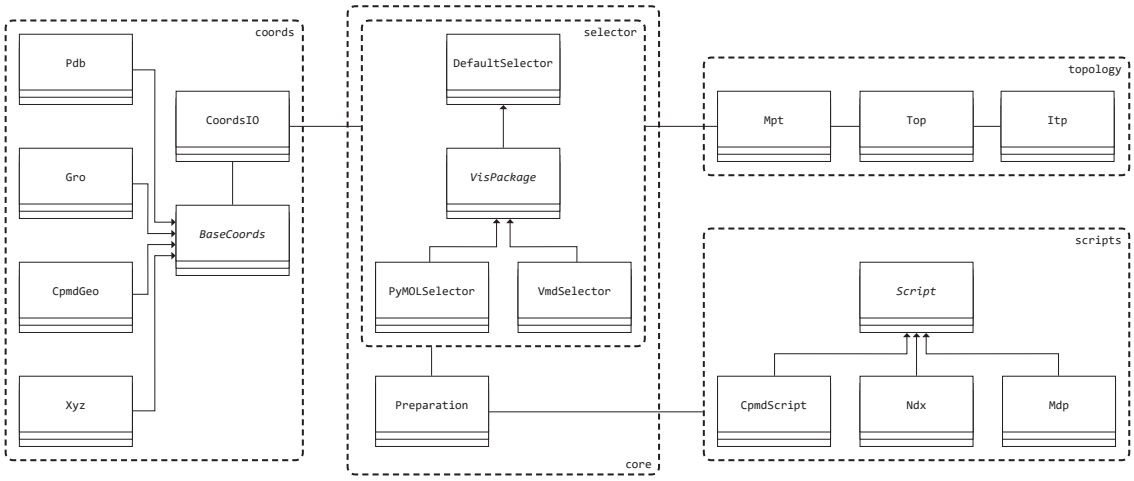


Figure 5.2. Organization of the main classes in MiMiCPy. Adapted from Ref. [A].

All functionalities are centered around, and exposed to the user through the `core` module. Specifically, the user is meant to interact with MiMiCPy mainly through the `Preparation` class, which is the 'central' class that keeps track of all the selected QM atoms, and provides functionalities to generate the required MiMiC-complaint input files. This was demonstrated in line 3 of Listing 5.17. In reality, we can split this into two lines of code:

```
1 selector = mp.DefaultSelector("topol.top", "coords.gro")
2 prep = mp.Preparation(selector)
```

Listing 5.20 Initializing a selector and a preparation object within the MiMiCPy Python library.

Notice that the variable `selector` was passed as a required parameter to the `Preparation` constructor. This is an instance of the "Selector"-type classes (contained within in the `selector` module). These classes combine coordinate and topology information of the system, and actually handle the selection of the atoms. Different classes are provided depending on the specific selection method required. In fact, the `Preparation` class can be essentially thought of as a decorator for the selector classes, aggregating them and attaching the new behavior of input file generation. All "Selector"-type classes expose two functions: (i) `select()` and (ii) `mm_box()`. These are functionalities required by the `Preparation` classes, the former performing the parsing of the selection query of the QM atoms (using the chosen method) and the latter including calculating the total MM box size.

Listing 5.20 uses the default MiMiCPy selection language (as described in 5.2.1), and hence an object of the `DefaultSelector` class is instantiated. To use the selection languages of VMD and/or PyMOL, instances of the `VmdSelector` or `PyMOLSelector` classes can be created using a similar syntax. This is not meant to be accessed directly by the user, but rather internally by the VMD and PyMOL PrepQM plugins. `VmdSelector` and `PyMOLSelector` provide a simplified façade to the VMD and PyMOL software packages in order to work with the `Preparation` class. These classes inherit from the abstract `VisPackage` class (which in turn inherits from `DefaultSelector`). This ensures that a common set of rules are laid down for all these façade classes. A new façade class can be easily added to allow MiMiCPy to interface with other molecular visualization packages. `VisPackage` primarily includes abstract methods (to be overloaded by the children) that handle the conversion of the selected atoms (including coordinates/distance units) to the internal MiMiCPy format.

The coordinate and topology data are loaded into the "Selector"-type classes by using dedicated `CoordsIO` and `Mpt` (MiMiCPy topology) classes. These handle different coordinate and topology formats by passing the information to dedicated parser classes. The coords.gro or coords.pdb coordinate file can be read within MiMiCPy:

```
1 gro_hndl = mp.CoordsIO("coords.gro")
2 pdb_hndl = mp.CoordsIO("coords.pdb")
```

Listing 5.21 Initializing a coordinate IO object within the MiMiCPy Python library.

Multiple coordinate formats (GRO, PDB, CPMD GEOMETRY, XYZ) are supported by the same `CoordsIO` class. All these classes are housed within the dedicated `coords` module. In reality, the parsing of these formats are handled by dedicated classes, including `Gro`, `Pdb`, `CPMDGeo` and `Xyz` classes to handle the respective formats. These which inherits from the abstract `BaseCoords` class, providing the skeleton of a coordinate parser. The CoordsIO class acts as an adapter that aggregates and wraps the different parser classes, and exposes only the coordinate information as a Pandas `DataFrame` (through `coords_hndl.coords`), the box size (through `coords_hndl.box`), and the write function (through `coords_hndl.write()`) to the user.

The `Mpt` class functions in a similar way, as an adapter interfacing multiple topology parser classes. The `Mpt` class provides a common framework to deal with disparate topology formats. Mainly, it exposes methods for selecting specific atoms from the topology. Currently, only the GROMACS topology format (.top) is supported. Other formats may easily be supported by adding new classes that interface with the `Mpt` class. All topology-related classes are housed within the `topology` module. The topol.top topology file can be read within MiMiCPy:

```
1 mpt_hndl = mp.Mpt.from_file("topol.top")
2 sele = mpt_hndl.select("resname is ACT")
```

Listing 5.22 Initializing a topology object within the MiMiCPy Python library.

The main function exposed by `Mpt` instances to the user is the `select()` function (line 2 of 5.22), which actually implementing the default MiMiCPy selection language exposed to the user through `DefaultSelector`. Here `sele` is a Pandas `DataFrame` containing: the atom numbers (`number` column), the atom types (`type`), the residue IDs (`resid` column), the residue names (`resname` column), the atom names (`name` column), the atomic charges (`charge` column), the atomic elements (`element` column), and the atomic masses (`mass` column) of the atoms corresponding to the selection passed (`resname is ACT` in this case). Since the rest of the Listing within MiMiCPy interfaces with the coordinates and topology parser functionalities through only the common `CoordsIO` and `Mpt` classes, further topology and coordinates parser classes can be seamlessly added without disturbing the rest of the code. This modular structure is what allows MiMiCPy to be seamlessly extended and support new file formats. This allows it to quickly keep up with new developments in the MiMiC framework (like support for new MM/QM engines).

All scripts in MiMiCPy are represented as children of the abstract `Script` class (placed within the `scripts` module). This includes the `ndx` and `cpmd` objects in line 5 of Listing 5.17, which are of type `Ndx` and `CpmdScript` respectively. This allows for 'pythonic' interactions with these script instances, i.e., using the dot operator for

setting and getting of script properties. For example, the total net charge of the QM region, reported as the `CHARGE` parameter in the `&SYSTEM` section of the CPMD input file can be accessed through `cpmd.SYSTEM.CHARGE`. All `Script` instances can be converted to and from string, and in turn, into text files:

```
1 with f as open("cpmd.inp", "w"):
2     f.write(str(cpmd))
3
4 with f as open("index.ndx", "w"):
5     f.write(str(ndx))
```

Listing 5.23 Writing script objects to file within the MiMiCPy Python library.

The module approach of MiMiCPy allows it to be extremely flexible in adding new features. This includes quickly supporting new subprograms of MiMiC; MiMiC is expected to add support for new subprograms in the future (for e.g., the currently work-in-progress CFOUR interface[211]). MiMiCPy will need to support the new coordinates, topology and scripts file types associated with the new subprograms. This can be easily achieved by: creating a child of the `BaseCoords` class that interfaces with the `CoordsIO` class, a topology parser interfacing with the `Mpt` class, and a script parser that inherits from the `Scripts` class. Furthermore, the `VisPackage` class can be extended to create plugins for new molecular visualization packages. The key point to note it that all these new features can be incorporated without disturbing the functioning of the rest of the code, allowing for faster development cycles and easier code maintenance.

5.5. Conclusion

This chapter presented MiMiCPy, a companion tool of or a front-end tool for MiMiC. The code simplifies the preparation and debugging of MiMiC-compliant input files via a user-friendly interface. An an extensive list of command-line tools are provided. This includes, PrepQM for the generation of CPMD input files and GROMACS tpr files from the GROMACS topology and coordinate files. An easy-to-use selection language allows the selection and design of QM regions, with the correct guessing of atomic species from the MM topology. Further tools to facilitate running MiMiC simulations for biomolecules are also available.

A plugin version of PrepQM for PyMOL and VMD is provided for the selection of visually complex QM regions. MiMiCPy can also be used as a Python library, allowing one to develop complex drug design workflows to with MiMiC-based QM/MM simulations. The package has been designed with a modular and object-oriented ap-

proach. This allows the package to easily support new subprograms with a multi-scale framework, when they become available in MiMiC.

Figure 5.3. Updated QHPC–VS protocol, which includes the input preparation step with MiMiCPy.

The development of MiMiCPy, solves the problem 1 (discussed in Section 1.1.2) with the proposed QHPC–VS protocol of Figure 1.3. We can now update the protocol to include the use of MiMiCPy as in Figure 5.3. This is used in Chapter 6 to tackle problem 2 mentioned in Section 1.1.2.

6. QM/MM MD Simulations of Wild Type IDH1

The results in this chapter are based on publication [B] (see Section 1.3).

MiMiC has been previously used to study various interesting biological systems.[19, 24, 21, 23] However, its performance and scaling for larger, therapeutically-relevant biomolecules has not been so thoroughly studied. The applicability of MiMiC for drug design cannot be demonstrated without achieving QM/MM simulations of such systems in a reasonable amount of time. This is problem 2 as discussed in Section 1.1.2. I intend to tackle this in this chapter by selecting a suitable drug target as a test case. From the discussions in Chapter 4, IDH1 is an interesting drug target for glioma warranting further study. Furthermore, given the complex active site (consisting of the divalent Mg^{2+} ion, charged small molecule ICT, and large $NADP^{+}$ cofactor) a quantum description is required to sufficiently describe the systems. Experimental and computational literature on the dynamics and catalysis of wt-IDH1 already exists. In order to test the performance of MiMiC, I simulate and analyze the behavior of wt-IDH1 in this chapter. The subsequent chapter (Chapter 7) will focus on the simulation of the less-known mut-IDH1.

6.1. cMD Equilibriation

Before performing QM/MM simulations, the first step of the QHPC-VS protocol is to perform cMD simulation starting from the crystal structure. This provides a description of the Michaelis complex of the enzyme with clues on the possible base.

6.1.1. Methods

Computational Details. The crystal structure of the human IDH1 enzyme (PDB ID: 1T0L)[25] was solvated with TIP3P waters [212] and the total charge of the simulation box was neutralized by adding Na^{+} ions. The system (130,828 atoms in total) was then equilibriated using the GROMACS package [167] by running a cumulative

1 µs of classical MD (cMD) using the Amber99sb*-ildn force field.[213, 214] This is also the force-field that will be used to describe the MM region in MiMiC QM/MM simulations of Section 6.2. These simulations were performed on the CLAIX18 cluster of the RWTH Aachen University. The force field parameters for $NADP^+$ were obtained from a previous study[215] and the parameters for isocitrate were generated using the Generalised Amber Force Field [216], and partial charges were parameterized using the RESP method at HF/6-31G* level of theory. This was done using the ANTECHAMBER software package and the python package ACPYPE. [217, 210]

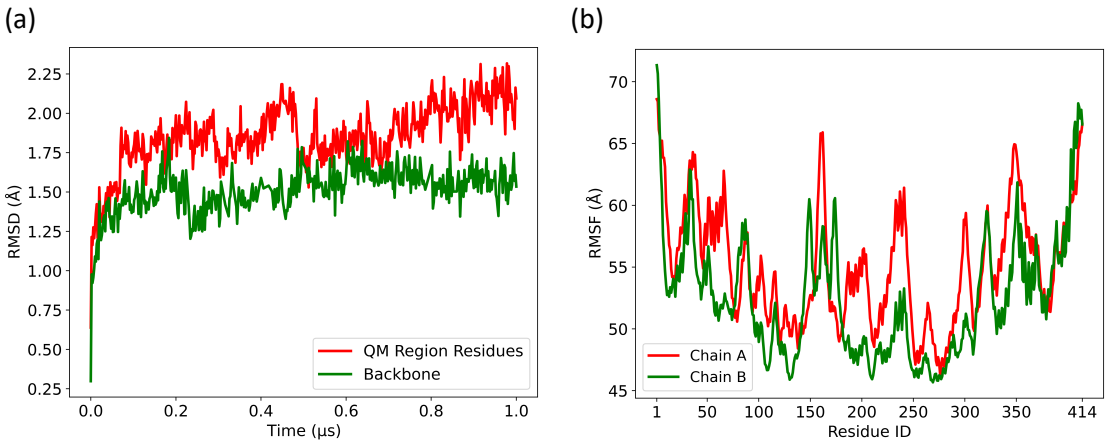


Figure 6.1. The Root Mean Squared (a) Deviation (RMSD) of backbone and important QM residues in active sites, and (b) Fluctuation (RMSF) of subunit A and subunit B of wt-IDH1 throughout the cMD simulation. Adapted from Ref. [B].

Equilibriation. The crystal structure and solvent must be equilbriated to standard conditions before production simulations. For this, geometry optimization, NVT and NPT ensemble simulations with heavy atoms constrained were performed in subsequent steps to smoothly ramp up the temperature to 300 K and pressure to 1 bar. The constraints were released, and the simulation was extended in the NPT ensemble. To confirm that the system is equilibrated, the root mean squared deviation (RMSD) of the backbone carbons and those of residues in what will be assigned to the QM region (see Section 6.2.1) during the course of the cMD simulations was calculated (Figure 6.1a). The RMSD of the backbone carbons stabilizes around a value of ~1.5 Å after 0.2 µs, showing minimal deviation from the crystal structure. The RMSD of the QM region residues stabilized at ~2 Åat around the same time. Thus, the production run and further analysis is considered after discarding the first 0.2 µs of equilibriation.

6.1.2. Analysis

RMSF. The root mean squared fluctuations (RMSF) of the backbone carbons of the protein residues for each subunit/chain was calculated (Figure 6.1b). Arg100, Arg109, Arg132, Lys212, Asp252, Asp275, Asp279 and Glu306 (from both chains) represent a local minimum in the RMSF. These are key residues involved in binding of the ligand and cofactor, and underscore their conserved nature within the IDH1 active site and importance in protein functioning.

Michaelis Complex. Figure 6.2a shows the structure of the Michaelis complex of IDH1 with protonated Lys212^{B} as obtained from our simulations: the network of interaction involving the Mg^{2+} ion, ICT, $NADP^{+}$, and the protein residues in the active site are largely similar to the X-ray structure described in Section 4.3. Nevertheless, there are certain key differences:

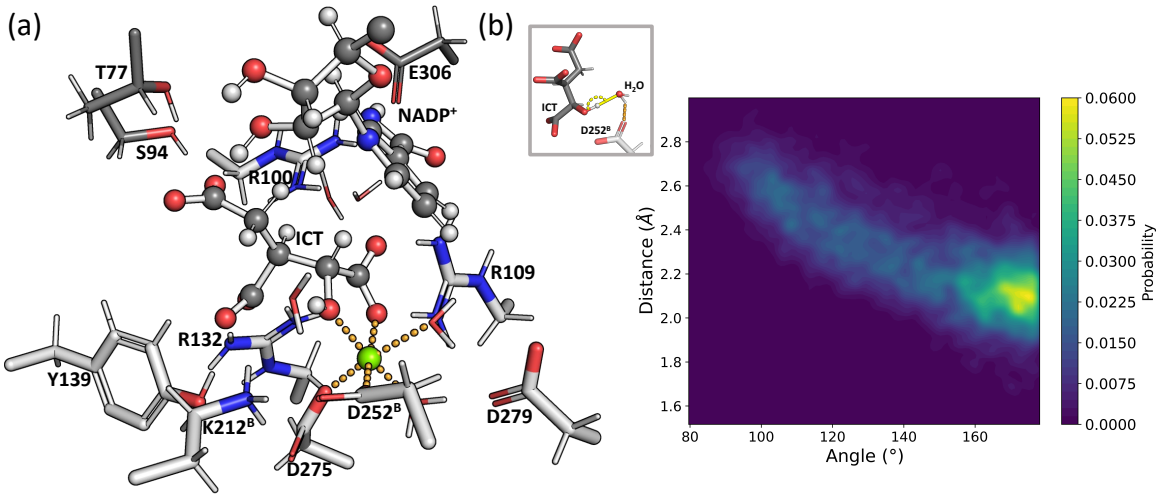


Figure 6.2. (a) Schematic of the Michaelis complex of the wt-IDH1 active site as obtained from cMD simulations. ICT, part of the $NADP^{+}$ pictured, and all residues in light gray are placed in the QM region in our MiMiC-QM/MM simulations. (b) Histogram of the bonding distance versus angle of the ICT Cα alcohol-water interaction for various points on the cMD trajectory. The inset depicts the bonding distance and angle measured. Adapted from Ref. [B].

1. A water molecule forms an H-bond with Asp252^{B} and with the α-alcohol of ICT.

2. Arg100 has moved away from the α-carboxylate of ICT, establishing a weaker water-mediated interaction. Neves et. al found a similar result in their cMD simulations.[203]

3. Thr77 has moved closer to and interacts directly with the γ-carboxylate of ICT.

Point 1 suggests that Asp252^{B} is well positioned to abstract a proton from the Cα hydroxyl of ICT through the H-bonded water molecule. Asp279, on the other hand, interacts with Mg^{2+} through one of the water molecules coordinating with the metal ion. This mediated interaction moves the residue farther away from the ICT Cα alcohol, implying that it is an unlikely candidate for the base. The water H-bonded to Asp252^{B} instead, closely interacts with the Cα alcohol. The histogram of the ICT Cα alcohol-water interaction, i.e., bonding distance versus angle, in the active site of subunit A for various points across the cMD simulations is shown in Figure 6.2b. The ensemble has a significant probability of existing in the space where the angle of bonding is between 170° and 180°, and the distance is between 2 Å and 2.2 Å. This corresponds to configurations where ICT and the water-Asp252^{B} interact through a strong hydrogen bond, and hence the Asp252^{B}-water pair can function as a base. This leads us to suggest that Asp252^{B} is the most likely base candidate, in line with the hypothesis of Hurley et. al. discussed in Section 4.3.1. A point from this configuration is chosen as the initial coordinates for QM/MM MD, where the free energy barrier for this pathway is investigated.

6.2. QM/MM MD with MiMiC

6.2.1. Selecting the QM Region with MiMiCPy

The QM region should include naturally all necessary residues to completely describe the reaction. At the same time, too large a QM region (both in terms of number of atoms and the minimum bounding QM box size) will lead to an extremely expensive calculation. Furthermore, a total negative QM charge should be avoided as they are difficult to handle within the adopted plane wave approach. The optimization of the QM region thus requires a delicate balance of these factors. The resulting QM region was designed with the PrepQM subcommand of MiMiCPy, as described in Chapter 5. The subunit A active site was placed under the QM region. The selection used in the interactive environment is given below:

```
Please enter selection below. For more information type 'help'
> add resid is 833 or resid is 829
> add resid is 835 or resid is 840
> add resid is 831 and (name is C5D or name is C4D or name is H14 or
    name is O4D or name is C1D or name is H19 or name is C2D or name
    is H17 or name is O2D or name is H18 or name is C3D or name is
   H15 or name is O3D or name is H16 or name is N1N or name is C6N
   or name is C5N or name is H25 or name is C4N or name is H23 or
   name is H24 or name is C3N or name is C2N or name is H20 or name
   is C7N or name is O7N or name is N7N or name is H22 or name is
   H21)
```

```
> add (resid is 275 or resid is 279 or resid 666 or resid is 139 or
   resid is 626 or resid is 100 or resid is 109 or resid is 132)
> delete name is N or name is H or name is HA or name is C or name
   is O
> delete (resid is 626 or resid is 100 or resid is 109 or resid is
   132) and name is CA
> delete (resid is 100 or resid is 109 or resid 132) and (name is CB
    or name is HB1 or name is HB2 or name is HG1 or name is HG2)
> add resid is 25198
```

Listing 6.1 An example of a complex selection query.

The first two lines select ICT (residue index 833), Mg^{2+} (residue index 829) and the two water molecules coordinating with the ion (residue indices 835 and 840). The third line selects the $NADP^+$ nicotinamide ring. The total $NADP^+$ molecule is extremley large, and contains extremely negatively charged phosophate groups. Thus the decision was made to include only the nicotinamde ring, since only this participates in the reaction. This not only involves placing the QM-MM boundary within the cofactor, but also unoptimally cutting across a C-O bond. Nevertheless, this was deemed necessary to allow for efficient simulations. Arg100/109/132, Asp275/279/252^B (internally numbered as residue 666 in the GROMACS tpr), $Lys212^B$ (internally numbered as residue 626 in the GROMACS topology), and Tyr139. The QM-MM boundaries are placed at the Cγ of the Args and Lys, and Cα of the Tyr and Asps. Finally, the water molecule (residue index 25198) H-bonding with $Asp252^B$ and the α-alcohol of ICT is included. The handling of boundary atoms were done automatically by PrepQM see (Listing 5.6). This results in 142 atoms in the QM region, with a required minimum QM box size of around 25.0 a.u. × 25.0 a.u. × 25.0 a.u. to bound all QM atoms.

6.2.2. Methods

Computational Details. Born-Oppenheimer MD was employed, with a timestep of 0.5 fs. Temperature was maintained around 300 K using a Nosé-Hoover thermostat. The mixed QM-MM electrostatic interactions are split into short and long range (see Section 3.3.3), where the cutoff radius of 32 a.u. was used. The long-range interactions were computed using Equation 3.9 up to the 5^{th} order multipole expansion. The Tuckerman method of the Poisson solver is used to decouple the long-range electrostatic interactions of the QM region with its periodic images.[218] This method has been proven to be the most effective for typical systems studied in biology. The method requires a sufficiently large padding between the the outermost atoms and the box walls. Here, the plane-wave basis set was expanded (see Equation 2.59 in Section 2.2.2.2.2) in a cubic QM box of length $\mathtt{X}$. To determine the ideal cutoff energy $\mathtt{E_{cut}}$ and the QM box length $\mathtt{X}$, they are varied and the combination that lead to force norm convergence were noted. From Figure 6.3a, $\mathtt{E_{cut}} = 100$ Ry and $\mathtt{X} = 46$ a.u. (box

of size 46.0 a.u. × 46.0 a.u. × 46.0 a.u.) were used for the simulations. This resulted in 294 real-space grid points for the electronic density along the x-axis. All QM/MM simulations were performed on the JUWELS cluster of the Jülich Supercomputing Center, where there are 48 cores per node.[219]

Equilibriation. The equilibrated structure from GROMACS served as the starting point for QM/MM MD. However, these are structures equilibriated at the MM level of theory, and a minimization/optimization procedure is equilibrate them according to the QM/MM energy surface. Due to various technical reasons, a simple geometry optimization procedure within MiMiC-QM/MM has not been implemented. Instead, an initial annealing of the MM structure at the QM/MM level is performed. Here the temperature of the system is progressively reduced to close to 0 K, and dynamics at this temperature is performed to minimize the structure. The system is then subsequently heated back to 300 K, which generates QM/MM velocities at this temperature.

6.2.3. Benchmarking

Nodes N	Cores	Task/Node p_N	Tasks P	CP Groups G	Planes/Task
7	336	6	42	1	7
147	7,056	6	882	6	2
294	14,112	6	1764	12	2
588	28,224	4	2352	24	3
882	42,336	6	5302	36	2
1176	56,448	6	7056	48	2
1470	70,560	6	8820	60	2
1764	84,672	6	10584	72	2

Table 6.1. The configurations used in CPMD for the B3LYP QM/MM MD benchmarks of the IDH1 system reported in Figure 6.3b.

With equilibrated structures, we can now test the scaling of CPMD within the MiMiC QM/MM MD scheme for the large, therapeutically relevant IDH1 enzyme. Only certain node/MPI task/core configurations lead to load balancing. For this, we recall Equation 3.6 and note that $M_x = 294$ (from Section 6.2.2). With only 1 node always assigned to GROMACS, the number of nodes N, number of processors/tasks p_N, and number of CP groups G are varied for CPMD (Table 6.1). The speedup (according to Equation 3.1), is calculated at each configuration for both the BLYP and the B3LYP functional, and shown in Figure 6.3b.

The strong scaling benchmarks performed using the more accurate B3LYP functional exhibits parallel efficiency ~70% (calculated using Equation 3.2) up to and

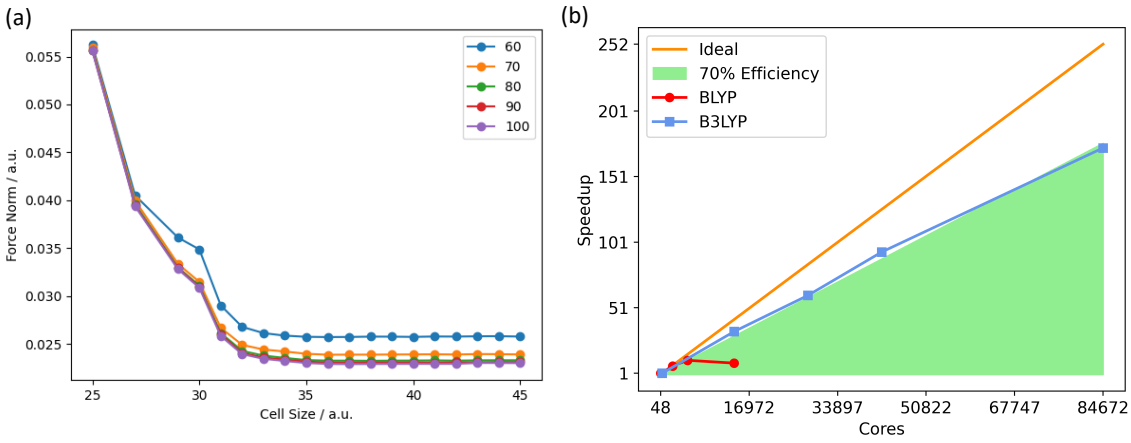


Figure 6.3. (a) Convergence of the force norm with varying E_{cut} and QM box length. (b) Strong scaling of MiMiC-based DFT QM/MM MD simulations at the BLYP and B3LYP level of theory of wt-IDH1 as a function of the number of cores assigned to CPMD.

incredible 84,672 cores (1764 JUWELS nodes). This is much greater than the previously demonstrated best perforrmance for MiMiC (up to ~13,000 cores).[18] This shows that a MiMiC-QM/MM simulation is able to take full advantage of the powerful scaling of CPMD, with minimal added overhead in communication. This validates the parallilzation techniques, discussed in Section 3.3, that are employed in MiMiC. Nevertheless, the performance at 84,672 cores is only 0.74 ps/day, which is not enough to simulate meaningful dynamics for drug design. Here we note that even at 84,672 cores, the strong scaling behavior maintains a largely linear trend. The scaling could not be extended, because subsequent node configurations allowing for load balancing in CPMD would exceed the size of the JUWELS cluster. With more resources available, the performance of MiMiC could be pushed even further, possibly allowing for nanosecond simulations at the B3LYP level. In this regard, the resources of an exascale supercomputer might be incredibly useful. Running MiMiC QM/MM at the exascale might allow us to produce nanosecond timescales at the B3LYP level for large biological targets, which would be a huge breakthrough in CADD. With the current configuration, however, it would require 2.7 Mcore-h/ps worth of computer time to run at the B3LYP level. This is not feasible with the computer time at our disposal on pre-exascale machines.

Using the cheaper BLYP functional, MiMiC-QM/MM MD simulations scale efficiently up to a more modest of 5,184 cores. However, we achieve a better performance of 5.4 ps/day. Running at this configuration would require around only 0.02 Mcore-hours/ps, with which relatively long-timescale simulations are accessible using current resources. We use the BLYP functional to investigate the catalysis of wt-IDH1.

The reason why B3LYP scales much better than BLYP must be mentioned:

- CPMD is specifically optimized for strong scaling, with techniques as discussed in Section 3.3.1. Furthermore, MiMiC itself adds minimum overhead to the overall QM/MM simulation.

- B3LYP is computationally heavier than BLYP. This is mainly due to the many more two-electrons integrals involved in the evaluation (see Equation 2.56), requiring a higher number of expensive FFTs. As $t_{B3LYP}(1) > t_{BLYP}(1)$, the speedup as given by Equation 3.2, can go much higher for B3LYP than BLYP.

6.2.3.1. Debugging MiMiC for Better Scaling

This was the first time that the scaling of MiMiC was tested beyond ~13,000 cores. A significant, hitherto undiscovered design flaw was found during this process. Moving beyond 294 nodes (=1764 cores) would lead to crashing of the simulations with an error from the MiMiC subroutines involved in calculating the mixed electrostatic interactions. Specifically, this would occur when attempting to increase the total number of MPI tasks beyond 2214. Discussion with the rest of the MiMiC development team (see list of authors in [A]), identified the problem as the following. Recalling from Section 3.3, MiMiC calculates the electrostatic interactions between the QM plane waves and the MM point charges by splitting the list of MM atoms into subsets, and distributing them across different MPI tasks (see Figure b). If the number of MPI tasks exceed the number of MM atoms being considered, some tasks will not get any MM atoms. These tasks would be idle, and would lead to load imbalances. Hence, it was reasonably assumed that this would negatively impact strong scaling, and should be avoided by the user. A prevention mechanism, to stop the user for making such decisions, was hard-coded into MiMiC. Specifically, the subroutines to compute the QM-MM electrostatic interaction contained the following preamble (in Fortran90):

```
1 if (at_st > at_end) then
2     call handle_error(SEVERITY_FATAL, &
3                       TYPE_INCORRECT_ARG, &
4                       "at_st should be smaller than at_end", &
5                       __FILE__, __LINE__)
6 endif
```

Listing 6.2 The preamble to the QM-MM electrostatic interaction subroutine with a mechanism to prevent idle tasks.

The variable `at_start` and `at_end` in line 1 of Listing 6.2, passed to the subroutine, are the start and end atom indices of the subset of MM atoms assigned to that particular MPI task. With this `if` condition, the error in line 4 would be triggered even if the number of MM atoms are assigned as 0. This is because, the atom-

sorting algorithm would always assigned a positive value to `at_start`, but assign 0 to `at_end=0` when no atoms are involved. In the wt-IDH1 system, 294 nodes is the inflection point where the number of tasks exceed the number of MM atoms, and this produced the error message. On a second look at the problem, it was realized that the computation of short-range and long-range QM/MM electrostatic interactions in MiMiC is not the bottleneck of the calculation. Having idle processes while computing the more expensive DFT subroutines, should not affect the scaling greatly. The latter is then mainly dictated by the QM calculation, which is being artificially prevented from scaling further by the approach in Listing 6.2. The QM sub-component can be run on more resources and gain in performance, even while the QM/MM electrostatic calculations are not efficiently distributed in terms of computational resources (since the total allocated amount is not used). The code was then fixed to allow for processes to sit idle with no MM atoms, but still not crash the program completely (by D. Mandelli of the MiMiC development team):

```
1 if (natoms < 0) then
2     call handle_error(SEVERITY_FATAL, &
3                       TYPE_INCORRECT_ARG, &
4                       "natoms should be non negative", &
5                       __FILE__, __LINE__)
6 endif
```

Listing 6.3 The edited preamble to the QM-MM electrostatic interaction subroutine to allow idle tasks for better scaling.

Instead of receiving `at_end`, the number of MM atoms assigned to the each MPI process is now passed to the subroutine as `natoms` as in Listing 6.3. This makes it easier to check for the case where only a negative number of MM atoms are received, while still silently allowing for idle processes. Furthermore, an explicit `if` condition to check for 0 MM atoms was added after Listing 6.3 in the code:

```
1 if (natoms == 0) then
2     call timer_stop
3     return
4 endif
```

Listing 6.4 An explicit check for 0 MM atoms in the QM-MM electrostatic interaction subroutine.

If there are indeed no MM atoms assigned to the process, the clock measuring the CPU wall time for debugging is stopped and the function is made to gracefully return. Occurrences of `at_end` are all replaced with `at_start+natoms-1` throughout the code. This lead to a significant improvement in the MiMiC code, that was previously

limiting the scaling.

6.2.4. Free Energy Barrier of the Normal Reaction

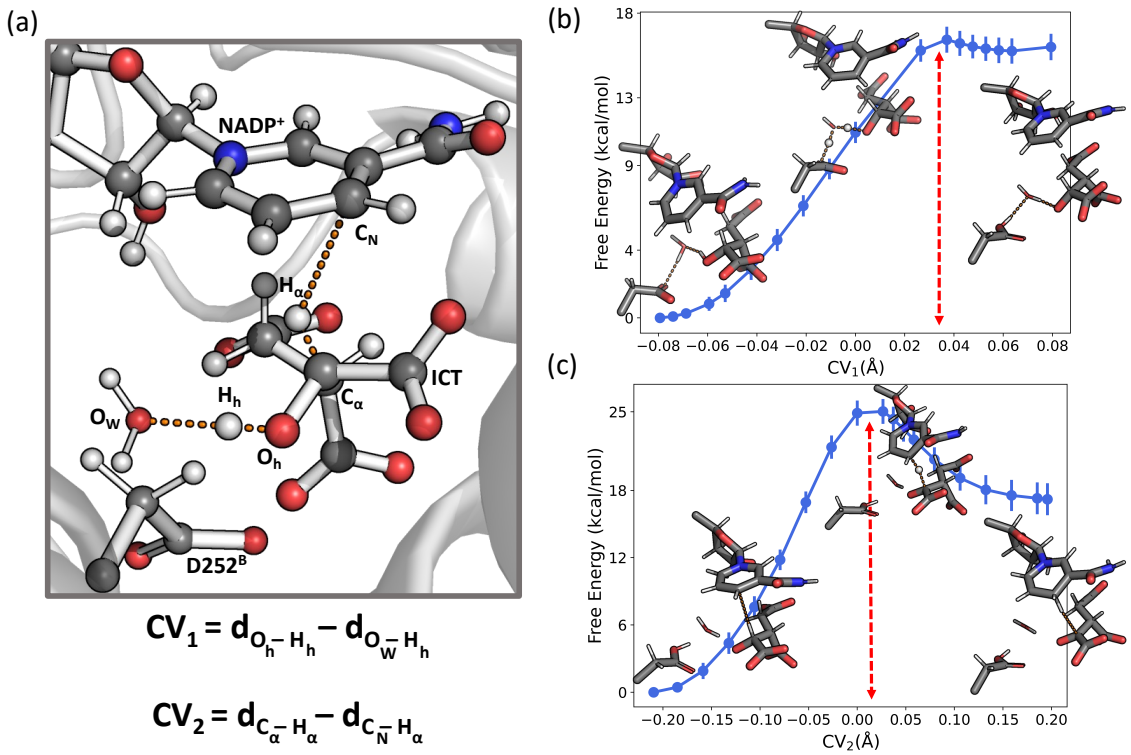


Figure 6.4. (a) Definition of the collective variables $CV_{1,2}$ used for thermodynamic integration. (b) Free energy of the ICT to OXS conversion with respect to CV_1. Adapted from Ref. [B].

Thermodynamic integration (TI) [220, 221] at the BLYP level is used to calculate the free energy associated with the first sub-step of first step of the normal reaction (conversion of ICT to OXS mediated by $Asp252^B$), followed by the second sub-step (reduction of the $NADP^+$ ring) as described in Section 4.3.1 and 4.4. The free energy is described using a collective variable (CV) that completely describes the progress of the reaction; CV_1 for the first substep and CV_2 for the second. The calculation was done in three steps:

1. TI is used to study the conversion of ICT to OXS mediated by the $Asp252^B$-water pair. 18 QM/MM MD simulations (677 fs each) constraining CV_1 are performed. CV_1 is defined as the difference $d_{O_h-H_h} - d_{H_h-O_W}$ between the distances of the proton from the two relevant oxygen atoms (see figure 6.4a).

2. This is followed by an unconstrainted QM/MM MD of 677 fs to equilibriate the

	Lys212^{B}	Asp279	Asp252^{B}
Deprotonation of ICT	1.5	12.2	16.6 (± 0.7)
Reduction of $NADP^{+}$	13.4	21.4	24.0 (± 1.6)

Table 6.2. Free energies (in kcal/mol) associated with the first step of the wt-IDH1 catalysis, for various base residues as initiators of the reaction. The Helmholtz free energy for Asp252^{B} as base is from this work, while the Gibbs free energies for the pathways with Lys212^{B} and Asp279 as base are from Ref. [203].

product from step 1.

3. Finally, TI is used to study the reduction of $NADP^{+}$ to NADPH. 19 QM/MM MD simulations (each of 1.4 ps) constraining CV_2 are performed. CV_2 is defined as the difference $d_{C_\alpha - H_\alpha} - d_{H_\alpha - C_N}$ (see figure 6.4a).

A cumulative 39 ps of MD were performed, which were obtained in the span of one week. Statistical errors of the free energies were computed from 24 (in step 1) and 54 (in step 3) independent estimates of the potential of mean force obtained dividing the simulation into 50 chunks after discarding the first 95 fs. The free energy profile, along with representative starting, transition state and final configurations, are show in Figure 6.4b and c. This gives the overall free energy barriers obtained for the two steps as ≈ 16 and ≈ 24 kcal/mol, respectively (Table 6.2). These values are fairly similar to those of the Asp279 pathway. Furthermore, both are significantly higher than the barrier the Lys212^{B} pathway. This indicates that deprotonated Lys212^{B} is most likely base for the reaction, and supports the conclusion of Ref. [203].

Further visually inspection of the nature of the bond breaking and formation during the reaction can be carried out. This can be done with Wannier center analysis (discussed in Appendix B.1).

6.3. Conclusions

The chapter presented the study of the therapeutically relevant, metal-containing wt-IDH1 catalysis at the QM/MM level with MiMiC. The main goal of this was to confirm the applicability of the first three steps of the QHPC–VS protocol of Figure 5.3. (i) An initial cMD simulation of wt-IDH1 was performed to understand the Michaelis complex, and identified residues that would play the most important role in the catalysis. A likely pathway was selected, and used as the starting sturcture for QM/MM MD. (ii) The MiMiCPy package, developed in Chapter 3, greatly simplified the selection of the QM atoms and MiMiC-input preparation for this very complex biological systems. (iii) Finally, the free energy barrier associated with the selected

pathway was successfully calculated with MiMiC-QM/MM MD. The energetic was compared with previous literature, and was confirmed to be in line with the current understanding of the mechanism. This supports the claim that MiMiC-QM/MM is capable of accurately capturing the transition state dynamics of the biological targets. This solves problem 2 mentioned in Section 1.1.2, opening the door to perform accurate drug design against metalloproteins, and design transition state analogs (as envisioned in Section 1.1.1).

The scaling of wt-IDH1 with BLYP and B3lYP functionals was also studied. Here the incredible scaling of MiMiC was confirmed up to thousands of CPU cores, exceeding previous results and achieving several ps/day in a single QM/MM MD run. In particular, the extreme scalability at the B3LYP level indicates viability for accurate description of enzymatic reactions when large computational resources are provided. Besides highlighting the efficient use of computational resources by the chosen QM layer (CPMD), these performances further demonstrate the effectiveness of a loose-coupling, multiple-program multiple-data paradigm for the development of extremely scalable first principle QM/MM interfaces. More importantly, the scaling efficiency did not drop below 70%, even when utilizing the complete pre-exascale JUWELS supercomputer. Access to more cores could allow us to greatly increase the performance (in terms of ps/day) of the code. This points to the strong impact that MiMiC QM/MM can have in biochemistry by profiting from upcoming exascale supercomputers.

In Chapter 7, we demonstrate how MiMiC can be used to simulate the 'non-simulable' mut-IDH1 enzyme mentioned as problem 3 in Section 1.1.2. This allows for proposing radiotracer candidates for glioma detection using the 'undruggable' mut-IDH1 active site (as per problem 4 in Section 1.1.2) in Chapter 8.

7. QM/MM MD Simulations of Mutant IDH1

In the previous two chapters, we have used the combination of MiMiC and MiMiCPy within the QHPC–VS protocol (as presented in Figure 5.3) to simulate wt-IDH1 at the QM/MM level. Mut-IDH1 poses a larger challenges for MD simulations and as it is a target not simulatable with the current QHPC-VS protocol (Section 7.1). This is an example of the problem 3 as discussed in Section 1.1.2. In this chapter, I bring in significant improvements to the QHPC–VS protocol to simulate mut-IDH1. This will bring us one step closer to performing drug design in Chapter 8, with this being the main deliverable of the entire thesis.

7.1. Mut-IDH1 not Simulatable with Current Protocol

To design selective radiotracers for mut-IDH1 PET imaging, the most important step is to perform MD simulations of this metal-containing protein. As it has been discussed throughout the thesis, classical force fields cannot provide a good description of metal coordination (especially in biological systems). In wt-IDH1 (Chapter 6), the specific conditions of the active site and/or the nature of the crystal structure allowed MM force fields to describe the system. Unfortunately, this is absolutely not the case in mut-IDH1, and cMD simulations completely fail (see Appendix B.3). This is due to the peculiarity of the system and the crystal structure.

The active site of the catalytically active mut-IDH1 (similar to wt-IDH1) contains αKG, NADPH and an Mg^{2+} ion in the catalytically active state. The αKG molecule is a non-standard small molecule, where the charge distribution and partial charges required to simulate its behavior are not part of standard MM force fields. This makes simulating its interactions with the rest of the protein and NADPH difficult. Furthermore, it is known that the αKG–Mg^{2+} coordination in the active site is much weaker that the corresponding ICT–Mg^{2+} case. [195] This indicates significant charge redistribution which is very difficult to capture without quantum effects. To crystallize the αKG–protein complex, most studies inactivated the enzyme by replacing the Mg^{2+} with a Ca^{2+} ion.[178, 204, 222, 223] This includes the best known starting structure (PDB ID: 3INM[178]), used in this work. The Ca^{2+} ion induces a heptacoordination

sphere, with seven ligands crystallized around the metal center (Figure 4.5a and B.3a). As Mg^{2+} is always hexacoordinated, the catalytically active site would only have six coordinating ligands around the metal. One of the seven ligands around the metal center in the crystal structure should be absent in the 'true' coordination sphere. It is not known *a priori* which ligand should leave, and this cannot be predicted by cMD. This can only be obtained by minimizing the active site at the quantum level with MiMiC-QM/MM. However, QM/MM MD requires a cMD equilibrated structure as depicted in the QHPC–VS protocol of Figure 5.3. Thus, this circular requirement renders this mut-IDH1 not simulatable with the QHPC–VS protocol (as described in problem 3 of Section 1.1.2).

As discussed in Section 2.1.2, accurate MD structures (whether at the classical or QM/MM level) are required to perform accurate CADD. The non-simulatability of mut-IDH1 greatly hampers drug design efforts and is a major reason why the active site of the protein is undruggable. A major goal of this thesis is to solve the non-simulatability of this target and unlock druggability.

7.2. Editing the Protocol for Mut-IDH1

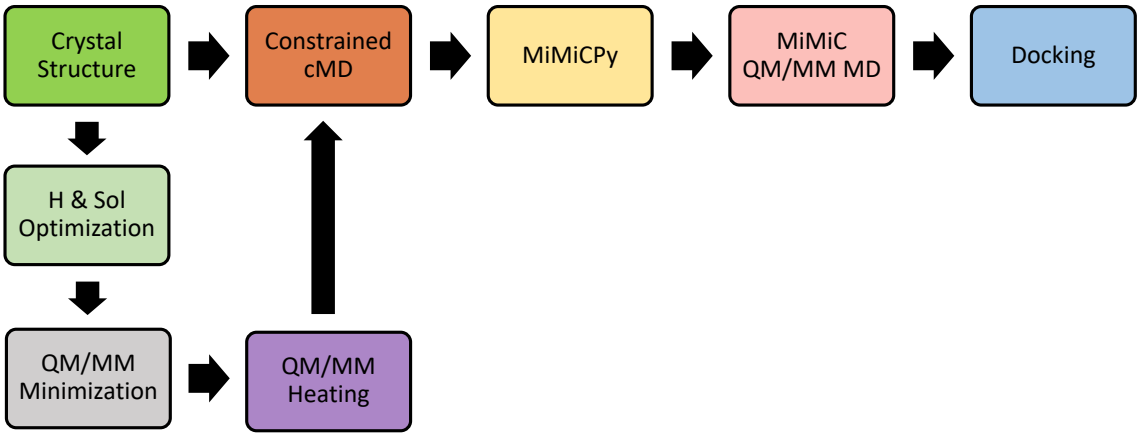


Figure 7.1. Updated QHPC–VS protocol to simulate non-simulatable biological targets like mut-IDH1.

To solve the circular requirement preventing the usage of the QHPC-VS protocol for mut-IDH1, three additional preparatory steps are added before running cMD (see Figure 7.1). These are mainly to minimize the coordination sphere of the metal ion at the QM/MM. A description is given in Sections 7.2.1–7.2.3.

7.2.1. Hydrogen and Solvent Optimization

Before any minimization at the QM/MM level is attempted, the positions of the hydrogens and solvent added to the crystal structure of the protein must be optimized at the MM level. To first optimize the former, a steepest descent minimization at the MM level is performed. To optimize the latter, the system (including all added hydrogens, waters and ions) are heated to 300 K from the current 0 K (again at the MM level). This imparts kinetic energy to the solvent and randomizes their positions. Both these steps are performed with the heavy atoms in the protein fixed in place, using the following conditions:

- Positions restraints on all heavy atoms, especially the heavy atoms involved in Mg^{2+} coordination.
- Restrain oxygen of the two crystal structure waters coordinating with Mg^{2+}.

7.2.2. QM/MM Minimization

With the solvent and hydrogens equilibriated at 300 K, minimization of the Mg^{2+} coordination sphere can be attempted. Minimization in MiMiC-QM/MM implies running dynamics at near 0 K. Since mut-IDH1 is a dimer with two Mg^{2+}-containing active sites, two subsequent QM/MM steps are performed, with the active site of two subunits separately treated at the QM level. During this process, the positions of all atoms in the active site not in the QM region are constrained, so to not undergo rearrangement under the MM force field. The CPMDid command from MiMiCPy is used to obtain the CPMD indices of the atoms of the active site in the MM region.

To minimize the active site of subunit A, annealing is carried out by progressively decreasing the temperature from 300 K to around 0 K. The system is then run for a significant amount of time near 0 K, until equilibration of the Mg^{2+} coordination is observed. Subsequently, we switch the QM region to subunit B and run an NVE simulation, where the temperature is expected to rise slowly from 0 K. The system is then annealed back to 0 K, where dynamics is run to obtain the minimized structure.

7.2.3. QM/MM Heating

Having obtained minimized structures at the QM/MM level at 0 K, we would need to reheat the system to 300 K and equilibrate the system by inducing dynamics with the NVT ensemble. Usually this is done at the MM level. But in this case, we have a structure minimized at the QM/MM level, and moving to the MM surface directly

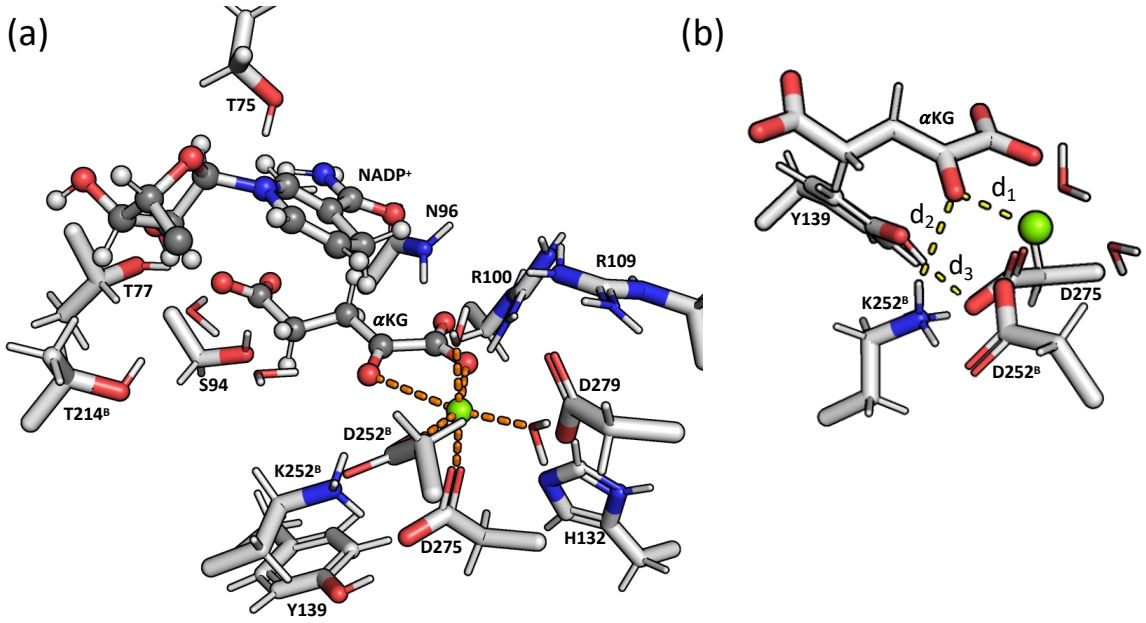


Figure 7.2. (a) Schematic of the Michaelis complex of the mut-IDH1 active site in the KH/D(B) configuration as obtained from QM/MM MD simulations. (b) Representation of the distances in the mut-IDH1 active site measured during QM/MM MD.

will cause instabilities. So, heating to 300 K is performed at the QM/MM level. As in the previous step, two subsequent QM/MM heating steps are performed, in which only the active site of one subunit at a time is treated at the QM level, while the other is constrained and treated at the MM level. Specifically, using the restart file (with coordinates, wavefunctions and velocities) from the previous step, the system in heated to 300 K with the active site of subunit B as the QM region. Then, the QM region is switched to active site of subunit A. The system is annealed from 300 K to 0 K, and heated back to 300 K. This results in a protein structure where both active sites have been heated to 300 K at the QM level.

7.2.4. Further Modifications

Subsequent steps can now be performed as usual. 5 ns of NVT, 5 ns of NPT (both with positions restraints), and a final production cMD is run for about 500 ns to get a MM equilibirated solvate mut-IDH1 complex. However, even with the active site minimized, the interactions of αKG exhibited in the crystal structure are not conserved during cMD. Appropriate constrains are applied throughout the cMD to maintain the same interaction network of αKG as that described in Section 4.4. This included the Mg^{2+} coordination sphere, where the interaction of Mg^{2+} with the C_α ketonic oxygen of αKG was not maintained by cMD (even after correcting the

coordination sphere using the procedure of Sections 7.2.1–7.2.3).

To study the multiple protonation states relevant to the catalysis as discussed in Section 4.4.1, the procedure to obtain a production QM/MM trajectory of mut-IDH1 (as discussed so far) is performed for both (i) protonated $Lys212^B$/deprotonated Asp275 (**KH/D**), and (ii) deprotonated $Lys212^B$/protonated Asp275 (**K/DH**). The final step of QM/MM MD with MiMiC is performed as two subsequent substeps of 20 ps each, with the active site of two subunits separately treated at the QM level. The active site not in the QM region is constrained. Performing this for each of the two protonation states assigned for each active site of the dimeric mut-IDH1, results in 4 QM/MM MD trajectories: **KH/D(A)**, **KH/D(B)**, **K/DH(A)**, **K/DH(B)**. During the simulation of KH/D(A), $Lys212^B$ and Asp275 were well-positions in multiple frames for the deprotonation of $Lys212^B$ to takes place. This was forced with a quick constrained simulation; a subsequent unconstrained equilibrium simulation of the active site with deprotonated $Lys212^B$ and protonated Asp275 was carried out. This is referred to as: **K/DH(A*)**. All 5 QM/MM MD trajectories were simulated for 20 ps, resulting in a total of 100 ps worth of data.

7.3. QM/MM Dynamics of αKG in Mut-IDH1

Throughout the 20 ps simulations time for each of the configurations, the interaction network of αKG with the protein is largely maintained as that of the crystal structure. Only that of **KH/D(B)** is different (Figure 7.2a). Asn96 has moved away from the γ-carboxylate of α, and towards α-carboxylate. This consequently has pushed away both Arg100 and 109 from the same α-carboxylate. This probably provides more flexibility to the α-carboxylate and β-ketone groups, allowing the later to interact more strongly with $Lys212^B$ (see discussion below).

The stability of the Mg^{2+} coordination sphere during the QM/MM MD, for all 5 configurations, is shown in Figure B.2 in the appendix. Most interactions stay within the 2 Å range. However, the interaction of Mg^{2+} with the C_α ketonic oxygen of αKG (see distance d_1 in Figure 7.2b) varies widely across the different configurations (Figure 7.3b). The average Mg^{2+}–αKG ketonic oxygen distance follows the pattern: KH/D(B) (2.6 Å) > KH/D(A) (2.4 Å) > K/DH(A*) (2.3 Å) ~ K/DH(A) (2.3 Å) > K/DH(B) (2.2 Å). Magnesium is known to have the a tight coordination sphere closest to the ideal octahedral geometry, with a typical Mg–O distance of 2.1 Å.[224] The severe deviations of the Mg^{2+}–αKG ketonic oxygen distance from this value seems to indicate that the coordination between Mg^{2+} and αKG is weak.

It is clear from the previous discussion that the configurations with deprotonated $Lys212^B$ exhibit a stronger Mg^{2+}–αKG ketone coordination, as opposed to those with protonated $Lys212^B$. The extra proton on the N of $Lys212^B$ forms a H-bond with C_α

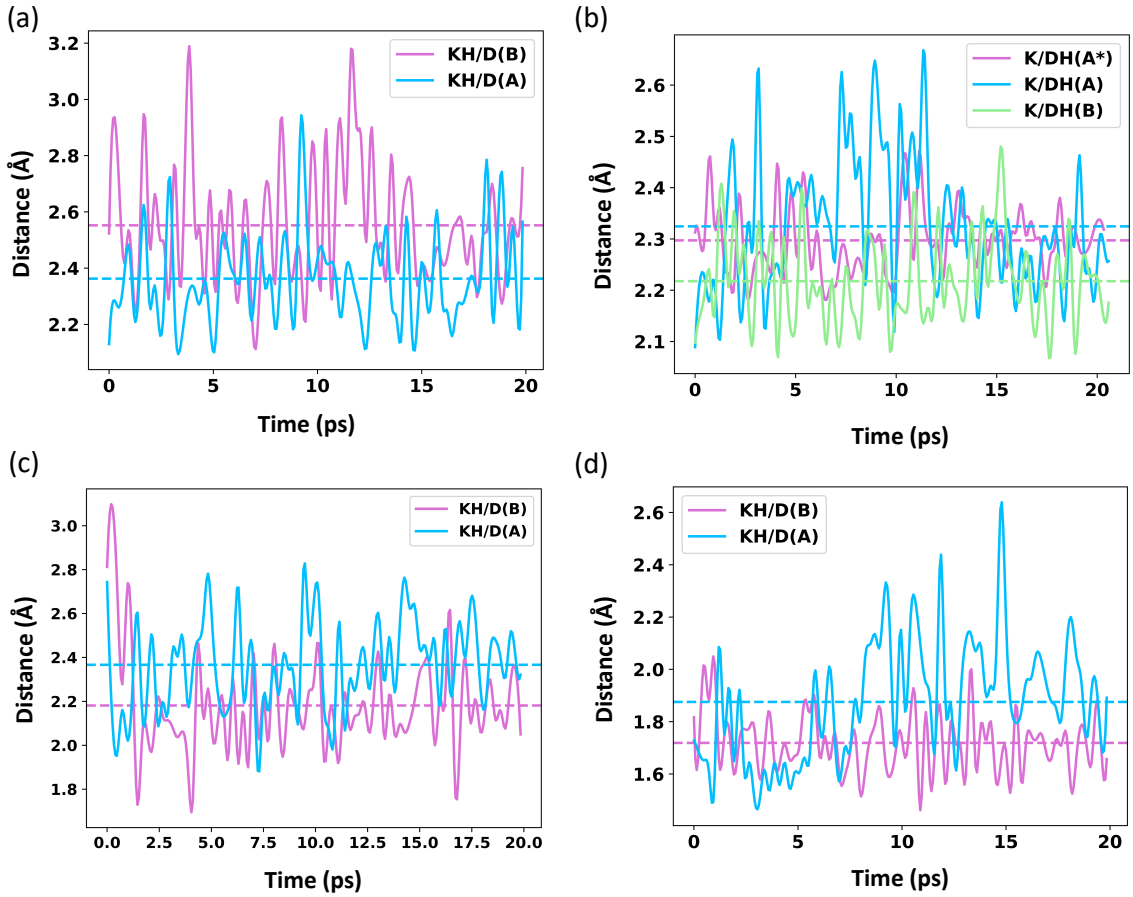


Figure 7.3. Plots of distances (a) d_1, (b) d_2 and (c) d_3 with respect to time for various configurations of the mut-IDH1 Michaelis complex during QM/MM MD. Distances labelled according to Figure 7.2b.

ketonic oxygen of αKG, and interferes with its coordination to Mg^{2+} (see distance d_2 in Figure 7.2b). In fact, the longer Mg^{2+}–αKG ketonic oxygen distance in KH/D(B), as compared to KH/D(A), can be explained by the closer interaction $Lys212^B$ establishes with the αKG ketone (Figure 7.3c). These configurations demonstrate the strong possibility of $Lys212^B$ acting as the acidic proton donor to initiate the reaction.

Indeed, it was noted in the beginning of this section that the different interactions of Asn96, Arg100 and Arg109 with αKG allows for more flexibility of the latter in KH/D(B). To further investigate why $Lys212^B$ forms a stronger interaction αKG ketone in KH/D(B) than KH/D(A), we note the primary impediment to a strong H-bond between $Lys212^B$ and αKG is the competing H-bond between the proton of $Lys212^B$ and O of Asp275. This O of Asp275 also forms a H-bond with the alcohol of Tyr139 (see distance d_3 in Figure 7.2b). This Tyr139–Asp275 distance is closer in KH/D(B) than KH/D(A) (Figure Figure 7.3d). A stronger Tyr139–Asp275 interaction then, results in a weaker interaction with $Lys212^B$ and pushes it away. This allows it to then form a stronger H-bond with αKG ketone. Tyr, although not the primary proton donor, discourages the $Lys212^B$–Asp275 interaction and promotes $Lys212^B$ to be in the protonated state. Thus, Tyr is important in supporting the catalysis.

Clearly, the K/DH configurations represent a more stable binding of Mg^{2+}–αKG, compared to the KH/D states. Simultaneously, the catalysis can likely only take place in the presence of a protonated $Lys212^B$. Thus, these two configurations might represent different stages of the αKG–IDH1 Michaelis complex. The K/DH states might represent a more 'early' Michaelis complex, closer to the substrate binding event, and the KH/D states might represent a 'later' Michaelis complex closer to the transition state.

7.4. Conclusions

This chapter presents QM/MM MD simulations of mut-IDH1, which is an example of a non-simulatable biological target. For this, the QHPC–VS protocol was significantly updated to a new version as shown in Figure 7.1. This updated protocol presents a solution to problem 3 of Section 1.1.2. This allowed us to obtain a total of 100 ps of QM/MM equilibrium dynamics of the mut-IDH1, and sample various configurations of the protein. This included investigating pathways with both the protonated $Lys212^B$/deprotonated Asp275 pair, and the deprotonated $Lys212^B$/protonated Asp275 pair. Through these simulations, we provide the following insights into the mut-IDH1 catalysis:

- Although the Mg^{2+}–αKG ketonic oxygen distance varies with time, on average the distance across all configurations investigated was greater than the typical

Mg–O distance of 2.1 Å. This indicates that the binding of Mg^{2+} with αKG is weak. This corroborates the findings of Lie et. al., where they argued this weakness is the primary mechanism of inhibition selectivity in current allosteric mut-IDH1 inhibitors.[195, 31]

- Protonated $Lys212^{B}$ was found to be the only residue well-positioned to act as the acid initiator of the reaction. No acid candidate was found when $Lys212^{B}$ was deprotonated. This indicates that $Lys212^{B}$ must be protonated at the Michaelis complex for the reaction to take place.

- Tyr139 H-bonding with Asp275 was found to decrease the tendency of $Lys212^{B}$ deprotonation by Asp275. This promotes $Lys212^{B}$ to exist in the protonated state, and donate this extra proton to αKG instead. Thus Tyr139, though not directly the proton donor, plays an important role in promoting the catalysis. This corroborates the finds of Rendina et al., where mutating Tyr139 in mut-IDH1 was found to reduce the rate of, but not totally eliminate, the catalysis.[204]

- The binding Mg^{2+} to αKG is stronger when $Lys212^{B}$ is deprotonated, compared to when it is protonated. This, coupled with the fact that only protonated $Lys212^{B}$ can initiate the reaction, indicates that this residue would be deprotonated closer to binding event of Mg^{2+}/αKG to the mut-IDH1 active site, or at the 'early' Michaelis complex. Closer to the catalytically active (or later) Michaelis complex then, $Lys212^{B}$ would be protonated to set-up the catalysis.

In summary, multiple conformations of mut-IDH1 have been explored, across the spectrum from the ground state to close to the transition state. This large spectrum gives us enough sampling to perform drug design for the mut-IDH1 active site. This is discussed in Chapter 8.

8. Radiotracer Design for Glioma Diagnosis

In the previous chapter, we modified the QHPC–VS protocol (as presented in Figure 7.1) to simulate mut-IDH1. This produces a spectrum of configurations of the mut-IDH1 Michaelis complex. This gives enough information to perform drug design. In this chapter, a molecular docking procedure to obtain mut-IDH1 selective PET radiotracer candidates for the non-invasive diagnosis of mut-IDH1 glioma is performed. This demonstrates the usefulness of MiMiC in CADD, as discussed in point 4 in Section 1.1.2.

The conditions for a PET radiotracer precursor for mut-IDH1 detection were discussed in points 1–4 of Section 4.2. Specifically, satisfying the condition under point 4 (the precursor should bind selectively only to mut-IDH1 and not to wt-IDH1) is the primary focus of the docking efforts of this chapter. Furthermore, point 3 (the precursor should cross the BBB so that it reaches mut-IDH1 within glial cells in the patient's brain) is also noteworthy. This is difficult to ensure in a computational set-up, and needs experimental validation. Nevertheless, to increase the probability of obtaining hits that cross the BBB, docking starting from a database of molecules known to cross the BBB would be beneficial. For this, the molecules marked as BBB+ (or those that cross the BBB) from the recently published **B3DB** database[225] are collected. This consists of 4956 molecules. Molecular docking was carried out with this BBB+ subset of the B3DB database using the procedure described in Section 8.1.

8.1. Steps for Molecular Docking

The B3DB database needs to be sieved to ensure that the resulting hits are selective only for the mut-IDH1 protein. This implies selecting for the following conditions:

- As explained in Section 4.2.1, the main point of differentiation between the mutant and wild-type isoforms is in the active site. Thus, it must be ensured that the small molecule has high binding affinity for the mut-IDH1 active site. Specifically, it must have higher binding affinity for the active site than αKG

itself. This way, the molecule can push out the natural substrate of mut-IDH1.

- To produce selectivity, this molecule must have low binding affinity to the wt-IDH1 active site. Specifically, it must have lower binding affinity to the wt-IDH1 active site than the natural substrate ICT.

- Even if this molecule is selective for the mutant active site over the wild type, it can still posses a high affinity for the allosteric IDH1 dimer interface (similar to the current the mut-IDH1 inhibitors). This would negate any selectivity. Thus, it must be ensured that this molecule has low affinity for the IDH1 dimer interface.

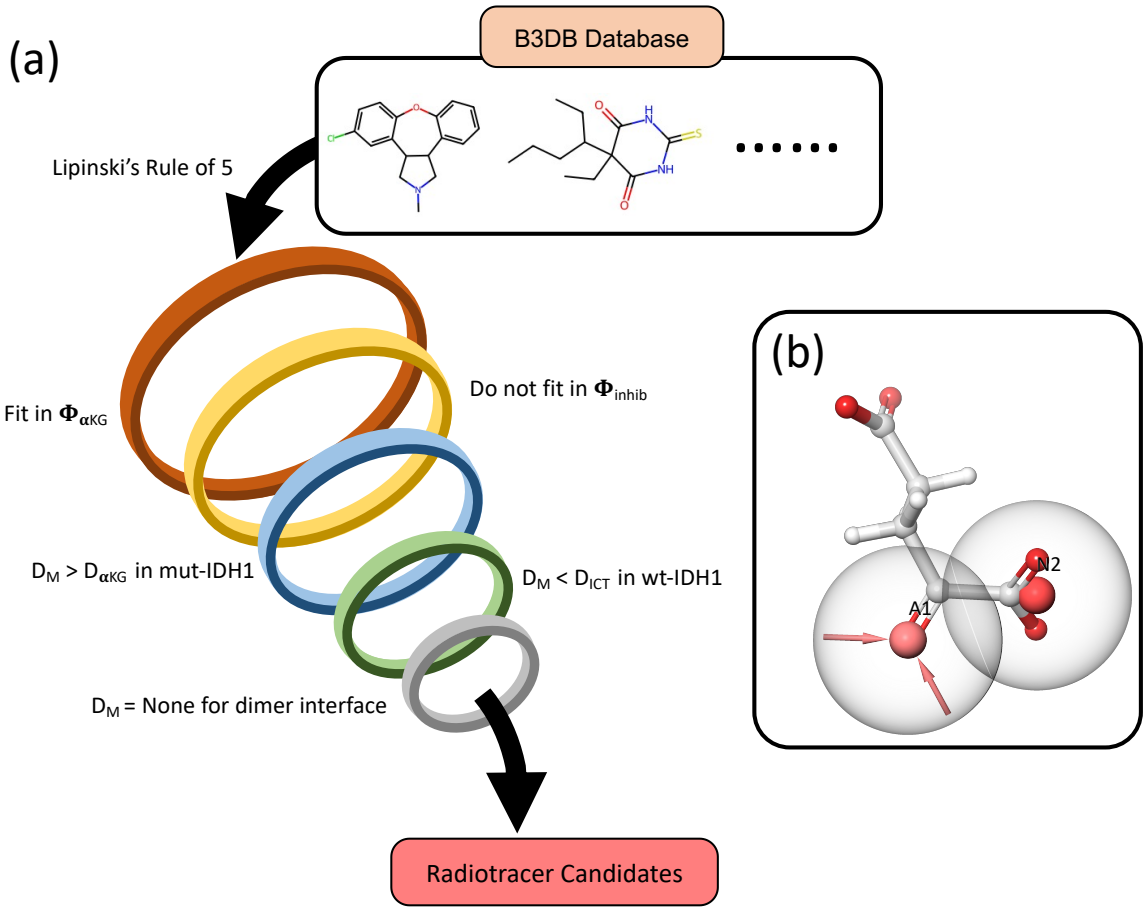


Figure 8.1. Five step molecular docking protocol proposed in this work to drug the undruggable mut-IDH1 active site, and result in mut-IDH1 selective radiotracer candidates.

The above three conditions are satisfied using a 5 step filtration process as shown in Figure 8.1a. The Schrödinger software package version 2022 was used to performing all of these steps, mainly involving Phase[226], and Glide (see Section 2.1.1 on more details on Glide). The B3DB BBB+ database is pre-filter with Lipinski's rule of

five the step to find drug-like molecules. The cutoff empirical parameters used are: molecular weight ≤ 500, hydrogen bond donors ≤ 5, hydrogen bond acceptors ≤ 10, and ClogP ≤ 5.[227]

1. Small molecules are first sieved based on their ability to binding with the Mg^{2+} in the binding site. For this, a 3D pharmacophore (see Section 2.1) of the chemical features of αKG that interact with the Mg^{2+} is developed (referred to as $\Phi_{\alpha KG}$ in Figure 8.1). This is shown in Figure 8.1b. Only small molecules that fit within this pharmacophore are selected.

2. Small molecules are sieved based on their dissimilarity with current inhibitors, which bind at the dimer interface. For this, a 3D pharmacophore of mut-IDH1 inhibitors at the dimer interface using the ensemble of inhibitor-IDH1 crystal structures is developed (referred to as Φ_{inhib} in Figure 8.1). Small molecules that do not fit in this pharmacophore are selected for further analysis.

3. With this smaller sub-set of molecules, molecular docking to the mut-IDH1 active site can be performed. This consists of the following sub-steps:

   a) A Glide XP docking of αKG to the ensemble of mut-IDH1 active sites from the different QM/MM configurations (from Chapter 7) is performed. The docking score is noted as $D^{mut}_{\alpha KG}$.

   b) A Glide XP docking of the small molecules to the ensemble of mut-IDH1 active sites is performed. Glide HTVS is skipped for the B3DB database, given its small size. The XP docking score of each molecule M is noted as D^{mut}_{M}.

   c) Molecules with docking score $D^{mut}_{M} > D^{mut}_{\alpha KG}$ are selected, the rest are discarded. Alternatively, if $R^{mut}_{M,\alpha KG}$ is the ratio of docking score of M to that of αKG, then $R^{mut}_{M,\alpha KG} > 1$.

   The mut-IDH1 structures used are obtained after 20 ps of QM/MM as per Chapter 7. Specifically, steps a–c are carried out with two ensembles of mut-IDH1 structures consisting of configurations selected from:

   - protonated $Lys212^{B}$/deprotonated Asp275 simulations, i.e., KH/D(A) and KH/D(B). This results in the docking ratio $R^{mut\text{-}KH/D}_{M,\alpha KG}$.
   - deprotonated $Lys212^{B}$/protonated Asp275 simulations, i.e., K/DH(A), K/DH(B), and K/DH(A*). This gives the docking ratio $R^{mut\text{-}K/DH}_{M,\alpha KG}$.

4. Docking to the wt-IDH1 active site is performed. This consists of the following steps:

a) A Glide XP docking of ICT to the wt-IDH1 active sites is performed. The docking score is noted as D^{wt}_{ICT}.

b) Glide XP docking of the small molecules to wt-IDH1 active site is performed. The score of each molecule M is noted as D^{wt}_{M}.

c) Only molecules with docking score $D^{wt}_{M} < D^{wt}_{ICT}$ are selected. Alternatively, if $R^{wt}_{M,ICT}$ is the ratio of docking score of M to that of αKG, then $R^{wt}_{M,ICT} < 1$.

Steps a–c are carried out with wt-IDH1 structures from:

- protonated $Lys212^{B}$/deprotonated Asp275 equilibrium QM/MM MD structure from Chapter 6. This results in the docking ratio $R^{wt\text{-}KH/D}_{M,ICT}$.
- deprotonated $Lys212^{B}$/protonated Asp275 cMD structure from Ref [203]. This gives the docking ratio $R^{wt\text{-}K/DH}_{M,ICT}$.

5. A Glide SP docking to the dimer interface from an ensemble of inhibitor-IDH1 crystal structures is performed for the remaining molecules. Those that cannot be docked to the dimer interface are selected. These are the small molecules reported to be potential precursor for PET radiotracers.

The ensemble of inhibitor-IDH1 crystal structures, used in steps 2 and 5, to obtain the structure of the dimer interface was obtained from crystal determinants with PDB IDs: 4UMX[189], 5DE1[190], 6ADG[191], 6U4J[192], and 5LGE[193].

To further rank the hits from step 5 in their order of biological activity, we should maximize the value of the docking ratios from step 3 and minimize that of step 4. A ranking function can be constructed as:

$$\text{Ranking} = R^{mut\text{-}K/DH}_{M,\alpha KG} + R^{mut\text{-}KH/D}_{M,\alpha KG} - \left(R^{wt\text{-}K/DH}_{M,ICT} + R^{wt\text{-}KH/D}_{M,ICT}\right) \tag{8.1}$$

A molecule with a higher ranking score would be more selective for the mut-IDH1 active site.

8.2. PET Radiotracer Candidates

On passing 4956 BBB+ molecules from the B3DB database through the docking procedure described in Section 8.1 resulted in: 4385 hits after pre-filtering with Lipinski's rule of five, 782 hits in step 1, 742 hits in step 2, 122 in step 3, 122 in step 4 (no

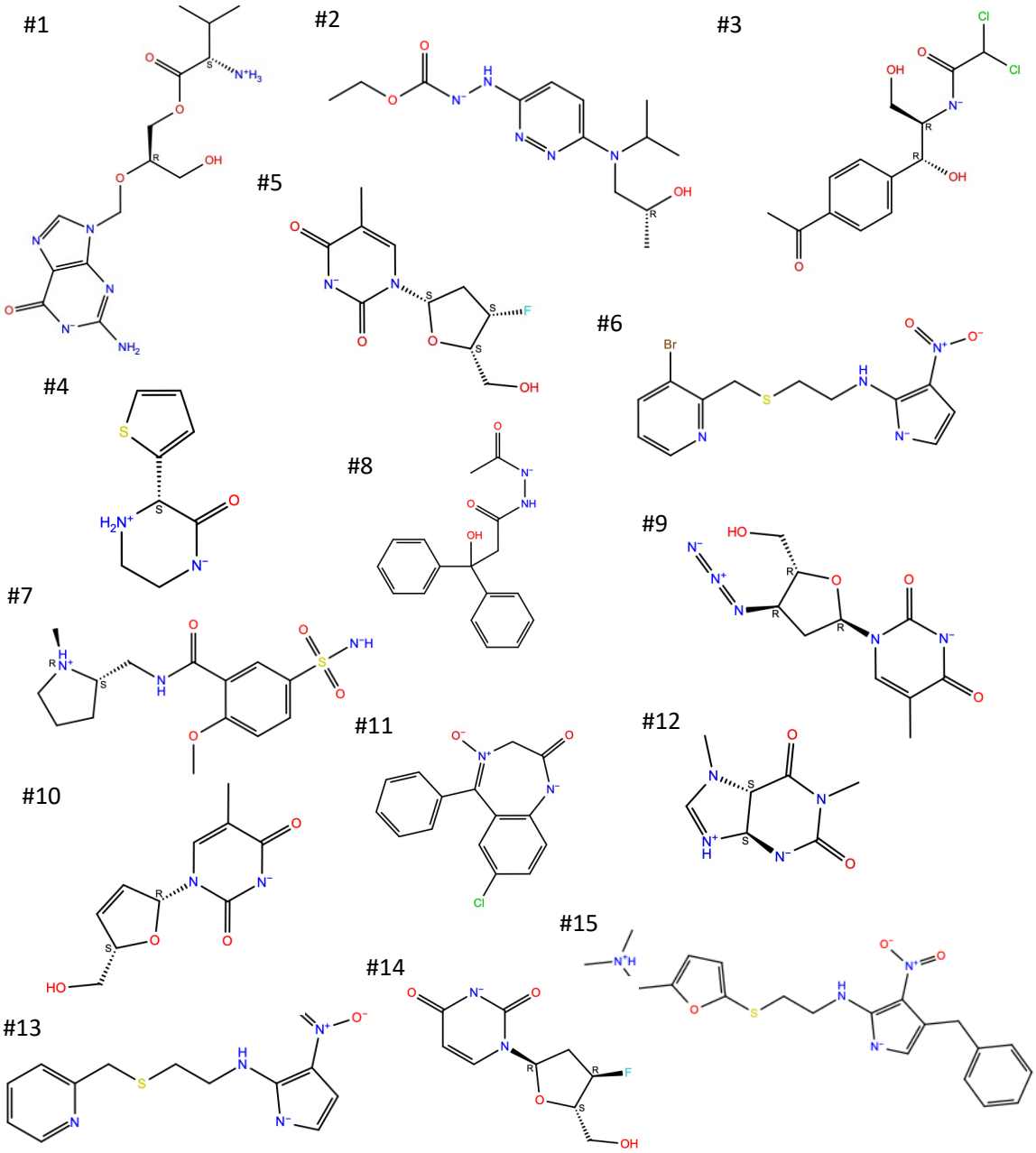


Figure 8.2. The list of hits, or potential radiotracer precursors, obtain from following the procedure in Figure 8.1.

molecules were found to dock better in the wt-IDH1 binding site than ICT), and finally 22 in step 5. These are ranked as per Equation 8.1. The final list contained multiple stereoisomers of the same compounds. Excluding all but the stereoisomers with the highest ranking, resulted in final tally of 15 potential PET radiotracer precursors (Figure 8.2).

Since, a PET radiotracer should posses a radioactive fluorine, precursors already containing fluorine groups will allow for simpler organic synthesis protocols for PET radiotracers. It is interesting to note that, in the group of hits: 2 molecules contain a fluorine (compound no. 5 and 14), 3 molecules contain other halogens (compound no. 3, 6, and 11), and rest contain no halogens.

8.2.1. [^{18}F]-Fluorothymidine

Among the top five scoring hits in Figure 8.2, compound no. 5 is fluorine-containing. This is a known, biologically-active compound called alovudine or fluorothymidine (FLT).[228, 229] FLT labelled with ^{18}F or [^{18}F]-FLT, has already been shown to act as a PET tracer for imaging of malignancy in humans.[230, 231] Our work indicates that [^{18}F]-FLT could also function as a PET tracer for mut-IDH1 glioma.

To demonstrate that FLT, and by extension [^{18}F]-FLT, is indeed a substrate analog of αKG, the 3-D schematic of the docked compound within the mut-IDH1 active site is shown in Figure 8.3. Alovudine is a fluorinated deoxythymidine, with a deoxyribose joined to the thymine (similar to a nucleoside). Alovudine coordinated with Mg^{2+} in a bidentate fashion, through the 3-N and 4-ketonic O of the thymine ring. The 4-ketone also interacts with the Arg109. The thymine ring also possess another ketone at the 2 position, which satisfies H-bond interactions with the protonated Lys212^{B}. This ketone shows no interaction partners when Lys212^{B}is deprotonated. The deoxyribose establishes strong interaction with residues that previously interacted with the γ-carboxlate of αKG, although the network differs significantly between KH/D and K/DH configurations. In KH/D, the 3'-F forms a halogen bond with Thr77, with an additional possible interactions with Thr75 through a water molecule. Ser94 is also well positioned to establish (possibly water-mediated) H-bond with 5'-O. Finally, the O of the 4-hydroxymethly group is well positioned to interact with the NADPH ribose and Thr214^{B}. In K/DH, on the other hand, the 3'-F forms a halogen bond with Ser94, with Thr77 also in the vicinity. The ring containing the 5'-O is rotated to establish an interaction with Asn96, and a possible water-mediated interaction with the NADPH ribose. This rotation also brings the 4-hydroxymethly group in close contact with Thr75.

FLT possess a methyl group in the 5 position of the thymine ring, that pushes away residues like Arg 100 and Glu 306. Thus, a possible route to optimize the binding

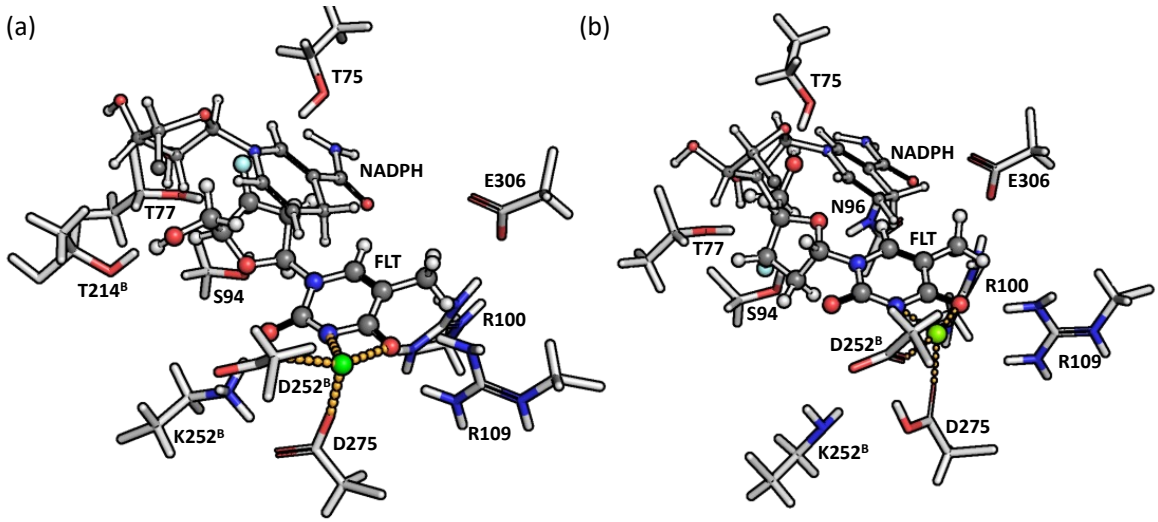


Figure 8.3. Docked structure of FLT in configuration (a) KH/D, and (b) K/DH of mut-IDH1. The fluorine atom is represented by a light blue sphere.

of FLT further, would be to replace this methyl group. In fact, compound no. 14 from Figure 8.2 is structurally similar to FLT, with the only difference being the lack of the methyl group in the 5 position. This allows it to fit better in the mut-IDH1 binding site, with a slightly higher $R_{M,\alpha KG}^{\text{mut-K/DH}}$ and $R_{M,\alpha KG}^{\text{mut-KH/D}}$ scores. But, it also docks well into the wt-IDH1 pocket, with higher $R_{M,ICT}^{\text{wt-K/DH}}$ and $R_{M,ICT}^{\text{wt-KH/D}}$ scores. This brings down the overall ranking of compound no. 14, making it less suitable as a selective radiotracer. Similarly, compound no. 9 and 10 bind better than FLT to the wt-IDH1 active site, reducing their overall ranking.

8.2.2. Clinical Significance of Other Compounds

In addition to [^{18}F]-FLT, the other hits are also promising candidates for experimental testing. Some of these have already been utilized in a clinical setting:

- Compound no. 1 or **valganciclovir** is an important treatment method against cytomegalovirus infection.[232]
- Compound no. 2 or **cadralazine** is a commerically available drug to treat hypertension.[233]
- Compound no. 3 or **cetophenicol** shows antibacterial and immunosuppressive effects.[234] It contains 2 Cl atoms.
- Compound no. 4 or **tenilsetam** show inhibitory effect on the Maillard reaction involved in Alzheimer's disease.[235]

- Compound no. 8 (or **diphoxazide**[236]) and compound no. 11 (or **demoxepam**[237]) show antianxiety or anticonvulsant properties. Demoxepam contains a Cl atom.

- Compound no. 9 (commercially known as **zidovudine**[238]) and 10 (commercially known as **stavudine**[239]) are primary treatments for human immunodeficiency virus. Stavudine is toxic at higher doses, and WHO has recommended its phase-out as a treatment strategy.[240]

- Compound no. 12 is known as **paraxanthine** and the main metabolite of caffine in humans.[241]

No clinical information was found for:

- Compound no. 14 is **2',3'-Dideoxyuridine**. This is a derivative of uridine, one of the 5 standard nucleosides involved in DNA/RNA synthesis.[242]

- Compound no. 6 (also known as **sulmepride**): IUPAC name is N-[2-[(3-bromopyridin-2 yl)methylsulfanyl]ethyl]-3-nitro-1H-pyrrol-2-amine

- Compound no. 7: IUPAC name is 2-methoxy-N-[(1-methylpyrrolidin-2-yl)methyl]-5-sulfamoylbenzamide

- Compound no. 13: IUPAC name is 3-nitro-N-[2-(pyridin-2-ylmethylsulfanyl)ethyl]-1H-pyrrol-2-amin

- Compound no. 15: InChI is InChI=1S/C20H24N4O3S/c1-23(2)14-17-8-9-18(27-17)28-11-10-21-20-19(24(25)26)16(13-22-20)12-15-6-4-3-5-7-15/h3-9,13,21-22H,10-12,14H2,1-2H3

8.3. Conclusions

This chapter, for the first time, presents small molecules that bind selectively to only the mut-IDH1 active site, over the wt-isoform. This is an essential requirement to designing PET radiotracers for the non-invasive diagnosis of glioma. Using QM/MM MD simulations from Chapter 6, a 5 step protocol was designed to sieve a database for mut-IDH1 selective molecules (see Figure 8.1). This resulted in 15 hits. Notably, replacing the QM/MM mut-1DH1 structures with just the crystal structure of mut-IDH1 (PDB ID: 3INM) resulted in 60% of these hits being erroneously classified as non-selective for the mut-IDH1 active site. This underlines the role played by QM/MM MD in this work. The structure of the metal-containing active site is very complex, and only MD simulations at the QM/MM level can accurately capture the different configuration required for drug design. Thus, here me show that QM/MM,

within MiMiC, allows us to unlock druggability for previously undruggable targets. This is the solution to final problem 4 as discussed in Section 1.1.2. By solving all 4 problems mentioned in 1.1.2, we have conclusively demonstrated the power of QM/MM MD within drug design.

Most of the PET radiotracer precursors suggested by the protocol are already biologically active compounds, tested in a clinical setting. This should expedite the experimental testing of the compounds. Excitingly, the radioactively marked version of one there hits (fluorothymidine), are already being used as radiotracer for various other cancers. Our studies suggest that this could be repurposed to selectively image mut-IDH1 glial cells as well. This discovery could provide a large impetus in the development of non-invasive diagnostics for IDH1-based glioma.

9. Conclusions

Classical MD simulations are currently limited to only a subset of biological targets, limiting the druggable space. This greatly increases the capital and time investment required to develop effective disease inhibitors. QM/MM MD simulations opens up the possibility of drugging more difficult targets like RNA, metalloproteins, enyzme transition states and covalent inhibitors. Here, I have demonstrated how high performance computing-based QM/MM molecular dynamics simulations can be successfully incorporated into a computational drug design protocol. This was shown using the case of the hard-to-drug active site of the IDH1 enyzme, an important therapeutic target of glioma. Specifically, this work provides the following two deliverables:

1. The Quantum HPC-based Virtual Screening protocol or the QHPC-VS protocol was developed for simulation of complex targets, where quantum phenomenon dominate. These include biomolecules that cannot be simulated with the standard combination of classical and QM/MM MD. This was developed through Chapters 3-8; a summarized version is shown in Figure 9.1.

2. The QHPC-VS protocol was utilized for the drug design of the mut-IDH1 active site. Previously, no known inhibitors were suggested for this pocket. I obtained a list of substrate analogs of αKG, which could potentially be used in precursor to synthesize PET radiotracers for the non-invasive detection of glioma. This will aid in the safer and more accurate diagnosis of these deadly diseases in patients.

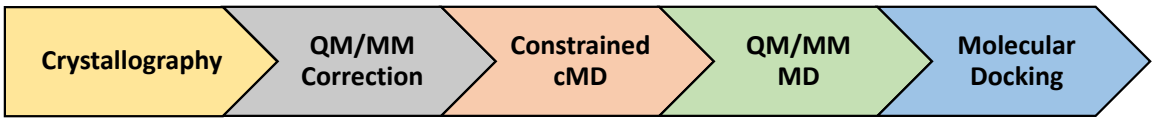


Figure 9.1. An overview of the QHPC-VS protocol developed in this thesis to incorporate quantum simulation within drug design.

In recent times, IDH1 has emerged as an important therapeutic target for the deadly disease of glioma.[26] Both the wild type and mutant isoforms are involved. Between these, mut-IDH1 has emerged as a potential predictive biomarker with great prognostic value, as patients expressing this have better survival rates and further be more sensitive to certain targeted therapies.[27, 28] The development of non-invasive imaging methods for the detection of mutant-associated glioma could greatly

help doctors' diagnosis. Positron emission tomography (PET) technique, involving a selectively-binding radiotracer that illuminates mut-IDH1 within glial cells in the brain, is an ideal candidate.[29, 30] However, this requires designing binders to the mut-IDH1 active site, which has proven difficult so far. This had up till now rendered the PET method of glioma diagnosis infeasible. The QHPC-VS protocol designed in this work lead to various substrate analogs which satisfy the selectivity requirement of a PET radiotracer. Many of these have already been tested in a clinical setting, with data on the adsorption, toxicity and other biologically relevant properties. Most excitingly, one of the top 5 hits was fluorothymidine, the radioactive form of which is already commonly utilized to image multiple cancers with PET.[230, 231] This work suggests that fluorothymidine may also be effective in differentiating between wt- and mut-IDH1 in glial cells.

Although these molecule leads are promising, experimental tests are needed to identify actual hits. This could be the next step of in this work by our experimental collaborator Prof. Bernd Neumaier, involving the synthesize and testing small molecules in the wet lab. If confirmed, it would be a big breakthrough in the field of non-invasive detection of mut-IDH1 associated glioma. PET method for early diagnosis of glioma would become more feasible, increasing the survivability of patients.

By utilizing MiMiC towards a major problem in the non-invasive detection of glioma, I successfully demonstrated the power of DFT QM/MM MD within a drug design pipeline. This work can be placed within the larger context of the digital revolution we are witnessing within the pharmaceutical industry. Spurred on by the advances in artificial intelligence (AI) of the previous decade, multiple venture capital-backed start ups and large pharmaceutical companies like BenovalentAI, IktosAI, Qubit, AstraZeneca and Merck are currently applying neural networks and other computational approaches like cloud and quantum computing to drug design.[243, 244, 245, 246] This is hoped to bring in major improvements to the overall drug design process.

9.1. Increasing QM/MM Performance in Drug Design

The drug design protocol suggested in Figure 9.1 is essentially a modified form of the general method of structure based-virtual screening discussed in Figure 2.2b. MD simulations at the quantum level were were followed by docking still conducted at the classical level. This limits the accuracy of the substrate analogs suggested here. First principle QM/MM MD simulations would be highly beneficial here by refining the binding poses as predicted by current docking methods.[247, 248, 249] However, the performance of the QM/MM method would preferably would need to be improved to perform this. Below, I describe some of the possible methods of increasing the performance of QM/MM MD.

Pushing the performance of MiMiC further would probably involve reaching the exascale regime. Modern supercomputers makes extensive usage of architectures consisting of heterogeneous nodes that combine multicore CPUs with GPUs. [250, 251, 15] This is especially true for exascale supercomputers. Achieving exascale in biochemistry would require coupling MM and QM software that are able to scale on many (≈100–1000) GPUS in addition to CPUs. Most cMD codes already heavily exploit GPUs to reach higher scaling[107, 252, 253, 254], including GROMACS [255] used in MiMiC. However, GPU support for the DFT component is still an ongoing process.[256, 257, 258, 259, 260] TeraChem is a very recent example of a QM software that has support for GPUs.[261, 262] Only very recently has scalability of DFT-based MD been achieved for over thousands of GPUs by exploiting innovative linear scaling approaches and sparse algebra methods within an extended tight-binding scheme [263]. GPU-ready QM/MM codes are also few, with the best examples being the recent TeraChem protocol buffers[148] and the QUICK-Amber interface[149]. This general lack of GPU-ready QM codes (compared to MM codes) will arguably hamper serious endeavours to port first principle QM/MM MD interfaces to exascale system, where GPUs provide the bulk of the computing power. This indicates that to go towards exascale DFT-QM/MM MD, there is a necessity to develop innovative algorithms that augment and go beyond standard MD approaches. These could involve statistical mechanics-based ensemble methods [264], path sampling [265] and path-integral-like approaches [266].

Machine learning (ML) and force matching techniques in the context of molecular dynamics has recently become a fast-emerging field.[267, 268] ML models work natively on GPUs, and as they normally rely on local interactions alone, they can be exceptionally scalable on distributed architectures.[269, 270] We can then expect DFT QM/MM MD simulations to tremendously profit from ML techniques, in the context of GPU utilization on exascale machines. Hybrid ML/MM models have been introduced, that enable the simulation of biological targets with an ML representation of a QM potential at near QM/MM accuracy, at a fraction of the computational cost.[271, 247, 272, 273, 274] Furthermore, ML and QM/MM can benefit from a synergy, as accurate training of ML models requires having in the first place dataset generated through QM/MM MD calculations.

A new doctoral student, Sachin Shivakumar, might implement some of these methods and provide a quantum-level prediction of the binding free energies of the mut-IDH1 radiotracer candidates suggested in this work.

A. MiMiC-Compliant Run Files with MiMiCPy PrepQM

Running the PrepQM command discussed in Section 5.2 generates the CPMD input script cpmd.inp and GROMACS index file index.ndx for a MiMiC run. We will explore these in sections A.1 and A.2.

A.1. GROMACS Run File

The GROMACS index file index.ndx contains the GROMACS indices of the QM atoms selected:

```
; Generated by MiMiCPy
[ QMatoms ]
   1    2    3    4    5    6    7    8    9   10
```

Listing A.1 Example of a GROMACS index file from PrepQM.

The file index.ndx should be used to generate the GROMACS .tpr run binary for the MiMiC-QM/MM simulation by passing it to the GROMACS preprocessor tool available from the local GROMACS installation. It is to be noted that the same coordinate and topology file used to run PrepQM should be passed to the GROMACS preprocessor. If the mdp file is passed to `mimicpy prepqm`, it can automatically generate the tpr file:

```
$ mimicpy prepqm -top topol.top -coords coords.gro -mdp mimic.mdp
   -gmx gmx_mpi_d -tpr mimic.tpr
```

Listing A.2 PrepQM command to directly generate a MiMiC-compatible TPR file.

Running the command in Listing A.2 will generate not only cpmd.inp and index.ndx, but also mimic.tpr by calling the GROMACS preprocessor (`gmx_mpi_d grompp`) in the

background. Passing mimic.mdp is also advantageous in that it performs checks to makes sure the GROMACS MM settings are compatible with a MiMiC run. Specifically, it automatically adds two mdp commands:

```
integrator = mimic
qmmm-grps = QMatoms
```

Listing A.3 GROMACS MDP options required for a MiMiC-QM/MM run.

The `integrator` set in GROMACS must always be `mimic` for a MiMiC run, and the appropriate name of the QM atoms groups in index.ndx (by default set to `QMatoms` as in Listing A.1) should be passed to `qmmm-grps`. The name of the QM atoms groups can be changed by passing it with the `-qma` option. This will change it both in index.ndx and mimic.mdp (if passed).

A.2. CPMD Input File

The contents of cpmd.inp for the QM atoms selected is shown below:

```
1  &MIMIC
2      PATHS
3          1
4  /path/to/tpr
5      OVERLAPS
6          10
7          2 1 1 1
8          2 2 1 2
9          2 3 1 3
10         2 5 1 4
11         2 6 1 5
12         2 7 1 6
13         2 8 1 7
14         2 9 1 8
15         2 10 1 9
16         2 4 1 10
17     BOX
18         47.0035358430417 47.0035358430417 47.0035358430417
19 &END
20
21 &CPMD
22     MIMIC
23 &END
24
25 &SYSTEM
```

```
26      CELL
27          5.8 1.0 1.3 0 0 0
28      CHARGE
29          0
30  &END
31
32  &ATOMS
33  *C_MT_BLYP.psp KLEINMAN-BYLANDER
34      LMAX=S
35      3
36      26.815213708440     45.561296864727     24.037316305240
37      25.416816376217     44.106207748765     22.034206613136
38      27.760076770753     44.068413226273     26.286090393544
39
40  *H_MT_BLYP.psp KLEINMAN-BYLANDER
41      LMAX=S
42      6
43      25.983734213604     44.673125586153     20.125583227264
44      26.078220519836     42.159789840401     21.920823045659
45      23.375912161621     44.295180361228     22.260973748092
46      28.156919256924     42.084200795416     25.832556123634
47      29.215165886714     45.013276288586     27.382131545827
48      26.078220519836     44.030618703780     27.476617852059
49
50  *O_MT_BLYP.psp KLEINMAN-BYLANDER
51      LMAX=S
52      1
53      27.041980843395     47.847865475525     23.942829999009
54
55  &END
```

Listing A.4 Basic CPMD commands for a MiMiC run outputted from MiMiCPy.

Listing A.4 is a barebones MiMiC-compliant CPMD input file, containing the most essential commands relevant for a MiMiC simulation. The most important section of cpmd.inp generated by PrepQM is `&ATOMS`. This contains the coordinates of the selected atoms, read from equilibrate.gro (as specified in Listing A.2). These are arranged into blocks of atomic elements according to the input format specifications of CPMD. Each atom block in CPMD is associated with pseudopotential details (like pseudopotential filenames, `LMAX`, `LOC`, etc.). PrepQM fills in default values for these, but a text file with these details for each element can be passed. This will allow PrepQM to automatically fill in these details for each atom in the CPMD input file. The data for each element is given in separate lines, and each line has the following format:

```
<element> <pp filename> <boundary pp filename> <labels> <lmax> <loc>
```

A "`-`" can be used to skip any field in this file. A specific example of such a file for a

system consisting of carbon, hydrogen and oxygen is given below:

```
C C_MT_BLYP.psp C_GIA_DUM_AN_BLYP KLEINMAN-BYLANDER P -
H H_MT_BLYP.psp - KLEINMAN-BYLANDER S -
O O_MT_BLYP.psp - KLEINMAN-BYLANDER P -
```

Listing A.5 Pseudopotential data file to automatically fill up the `&ATOMS` section with MiMiPy.

This can be stored in a file named pp_info.dat (to be used in Listing A.7). A `&MIMIC` section, consisting of the `PATH` to the GROMACS tpr, the MM `BOX` size, and the `OVERLAPS` section is also generated. The latter consists of a mapping of the atom IDs from GROMACS (according to the MM topology) to CPMD (according to the order of atoms appearing in `&ATOMS`). The QM box size and charge are also important information to be passed to CPMD. The former is specified with `CELL` under `&SYSTEM`. By default, MiMiCPy writes a cell size exactly bounding the QM region. This is often not enough to contain the plane waves of the QM region, and a padding needs to be added. This can be passed using the `-pad` option in `mimicpy prepqm`. The padding should be chosen wisely. The total QM charge is calculated by summing the MM charges read from the force field data. This is usually sufficient, but it may need to be cross checked for certain applications. The default behavior can be overridden by passing a charge value using the `-q` option.

The final CPMD input file template generated by, e.g., Listing A.4, can then be finally completed by hand by including any other parameter needed for the specific application. A more efficient way to achieve this is to pre-store any extra parameters in a separate template file template.inp, including the rest of the parameters within the `&MIMIC`, `&CPMD`, `&SYSTEM` and `&DFT` sections that are required for the MiMiC QM/MM run.

```
1  &MIMIC
2      LONG-RANGE COUPLING
3      FRAGMENT SORTING ATOM-WISE UPDATE
4          100
5      CUTOFF DISTANCE
6          20.0
7      MULTIPOLE ORDER
8          3
9  &END
10
11 &CPMD
12     MOLECULAR DYNAMICS CP
13     ISOLATED MOLECULE
14     QUENCH BO
```

```
15      ANNEALING IONS
16          0.99
17      TEMPERATURE
18          300
19      EMASS
20          600.
21      TIMESTEP
22          5.0
23      MAXSTEP
24          10000
25      TRAJECTORY SAMPLE
26          0
27      STORE
28          100
29      RESTFILE
30          1
31  &END
32
33  &SYSTEM
34      POISSON SOLVER TUCKERMAN
35      SYMMETRY
36          0
37      CUTOFF
38          70.
39  &END
40
41  &DFT
42      FUNCTIONAL BLYP
43  &END
```

Listing A.6 An example of a template CPMD input file for MiMiCPy.

As an example, Listing A.6 details the commands for a MiMiC-QM/MM simulated annealing. Having prepared such template file, MiMiCPy can be invoked as per Listing A.2, this time passing template.inp via the `-inp` option (Listing A.7). This command will result in a ready-to-use CPMD input file for a simulated annealing MiMiC run.

```
$ mimicpy prepqm -top topol.top -coords coords.gro -inp template.inp
    -pad 0.35 -pp pp_info.dat -out annealing.inp
```

Listing A.7 PrepQM command to generate a MiMiC-compatible CPMD input file from a template.

With this command, PrepQM automatically inserts the `&ATOMS` section, etc. into template.inp. The resulting CPMD input file (annealing.inp) is given below:

```
&MIMIC
    LONG-RANGE COUPLING
    FRAGMENT SORTING ATOM-WISE UPDATE
        100
    CUTOFF DISTANCE
        20.0
    MULTIPOLE ORDER
        3
    PATHS
        1
/path/to/tpr
    OVERLAPS
        10
        2 1 1 1
        2 2 1 2
        2 3 1 3
        2 5 1 4
        2 6 1 5
        2 7 1 6
        2 8 1 7
        2 9 1 8
        2 10 1 9
        2 4 1 10
    BOX
        47.0035358430417 47.0035358430417 47.0035358430417
&END

&CPMD
    MOLECULAR DYNAMICS CP
    ISOLATED MOLECULE
    QUENCH BO
    ANNEALING IONS
        0.99
    TEMPERATURE
        300
    EMASS
        600.
    TIMESTEP
        5.0
    MAXSTEP
        10000
    TRAJECTORY SAMPLE
        0
    STORE
        100
    RESTFILE
        1
&END

&SYSTEM
    CELL
```

```
52          19.1 1.0 1.1 0 0 0
53      CHARGE
54          0
55  &END
56
57  &DFT
58      FUNCTIONAL BLYP
59  &END
60
61  &ATOMS
62  *C_MT_BLYP.psp KLEINMAN-BYLANDER
63      LMAX=P
64      3
65      26.815213708440     45.561296864727     24.037316305240
66      25.416816376217     44.106207748765     22.034206613136
67      27.760076770753     44.068413226273     26.286090393544
68
69  *H_MT_BLYP.psp KLEINMAN-BYLANDER
70      LMAX=S
71      6
72      25.983734213604     44.673125586153     20.125583227264
73      26.078220519836     42.159789840401     21.920823045659
74      23.375912161621     44.295180361228     22.260973748092
75      28.156919256924     42.084200795416     25.832556123634
76      29.215165886714     45.013276288586     27.382131545827
77      26.078220519836     44.030618703780     27.476617852059
78
79  *O_MT_BLYP.psp KLEINMAN-BYLANDER
80      LMAX=P
81      1
82      27.041980843395     47.847865475525     23.942829999009
83
84  &END
```

Listing A.8 A complete CPMD input file for a MiMiC-QM/MM annealing run generated with MiMiCPy.

In the CPMD input file of Listing A.8, a padding of 0.35 nm has been added to the QM box size in all directions, and pp_info.dat (with contents from Listing A.5) was used to assign the right `LMAX` values to the atoms.

B. Extra Data on IDH1

B.1. Wannier Center Analysis of the Wt-IDH1 Catalysis

The discussion in this section is reproduced from publication [B] (see Section 1.3).

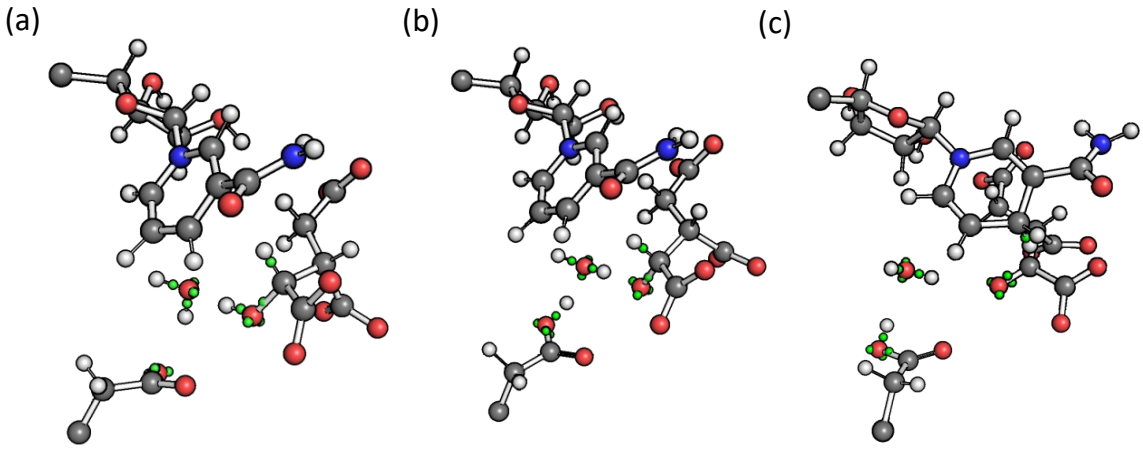


Figure B.1. Wannier centers depicted in green for select bonds at (a) $CV_1 \approx -0.08$ Å (b) $CV_1 \approx 0.08$ Å or $CV_2 = 0$ Å (c) $CV_2 \approx 0.2$ Å. Adapted from Ref. [B].

Wannier centers can be roughly conceptualized as representing a pair of electrons in the space. These are calculated at mutliple steps of the wt-IDH1 catalysis and shown in Figures B.1a–c. In the reactant state ($CV_1 \approx -0.08$ Å), the C_α-O_h bond length is equal to 1.6 Å, with a Wannier center located at ≈ 1.0 Å from C_α, indicating a single bond character. At $CV_1 = 0$ Å, closer to the transition state, the water molecule exists as a hydronium ion stabilized by $Asp252^B$. In this configuration, O_W interacts with H_h, while one of the hydrogen atoms bound to O_W interacts with the $Asp252^B$ side chain. The Wannier center along the O_h-H_h bond is located farther away from H_h than in the reactant state by ≈ 0.2 Å, indicating an increasingly higher polar character of the bond and the transfer of a proton to O_W. This Wannier center is more closely associated with O_h, indicating a developing negative charge on it. In the final product ($CV_1 \approx 0.08$), the C_α-O_h bond length decreases to ≈ 1.3 Å, and the Wannier center along the bond is ≈ 0.8 Å away from C_α. Furthermore, $Asp252^B$ is

protonated and the ICT C_α hydroxyl group is deprotonated with a negatively charged O_h, due to the extra third Wannier center associated with it.

Starting from the product state of the first sub-step, we calculated the free energy change with increasing CV_2. In the reactant state ($CV_1 \approx -0.2$ Å), the Wannier center along the C_α-H_α bond is ≈ 0.7 Å from C_α and ≈ 3.5 Å from C_N of the $NADP^+$ ring. At the transition state ($CV_2 \approx 0$ Å), the hydride transfer of H_α to C_N takes place. The Wannier center along the C_α-H_α bond is now ≈ 1.3 Å from C_α and ≈ 1.5 Å from C_N. Furthermore, the third Wannier center associated with O_h from the product of the previous step, has now moved closer to C_α (from ≈ 1.5 Å to ≈ 1.1 Å) and more along the C_α-O_h bond. This, together with the fact that the C_α-O_h bond length reduces to 1.3 Å, indicates the emergence of a partial double bond character along the C_α-O_h bond. At the product ($CV_2 \approx 0.2$ Å), this extra Wannier center moves to ≈ 0.8 Å from C_α. This results in two Wannier centers along the C_α-O_h bond, and indicates the establishment of a full double bond, i.e., the formation of a ketone. The Wannier center along the C_α-H_α bond moves ≈ 3.4 Å away from C_α, with this Wannier center falling along the newly formed C_N-H_α bond. The hydride transfer of H_α to the $NADP^+$ ring is complete.

B.2. Mg Coordination in Mut-IDH1 Active Site during QM/MM MD

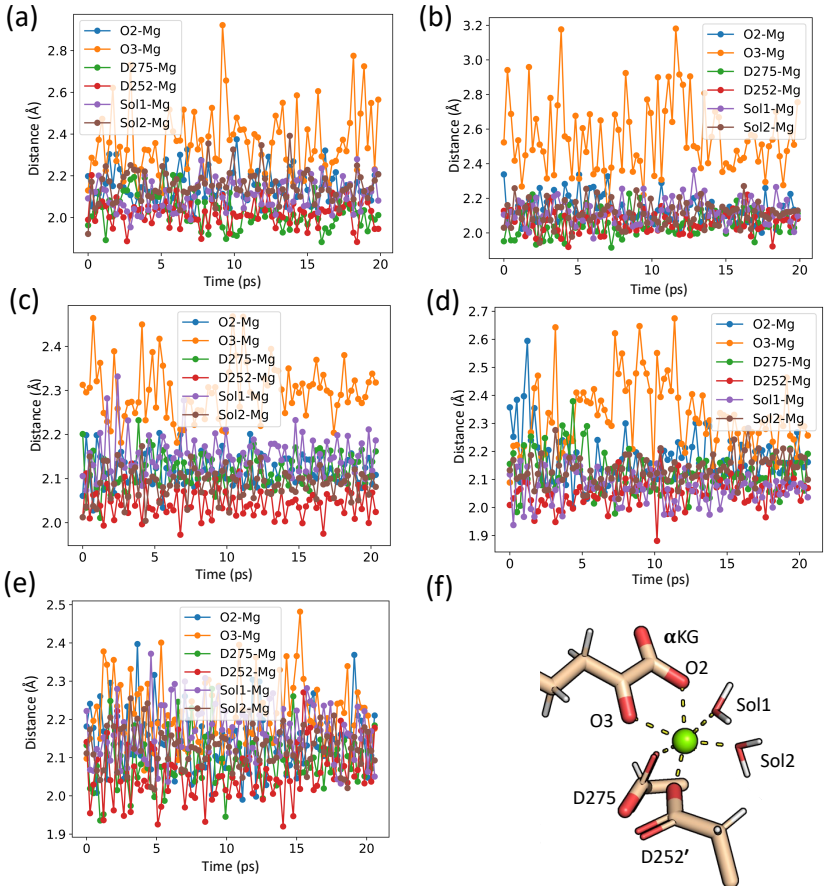


Figure B.2. The stability of the Mg^{2+} coordination sphere with respect to time for various configurations (a) KH/D(A), (b) KH/D(B), (c) K/DH(A*), (d) K/DH(A), and (e) K/DH(B) of mut-IDH1 during QM/MM MD. (f) Representation of the distances in the mut-IDH1 active site measured during QM/MM MD.

B.3. Unconstrained cMD Simulations of Mut-IDH1

As mention in Section 7.1 and shown in Figure B.3, cMD could not accurately capture the bidentate nature of αKG–Mg^{2+} binding in mut-IDH1. It is known that this bidentate binding is required for the catalysis, and is expected to be present at the Michaelis complex.[178] Thus, we can conclude that cMD cannot reproduce the Michaelis complex of mut-IDH1. The following force field parameters were tested:

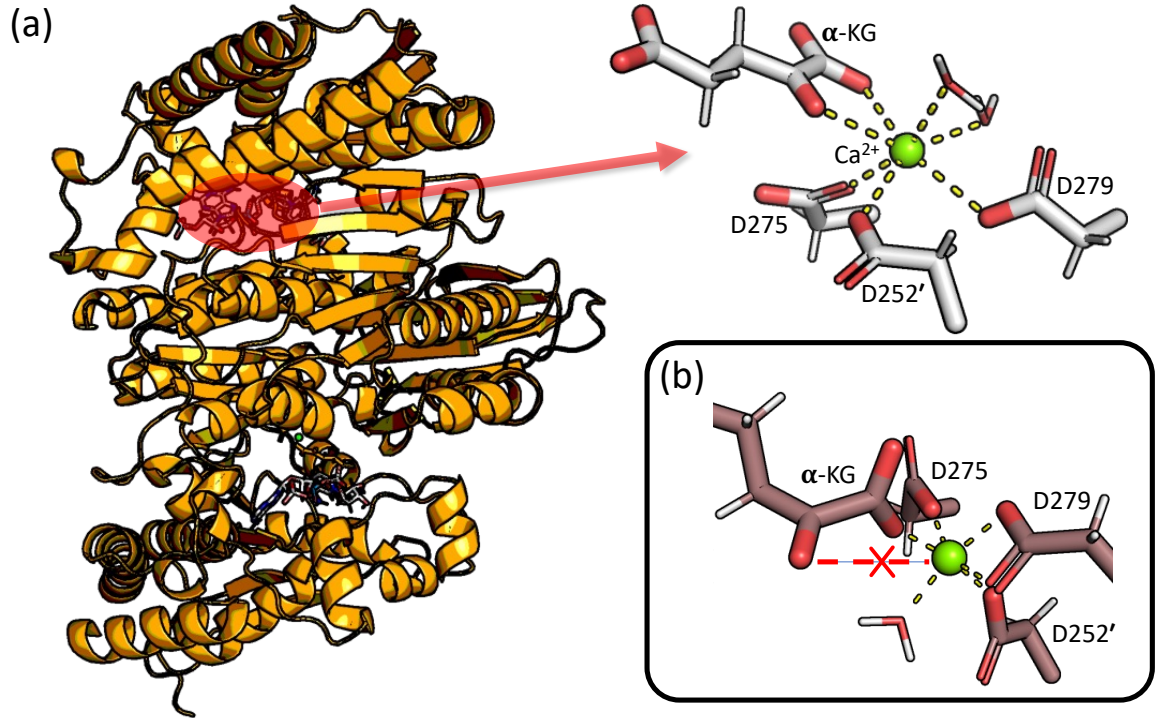


Figure B.3. (a) Cartoon representation of the mut-IDH1 with the active site containing the heptacoordinated Ca^{2+} coordination sphere. The αKG substrate is coordinated to Ca^{2+} in a bidendate fashion. (b) Loss of the bidendate coordination of αKG during cMD simulations.

1. The standard MG(II) parameters in the Amber99sb*-ildn force field used for wt-IDH1[213, 214]

2. 6-12 Lennard-Jones parameters of MG(II) from Grotz et. al.[275]

3. 6-12 Lennard-Jones parameters of MG(II) from Zhang et. al.[276]

The last two are some of the latest non-bonded parameters available in literature. More complicated parameters were not attempted for the following reasons:

1. We wish to promote a parameter-free approach as far as possible, so as to make the QHPC—DD pipeline applicable for a widest variety of drug targets.

2. As described in Section 7.1, the coordination of Mg(II) within to mut-IDH1 active site is not clear from the crystal structure. Usage of custom MM parameters is not possible without *a priori* knowledge of the coordination.

Around 10 ns of cMD simulations were attempted with each of the three chosen force fields parameters. All failed to capture the bidentate nature of the binding. In fact, in most simulations, The bidendate binding was broken in the preliminary step of cMD, i.e., at minimization (shown in Figure B.3b). This was one of the motivations for us to update the simulation pipeline by replacing the inadequate MM minimization with a QM/MM one as discussed in Section 7.2.2.

References

[1] David Austin and Tamara Hayford. *Research and development in the pharmaceutical industry*. Tech. rep. CBO, 2021.

[2] Duxin Sun et al. "Why 90% of clinical drug development fails and how to improve it?" In: *Acta Pharmaceutica Sinica. B* 12.7 (July 2022), p. 3049.

[3] Olivier J. Wouters, Martin McKee, and Jeroen Luyten. "Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018". In: *JAMA* 323.9 (Mar. 2020), p. 844.

[4] Kevin Dondarski and Kevin Lesser. *Seize the digital momentum Measuring the return from pharmaceutical innovation 2022 Contents*. Tech. rep. Deloitte, 2023.

[5] Anastasiia V. Sadybekov and Vsevolod Katritch. "Computational approaches streamlining drug discovery". In: *Nature 2023 616:7958* 616.7958 (Apr. 2023), pp. 673–685.

[6] Nalini Schaduangrat et al. "Towards reproducible computational drug discovery". In: *Journal of Cheminformatics 2020 12:1* 12.1 (Jan. 2020), pp. 1–30.

[7] Marco De Vivo et al. "Role of Molecular Dynamics and Related Methods in Drug Discovery". In: *Journal of Medicinal Chemistry* 59.9 (May 2016), pp. 4035–4061.

[8] Victor T. Sabe et al. "Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review". In: *European Journal of Medicinal Chemistry* 224 (Nov. 2021), p. 113705.

[9] Marco De Vivo. "Bridging quantum mechanics and structure-based drug design". In: *Frontiers in Bioscience* 16.1 (2011), p. 1619.

[10] Pavlína Pokorná et al. "QM/MM Calculations on Protein-RNA Complexes: Understanding Limitations of Classical MD Simulations and Search for Reliable Cost-Effective QM Methods". In: *Journal of Chemical Theory and Computation* 14.10 (Oct. 2018), pp. 5419–5433.

[11] Marc W. Van Der Kamp and Adrian J. Mulholland. "Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology". In: *Biochemistry* 52.16 (Apr. 2013), pp. 2708–2728.

[12] Elizabeth Brunk and Ursula Rothlisberger. “Mixed Quantum Mechanical/-Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States”. In: *Chemical Reviews* 115.12 (June 2015), pp. 6217–6263.

[13] David S. Hartsough and Kenneth M. Jr. Merz. “Dynamic Force Field Models: Molecular Dynamics Simulations of Human Carbonic Anhydrase II Using a Quantum Mechanical/Molecular Mechanical Coupled Potential”. In: *The Journal of Physical Chemistry* 99.28 (July 1995), pp. 11266–11275.

[14] Hans Martin Senn and Walter Thiel. “QM/MM methods for biomolecular systems”. In: *Angewandte Chemie (International ed. in English)* 48.7 (Feb. 2009), pp. 1198–1229.

[15] Ariana Remmel. “Welcome to exascale computing”. In: *C&EN Global Enterprise* 100.31 (Sept. 2022), pp. 29–33.

[16] J. Hutter et al. *CPMD, Copyright IBM Corp 1990-2022, Copyright MPI für Festkörperforschung Stuttgart 1997-2001.*

[17] David Van Der Spoel et al. “GROMACS: fast, flexible, and free”. In: *Journal of computational chemistry* 26.16 (Dec. 2005), pp. 1701–1718.

[18] Jógvan Magnus Haugaard Olsen et al. “MiMiC: A Novel Framework for Multiscale Modeling in Computational Chemistry”. In: *Journal of Chemical Theory and Computation* 15.6 (June 2019), pp. 3810–3823.

[19] Viacheslav Bolnykh et al. “Extreme Scalability of DFT-Based QM/MM MD Simulations Using MiMiC”. In: *Journal of Chemical Theory and Computation* 15.10 (Oct. 2019), pp. 5601–5613.

[20] Viacheslav Bolnykh et al. “MiMiC: Multiscale Modeling in Computational Chemistry”. In: *Frontiers in Molecular Biosciences* 7 (Mar. 2020).

[21] Maria Gabriella Chiariello et al. “Molecular Basis of CLC Antiporter Inhibition by Fluoride”. In: *Journal of the American Chemical Society* 142.16 (Apr. 2020), pp. 7254–7258.

[22] Maria Gabriella Chiariello et al. “Mechanisms Underlying Proton Release in CLC-type F-/H+Antiporters”. In: *Journal of Physical Chemistry Letters* 12.18 (May 2021), pp. 4415–4420.

[23] Florian Karl Schackert et al. “Mechanism of calcium permeation in a glutamate receptor ion channel”. In: (Sept. 2022).

[24] Mirko Paulikat et al. “Proton Transfers to DNA in Native Electrospray Ionization Mass Spectrometry: A Quantum Mechanics/Molecular Mechanics Study”. In: *The Journal of Physical Chemistry Letters* 13.51 (Dec. 2022), pp. 12004–12010.

[25] Xiang Xu et al. “Structures of human cytosolic NADP-dependent isocitrate dehydrogenase reveal a novel self-regulatory mechanism of activity”. In: *The Journal of biological chemistry* 279.32 (Aug. 2004), pp. 33946–33957.

[26] Craig Horbinski. “What do we know about IDH1/2 mutations so far, and how do we use it?” In: *Acta neuropathologica* 125.5 (May 2013), pp. 621–636.

[27] Hai Yan et al. “IDH1 and IDH2 Mutations in Gliomas ”. In: *New England Journal of Medicine* 360.8 (Feb. 2009), pp. 765–773.

[28] Jason Beiko et al. “IDH1 mutant malignant astrocytomas are more amenable to surgical resection and have a survival benefit associated with maximal surgical resection”. In: *Neuro-Oncology* 16.1 (Jan. 2014), pp. 81–91.

[29] Satish K. Chitneni. “IDH1 Mutations in Glioma: Considerations for Radiotracer Development”. In: *SM radiology journal* 2.1 (2016).

[30] Michael M. Wollring et al. “Clinical applications and prospects of PET imaging in patients with IDH-mutant gliomas”. In: *Journal of Neuro-Oncology* 162.3 (May 2023), pp. 481–488.

[31] Shuang Liu et al. “Differentiating Inhibition Selectivity and Binding Affinity of Isocitrate Dehydrogenase 1 Variant Inhibitors”. In: *Journal of Medicinal Chemistry* 66.7 (Apr. 2023), pp. 5279–5288.

[32] Satish K. Chitneni et al. “Synthesis and evaluation of radiolabeled AGI-5198 analogues as candidate radiotracers for imaging mutant IDH1 expression in tumors”. In: *Bioorganic & Medicinal Chemistry Letters* 28.4 (Feb. 2018), pp. 694–699.

[33] Christina Eleftheria Tzeliou, Markella Aliki Mermigki, and Demeter Tzeli. “Review on the QM/MM Methodologies and Their Application to Metalloproteins”. In: *Molecules* 27.9 (May 2022).

[34] Andrew J. Thomson and Harry B. Gray. “Bio-inorganic chemistry”. In: *Current Opinion in Chemical Biology* 2.2 (Apr. 1998), pp. 155–158.

[35] Matthieu Rouffet and Seth M. Cohen. “Emerging Trends in Metalloprotein Inhibition”. In: *Dalton transactions (Cambridge, England : 2003)* 40.14 (Apr. 2011), p. 3445.

[36] A. I. Anzellotti and N. P. Farrell. “Zinc metalloproteins as medicinal targets”. In: *Chemical Society Reviews* 37.8 (July 2008), pp. 1629–1651.

[37] David P. Martin, David T. Puerta, and Seth M. Cohen. “Metalloprotein Inhibitors”. In: *Ligand Design in Medicinal Inorganic Chemistry* 9781118488522 (June 2014), pp. 375–403.

[38] Jessica L. Childs-Disney et al. “Targeting RNA structures with small molecules”. In: *Nature Reviews Drug Discovery 2022 21:10* 21.10 (Aug. 2022), pp. 736–762.

[39] Mattia Bernetti et al. “Computational drug discovery under RNA times”. In: *QRB Discovery* 3 (Nov. 2022), e22.

[40] Jiří Šponer et al. “RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview”. In: *Chemical Reviews* 118.8 (Apr. 2018), pp. 4177–4338.

[41] Julie Puyo-Fourtine et al. “Consistent Picture of Phosphate-Divalent Cation Binding from Models with Implicit and Explicit Electronic Polarization”. In: *Journal of Physical Chemistry B* 126.22 (June 2022), pp. 4022–4034.

[42] Pavel Hobza et al. “C-HO contacts in the adenineuracil Watson-Crick and uraciluracil nucleic acid base pairs: Nonempirical ab initio study with inclusion of electron correlation effects”. In: *Journal of Physical Chemistry B* 104.26 (June 2000), pp. 6286–6292.

[43] Fandi Sutanto, Markella Konstantinidou, and Alexander Dömling. “Covalent inhibitors: a rational approach to drug discovery”. In: *RSC Medicinal Chemistry* 11.8 (Aug. 2020), p. 876.

[44] Stephane De Cesco et al. “Covalent inhibitors design and discovery”. In: *European Journal of Medicinal Chemistry* 138 (Sept. 2017), pp. 96–114.

[45] Shaloam Dasari and Paul Bernard Tchounwou. “Cisplatin in cancer therapy: Molecular mechanisms of action”. In: *European Journal of Pharmacology* 740 (Oct. 2014), pp. 364–378.

[46] Paolo Carloni, Michiel Sprik, and Wanda Andreoni. “Key Steps of the cis-Platin-DNA Interaction: Density Functional Theory-Based Molecular Dynamics Simulations”. In: *Journal of Physical Chemistry B* 104.4 (Feb. 2000), pp. 823–835.

[47] Elizabeth R. Jamieson and Stephen J. Lippard. “Structure, Recognition, and Processing of CisplatinDNA Adducts”. In: *Chemical Reviews* 99.9 (1999), pp. 2467–2498.

[48] Vania Calandrini et al. “Computational metallomics of the anticancer drug cisplatin”. In: *Journal of Inorganic Biochemistry* 153 (Dec. 2015), pp. 231–238.

[49] Lamei Huang et al. “KRAS mutation: from undruggable to druggable in cancer”. In: *Signal Transduction and Targeted Therapy 2021 6:1* 6.1 (Nov. 2021), pp. 1–20.

[50] Pavel Banáš et al. “Theoretical studies of RNA catalysis: Hybrid QM/MM methods and their comparison with MD and QM”. In: *Methods (San Diego, Calif.)* 49.2 (Oct. 2009), p. 202.

[51] Mark A. Ditzler et al. “Molecular dynamics and quantum mechanics of RNA: Conformational and chemical change we can believe in”. In: *Accounts of Chemical Research* 43.1 (Jan. 2010), pp. 40–47.

[52] Gary B. Evans, Vern L. Schramm, and Peter C. Tyler. “The transition to magic bullets – transition state analogue drug design”. In: *MedChemComm* 9.12 (Dec. 2018), pp. 1983–1993.

[53] Vern L. Schramm. “Transition states, analogues, and drug development”. In: *ACS Chemical Biology* 8.1 (Jan. 2013), pp. 71–81.

[54] Richard Wolfenden. “Transition state analog inhibitors and enzyme catalysis”. In: *Annual review of biophysics and bioengineering* 5.1 (1976), pp. 271–306.

[55] Gregory Sliwoski et al. “Computational Methods in Drug Discovery”. In: *Pharmacological Reviews* 66.1 (Jan. 2014), p. 334.

[56] Thomas Seidel et al. “The Pharmacophore Concept and Its Applications in Computer-Aided Drug Design”. In: *Progress in the Chemistry of Organic Natural Products* 110 (2019), pp. 99–141.

[57] C. G. Wermuth et al. “Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998)”. In: *Pure and Applied Chemistry* 70.5 (Jan. 1998), pp. 1129–1143.

[58] Deborah Giordano et al. “Drug Design by Pharmacophore and Virtual Screening Approach”. In: *Pharmaceuticals* 15.5 (May 2022).

[59] Maria Batool, Bilal Ahmad, and Sangdun Choi. “A Structure-Based Drug Discovery Paradigm”. In: *International Journal of Molecular Sciences* 20.11 (June 2019).

[60] Evanthia Lionta et al. “Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances”. In: *Current topics in medicinal chemistry* 14.16 (2014), pp. 1923–1938.

[61] Xuan-Yu Meng et al. “Molecular Docking: A powerful approach for structure-based drug discovery”. In: *Current computer-aided drug design* 7.2 (June 2011), p. 146.

[62] Pedro H.M. Torres et al. “Key Topics in Molecular Docking for Drug Design”. In: *International Journal of Molecular Sciences* 20.18 (Sept. 2019).

[63] N. Moitessier et al. “Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go”. In: *British Journal of Pharmacology* 153.S1 (Mar. 2008), S7–S26.

[64] Isabella A. Guedes, Camila S. de Magalhães, and Laurent E. Dardenne. “Receptor-ligand molecular docking”. In: *Biophysical Reviews* 6.1 (Mar. 2014), pp. 75–87.

[65] Veronica Salmaso and Stefano Moro. “Bridging Molecular Docking to Molecular Dynamics in Exploring Ligand-Protein Recognition Process: An Overview”. In: *Frontiers in Pharmacology* 9 (Aug. 2018).

[66] Ashutosh Tripathi and Vytas A Bankaitis. “Molecular Docking: From Lock and Key to Combination Lock”. In: *Journal of molecular medicine and clinical applications* 2.1 (2017).

[67] Zhe Wang et al. “Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power”. In: *Physical Chemistry Chemical Physics* 18.18 (May 2016), pp. 12964–12975.

[68] Camila Silva De Magalhães et al. “A dynamic niching genetic algorithm strategy for docking highly flexible ligands”. In: *Information Sciences* 289.1 (Dec. 2014), pp. 206–224.

[69] Thomas A. Halgren. “Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94”. In: *Journal of Computational Chemistry* 17.5-6 (Apr. 1996), pp. 490–519.

[70] Garrett M. Morris et al. “Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function”. In: *Journal of Computational Chemistry* 19.14 (Nov. 1998), pp. 1639–1662.

[71] Richard A. Friesner et al. “Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy”. In: *Journal of Medicinal Chemistry* 47.7 (Mar. 2004), pp. 1739–1749.

[72] Thomas A. Halgren et al. “Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening”. In: *Journal of Medicinal Chemistry* 47.7 (Mar. 2004), pp. 1750–1759.

[73] *How well does Glide deal with binding sites containing metal ions for docking? | Schrödinger*.

[74] Richard A. Friesner et al. “Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes”. In: *Journal of Medicinal Chemistry* 49.21 (Oct. 2006), pp. 6177–6196.

[75] Hongjian Li et al. “Machine-learning scoring functions for structure-based virtual screening”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 11.1 (Jan. 2021), e1478.

[76] Qurrat Ul Ain et al. “Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 5.6 (Nov. 2015), pp. 405–424.

[77] Holger Gohlke, Manfred Hendlich, and Gerhard Klebe. “Knowledge-based scoring function to predict protein-ligand interactions”. In: *Journal of Molecular Biology* 295.2 (Jan. 2000), pp. 337–356.

[78] Ingo Muegge. “PMF Scoring Revisited”. In: *Journal of Medicinal Chemistry* 49.20 (Oct. 2006), pp. 5895–5902.

[79] Douglas B. Kitchen et al. “Docking and scoring in virtual screening for drug discovery: methods and applications”. In: *Nature Reviews Drug Discovery 2004 3:11* 3.11 (Nov. 2004), pp. 935–949.

[80] Heather A Carlson. “Protein flexibility and drug design: how to hit a moving target”. In: *Current Opinion in Chemical Biology* 6.4 (Aug. 2002), pp. 447–452.

[81] Katrina W. Lexa and Heather A. Carlson. “Protein flexibility in docking and surface mapping”. In: *Quarterly Reviews of Biophysics* 45.3 (Aug. 2012), pp. 301–343.

[82] Marcus Fischer et al. “Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery”. In: *Nature Chemistry* 6.7 (July 2014), pp. 575–583.

[83] Andrew R. Leach. “Ligand docking to proteins with discrete side-chain flexibility”. In: *Journal of Molecular Biology* 235.1 (Jan. 1994), pp. 345–356.

[84] Ian L. Alberts, Nikolay P. Todorov, and Philip M. Dean. “Receptor flexibility in de novo ligand design and docking”. In: *Journal of Medicinal Chemistry* 48.21 (Oct. 2005), pp. 6585–6596.

[85] Jens Meiler and David Baker. “ROSETTALIGAND: Protein–small molecule docking with full side-chain flexibility”. In: *Proteins: Structure, Function, and Bioinformatics* 65.3 (Nov. 2006), pp. 538–548.

[86] Woody Sherman, Hege S. Beard, and Ramy Farid. “Use of an Induced Fit Receptor Structure in Virtual Screening”. In: *Chemical Biology & Drug Design* 67.1 (Jan. 2006), pp. 83–84.

[87] Jacob D. Durrant and J. Andrew McCammon. “Molecular dynamics simulations and drug discovery”. In: *BMC Biology* 9.1 (Oct. 2011), pp. 1–9.

[88] Paweł Śledź and Amedeo Caflisch. “Protein structure-based drug design: from docking to molecular dynamics”. In: *Current Opinion in Structural Biology* 48 (Feb. 2018), pp. 93–102.

[89] Dario Gioia et al. “Dynamic Docking: A Paradigm Shift in Computational Drug Discovery”. In: *Molecules 2017, Vol. 22, Page 2029* 22.11 (Nov. 2017), p. 2029.

[90] Jung Hsin Lin et al. “Computational drug design accommodating receptor flexibility: The relaxed complex scheme”. In: *Journal of the American Chemical Society* 124.20 (May 2002), pp. 5632–5633.

[91] Rommie E. Amaro, Riccardo Baron, and J. Andrew McCammon. “An improved relaxed complex scheme for receptor flexibility in computer-aided drug design”. In: *Journal of Computer-Aided Molecular Design* 22.9 (Jan. 2008), pp. 693–705.

[92] Xiaoyu Zhao et al. “Molecular investigation of the dual inhibition mechanism for targeted P53 regulator MDM2/MDMX inhibitors”. In: *Physical Chemistry Chemical Physics* 24.27 (July 2022), pp. 16799–16815.

[93] David M. Ferguson, Randall J. Radmer, and Peter A. Kollman. “Determination of the relative binding free energies of peptide inhibitors to the HIV-1 protease”. In: *Journal of Medicinal Chemistry* 34.8 (Aug. 1991), pp. 2654–2659.

[94] Andrea Cavalli et al. “A computational study of the binding of propidium to the peripheral anionic site of human acetylcholinesterase”. In: *Journal of Medicinal Chemistry* 47.16 (July 2004), pp. 3991–3999.

[95] Kimichi Suzuki, Satoshi Maeda, and Keiji Morokuma. “Roles of Closed- and Open-Loop Conformations in Large-Scale Structural Transitions of l -Lactate Dehydrogenase”. In: *ACS Omega* 4.1 (Jan. 2019), pp. 1178–1184.

[96] Buyong Ma and Ruth Nussinov. “Enzyme dynamics point to stepwise conformational selection in catalysis”. In: *Current opinion in chemical biology* 14.5 (Oct. 2010), p. 652.

[97] Dominik Marx and Jürg Hutter. *Ab initio molecular dynamics: Basic theory and advanced methods*. Cambridge University Press, 2009, pp. 1–567.

[98] M. Born and R. Oppenheimer. “Zur Quantentheorie der Molekeln”. In: *Annalen der Physik* 389.20 (1927), pp. 457–484.

[99] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation*. Academic Press, 2002.

[100] M. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, 2010.

[101] Michael P. Allen and Dominic J. Tildesley. *Computer simulation of liquids: Second edition*. Oxford University Press, 2017, pp. 1–626.

[102] Shuichi Nosé. “A unified formulation of the constant temperature molecular dynamics methods”. In: *The Journal of Chemical Physics* 81.1 (1984), pp. 511–519.

[103] William G. Hoover. “Canonical dynamics: Equilibrium phase-space distributions”. In: *Physical Review A* 31.3 (1985), pp. 1695–1697.

[104] Shuichi Nosé. “A unified formulation of the constant temperature molecular dynamics methods”. In: *J. Chem. Phys.* 81.1 (Aug. 1998), p. 511.

[105] Michele Parrinello and Aneesur Rahman. “Polymorphic transitions in single crystals: A new molecular dynamics method”. In: *Journal of Applied Physics* 52.12 (1981), pp. 7182–7190.

[106] Shuichi Nosé and Michael L. Klein. “Constant pressure molecular dynamics for molecular systems”. In: *Molecular Physics* 50.5 (1983), pp. 1055–1076.

[107] Marcelo C.R. Melo et al. “CHARMM: The biomolecular simulation program”. In: *J. Comput. Chem.* 30.10 (2009), pp. 1545–1614.

[108] Viktor Hornak et al. “Comparison of multiple Amber force fields and development of improved protein backbone parameters”. In: *Proteins* 65.3 (Nov. 2006), pp. 712–725.

[109] Pierre Hohenberg and Walter Kohn. “Inhomogeneous Electron Gas”. In: *Physical Review* 136.3B (1964), pp. 864–871.

[110] Walter Kohn and Lu Jeu Sham. "Self-consistent equations including exchange and correlation effects". In: *Physical Review* 140.4A (1965).

[111] Eberhard Engel and Reiner M. Dreizler. *Density Functional Theory: An Advanced Course*. Springer Berlin Heidelberg, 2011.

[112] Jianmin Tao et al. "Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids". In: *Physical Review Letters* 91.14 (2003).

[113] Kieron Burke. "Perspective on density functional theory". In: *Journal of Chemical Physics* 136.15 (2012), p. 150901.

[114] Axel D. Becke. "Density-functional exchange-energy approximation with correct asymptotic behavior". In: *Physical Review A* 38.6 (1988), pp. 3098–3100.

[115] Chengteh Lee, Weitao Yang, and Robert G. Parr. "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density". In: *Physical Review B* 37.2 (1988), pp. 785–789.

[116] Axel D Becke. "Density-functional thermochemistry. I. The effect of the exchange-only gradient correction". In: *J. Chem. Phys.* 96.3 (1992), pp. 2155–2160.

[117] Axel D. Becke. "A new mixing of Hartree-Fock and local density-functional theories". In: *The Journal of Chemical Physics* 98.2 (1993), pp. 1372–1377.

[118] Axel D Becke. "Density-functional thermochemistry. II. The effect of the Perdew–Wang generalized-gradient correlation correction". In: *J. Chem. Phys.* 97.12 (1992), pp. 9173–9177.

[119] P. J. Stephens et al. "Ab Initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields". In: *Journal of Physical Chemistry* 98.45 (1994), pp. 11623–11627.

[120] Seymour H. Vosko, L. Wilk, and M. Nusair. "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis". In: *Canadian Journal of Physics* 58.8 (1980), pp. 1200–1211.

[121] Axel D. Becke. "The Challenge of d and f Electrons. Density Functional Theories in Quantum Chemistry". In: *The Challenge of d and f Electrons*. ACS Symposium Series, 1988. Chap. 12, pp. 165–179.

[122] Axel D. Becke. "Density-functional thermochemistry. III. The role of exact exchange". In: *The Journal of Chemical Physics* 98.7 (1993), pp. 5648–5652.

[123] N. Troullier and José Luriaas Martins. "Efficient pseudopotentials for plane-wave calculations". In: *Phys. Rev. B* 43 (3 Jan. 1991), pp. 1993–2006.

[124] A. Warshel and M. Levitt. "Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme". In: *Journal of Molecular Biology* 103.2 (May 1976), pp. 227–249.

[125] Gerrit Groenhof. "Introduction to QM/MM Simulations". In: (2013), pp. 43–66.

[126] O. Anatole Von Lilienfeld et al. “Variational optimization of effective atom centered potentials for molecular properties”. In: *The Journal of Chemical Physics* 122.1 (Dec. 2004), p. 014113.

[127] Lung Wa Chung et al. “The ONIOM Method and Its Applications”. In: *Chemical Reviews* 115.12 (June 2015), pp. 5678–5796.

[128] Sven Roßbach and Christian Ochsenfeld. “Influence of Coupling and Embedding Schemes on QM Size Convergence in QM/MM Approaches for the Example of a Proton Transfer in DNA”. In: *Journal of Chemical Theory and Computation* 13.3 (Mar. 2017), pp. 1102–1107.

[129] Richard A. Friesner and Victor Guallar. “AB INITIO QUANTUM CHEMICAL AND MIXED QUANTUM MECHANICS/MOLECULAR MECHANICS (QM/MM) METHODS FOR STUDYING ENZYMATIC CATALYSIS”. In: *Annual Review of Physical Chemistry* 56.1 (May 2005), pp. 389–427.

[130] Jiali Gao et al. “Mechanisms and Free Energies of Enzymatic Reactions”. In: *Chemical Reviews* 106.8 (Aug. 2006), pp. 3188–3209.

[131] Hans Martin Senn and Walter Thiel. “QM/MM studies of enzymes”. In: *Current Opinion in Chemical Biology* 11.2 (Apr. 2007), pp. 182–187.

[132] Maria J. Ramos and Pedro A. Fernandes. “Computational Enzymatic Catalysis”. In: *Accounts of Chemical Research* 41.6 (June 2008), pp. 689–698.

[133] Marc W. van der Kamp and Adrian J. Mulholland. “Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology”. In: *Biochemistry* 52.16 (Apr. 2013), pp. 2708–2728.

[134] Alexandra T.P. Carvalho et al. “Challenges in computational studies of enzyme structure, function and dynamics”. In: *Journal of Molecular Graphics and Modelling* 54 (Nov. 2014), pp. 62–79.

[135] Chandan Kumar Das and Nisanth N. Nair. “Elucidating the Molecular Basis of Avibactam-Mediated Inhibition of ClassA β-Lactamases”. In: *Chemistry – A European Journal* 26.43 (Aug. 2020), pp. 9639–9651.

[136] Marten Prieß et al. “Molecular Mechanism of ATP Hydrolysis in an ABC Transporter”. In: *ACS Central Science* 4.10 (Oct. 2018), pp. 1334–1343.

[137] Rui P. P. Neves, Pedro A. Fernandes, and Maria J. Ramos. “Role of Enzyme and Active Site Conformational Dynamics in the Catalysis by α-Amylase Explored with QM/MM Molecular Dynamics”. In: *Journal of Chemical Information and Modeling* 62.15 (Aug. 2022), pp. 3638–3650.

[138] Victor A. Streltsov et al. “Discovery of processive catalysis by an exo-hydrolase with a pocket-shaped active site”. In: *Nature Communications* 10.1 (May 2019), p. 2222.

[139] Dmitriy A. Lukoyanov et al. “Electron Redistribution within the Nitrogenase Active Site FeMo-Cofactor During Reductive Elimination of H ₂ to Achieve NN Triple-Bond Activation”. In: *Journal of the American Chemical Society* 142.52 (Dec. 2020), pp. 21679–21690.

[140] Ming Hsun Ho et al. “Unraveling the catalytic pathway of metalloenzyme farnesyltransferase through QM/MM computation”. In: *Journal of Chemical Theory and Computation* 5.6 (June 2009), pp. 1657–1666.

[141] Ruibo Wu et al. “Flexibility of catalytic zinc coordination in thermolysin and HDAC8: A Born-Oppenheimer ab initio QM/MM molecular dynamics study”. In: *Journal of Chemical Theory and Computation* 6.1 (Jan. 2010), pp. 337–343.

[142] Ravi Tripathi, Harald Forbert, and Dominik Marx. “Settling the Long-Standing Debate on the Proton Storage Site of the Prototype Light-Driven Proton Pump Bacteriorhodopsin”. In: *The Journal of Physical Chemistry B* 123.45 (Nov. 2019), pp. 9598–9608.

[143] Omar Valsson et al. “Rhodopsin absorption from first principles: Bypassing common pitfalls”. In: *Journal of Chemical Theory and Computation* 9.5 (May 2013), pp. 2441–2454.

[144] Paul Benjamin Woiczikowski et al. “Nonadiabatic QM/MM simulations of fast charge transfer in Escherichia coli DNA photolyase”. In: *Journal of Physical Chemistry B* 115.32 (Aug. 2011), pp. 9846–9863.

[145] Matteo Guglielmi et al. “Photodynamics of Lys+-Trp protein motifs: Hydrogen bonds ensure photostability”. In: *Faraday Discussions* 163.0 (July 2013), pp. 189–203.

[146] Ashlyn R. Murphy, Mark A. Hix, and Alice Walker. “Exploring the effects of mutagenesis on FusionRed using excited state QM/MM dynamics and classical force field simulations”. In: *ChemBioChem* (Feb. 2023).

[147] Kenneth M. Merz. “Using quantum mechanical approaches to study biological systems”. In: *Accounts of Chemical Research* 47.9 (Sept. 2014), pp. 2804–2811.

[148] Vinícius Wilian D. Cruzeiro et al. “TeraChem protocol buffers (TCPB): Accelerating QM and QM/MM simulations with a client–server model”. In: *The Journal of Chemical Physics* 158.4 (Jan. 2023).

[149] Madushanka Manathunga et al. “Quantum Mechanics/Molecular Mechanics Simulations on NVIDIA and AMD Graphics Processing Units”. In: *Journal of Chemical Information and Modeling* (Jan. 2023).

[150] Viacheslav Bolnykh et al. *MiMiC Communication Library (v2.0.1)*. 2022.

[151] Jógvan Magnus Haugaard Olsen et al. *MiMiC: A Framework for Multiscale Modeling in Computational Chemistry (v0.2.0)*. 2022.

[152] Sander Pronk et al. “GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit”. In: *Bioinformatics* 29.7 (Apr. 2013), pp. 845–854.

[153] Per Larsson, Berk Hess, and Erik Lindahl. “Algorithm improvements for molecular dynamics simulations”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.1 (Jan. 2011), pp. 93–108.

[154] David E. Shaw et al. “Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer”. In: *International Conference for High Performance Computing, Networking, Storage and Analysis, SC* 2015-January.January (Jan. 2014), pp. 41–53.

[155] Martin Kotev and Constantino Diaz Gonzalez. “Molecular Dynamics and Other HPC Simulations for Drug Discovery”. In: *Methods in molecular biology (Clifton, N.J.)* 2716 (2024), pp. 265–291.

[156] Viacheslav Bolnykh, Ursula Rothlisberger, and Paolo Carloni. “Biomolecular Simulation: A Perspective from High Performance Computing”. In: *Israel Journal of Chemistry* 60.7 (July 2020), pp. 694–704.

[157] Message Passing Interface Forum. *MPI: A Message-Passing Interface Standard Version 4.0*. Tech. rep. 2021.

[158] Tanmoy Chakraborty and Ramon Carbo-Dorca, eds. *Theoretical and Quantum Chemistry at the Dawn of the 21st Century*. Apple Academic Press, June 2018.

[159] Jürg Hutter and Alessandro Curioni. “Dual-level parallelism for ab initio molecular dynamics: Reaching teraflop performance with the CPMD code”. In: *Parallel Computing* 31.1 (Jan. 2005), pp. 1–17.

[160] Dominik Marx and Michele Parrinello. “Ab initio path-integral molecular dynamics”. In: *Zeitschrift für Physik B Condensed Matter* 95.2 (June 1994), pp. 143–144.

[161] James W. Cooley and John W. Tukey. “An algorithm for the machine calculation of complex Fourier series”. In: *Mathematics of Computation* 19.90 (1965), pp. 297–301.

[162] Jürg Hutter and Alessandro Curioni. “Car–Parrinello Molecular Dynamics on Massively Parallel Computers”. In: *ChemPhysChem* 6.9 (Sept. 2005), pp. 1788–1793.

[163] Valery Weber et al. “Shedding Light on Lithium/Air Batteries Using Millions of Threads on the BG/Q Supercomputer”. In: *2014 IEEE 28th International Parallel and Distributed Processing Symposium*. IEEE, May 2014, pp. 735–744.

[164] Tom Darden, Darrin York, and Lee Pedersen. “Particle mesh Ewald: An N log(N) method for Ewald sums in large systems”. In: *The Journal of Chemical Physics* 98.12 (June 1993), pp. 10089–10092.

[165] Berk Hess et al. “GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation”. In: *Journal of Chemical Theory and Computation* 4.3 (Mar. 2008), pp. 435–447.

[166] M. R.S. Pinches, D. J. Tildesley, and D. J. Tildesley. "Large Scale Molecular Dynamics on Parallel Computers using the Link-cell Algorithm". In: *Molecular Simulation* 6.1-3 (1991), pp. 51–87.

[167] Mark James Abraham et al. "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". In: *SoftwareX* 1-2 (Sept. 2015), pp. 19–25.

[168] Alessandro Laio, Joost VandeVondele, and Ursula Rothlisberger. "A Hamiltonian electrostatic coupling scheme for hybrid Car–Parrinello molecular dynamics simulations". In: *The Journal of Chemical Physics* 116.16 (Apr. 2002), p. 6941.

[169] Susana Gonçalves et al. "Induced fit and the catalytic mechanism of isocitrate dehydrogenase". In: *Biochemistry* 51 (36 Sept. 2012). Paper for PDB ID: 4AJ3, pp. 7098–7115.

[170] Christine E. Quartararo et al. "Structural, kinetic and chemical mechanism of isocitrate dehydrogenase-1 from mycobacterium tuberculosis". In: *Biochemistry* 52 (10 Mar. 2013), pp. 1765–1775.

[171] Ho Jin Koh et al. "Cytosolic NADP+-dependent isocitrate dehydrogenase plays a key role in lipid metabolism". In: *Journal of Biological Chemistry* 279.38 (Sept. 2004), pp. 39968–39974.

[172] Woo Suk Nam, Kwon Moo Park, and Jeen Woo Park. "RNA interference targeting cytosolic NADP+-dependent isocitrate dehydrogenase exerts anti-obesity effect in vitro and in vivo". In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1822.8 (Aug. 2012), pp. 1181–1188.

[173] Elena Bogdanovic et al. "IDH1 regulates phospholipid metabolism in developing astrocytes". In: *Neuroscience Letters* 582 (Oct. 2014), pp. 87–92.

[174] Seung Hee Jo et al. "Control of Mitochondrial Redox Balance and Cellular Defense against Oxidative Damage by Mitochondrial NADP+-dependent Isocitrate Dehydrogenase". In: *Journal of Biological Chemistry* 276.19 (May 2001), pp. 16168–16176.

[175] Su Min Lee et al. "Cytosolic NADP+-dependent isocitrate dehydrogenase status modulates oxidative damage to cells". In: *Free Radical Biology and Medicine* 32.11 (June 2002), pp. 1185–1196.

[176] D. A. Tennant et al. "Reactivating HIF prolyl hydroxylases under hypoxia results in metabolic catastrophe and cell death". In: *Oncogene 2009 28:45* 28.45 (Aug. 2009), pp. 4009–4021.

[177] Bryce W. Carey et al. "Intracellular α-ketoglutarate maintains the pluripotency of embryonic stem cells". In: *Nature 2014 518:7539* 518.7539 (Dec. 2014), pp. 413–416.

[178] Lenny Dang et al. "Cancer-associated IDH1 mutations produce 2-hydroxyglutarate". In: *Nature* 462.7274 (Dec. 2009), pp. 739–744.

[179] Michael Weller et al. “Glioma”. In: *Nature Reviews Disease Primers 2015 1:1* 1.1 (July 2015), pp. 1–18.

[180] Gabriel Alzial et al. “Wild-type isocitrate dehydrogenase under the spotlight in glioblastoma”. In: *Oncogene 2021 41:5* 41.5 (Nov. 2021), pp. 613–621.

[181] David N. Louis et al. “The 2021 WHO Classification of Tumors of the Central Nervous System: a summary”. In: *Neuro-Oncology* 23.8 (Aug. 2021), p. 1231.

[182] Michael Weller et al. “EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood”. In: *Nature Reviews Clinical Oncology 2020 18:3* 18.3 (Dec. 2020), pp. 170–186.

[183] Craig Horbinski. “What do we know about IDH1/2 mutations so far, and how do we use it?” In: *Acta Neuropathol.* 125.5 (May 2013), pp. 621–636.

[184] Hao Wu and Yi Zhang. “Reversing DNA methylation: mechanisms, genomics, and biological functions”. In: *Cell* 156.1-2 (2014), pp. 45–68.

[185] Junjie U. Guo et al. “Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain”. In: *Cell* 145.3 (Apr. 2011), pp. 423–434.

[186] Danielle Golub et al. “Mutant isocitrate dehydrogenase inhibitors as targeted cancer therapeutics”. In: *Frontiers in Oncology* 9.MAY (2019).

[187] Rachel Bar-Shalom, Ana Y. Valdivia, and M. Donald Blaufox. “PET imaging in oncology”. In: *Seminars in Nuclear Medicine* 30.3 (July 2000), pp. 150–185.

[188] M. Unterrainer et al. “Recent advances of PET imaging in clinical radiation oncology”. In: *Radiation Oncology 2020 15:1* 15.1 (Apr. 2020), pp. 1–15.

[189] Gejing Deng et al. “Selective inhibition of mutant isocitrate dehydrogenase 1 (IDH1) via disruption of a metal binding network by an allosteric small molecule”. In: *Journal of Biological Chemistry* 290.2 (Sept. 2015), pp. 762–774.

[190] Ujunwa C. Okoye-Okafor et al. “New IDH1 mutant inhibitors for treatment of acute myeloid leukemia”. In: *Nature Chemical Biology 2015 11:11* 11.11 (Oct. 2015), pp. 878–886.

[191] Rui Ma and Cai Hong Yun. “Crystal structures of pan-IDH inhibitor AG-881 in complex with mutant human IDH1 and IDH2”. In: *Biochemical and Biophysical Research Communications* 503.4 (Sept. 2018), pp. 2912–2917.

[192] Justin A. Caravella et al. “Structure-Based Design and Identification of FT-2102 (Olutasidenib), a Potent Mutant-Selective IDH1 Inhibitor”. In: *Journal of Medicinal Chemistry* 63.4 (Feb. 2020), pp. 1612–1623.

[193] Stefan Pusch et al. “Pan-mutant IDH1 inhibitor BAY 1436032 for effective treatment of IDH1 mutant astrocytoma in vivo”. In: *Acta Neuropathologica* 133.4 (Apr. 2017), pp. 629–644.

[194] Janeta Popovici-Muller et al. “Discovery of the first potent inhibitors of mutant IDH1 that lower tumor 2-HG in vivo”. In: *ACS Medicinal Chemistry Letters* 3.10 (Oct. 2012), pp. 850–855.

[195] Shuang Liu et al. “Roles of metal ions in the selective inhibition of oncogenic variants of isocitrate dehydrogenase 1”. In: *Communications Biology 2021 4:1* 4.1 (Nov. 2021), pp. 1–16.

[196] Niamh Coleman and Jordi Rodon. “Taking Aim at the Undruggable”. In: *American Society of Clinical Oncology Educational Book* 41 (June 2021), e145–e152.

[197] Xin Xie et al. “Recent advances in targeting the “undruggable” proteins: from drug discovery to clinical trials”. In: *Signal Transduction and Targeted Therapy 2023 8:1* 8.1 (Sept. 2023), pp. 1–71.

[198] Zhong Yin Zhang. “Drugging the undruggable: Therapeutic potential of targeting protein tyrosine phosphatases”. In: *Accounts of Chemical Research* 50.1 (Jan. 2017), pp. 122–129.

[199] Jill M. Bolduc et al. “Mutagenesis and laue structures of enzyme intermediates: Isocitrate dehydrogenase”. In: *Science* 268.5215 (1995), pp. 1312–1318.

[200] James H. Hurley et al. “Catalytic mechanism of NADP(+)-dependent isocitrate dehydrogenase: implications from the structures of magnesium-isocitrate and NADP+ complexes”. In: *Biochemistry* 30.35 (Sept. 1991), pp. 8671–8678.

[201] Neil B. Grodsky, Sambanthamurthy Soundar, and Roberta F. Colman. “Evaluation by site-directed mutagenesis of aspartic acid residues in the metal site of pig heart NADP-dependent isocitrate dehydrogenase”. In: *Biochemistry* 39.9 (Mar. 2000), pp. 2193–2200.

[202] Tae Kang Kim, Peychii Lee, and Roberta F. Colman. “Critical role of Lys212 and Tyr140 in porcine NADP-dependent isocitrate dehydrogenase.” In: *J. Biol. Chem.* 278.49 (Dec. 2003), pp. 49323–49331.

[203] Rui P.P. Neves, Pedro A. Fernandes, and Maria J. Ramos. “Unveiling the Catalytic Mechanism of NADP+-Dependent Isocitrate Dehydrogenase with QM/MM Calculations”. In: *ACS Catal.* 6.1 (Jan. 2016), pp. 357–368.

[204] Alan R. Rendina et al. “Mutant IDH1 Enhances the Production of 2-Hydroxyglutarate Due to Its Kinetic Mechanism”. In: *Biochemistry* 52 (26 July 2013), pp. 4563–4577.

[205] Katarzyna widerek et al. “Protein conformational landscapes and catalysis. Influence of active site conformations in the reaction catalyzed by L-lactate dehydrogenase”. In: *ACS Catalysis* 5.2 (Feb. 2015), pp. 1172–1185.

[206] Huo-Lei Peng et al. “Energy Landscape of the Michaelis Complex of Lactate Dehydrogenase: Relationship to Catalytic Mechanism”. In: *Biochemistry* 53.11 (Mar. 2014), pp. 1849–1857.

[207] Bei Yang et al. “Molecular mechanisms of “off-on switch” of activities of human IDH1 by tumor-associated mutation R132H”. In: *Cell Research* 20.11 (Nov. 2010), pp. 1188–1200.

[208] William Humphrey, Andrew Dalke, and Klaus Schulten. “VMD – Visual Molecular Dynamics”. In: *Journal of Molecular Graphics* 14 (1996), pp. 33–38.

[209] Schrödinger LLC. *The PyMOL Molecular Graphics System, Version 2.3.4*. Nov. 2015.

[210] Alan W. Sousa Da Silva and Wim F. Vranken. “ACPYPE - AnteChamber PYthon Parser interfacE”. In: *BMC Research Notes* 5.1 (July 2012), pp. 1–8.

[211] Till Kirsch et al. “Wavefunction-Based Electrostatic-Embedding QM/MM Using CFOUR through MiMiC”. In: *Journal of Chemical Theory and Computation* 18.1 (Jan. 2022), pp. 13–24.

[212] Pekka Mark and Lennart Nilsson. “Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K”. In: *J. Phys. Chem. A* 105.43 (2001).

[213] Robert B. Best and Gerhard Hummer. “Optimized Molecular Dynamics Force Fields Applied to the HelixCoil Transition of Polypeptides”. In: *J. Phys. Chem. B* 113 (26 July 2009), pp. 9004–9015.

[214] Kresten Lindorff-Larsen et al. “Improved side-chain torsion potentials for the Amber ff99SB protein force field”. In: *Proteins* 78 (8 June 2010), pp. 1950–1958.

[215] Niklas Holmberg, Ulf Ryde, and Leif Bülow. “Redesign of the coenzyme specificity in l-Lactate dehydrogenase from Bacillus stearothermophilus using site-directed mutagenesis and media engineering”. In: *Protein Eng. Des. Sel.* 12 (10 Oct. 1999), pp. 851–856.

[216] Junmei Wang et al. “Development and testing of a general amber force field”. In: *J. Comput. Chem.* 25 (9 July 2004), pp. 1157–1174.

[217] Junmei Wang et al. “Automatic atom type and bond type perception in molecular mechanical calculations”. In: *J. Mol. Graphics Modell.* 25 (2 Oct. 2006), pp. 247–260.

[218] Glenn J. Martyna and Mark E. Tuckerman. “A reciprocal space based method for treating long range interactions in ab initio and force-field-based calculations in clusters”. In: *The Journal of Chemical Physics* 110.6 (Feb. 1999), pp. 2810–2821.

[219] Damian Alvarez. “JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre”. In: *Journal of large-scale research facilities JLSRF* 7 (Oct. 2021).

[220] E. A. Carter et al. “Constrained reaction coordinate dynamics for the simulation of rare events”. In: *Chem. Phys. Lett.* 156 (5 Apr. 1989), pp. 472–477.

[221] Paolo Carloni, Michiel Sprik, and Wanda Andreoni. “Key Steps of the cis-Platin-DNA Interaction: Density Functional Theory-Based Molecular Dynamics Simulations”. In: *J. Phys. Chem. B* 104 (4 Feb. 2000), pp. 823–835.

[222] Raphael Reinbold et al. “Resistance to the isocitrate dehydrogenase 1 mutant inhibitor ivosidenib can be overcome by alternative dimer-interface binding inhibitors”. In: *Nature Communications 2022 13:1* 13.1 (Aug. 2022), pp. 1–12.

[223] Xiao Liu et al. “Natural and synthetic 2-oxoglutarate derivatives are substrates for oncogenic variants of human isocitrate dehydrogenase 1 and 2”. In: *Journal of Biological Chemistry* 299.2 (Feb. 2023), pp. 102873–102874.

[224] Heping Zheng et al. “CheckMyMetal: a macromolecular metal-binding validation tool”. In: *Acta Crystallographica. Section D, Structural Biology* 73.Pt 3 (Mar. 2017), p. 223.

[225] Fanwang Meng et al. “A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors”. In: *Scientific Data 2021 8:1* 8.1 (Oct. 2021), pp. 1–11.

[226] Steven L Dixon, Alexander M Smondyrev, and Shashidhar N Rao. “PHASE: a novel approach to pharmacophore modeling and 3D database searching”. In: *Chemical biology & drug design* 67.5 (2006), pp. 370–372.

[227] Christopher A. Lipinski et al. “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings”. In: *Advanced Drug Delivery Reviews* 46.1-3 (Mar. 2001), pp. 3–26.

[228] Stefano Rusconi. “Alovudine Medivir”. In: *Current opinion in investigational drugs (London, England : 2000)* 4.2 (Feb. 2003), pp. 219–223.

[229] Dana Yehudai et al. “The thymidine dideoxynucleoside analog, alovudine, inhibits the mitochondrial DNA polymerase γ, impairs oxidative phosphorylation and promotes monocytic differentiation in acute myeloid leukemia”. In: *Haematologica* 104.5 (Apr. 2019), p. 963.

[230] Lukas B. Been et al. “[18F]FLT-PET in oncology: Current status and opportunities”. In: *European Journal of Nuclear Medicine and Molecular Imaging* 31.12 (Dec. 2004), pp. 1659–1672.

[231] V. R. Bollineni et al. “A systematic review on [18F]FLT-PET uptake as a measure of treatment response in cancer patients”. In: *European Journal of Cancer* 55 (Mar. 2016), pp. 81–97.

[232] Bénédicte Franck et al. “Pharmacokinetics, Pharmacodynamics, and Therapeutic Drug Monitoring of Valganciclovir and Ganciclovir in Transplantation”. In: *Clinical Pharmacology & Therapeutics* 112.2 (Aug. 2022), pp. 233–276.

[233] Donna McTavish, Ronald A. Young, and Stephen P. Clissold. “Cadralazine: A Review of its Pharmacodynamic and Pharmacokinetic Properties, and Therapeutic Potential in the Treatment of Hypertension”. In: *Drugs* 40.4 (Oct. 1990), pp. 543–560.

[234] Henry H. Freedman, Alfred E. Fox, and R. Suzanne Willis. “Influence of Chloramphenicol and Cetophenicol on Antibody Formation in Mice”. In: *Exp. Biol. Med. (Maywood)* 129.3 (Dec. 1968), pp. 796–799.

[235] H. Shoda et al. “Inhibitory Effects of Tenilsetam on the Maillard Reaction”. In: *Endocrinology* 138.5 (May 1997), pp. 1886–1892.

[236] J A Milligan and D S O’Doherty. “Experiences with diphoxazide, a new anticonvulsant drug”. In: *Med. Ann. Dist. Columbia* 30 (Sept. 1961), pp. 513–515.

[237] Lorton A. Schwartz and Edward Postma. “Metabolites of demoxepam, a chlordiazepoxide metabolite, in man”. In: *Journal of Pharmaceutical Sciences* 61.1 (Jan. 1972), pp. 123–125.

[238] Michelle I. Wilde and Heather D. Langtry. “Zidovudine: An Update of its Pharmacodynamic and Pharmacokinetic Properties, and Therapeutic Efficacy”. In: *Drugs* 46.3 (Sept. 1993), pp. 515–578.

[239] Andrew P. Lea and Diana Faulds. “Stavudine: A Review of its Pharmacodynamic and Pharmacokinetic Properties and Clinical Potential in HIV Infection”. In: *Drugs* 51.5 (Oct. 1996), pp. 846–864.

[240] Nombulelo Magula and Martin Dedicoat. “Low dose versus high dose stavudine for treating people with HIV infection”. In: *Cochrane Database of Systematic Reviews* 2017.6 (Jan. 2015).

[241] Marco Orrú et al. “Psychostimulant pharmacological profile of paraxanthine, the main metabolite of caffeine in humans”. In: *Neuropharmacology* 67 (Apr. 2013), pp. 476–484.

[242] Arpad Dobolyi et al. “Uridine Function in the Central Nervous System”. In: *Current Topics in Medicinal Chemistry* 11.8 (Apr. 2011), pp. 1058–1067.

[243] Jen Brogan. *Merck announces partnerships worth $1.3bn with BenevolentAI and Exscientia*. Sept. 2023.

[244] Alex McFarland. *Iktos Secures €15.5 Million in Funding to Accelerate AI Drug Discovery*. Mar. 2023.

[245] Craig Rhodes. “Qubit Pharmaceuticals Accelerates Drug Discovery With Hybrid Quantum Computing”. In: *NVIDIA Blogs* (Nov. 2022).

[246] Michael Peel. *AstraZeneca ties up with AI biologics company to develop cancer drug*. Dec. 2023.

[247] Dominic A Rufa et al. “Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning / molecular mechanics potentials”. In: *bioRxiv* 0.Mm (2020), p. 2020.07.29.227959.

[248] Phillip S. Hudson et al. “Obtaining QM/MM binding free energies in the SAMPL8 drugs of abuse challenge: indirect approaches”. In: *J. Comput.-Aided Mol. Des.* 36.4 (Apr. 2022), pp. 263–277.

[249] Andrea Rizzi, Paolo Carloni, and Michele Parrinello. “Multimap targeted free energy estimation”. In: Cv (2023).

[250] David Schneider. “The Exascale Era is Upon Us: The Frontier supercomputer may be the first to reach 1,000,000,000,000,000,000 operations per second”. In: *IEEE Spectrum* 59.1 (Jan. 2022), pp. 34–35.

[251] Charles Q. Choi. “The Beating Heart of the World’s First Exascale Supercomputer”. In: *IEEE Spectrum* (June 2022).

[252] James C. Phillips et al. “Scalable molecular dynamics on CPU and GPU architectures with NAMD”. In: *J. Chem. Phys.* 153.4 (July 2020), p. 044130.

[253] Andreas W. Götz et al. “Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born”. In: *J. Chem. Theory Comput.* 8.5 (May 2012), pp. 1542–1555.

[254] Romelia Salomon-Ferrer et al. “Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald”. In: *J. Chem. Theory Comput.* 9.9 (Sept. 2013), pp. 3878–3888.

[255] Szilárd Páll et al. “Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS”. In: *Solving Software Challenges for Exascale*. Ed. by Stefano Markidis and Erwin Laure. Cham: Springer International Publishing, June 2015, pp. 3–27.

[256] Thomas D. Kühne et al. “CP2K: An electronic structure and molecular dynamics software package -Quickstep: Efficient and accurate electronic structure calculations”. In: *J. Chem. Phys.* 152.19 (2020).

[257] “From NWChem to NWChemEx: Evolving with the Computational Chemistry Landscape”. In: *Chem. Rev.* 121.8 (Apr. 2021), pp. 4962–4998.

[258] Paolo Giannozzi et al. “Quantum ESPRESSO toward the exascale”. In: *J. Chem. Phys.* 152.15 (Apr. 2020), p. 154105.

[259] M. Manathunga et al. *QUICK-22.03, University of California San Diego, CA and Michigan State University, East Lansing, MI, 2022.*

[260] “Roadmap on Electronic Structure Codes in the Exascale Era”. In: (2022).

[261] Ivan S. Ufimtsev and Todd J. Martínez. “Graphical Processing Units for Quantum Chemistry”. In: *Comput. Sci. Eng.* 10.6 (Nov. 2008), pp. 26–34.

[262] “A direct-compatible formulation of the coupled perturbed complete active space self-consistent field equations on graphical processing units”. In: *J. Chem. Phys.* 146.17 (May 2017), p. 174113.

[263] Robert Schade et al. “Towards electronic structure-based ab-initio molecular dynamics simulations with hundreds of millions of atoms”. In: *Parallel Comput.* 111.January (July 2022), p. 102920.

[264] Yuji Sugita and Yuko Okamoto. “Replica-exchange molecular dynamics method for protein folding”. In: *Chem. Phys. Lett.* 314.1-2 (Nov. 1999), pp. 141–151.

[265] Sander Pronk et al. “Molecular Simulation Workflows as Parallel Algorithms: The Execution Engine of Copernicus, a Distributed High-Performance Computing Platform”. In: *J. Chem. Theory Comput.* 11.6 (June 2015), pp. 2600–2608.

[266] Davide Mandelli, Barak Hirshberg, and Michele Parrinello. “Metadynamics of Paths”. In: *Phys. Rev. Lett.* 125.2 (July 2020), p. 026001.

[267] William Martin et al. “Interpretable artificial intelligence and exascale molecular dynamics simulations to reveal kinetics: Applications to Alzheimer’s disease”. In: *Curr. Opin. Struct. Biol.* 72 (Feb. 2022), pp. 103–113.

[268] Frank Noé et al. “Machine learning for molecular simulation”. In: *Annu. Rev. Phys. Chem.* 71 (2020), pp. 361–390.

[269] Denghui Lu et al. “86 PFLOPS Deep Potential Molecular Dynamics simulation of 100 million atoms with ab initio accuracy”. In: *Comput. Phys. Commun.* 259 (2021), p. 107624.

[270] Albert Musaelian et al. “Learning local equivariant representations for large-scale atomistic dynamics”. In: *Nat. Commun.* 14.1 (2023), p. 579.

[271] Shae-Lynn J Lahey and Christopher N Rowley. “Simulating protein–ligand binding with neural network potentials”. In: *Chem. Sci.* 11.9 (2020), pp. 2362–2368.

[272] Xiaoliang Pan et al. “Machine-learning-assisted free energy simulation of solution-phase and enzyme reactions”. In: *J. Chem. Theory Comput.* 17.9 (2021), pp. 5745–5758.

[273] Mingyuan Xu, Tong Zhu, and John ZH Zhang. “Automatically constructed neural network potentials for molecular dynamics simulation of zinc proteins”. In: *Front. Chem.* 9 (2021), p. 692200.

[274] Raimondas Galvelis et al. “NNP/MM: Fast molecular dynamics simulations with machine learning potentials and molecular mechanics”. In: *arXiv* (2022).

[275] Kara K Grotz, Sergio Cruz-León, and Nadine Schwierz. “Optimized magnesium force field parameters for biomolecular simulations with accurate solvation, ion-binding, and water-exchange properties”. In: *Journal of chemical theory and computation* 17.4 (2021), pp. 2530–2540.

[276] Yongguang Zhang et al. “Rational design of nonbonded point charge models for divalent metal cations with Lennard-Jones 12-6 potential”. In: *Journal of Chemical Information and Modeling* 61.8 (2021), pp. 4031–4044.

[277] Shuang Liu et al. “Differentiating Inhibition Selectivity and Binding Affinity of Isocitrate Dehydrogenase 1 Variant Inhibitors”. In: *Journal of Medicinal Chemistry* 66.7 (Apr. 2023), pp. 5279–5288.

[278] Bharath Raghavan et al. “MiMiCPy: An Efficient Toolkit for MiMiC-Based QM/MM Simulations”. In: *Journal of Chemical Information and Modeling* 63.5 (Mar. 2023), pp. 1406–1412.

[279] Navraj S. Nagra et al. “The company landscape for artificial intelligence in large-molecule drug discovery”. In: *Nature Reviews Drug Discovery* 22.12 (Dec. 2023), pp. 949–950.

[280] Anastasiia V. Sadybekov and Vsevolod Katritch. “Computational approaches streamlining drug discovery”. In: *Nature* 616.7958 (Apr. 2023), pp. 673–685.

[281] Riccardo Capelli et al. “Accuracy of Molecular Simulation-Based Predictions of k off Values: A Metadynamics Study”. In: *J. Phys. Chem. Lett.* 11.15 (Aug. 2020), pp. 6373–6381.

[282] Katya Ahmad et al. “Enhanced-Sampling Simulations for the Estimation of Ligand Binding Kinetics: Current Status and Perspective”. In: *Front. Mol. Biosci* 9.June (June 2022), pp. 1–17.