

Advancing first principle-based molecular dynamics of biological systems with machine learning

Présentée le 22 septembre 2023

Faculté des sciences de base
Laboratoire de chimie et biochimie computationnelles
Programme doctoral en physique

pour l'obtention du grade de Docteur ès Sciences

par

François Louis MOUVET

Acceptée sur proposition du jury

Prof. J. H. Dil, président du jury
Prof. U. Röthlisberger, directrice de thèse
Prof. M. Tuckerman, rapporteur
Prof. J. M. Hugaard Olsen, rapporteur
Prof. Ph. Schwaller, rapporteur

Acknowledgements

First, I wish to express my gratitude towards Professor Ursula Röthlisberger, my supervisor, for welcoming me into her group. Thank you for the insightful discussions and for creating an environment where I was allowed to explore my ideas. I also want to thank the LCBC postdocs, Ruby, Quanta and Mocca, for never missing a project update meetings with their comforting presence.

I would like to express my gratitude to the members of my thesis committee, Prof. Mark Tuckerman, Prof. J. Magnus Haugaard Olsen, Prof. Philippe Schwaller and Prof. Hugué Dil. Thank you for providing critical feedback to my work and the engaging discussions during the defence.

This was a long PhD and I was lucky to cross the path of many individuals that contributed to the vibrant atmosphere at LCBC. I would like to thank all LCBC members, old and new, for the great times we shared together. I would like to give a special thanks to Karin Pasche for her efficient and constant support that often went beyond my administrative struggles. I would like to thank Martin Bircher for the great times during my first years. Our discussions taught me a lot about science and how important coffee breaks are. I also want to thank Murat Kılıç for his warm welcome into our BCH 4111 and for his help during my first steps into this field. A HUGE thank goes to Justin Villard, the other Dupont (according to Karin). Thank you for the daily laughs in the office, the always spot-on scientific advice (*"J'ai bien vu que ça allait pas marcher... Mais j'ai rien dit!"*) and your unwavering support during challenging times.

Fortunately, there are people that reminded me that life continues outside of working hours. I want to thank the Agone crew, Loic, Quentin, Guillaume and Jeremy for our long evenings of wandering the Harmonde. Joining your group was a critical success, to the point that there seems to be no good dice rolls left for my character. Alban for carrying our gaming nights with his *support*. Finally, the old gang, Nicho, Lisa, Gael and Marija, with whom I shared my worries over countless binouses and aperos over the last 16 years. Thank you all!

Many thanks to the Medici family, Sonya, Giovanni, Vasco, Elisa and Matilde, for making me feel at home in Ticino during much needed breaks. Grazie!

I want to give my warmest thanks to my family for their invaluable love and support during these many years. To my parents, Lydia and Laurent (Mut' and Pap'), for providing their

Acknowledgements

unrivaled *soutien au sportif* and to my sister Vic and her husband Thibaud for their support. I want to express all my love to my nephew Étienne that brings so much life and joy into every room he enters.

Finally, I can hardly express enough gratitude towards Siria as she brilliantly fulfills so many crucial roles in this project and in my life. Your love, support and advice during these long years made this PhD possible. Thank you for everything. ♡

Lausanne, August 31, 2023

François Mouvet

Abstract

Molecular dynamics (MD) simulations have emerged as a transformative approach to analyse molecular systems at the atomic level, offering valuable insights into complex biological processes. Many biological phenomena can only accurately be described by incorporating a quantum-mechanical (QM) description of atomic interactions, known as first-principles MD (FPMD). However, their computational cost precludes the simulation of large systems without compromising simulation time or accuracy. In MD simulations, the time step is limited by the fastest motions of the system. Multiple time step (MTS) algorithms mitigate this limitation by integrating the fast and slow force components with different time steps. In FPMD, two distinct QM methods can be used to capture these force contributions. A low-level method defines the fast force components and its difference with forces computed with a high-level method serves as the slow components.

Throughout this thesis project, we have used and developed diverse MD techniques, with a specific emphasis on MTS and biological applications.

The first project covers a preclinical investigation of drug candidates against the infection schistosomiasis. In close collaboration with experimental chemists and biologists, we provided computational insights on the mode of action of these putative drugs. Our simulations revealed their diverse binding poses in the target proteins resulting in different frequency of near-attack configurations of the reactive groups activating the drug. This finding could explain the different *in vitro* activities against schistosome species. However, all drugs proved unstable in acidic environments, precluding *in vivo* activity.

Then, we switched to the further development of MD methods by pursuing an ongoing project where fewest-switches surface hopping was combined with MTS to accelerate non-adiabatic MD simulations. This method computes Tully's transition probabilities at the outer steps and the Landau-Zener formula is used to detect transitions during the inner steps, that are then confirmed with a high-level fewest switches calculation. The method was successfully tested on a small prototypical system, the photorelaxation of protonated formalimine.

Next, we focus on using machine learning (ML) to infer forces during MTS simulations. We investigate two schemes. In the first, ML provides an estimate of the slow force components to bypass the high-level calculations. This method yields large speedups of ~ 163 at the cost of

Abstract

sampling phase space according to an approximation of the high-level method. In the second, ML infers a correction to the fast force components to reduce the gap between the two levels and thus allowing large increases of the outer time step for speedups of ~ 7 while sampling phase space according to the high-level method is guaranteed. Both schemes accurately reproduced the structure of liquid water.

Finally, we expand the ML-MTS approach by adding two significant improvements. First, we successfully incorporate the second ML-MTS scheme into a QM/MM framework. In addition, we develop an adaptive ML-MTS algorithm which enables on-the-fly retraining of the ML model based on the kernel-induced distance between new and current training configurations. The MTS ratio is then dynamically adjusted to optimize the use of high-level calculations. We successfully test this method on a molecule of acetone solvated in water and a small metalloprotein in aqueous solution.

Keywords: computational chemistry, multiple time step integration, machine learning, drug discovery, density functional theory, Born-Oppenheimer molecular dynamics, classical molecular dynamics, QM/MM.

Résumé

Les simulations de dynamique moléculaire (MD) sont apparues comme une approche novatrice permettant d'analyser les systèmes moléculaires au niveau atomique, offrant ainsi des informations précieuses sur des processus biologiques complexes. De nombreux phénomènes biologiques ne peuvent être décrits avec précision qu'en incorporant une description quantique (QM) des interactions atomiques, connue sous le nom de *ab initio* MD (AIMD). Cependant, le coût des calculs engendré par ces équations empêche la simulation de grands systèmes sans faire de compromis sur le temps total de la simulation ou sa précision. Dans les simulations MD, le pas de temps utilisé pour intégrer les équations du mouvement est limité par les mouvements les plus rapides du système. Les algorithmes d'intégration à pas de temps multiples (MTS) atténuent cette limitation en intégrant les composantes rapides et lentes de la force en utilisant des pas de temps différents. Dans le contexte de l'AIMD, deux méthodes QM distinctes peuvent être utilisées pour modéliser ces contributions de force. Une méthode de bas niveau définit la composante rapide de la force, tandis que sa différence avec les forces calculées avec une méthode de haut niveau sert de composante lente.

Au long de ce projet de thèse, nous avons utilisé et développé diverses techniques de MD, en mettant l'accent sur les algorithmes MTS et l'étude de systèmes biologiques.

Le premier projet traite d'une étude préclinique de potentiels nouveaux médicaments contre la schistosomiase, une maladie infectieuse parasitaire négligée. En étroite collaboration avec des chimistes expérimentaux et des biologistes, nous avons fourni des informations computationnelles sur le mode d'action de ces médicaments. Nos simulations ont révélé des variations dans la manière dont ces molécules se lient à leurs protéines cibles, se traduisant par des configurations plus ou moins favorables à la réaction nécessaire pour l'activation du médicament. Cette constatation pourrait expliquer les différences d'activités observées *in vitro* contre les différentes espèces de schistosomes. Cependant, tous les médicaments se sont révélés instables dans des environnements acides, ce qui entrave leur activité *in vivo*.

Nous nous sommes ensuite concentrés sur le développement de nouvelles approches pour accélérer les simulations MD. Nous avons commencé par poursuivre un projet en cours couplant l'algorithme MTS avec la méthode "fewest-switches surface hopping" pour accélérer les simulations MD non adiabatiques. Notre méthode calcule les probabilités de transition d'un état d'excitation à l'autre avec la méthode de Tully lors de l'intégration des composantes lentes

Résumé

de la force et utilise la formule de Landau-Zener pour détecter les transitions au cours des étapes intermédiaires. Ces transitions sont ensuite confirmées par un calcul de haut niveau. La méthode a été testée avec succès sur un petit système prototype, la photorelaxation de la forme protonée de la formaldimine.

Ensuite, nous explorons l'utilisation d'une méthode d'apprentissage automatique (ML) pour inférer des composantes de forces pendant les simulations MTS (ML-MTS). Nous considérons deux schémas. Dans le premier, l'apprentissage automatique fournit une estimation des composantes lentes de la force afin de contourner les calculs de haut niveau. Cette méthode permet d'obtenir des accélérations importantes de ~ 163 au prix d'un échantillonnage de l'espace des phases selon une approximation de la méthode de haut niveau. Dans la seconde, la méthode ML déduit une correction des composantes rapides de la force pour réduire l'écart entre les deux niveaux, permettant ainsi de grandes augmentations du pas de temps. Ceci permet d'accélérer les calculs d'un facteur ~ 7 tout en garantissant l'échantillonnage de l'espace de phase selon la méthode de haut niveau. Les deux schémas ont reproduit avec précision la structure de l'eau liquide.

Finalement, nous apportons deux améliorations significatives à l'approche ML-MTS. Tout d'abord, nous intégrons avec succès le deuxième schéma ML-MTS dans un modèle hybride traitant une partie du système avec des méthodes issues de la mécanique quantique et le reste avec de la MD classique. En outre, nous développons un algorithme ML-MTS adaptatif qui permet de ré-entraîner le modèle ML au cours d'une simulation. Le critère définissant la nécessité d'un ré-entraînement est basé sur la différence entre les nouvelles configurations d'entraînement et la configuration actuelle. Cette différence est mesurée par une métrique basée sur la fonction noyau des descripteurs de l'environnement atomique. Le pas d'intégration des composantes lentes de la force est alors ajusté dynamiquement pour optimiser l'utilisation des calculs de haut niveau. Nous montrons l'efficacité de cette méthode en simulant une molécule d'acétone dans l'eau et une petite métalloprotéine en solution aqueuse.

Contents

Acknowledgements	i
Abstract (English/Français)	iii
1 Introduction	1
1.1 Motivation	3
1.2 Thesis layout	3
2 Theory	7
2.1 Describing the motion of N interacting bodies	8
2.1.1 Hamiltonian classical mechanics	8
2.1.2 Poisson brackets and Liouville operator	10
2.1.3 Generating integration methods	11
2.1.4 Deriving the velocity Verlet algorithm	13
2.2 Classical molecular dynamics	15
2.3 Born-Oppenheimer molecular dynamics	16
2.4 Density Functional Theory	20
2.4.1 Kohn-Sham DFT	22
2.5 Approximations of the exchange-correlation functional	24
2.5.1 Local Density Approximation (LDA)	24
2.5.2 Generalized Gradient Approximation (GGA)	24
2.5.3 Hybrid functionals	25
3 Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis	27
3.1 Abstract	28
3.2 Introduction	28
3.3 Results and Discussion	30
3.3.1 In vitro studies	30
3.3.2 Studies on juvenile <i>S. mansoni</i> in the mouse model	31
3.3.3 In vivo studies on adult <i>S. haematobium</i>	31
3.3.4 Computational studies	32
3.3.5 Stability of OXA Analogues in acidic environments and in the presence of microsomes	34

Contents

3.3.6	Lipid nanocapsules loaded with Ph-CH ₂ -OXA	36
3.4	Conclusion	37
3.5	Experimental and computational details	37
3.6	Supplementary Figures	44
4	A multiple time step algorithm for trajectory surface hopping simulations	53
4.1	Abstract	54
4.2	Introduction	54
4.3	Theory	55
4.3.1	Trajectory surface hopping in CPMD	55
4.3.2	Trajectory surface hopping with multiple time step scheme	57
4.3.3	Standard MTS algorithms	58
4.3.4	MTS algorithm for trajectory surface hopping dynamics	60
4.4	Results and Discussion	62
4.4.1	Investigating different approximations for the non-adiabatic couplings	62
4.4.2	Comparing single trajectories via deterministic surface hopping	64
4.4.3	Stochastic surface hopping	69
4.5	Conclusions and outlook	71
4.6	Supplementary figures	73
5	Machine Learning-Enhanced Multiple Time-Step <i>Ab Initio</i> Molecular Dynamics	75
5.1	Abstract	76
5.2	Introduction	76
5.3	Theory	78
5.3.1	<i>Ab Initio</i> Nuclear Forces from Machine Learning	78
5.3.2	Multiple Time Step Molecular Dynamics	80
5.3.3	Implementation	81
5.4	Computational Details	82
5.4.1	Training Data: Sample Selection	83
5.4.2	ML Model: Kernel Function and FCHL19 Parameters	83
5.5	Results and Discussion	84
5.5.1	Model Performance	84
5.5.2	Energy Conservation	87
5.5.3	Structural Properties	88
5.5.4	Efficiency assessment	90
5.6	Conclusion	91
6	Multiple Time Step QM/MM Molecular Dynamics Enhanced With On-The-Fly Trained Machine Learning	93
6.1	Introduction	94
6.2	Theory	96
6.2.1	Machine learning-enhanced multiple time step MD	96
6.2.2	QM/MM MD	96

6.3	Methods	97
6.3.1	ML-MTS in QM/MM MD	97
6.3.2	Description of atomic environments	98
6.3.3	Adaptive scheme	100
6.3.4	Training Data Selection	100
6.3.5	Confidence criterion	102
6.3.6	Software	103
6.4	Results and Discussion	103
6.4.1	Simulation setup	103
6.4.2	ML-MTS for QM/MM	105
6.4.3	Adaptive ML-MTS	107
6.5	Conclusions and outlook	110
7	Conclusions and outlooks	113
7.1	Main results and outlooks	113
7.1.1	Computational study of drug candidates	113
7.1.2	Multiple time step algorithms	114
7.2	Final word	116
	Bibliography	117
	Curriculum Vitae	135

1 Introduction

I knew exactly what to do. But in a much more real sense, I had no idea what to do.

Michael G. Scott, *The Office*

The pursuit of discovering new drugs is an arduous and complex task that has long captivated the minds of scientists and researchers. The urgent need for effective therapies to combat diseases, coupled with the escalating challenges associated with traditional drug discovery methods, has spurred the rapid advancement of computational approaches in the domain of biochemistry. This has paved the way for a new era of drug discovery, where the power of computational methods and molecular simulations holds great promise in revolutionizing the way we identify and develop novel bioactive agents.

The process of finding new drugs typically involves the identification of biologically active compounds that exhibit desirable pharmacological properties, such as potency, selectivity, and safety [1]. However, the journey from initial target identification to the development of a clinically viable drug is full of obstacles, including high costs, lengthy timelines, and limited success rates [2]. The traditional drug discovery paradigm heavily relies on high-throughput screening, chemical synthesis, and experimental testing of large compound libraries, presenting formidable challenges in terms of time, resources, and ethical considerations.

Computational chemistry has revolutionized our understanding of biological processes by providing the means to simulate and analyze molecular systems at the atomic level thus unveiling valuable insights and empowering researchers to engage in rational drug design [3, 4]. By harnessing the power of computational algorithms, researchers can expedite the drug discovery process by virtually screening vast libraries of compounds, predicting their interactions with target biomolecules, and optimizing their properties to enhance efficacy and safety [5].

Over the last few decades, the field of computational chemistry has evolved at a rapid pace.

Introduction

Thanks to advancements in computational power and techniques, there has been an immense increase in the size of systems that can be studied with very accurate methods, starting from a handful of heavy atoms in the 90s to current complicated systems, such as full protein complexes involving metal centers. This significant expansion in the size of systems under study has opened up new avenues for scientific investigation, allowing researchers to delve into intricate molecular structures and to better understand the underlying phenomena.

Molecular dynamics (MD) is a powerful technique that enables the study of biological processes occurring under physiological conditions at finite temperature. It allows to draw an accurate picture of the motion of atoms, thus providing real-time monitoring of all atomic interactions. As a result, MD offers valuable insights into the intricate interactions that are often beyond the reach of traditional microscopy techniques due to their scale or speed limitations.

One of the challenges in molecular dynamics is the selection of an appropriate level of theory to describe the molecular system. The accuracy of the simulations heavily depends on the level of theory chosen, but more accurate methods typically come at a higher computational cost. Large systems are generally modelled by classical force fields, that offer computational efficiency but fail to capture the intricate quantum mechanical effects that are sometimes crucial for accurately describing biochemical systems. On the other hand, *ab initio* methods (based on quantum physics) such as Density Functional Theory (DFT) provide a more accurate treatment of the electronic structure but can become prohibitively computationally demanding, especially for large systems. Therefore, there is a need to develop efficient and reliable strategies to strike a balance between accuracy and computational cost. This includes the development of hybrid models that combine accurate quantum mechanical methods for specific regions of interest with computationally cheaper classical force fields for the rest of the system. Such hybrid quantum mechanical/molecular mechanical (QM/MM) approaches can significantly reduce the computational expense while still capturing the essential quantum mechanical effects [6].

An additional path to reduce the cost of *ab initio* molecular dynamics is to act on the integration time step. This time step is usually limited by the fastest vibrational motion in the system, which imposes very frequent expensive force computations using quantum methods. To overcome this limitation, multiple time step algorithms (MTS) have been developed [7]. In this approach, the forces acting on a system are partitioned according to their characteristic frequencies, and different time steps are used to integrate the motion induced by these forces. These methods were originally developed in the context of classical molecular dynamics, where the segmentation of forces is more straightforward. In *ab initio* molecular dynamics, this separation is much less obvious and a plethora of variations flourished during the last years [8, 9, 10, 11, 12].

In parallel, machine learning (ML) methods have emerged as powerful tools also in the field of *ab initio* molecular dynamics, particularly for the computation of nuclear forces. Machine learning approaches, such as neural networks and Gaussian process regression,

offer an alternative to traditional quantum methods by learning the mapping between atomic configurations and corresponding forces from a training dataset obtained from quantum mechanical calculations [13, 14, 15]. Once trained, these models can rapidly predict forces for new (sufficiently similar) atomic configurations that can then be used in simulations [16, 17, 18, 19].

1.1 Motivation

The original goal of this PhD project was to utilize computational methods to contribute to the understanding of novel mechanisms in large biosystems, with the potential to rationalise the action of different drugs and help in the discovery of new ones. After gaining initial experience in this field of application through a preclinical investigation of a group of drug candidates, the methods and their limitations gained more of my attention, prompting subsequent projects that prioritized the development of innovative algorithms for simulating biosystems using approaches based on quantum mechanics.

In particular, we have chosen to concentrate on harnessing the potential of multiple time step integrators in various complex simulation scenarios that are commonly encountered in biological systems. These scenarios may involve simulations that describe non-adiabatic effects or contain a significant number of atoms requiring accurate and expensive quantum mechanics-based methods for their description.

Building upon the remarkable accomplishments of artificial intelligence (AI) algorithms in the field of chemistry, we aim to integrate some of these methods into the MTS algorithms. A first tempting approach would be to directly substitute high-level calculations entirely with an ML force inference thus achieving very large accelerations. Nevertheless, it is crucial to acknowledge that results derived from AI-based simulations may raise concerns due to the absence of physics-based calculations and the associated limitations and underlying assumptions. Our approach proposed here involves utilizing the MTS framework as a mean to prioritize physics as the primary driving force behind the simulations and ensure an automatic quality control. Here, the incorporation of ML forces is intended solely to enhance the time step, ensuring a balance between computational efficiency and maintaining physical accuracy. In this combination, ML and MTS algorithms can greatly improve the possibilities for simulating complex biological systems.

1.2 Thesis layout

This thesis begins with an introduction to the theoretical concepts (presented in chapter two) that serve as the fundamental basis for the research presented. Each subsequent chapter will then delve into a distinct research project, allowing for a focused exploration of different aspects within the broader scope of the thesis.

Introduction

The third chapter of this thesis focuses on a preclinical investigation of drug candidates for the treatment of schistosomiasis, a deadly neglected tropical infectious disease caused by three different species of schistosomes. This infection affects over 200 million people worldwide [20, 21, 22], mostly in developing countries. Schistosomiasis impairs both the physical and cognitive development of children, provokes organ failures in adults and can ultimately cause death. Currently, only one drug (Praziquantel) is actively used due to its effect on all three species of schistosomes. However, resistances are emerging and alternative drugs are needed. Some organometallic derivatives of the drug oxamniquine were designed as novel candidates by experimentalists from the groups of Prof. Gilles Gasser (PSL University, Paris) and Prof. Jennifer Keiser (Swiss TPH, Basel) with the aim of improving the efficacy of oxamniquine and broadening its use against the different species of schistosomes. Their research yielded promising results, as they identified molecules that demonstrated enhanced *in vitro* activity against all species of schistosomes [23]. With a collaborative project, comprehensive *in vitro*, *in vivo*, and computational investigations of these potential drugs were conducted to characterize and rationalize their biological activity. Our primary contribution was to explore the mode of action of these drug candidates using classical molecular dynamics simulations. By applying this computational approach, we gained insights into how these candidates interact with their target molecules and were able to suggest reasons for the observed differences in their potencies.

Following this experience, we moved towards the development of novel methods using multiple time step integration to reduce the cost of *ab initio* and *ab initio*-based QM/MM molecular dynamics.

In this vein, the fourth chapter of this thesis introduces an innovative algorithm that combines multiple time step integration with trajectory surface hopping (MTS-TSH). This approach aims to accelerate simulations of systems where non-adiabatic phenomena play a crucial role, such as photo-induced processes in biomolecules. This method builds upon the idea of combining two approaches to compute the transition probabilities. The Landau-Zener approximation is used as a fast screening measure to detect possible jumps during the inner MTS loop. If a transition is detected, a re-evaluation of the nonadiabatic coupling is performed using the high level method and Tully's fewest switches algorithm.

In the fifth chapter, we introduce two schemes for integrating machine learning force predictions into a multiple time step algorithm (ML-MTS). The first scheme focuses on significantly speeding up the calculations by entirely bypassing heavy computations, although it comes at the expense of reliability. On the other hand, the second scheme utilizes machine learning force inference to mitigate the loss of energy conservation that arise when the time step ratio is increased. By doing so, it is possible to reduce the frequency of costly computations while ensuring the accuracy of the resulting simulation. This approach provides a balance between computational efficiency and maintaining the desired level of accuracy.

In the sixth chapter, we discuss two major extensions of the second scheme of the ML-MTS

algorithm introduced in the previous chapter. Firstly, we adapt the method to enable QM/MM simulations, combining the strengths of quantum mechanics and molecular mechanics approaches thus enabling efficient ML-MTS-QM/MM simulations of biological material. Additionally, we introduce a novel adaptive ML-MTS scheme, in which the ML model is trained on-the-fly, thus removing the tedious requirement of building up a prior training set that imposes numerous and costly high level calculations. To exploit all computations efficiently, the time step used for integrating the slow force components can be dynamically adjusted to conform to the retraining events. With these extensions, the ML-MTS algorithm becomes a powerful tool for simulating complex biological systems.

Finally, the thesis concludes with a comprehensive summary of the main findings and contributions presented throughout this research. Additionally, potential future directions and possible outlooks are explored, highlighting opportunities for further research and advancements in this promising field.

2 Theory

I'm doing stuff, Lori... Things.

Rick Grimes, The Walking Dead

The theory presented in this chapter is inspired by different sections of the following sources

- Statistical Mechanics: Theory and Molecular Simulation, by Mark E. Tuckerman
- Ab initio molecular dynamics: Theory and Implementation, by Dominik Marx and Jürg Hutter
- Introduction to Electronic Structure Methods, by Ursula Rothlisberger

2.1 Describing the motion of N interacting bodies

The simulation of the dynamics of molecular systems, often referred to as molecular dynamics (MD), is a complex field that has roots both in classical and quantum physics. When confronted to a new (scary) problem, my high school physics teacher M. Claude Montandon always repeated: “If you do not know how to solve a problem, write Newton’s second law of motion”. So let us follow his wisdom and start with

$$\mathbf{F} = m\mathbf{a}. \quad (2.1)$$

Formulated in 1687 by the English physicist and mathematician Sir Isaac Newton, this equation implies that if a body is subjected to a force \mathbf{F} , it will start moving with an acceleration \mathbf{a} . Even though Newton’s interest in this equation was mostly revolving around describing the movement of celestial bodies, this equation found its way into almost every scientific field, including ours.

In an isolated system, a group of N atoms can be viewed as a group of interacting classical particles. If an atom i is located at position \mathbf{r}_i , the force acting on it can \mathbf{F}_i depend on the position of all other atoms. would be

$$\mathbf{F}_i = \mathbf{F}_i(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) \quad (2.2)$$

We can now write the equation of motion of all N atoms as

$$\left\{ \begin{array}{l} m_1 \ddot{\mathbf{r}}_1 = \mathbf{F}_1(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) \\ m_2 \ddot{\mathbf{r}}_2 = \mathbf{F}_2(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) \\ \vdots \\ m_N \ddot{\mathbf{r}}_N = \mathbf{F}_N(\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N) \end{array} \right. \quad (2.3)$$

And this is where the problems start. This seemingly easy problem suddenly became a system of N differential equations and all variables depend on each other. In a realistic molecular system, the interactions between particle are not linear and finding an analytical solution to Eq. (2.3) is impossible. Fortunately, mathematics did not stop after Newton.

2.1.1 Hamiltonian classical mechanics

The research in mathematics used to describe the temporal evolution of dynamical systems made great strides in the late 18th century and early 19th century, with the release in 1788 of Joseph-Louis Lagrange’s treaties on analytical mechanics [24] unifying the different co-existing theories and introducing Lagrange’s formulation of mechanics. This formalism was

2.1 Describing the motion of N interacting bodies

reformulated in 1833 by Sir William Rowan Hamilton and notably replaces the generalized velocities \dot{q} of the Lagrangian formalism with generalized momenta \mathbf{p} . The details of these formalisms are fascinating but go beyond the scope of this thesis.

In Hamiltonian mechanics, a mechanical system is described by the evolution of a phase space element (\mathbf{p}, \mathbf{q}) , where $\mathbf{p} = m\dot{\mathbf{q}}$ are the momenta of the particles in the system and \mathbf{q} their positions. The system is then described by the evolution of the Hamiltonian function, that in the cases treated in this thesis can be written in Cartesian coordinates for a system of N particles as

$$H(\mathbf{p}, \mathbf{q}) = T + U(\mathbf{r}_1, \dots, \mathbf{r}_N) \quad (2.4)$$

$$= \sum_{i=1}^N \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{q}_1, \dots, \mathbf{q}_N), \quad (2.5)$$

where T is the kinetic energy and U is the potential energy. The Hamiltonian is then the total energy of the system, expressed as a function of positions and momenta. Given the Hamiltonian, the equations of motions can be computed directly by Hamilton's equations of motion

$$\dot{q}_\alpha = \frac{\partial H}{\partial p_\alpha}, \quad \dot{p}_\alpha = -\frac{\partial H}{\partial q_\alpha}, \quad (2.6)$$

where the subscript α refers to each coordinate component of the positions q_1, \dots, q_{3N} or momenta p_1, \dots, p_{3N} .

One key property of Hamilton's equations of motion is that they conserve the total Hamiltonian

$$\frac{\partial H}{\partial t} = \sum_{\alpha=1}^3 N \left(\frac{\partial H}{\partial q_\alpha} \dot{q}_\alpha + \frac{\partial H}{\partial p_\alpha} \dot{p}_\alpha \right) \quad (2.7)$$

$$= \sum_{\alpha=1}^3 N \left(\frac{\partial H}{\partial q_\alpha} \frac{\partial H}{\partial p_\alpha} - \frac{\partial H}{\partial p_\alpha} \frac{\partial H}{\partial q_\alpha} \right) \quad (2.8)$$

$$= 0. \quad (2.9)$$

Since the Hamiltonian is equivalent to the total energy, this property is the law of energy conservation.

2.1.2 Poisson brackets and Liouville operator

Now let us consider a function that describes the evolution of a quantity in the phase space, $f(\mathbf{q}, \mathbf{p})$. The time evolution can be written as

$$\begin{aligned} \frac{df}{dt} &= \frac{\partial f}{\partial \mathbf{q}} \dot{\mathbf{q}} + \frac{\partial f}{\partial \mathbf{p}} \dot{\mathbf{p}} \\ &= \sum_{\alpha=1}^{3N} \left(\frac{\partial f}{\partial q_{\alpha}} \frac{\partial H}{\partial p_{\alpha}} - \frac{\partial f}{\partial p_{\alpha}} \frac{\partial H}{\partial q_{\alpha}} \right) \\ &\equiv \{f, H\}, \end{aligned} \quad (2.10)$$

where Hamilton's equation of motion Eq. (2.6) have been used. $\{f, H\}$ is the Poisson bracket of f and H . The relation Eq. (2.10) implies that the Poisson bracket $\{\bullet, H\}$ is a generator of the time evolution of any function evolving in the space described by the Hamiltonian H . For example, the coordinate of a system \mathbf{q} at time t can be directly computed from $\mathbf{q}(t=0)$ by solving Eq. (2.10) and using the anticommutativity property of the Poisson bracket ($\{f, H\} = -\{H, f\}$),

$$\begin{aligned} \frac{df}{dt} &= \{f, H\} \\ &= -\{H, \bullet\}f. \end{aligned} \quad (2.11)$$

Another popular formulation of the time evolution operator $\{H, \bullet\}$ is the Liouville operator L , defined as

$$iL = \{H, \bullet\} \quad (2.12)$$

$$= \sum_{\alpha=1}^{3N} \left(\frac{\partial H}{\partial p_{\alpha}} \frac{\partial}{\partial q_{\alpha}} - \frac{\partial H}{\partial q_{\alpha}} \frac{\partial}{\partial p_{\alpha}} \right) \quad (2.13)$$

where $i = \sqrt{-1}$ and Eq. (2.11) can be solved for any time t ,

$$f(t) = e^{iLt} f(0). \quad (2.14)$$

Unfortunately, the action of e^{iLt} on an element of phase space cannot be evaluated. If it could, that would mean that any mechanical problem could be solved exactly with an analytical solution.

2.1.3 Generating integration methods

Despite the deception of the inability of the Liouville operator to solve instantly every problem in mechanics, it can serve as a good base to create algorithms that approximate Hamilton's equation.

The Liouville operator can be separated into two parts, iL_1 and iL_2 , corresponding to

$$iL \equiv iL_1 + iL_2 \quad (2.15)$$

with

$$iL_1 = \sum_{\alpha=1}^{3N} \frac{\partial H}{\partial p_\alpha} \frac{\partial}{\partial q_\alpha}, \quad iL_2 = - \sum_{\alpha=1}^{3N} \frac{\partial H}{\partial q_\alpha} \frac{\partial}{\partial p_\alpha}. \quad (2.16)$$

These two separate operators are more convenient than their sum counterparts, in the sense that the actions of the exponential of iL_1 or iL_2 on an element of phase space can usually be evaluated exactly.

A property of these operators L_1 and L_2 that will be mentioned in the next chapters is that they do not commute, in general. This means that the order in which the operator is applied changes the results. This can be verified using a simple 1D example where the Hamiltonian can be expressed in Cartesian coordinates as

$$H = \frac{p^2}{2m} + U(q). \quad (2.17)$$

Using equation Eq. (2.16), the operators becomes

$$iL_1 = \frac{p}{m} \frac{\partial}{\partial q}, \quad iL_2 = - \frac{\partial U}{\partial q} \frac{\partial}{\partial p} = F(q) \frac{\partial}{\partial p}. \quad (2.18)$$

Applying the operator $iL_1 iL_2$ to an arbitrary function $f(q, p)$ gives

$$\begin{aligned} iL_1 iL_2 f(q, p) &= \frac{p}{m} \frac{\partial}{\partial q} F(q) \frac{\partial}{\partial p} f(q, p) \\ &= \frac{p}{m} F(q) \frac{\partial^2 f}{\partial p \partial q} + \frac{p}{m} \frac{\partial F}{\partial q} \frac{1}{m} \frac{\partial f}{\partial p}. \end{aligned} \quad (2.19)$$

Conversely, applying $iL_2 iL_1$ gives

$$\begin{aligned} iL_2 iL_1 f(q, p) &= F(q) \frac{\partial}{\partial p} \frac{p}{m} \frac{\partial}{\partial q} f(q, p) \\ &= F(q) \frac{p}{m} \frac{\partial^2 f}{\partial p \partial q} + F(q) \frac{1}{m} \frac{\partial f}{\partial q}. \end{aligned} \quad (2.20)$$

Therefore, the commutator applied to $f(q, p)$ is

$$[iL_1, iL_2]f(q, p) = \frac{p}{m} \frac{\partial F}{\partial q} \frac{\partial f}{\partial p} - \frac{F(q)}{m} \frac{\partial f}{\partial q} \neq 0. \quad (2.21)$$

Since no hypothesis was made on $f(q, p)$, this shows that iL_1 and iL_2 do not commute. This property is particularly inconvenient for Eq. (2.14), where the two operators cannot be applied successively as

$$f(t) = e^{iLt} f(0) = e^{(iL_1+iL_2)t} f(0) \neq e^{iL_1 t} e^{iL_2 t} f(0) \quad (2.22)$$

However, this separation can be achieved using Trotter's theorem as

$$e^{iLt} = e^{(iL_1+iL_2)t} = \lim_{P \rightarrow \infty} \left(e^{iL_2 t/2P} e^{iL_1 t/P} e^{iL_2 t/2P} \right)^P. \quad (2.23)$$

If we define an integration time step $\Delta t = t/P$, Eq. (2.23) can be expressed in a more intuitive way,

$$e^{iLt} = e^{(iL_1+iL_2)t} = \lim_{P \rightarrow \infty, \Delta t \rightarrow 0} \left(e^{iL_2 \Delta t/2} e^{iL_1 \Delta t} e^{iL_2 \Delta t/2} \right)^P. \quad (2.24)$$

The introduction of the decomposition with the symmetric Trotter theorem transformed the problem. Originally, the goal was to compute the time evolution after an arbitrarily long time t with a single application of the operator. Now we are computing P smaller time evolutions with an increment Δt that will eventually sum to the original time objective t . However, the limits $P \rightarrow \infty$ and $\Delta t \rightarrow 0$ is an obvious setback, as applying an operator an infinite number of time does not solve our problem. Therefore, we introduce a first approximation: P is finite. Eq. (2.24) then becomes

$$e^{iLt} = e^{(iL_1+iL_2)t} \approx \left(e^{iL_2 \Delta t/2} e^{iL_1 \Delta t} e^{iL_2 \Delta t/2} \right)^P + \mathcal{O}(P \Delta t^3). \quad (2.25)$$

The last term $\mathcal{O}(P \Delta t^3)$ indicates that the leading order error related to this approximation is proportional to $P \Delta t^3$. We remind that $P = t/\Delta t$, meaning that the error is proportional to Δt^2 .

2.1 Describing the motion of N interacting bodies

Isolating an individual small time increment, the operator is

$$e^{iL\Delta t} \approx e^{iL_2\Delta t/2} e^{iL_1\Delta t} e^{iL_2\Delta t/2} + \mathcal{O}(\Delta t^2) \quad (2.26)$$

2.1.4 Deriving the velocity Verlet algorithm

The application of the operator $e^{iL_1 t}$ and $e^{iL_2 t}$ to an element of phase space can be explicitly computed. Let us demonstrate this on a single particle moving in 1D with the Hamiltonian in Eq. (2.17). Using the operators L_1 and L_2 obtained in Eq. (2.18), Eq. (2.26) becomes

$$\exp(iL\Delta t) \approx \exp\left[\frac{\Delta t}{2} F(q) \frac{\partial}{\partial p}\right] \exp\left[\Delta t \frac{p}{m} \frac{\partial}{\partial q}\right] \exp\left[\frac{\Delta t}{2} F(q) \frac{\partial}{\partial p}\right]. \quad (2.27)$$

Starting from the initial conditions (q_0, p_0) , the evolution of this phase space element can be approximated by

$$\begin{bmatrix} q(\Delta t) \\ p(\Delta t) \end{bmatrix} \approx \exp\left[\frac{\Delta t}{2} F(q_0) \frac{\partial}{\partial p_0}\right] \exp\left[\Delta t \frac{p_0}{m} \frac{\partial}{\partial q_0}\right] \exp\left[\frac{\Delta t}{2} F(q_0) \frac{\partial}{\partial p_0}\right] \cdot \begin{bmatrix} q_0 \\ p_0 \end{bmatrix} \quad (2.28)$$

The application of these operators is not straightforward. Let us first study the action of an operator that has the same form as our operators,

$$\exp\left[c \frac{\partial}{\partial x}\right] \quad (2.29)$$

acting on a function $f(x)$ and where c is independent of x . This operator can be expanded in a Taylor series

$$\begin{aligned} \exp\left[c \frac{\partial}{\partial x}\right] f(x) &= \sum_{k=0}^{\infty} \frac{1}{k!} \left(c \frac{d}{dx}\right)^k f(x) \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} c^k \frac{d^k f}{dx^k}(x). \end{aligned} \quad (2.30)$$

We recognize here the Taylor expansion of the function $f(x+c)$ around $c=0$ and therefore

$$\exp\left[c \frac{\partial}{\partial x}\right] f(x) = f(x+c). \quad (2.31)$$

Chapter 2. Theory

We can use this identity in Eq. (2.28) to compute the successive application of all three operators on an element of phase space. First,

$$\exp \left[\frac{\Delta t}{2} F(q_0) \frac{\partial}{\partial p_0} \right] \begin{bmatrix} q_0 \\ p_0 \end{bmatrix} = \begin{bmatrix} q_0 \\ p_0 + \frac{\Delta t}{2} F(q_0) \end{bmatrix} \quad (2.32)$$

which is followed by

$$\exp \left[\Delta t \frac{p_0}{m} \frac{\partial}{\partial q_0} \right] \begin{bmatrix} q_0 \\ p_0 + \frac{\Delta t}{2} F(q_0) \end{bmatrix} = \begin{bmatrix} q_0 + \Delta t \frac{p_0}{m} \\ p_0 + \frac{\Delta t}{2} F(q_0 + \Delta t \frac{p_0}{m}) \end{bmatrix} \quad (2.33)$$

and finally

$$\begin{aligned} \begin{bmatrix} q(\Delta t) \\ p(\Delta t) \end{bmatrix} &\approx \exp \left[\frac{\Delta t}{2} F(q_0) \frac{\partial}{\partial p_0} \right] \begin{bmatrix} q_0 + \Delta t \frac{p_0}{m} \\ p_0 + \frac{\Delta t}{2} F(q_0 + \Delta t \frac{p_0}{m}) \end{bmatrix} \\ &\approx \begin{bmatrix} q_0 + \frac{\Delta t}{m} (p_0 + \frac{\Delta t}{2} F(q_0)) \\ p_0 + \frac{\Delta t}{2} F(q_0) + \frac{\Delta t}{2} F \left[q_0 + \frac{\Delta t}{m} (p_0 + \frac{\Delta t}{2} F(q_0)) \right] \end{bmatrix} \end{aligned} \quad (2.34)$$

This results give a recipe to follow to update the positions and momenta to compute the time evolution of a system. Moreover, using $v = p/m$, $q(\Delta t)$ can be written as

$$q(\Delta t) = q_0 + v_0 \Delta t + \frac{\Delta t^2}{2m} F(q_0). \quad (2.35)$$

The momenta part of Eq. (2.34) can be combined with Eq. (2.35) to obtain a simple form for the velocity update,

$$v(\Delta t) = v_0 + \frac{\Delta t}{2m} [F(q_0) + F(q(\Delta t))]. \quad (2.36)$$

This time propagation scheme is known as the velocity Verlet algorithm, one of the most used integration schemes in molecular dynamics. The algorithm is represented in Algorithm 2.1. This scheme can be derived more easily using Taylor expansions for the positions and velocities, but this development shows the capability of the Liouville formalism to build complex integration schemes.

-
- 1: Initialize positions, velocities, forces: $\mathbf{q}, \mathbf{p}, \mathbf{F}$
 - 2: **for** $i = 1$, maxiter **do** (MD loop)
 - 3: Momenta update: $\mathbf{p} \leftarrow \mathbf{p} + 0.5 \cdot \Delta t \cdot \mathbf{F}$
 - 4: Position update: $\mathbf{q} \leftarrow \mathbf{q} + \Delta t \cdot \mathbf{p}/m$
 - 5: Compute forces
 - 6: Momenta update $\mathbf{p} \leftarrow \mathbf{p} + 0.5 \cdot \Delta t \cdot \mathbf{F}$
 - 7: **end for** (MD loop)
-

ALG. 2.1: Standard velocity Verlet algorithm.

2.2 Classical molecular dynamics

The dynamical properties of molecules can often be well described by equations from classical physics. Molecules are usually represented as a group of atoms described as point-like particles of mass m and charge q connected by a complex network of spring-like potentials. In most classical MD methods, the potential energy is modelled by a variation of

$$\begin{aligned}
 U(\mathbf{r}_1, \dots, \mathbf{r}_N) = & \sum_{\text{bonds}} \frac{1}{2} k_{\text{bond}} (r - r_0)^2 + \sum_{\text{bends}} \frac{1}{2} k_{\text{bend}} (\theta - \theta_0)^2 \\
 & + \sum_{\text{tors}} \sum_{n=0}^6 A_n [1 + \cos(C_n \phi + \delta_n)] \\
 & + \sum_{i < j} \left\{ 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \right\}. \quad (2.37)
 \end{aligned}$$

In this equation, the first term represents the covalent bonds, modelled by a spring of rigidity k_{bond} and an equilibrium bond length r_0 . The second term represents the angular stretch between atom triplets, that is also modelled by a harmonic potential. The third term adds information on the preferential dihedral angle formed by 4 atoms. The next two terms are the non-bonded interactions: the van der Waals interaction and the Coulomb potential. Fig. 2.1 shows a graphical representation of the bonded terms in Eq. (2.37).

In general, all atoms can have different parameters for all these interactions. Therefore, this potential function contains a high number of inter-dependent parameters. The combination of the form of the energy potential and the set of parameters used to describe a molecule is generally referred to as the “force field”.

The most attractive property of molecular dynamics based on a classical potential is their affordable cost, allowing long simulations of large systems. However, a lot of properties of molecular systems can only be described by including the quantum nature of the interactions between atoms.

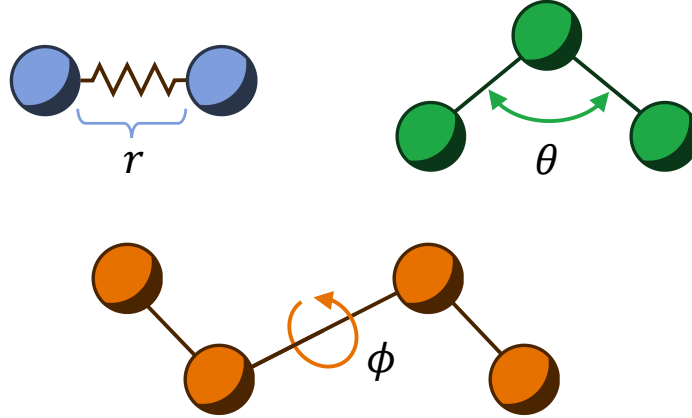


Figure 2.1: Representation of the bonded variables generally used in classical molecular dynamics.

2.3 Born-Oppenheimer molecular dynamics

Going beyond classical models for the interactions between atoms, we would want to use interactions derived directly from quantum mechanics to compute the displacements of the atoms in a molecule. In this section, we show under which hypotheses this can be achieved.

In non-relativistic quantum mechanics, the many-body system consisting of interacting nuclei at positions \mathbf{R}_I and electrons \mathbf{r}_i should ideally be solved by the time-dependent Schrödinger equation

$$i\hbar \frac{\partial}{\partial t} \Phi(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}; t) = \hat{H} \Phi(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}; t). \quad (2.38)$$

In its position representation, the Hamiltonian H containing all interactions between nuclei and electrons can be written as

$$\hat{H} = -\sum_I \frac{\hbar^2}{2M_I} \nabla_I^2 - \sum_i \frac{\hbar^2}{2m_e} \nabla_i^2 + \sum_{i < j} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|} + \sum_{I,i} \frac{e^2 Z_I}{|\mathbf{R}_I - \mathbf{r}_i|} + \sum_{I < J} \frac{e^2 Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} \quad (2.39)$$

$$= -\sum_I \frac{\hbar^2}{2M_I} \nabla_I^2 + \hat{H}_{\text{el}}(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}) \quad (2.40)$$

where the lowercase indices refer to electrons and uppercase indices to nuclei, M and m_e are nuclear and electronic masses, respectively, and Z are atomic numbers. The first two terms of Eq. (2.39) correspond to the kinetic energy operator and the last three are the Coulomb interactions between electron pairs, electron-nuclei pairs and nuclei pairs, respectively. Finally, Eq. (2.40) gathers all terms that do not depend on the nuclear momenta in the electronic

2.3 Born-Oppenheimer molecular dynamics

Hamiltonian H_{el} . To solve Eq. (2.38), let us consider the electronic part of the Hamiltonian for fixed nuclei. The time-independent Schrödinger equation for this system is

$$\hat{H}_{\text{el}}(\{\mathbf{r}_i\}; \{\mathbf{R}_I\})\Psi_k = E_k(\{\mathbf{R}_I\})\Psi_k \quad (2.41)$$

where k corresponds to the k -th electronic state. Then, using an Ansatz for the exact solution of this problem, the total wave function Φ can be expanded using the complete set of eigenfunctions $\{\Psi_l\}$ of \hat{H}_{el} along with the nuclear wave functions $\chi(\{\mathbf{R}_I\}, t)$ that can be viewed as time-dependent coefficients,

$$\Phi(\{\mathbf{r}_i\}, \{\mathbf{R}_I\}; t) = \sum_{l=0}^{\infty} \Psi_l(\{\mathbf{r}_i\}, \{\mathbf{R}_I\})\chi_l(\{\mathbf{R}_I\}; t). \quad (2.42)$$

This expression can be substituted in Eq. (2.38),

$$i\hbar \frac{\partial}{\partial t} \sum_{l=0}^{\infty} \Psi_l(\{\mathbf{r}_i\}, \{\mathbf{R}_I\})\chi_l(\{\mathbf{R}_I\}; t) = \hat{H} \sum_{l=0}^{\infty} \Psi_l(\{\mathbf{r}_i\}, \{\mathbf{R}_I\})\chi_l(\{\mathbf{R}_I\}; t). \quad (2.43)$$

Then, using the Hamiltonian \hat{H} from Eq. (2.40) in Eq. (2.43), multiplying from the left by Ψ_k^* , and integrating over all electronic coordinates \mathbf{r} , we obtain the coupled differential equations

$$i\hbar \frac{\partial}{\partial t} \chi_k = \left[-\sum_I \frac{\hbar^2}{2M_I} \nabla_I^2 + E_k(\{\mathbf{R}_I\}) \right] \chi_k + \sum_l C_{kl} \chi_l \quad (2.44)$$

where

$$C_{kl} = \int \Psi_k^* \left[-\sum_I \frac{\hbar^2}{2M_I} \nabla_I^2 \right] \Psi_l d\mathbf{r} + \sum_I \frac{1}{M_I} \int \Psi_k^* (-i\hbar \nabla_I) \Psi_l d\mathbf{r} (-i\hbar \nabla_I). \quad (2.45)$$

This operator is the exact nonadiabatic coupling between the k -th and l -th electronic states. The diagonal contribution C_{kk} only depends on a single adiabatic wave function Ψ_k and therefore represents a correction to the k -th adiabatic eigenvalue E_k of the electronic Schrödinger equation (Eq. (2.41)). The second term of Eq. (2.45) is equal to zero when the electronic wave function Ψ_k is real and the adiabatic approximation is obtained by only considering the diagonal terms

$$C_{kk} = -\sum_I \frac{\hbar^2}{2M_I} \int d\mathbf{r} \Psi_k^* \nabla_I^2 \Psi_k. \quad (2.46)$$

Chapter 2. Theory

The coupled differential equations in Eq. (2.44) are now completely decoupled as

$$i\hbar \frac{\partial}{\partial t} \chi_k = \left[\frac{\hbar^2}{2M_I} \nabla_I^2 + E_k(\{\mathbf{R}_I\}) + C_{kk}(\{\mathbf{R}_I\}) \right] \chi_k. \quad (2.47)$$

This decoupling has a consequence: the movements of the nuclei cannot change the quantum state, k , of the electronic subsystem during time evolution. Therefore, the coupled nuclear and electronic wavefunction in Eq. (2.42) can also be decoupled as

$$\Phi(\{\mathbf{r}_i, \{\mathbf{R}_I\}; t) \approx \Psi_k(\{\mathbf{r}_i, \{\mathbf{R}_I\}) \chi_k(\{\mathbf{R}_I\}; t). \quad (2.48)$$

Finally, the last approximation is to also neglect the diagonal terms C_{kk} and Eq. (2.47) becomes

$$i\hbar \frac{\partial}{\partial t} \chi_k = \left[\frac{\hbar^2}{2M_I} \nabla_I^2 + E_k(\{\mathbf{R}_I\}) \right] \chi_k. \quad (2.49)$$

This hypothesis defines the Born-Oppenheimer (BO) approximation. The BO approximation is often used by invoking the mass difference between nuclei and electrons, thus allowing to decouple the motion of the nuclei from the motion of the electrons. This approximation can be applied safely in most cases, but it breaks down when the energy difference between two states is low or when the system is subject to a strong electromagnetic field such as a laser, in which the electronic and nuclear degrees of freedom evolve on a similar time scale.

To use a classical MD propagation algorithm, we want to approximate the nuclei as classical point particles. Let us start by writing the the nuclear wave function as a function of an amplitude $A_k > 0$ and a phase S_k , which can both be considered real,

$$\chi_k = A_k e^{iS_k/\hbar}. \quad (2.50)$$

Plugging this expression into Eq. (2.49) and separating the real and imaginary parts, the equation for the nuclei can be expressed as

$$\frac{\partial S_k}{\partial t} + \sum_I \frac{1}{2M_I} (\nabla_I S_k)^2 + E_k = \hbar \sum_I \frac{1}{2M_I} \frac{\nabla_I^2 A_k}{A_k} \quad (2.51)$$

$$\frac{\partial A_k}{\partial t} + \sum_I \frac{1}{M_I} (\nabla_I A_k)(\nabla_I S_k) + \sum_I \frac{1}{2M_I} A_k (\nabla_I^2 S_k) = 0. \quad (2.52)$$

If the nuclear probability density is given by $\rho_k = |\chi_k|^2 = A_k^2$, multiplying Eq. (2.52) by A_k from the left gives

2.3 Born-Oppenheimer molecular dynamics

$$\frac{\partial A_k^2}{\partial t} + \sum_I \frac{1}{M_I} \nabla_I (A_k^2 \nabla_I S_k) = 0 \quad (2.53)$$

Studying the phase S_k of the nuclear wave function associated to the k -th eigenstate in Eq. (2.51), one term depends explicitly on \hbar . In the classical limit $\hbar \rightarrow 0$ this term vanishes and

$$\frac{\partial S_k}{\partial t} + \sum_I \frac{1}{2M_I} (\nabla_I S_k)^2 + E_k = 0. \quad (2.54)$$

Interestingly, this equation is isomorphic to Hamilton-Jacobi's equation of motion

$$\frac{\partial S_k}{\partial t} + H_k(\{\mathbf{R}_I\}, \{\nabla_I S_k\}) = 0 \quad (2.55)$$

with a classical Hamiltonian function given by

$$H_k(\{\mathbf{R}_I, \{\mathbf{P}_I\}) = T(\{\mathbf{P}_I\}) + V_k(\{\mathbf{R}_I\}) \quad (2.56)$$

expressed using the generalized coordinates $\{\mathbf{R}_I\}$ and momenta $\{\mathbf{P}_I\}$ of the nuclei. Using the connection transformation

$$\mathbf{P}_I \equiv \nabla_I S_k, \quad (2.57)$$

a Newtonian equation of motion can be written as

$$\dot{\mathbf{P}}_I = -\nabla_I V_k(\{\mathbf{R}_I\}) \quad (2.58)$$

$$= -\nabla_I E_k. \quad (2.59)$$

Finally, this equation can be written in a more familiar form

$$M_I \ddot{\mathbf{R}}_I(t) = -\nabla_I V_k^{BO}(\{\mathbf{R}_I(t)\}). \quad (2.60)$$

This equation signifies that within the BO approximation, the nuclei can be treated as classical particles subject to an effective V_k^{BO} which is given by the BO potential energy surface of the k -th electronic state. It is then possible to use the time-independent electronic Schrödinger equation (Eq. (2.41)) to compute the forces acting on the nuclei and simultaneously use

algorithms from classical molecular dynamics to compute the movement of the nuclei,

$$M_I \ddot{\mathbf{R}}_I(t) = -\nabla_I \min_{\Psi_0} \left\{ \langle \Psi_0 | \hat{H}_{\text{el}} | \Psi_0 \rangle \right\} \quad (2.61)$$

$$\hat{H}_{\text{el}} \Psi_0 = E_0 \Psi_0. \quad (2.62)$$

We now need a way to solve Eq. (2.62) efficiently.

2.4 Density Functional Theory

The problem now is to solve the non-relativistic time-independent Schrödinger equation

$$\hat{H}\Psi = E\Psi \quad (2.63)$$

with the electronic Hamiltonian \hat{H}_{el} in Eq. (2.40). To avoid cumbersome notations, let us define \hat{T}_e , \hat{V}_{ee} , \hat{V}_{eN} and \hat{V}_{NN} corresponding to each terms, so that the electronic Hamiltonian reads

$$\hat{H}_{\text{el}} = \hat{T}_e + \hat{V}_{eN} + \hat{V}_{ee} + \hat{V}_{NN}. \quad (2.64)$$

The the time-independent Schrödinger equation with the electronic wave function is

$$\hat{H}_{\text{el}} \Psi_{\text{el}}(\mathbf{r}_1, \dots, \mathbf{r}_N) = E_{\text{el}} \Psi_{\text{el}}(\mathbf{r}_1, \dots, \mathbf{r}_N). \quad (2.65)$$

However, finding Ψ_{el} is a complex task, mostly due to the electron-electron interaction term in Eq. (2.64). This is a many-body problem where all electrons are considered simultaneously.

Density functional theory (DFT) is one of the most popular approaches for investigating the electronic structure of such many-body systems. The central quantity of DFT is the electron density $\rho(\mathbf{r})$ that describes the probability to find an electron at position \mathbf{r} and can be expressed as

$$\rho(\mathbf{r}) = N \int \dots \int \Psi(\mathbf{r}, \mathbf{r}_2, \dots, \mathbf{r}_N) \Psi^*(\mathbf{r}, \mathbf{r}_2, \dots, \mathbf{r}_N) d\mathbf{r}_2 \dots d\mathbf{r}_N. \quad (2.66)$$

Compared to the total many-electron wave function $\Psi(\mathbf{r}_1, \dots, \mathbf{r}_N)$, $\rho(\mathbf{r})$ is only a function of a single set of three Cartesian coordinates, making this quantity much more comfortable to work with.

The base of DFT lies in the two Hohenberg-Kohn (HK) theorems that can be summarized as follows. The first HK theorem suggests that Eq. (2.66) can be reversed, meaning that a given predefined potential $v_{\text{ext}}(\mathbf{r})$ generated by the nuclei uniquely defines the density of the ground state ρ_0 . Consequently, it also defines the ground state electronic wave function Ψ_0 as a unique functional of the ground state density $\rho_0(\mathbf{r})$,

$$\Psi_0(\mathbf{r}, \mathbf{r}_2, \dots, \mathbf{r}_N) = \Psi[\rho_0(\mathbf{r})]. \quad (2.67)$$

Therefore, the expectation value of any (ground-state) observable \hat{O} is a functional of $\rho_0(\mathbf{r})$,

$$O_0 = O[\rho_0] = \langle \Psi[\rho_0] | \hat{O} | \Psi[\rho_0] \rangle. \quad (2.68)$$

In the specific example of the energy,

$$E_{v,0} = E_v[\rho_0] = \langle \Psi[\rho_0] | \hat{H}_{\text{el}} | \Psi[\rho_0] \rangle. \quad (2.69)$$

The second HK theorem says that if $\rho'(\mathbf{r})$ is another density,

$$E_v[\rho_0] \leq E_v[\rho']. \quad (2.70)$$

Coming back to Eq. (2.69), one can write

$$\begin{aligned} E_v[\rho_0] &= \langle \Psi[\rho_0] | \hat{H}_{\text{el}} | \Psi[\rho_0] \rangle \\ &= \langle \Psi[\rho_0] | \hat{T}_e + \hat{V}_{\text{ee}} + \hat{V}_{\text{eN}} | \Psi[\rho_0] \rangle \\ &= \langle \Psi[\rho_0] | \hat{T}_e | \Psi[\rho_0] \rangle + \langle \Psi[\rho_0] | \hat{V}_{\text{ee}} | \Psi[\rho_0] \rangle + \langle \Psi[\rho_0] | \hat{V}_{\text{eN}} | \Psi[\rho_0] \rangle \\ &= T_e[\rho] + V_{\text{ee}}[\rho] + V_{\text{eN}}[\rho] \\ &= F[\rho] + V_{\text{eN}}[\rho] \end{aligned} \quad (2.71)$$

where we have introduced an internal energy functional $F[\rho]$. We note that $T_e[\rho]$ and $V_{\text{ee}}[\rho]$ are universal functionals that do not depend on the considered system, defined by the nuclei. These terms have been merged in $F[\rho]$ in Eq. (2.71). By opposition, the electron-nuclei term

$$\hat{V}_{\text{eN}}(\mathbf{r}) = \hat{V}_{\text{ext}}(\mathbf{r}) = \sum_I \frac{Z_I}{|\mathbf{r} - \mathbf{R}_I|} \quad (2.72)$$

is system-dependent and can be viewed as the nucleus potential $v_{\text{ext}}(\mathbf{r})$ and can thus be written as

$$V_{\text{ext}}[\rho] = \int d\mathbf{r} \rho(\mathbf{r}) v_{\text{ext}}(\mathbf{r}). \quad (2.73)$$

Overall, the density ρ_0 that minimises the energy $E_v[\rho]$ is indeed the ground state,

$$E_v[\rho] = \min_{\Psi \rightarrow \rho} \langle \Psi | \hat{T} + \hat{V}_{\text{ee}} | \Psi \rangle + \int d\mathbf{r} \rho(\mathbf{r}) v_{\text{ext}}(\mathbf{r}). \quad (2.74)$$

Theoretically, this formalism should allow us to compute all observables for any system in their electronic ground state. However, the exact functional $F[\rho]$ is not known and approximating it is complex.

2.4.1 Kohn-Sham DFT

The most widely used way of implementing DFT was proposed in 1965 by Walter Kohn and Lu Jeu Sham [25]. They proposed to split the kinetic energy functional $T_e[\rho]$ into two parts: a first part, $T_s[\rho]$, representing the kinetic energy of non-interacting electrons of density ρ and a remainder $T_c[\rho]$,

$$T_e[\rho] = T_s[\rho] + T_c[\rho]. \quad (2.75)$$

The kinetic energy of non-interacting electrons $T_s[\rho]$ is not known directly as a functional of ρ . However, it can be computed as the sum of individual kinetic energies. Therefore, $T_s[\rho]$ can be expressed as a combination of single-particle orbitals $\phi_i(\mathbf{r})$ of the non-interacting system of density ρ ,

$$T_s[\rho] = -\frac{1}{2} \sum_i^N \int d\mathbf{r} \phi_i^*(\mathbf{r}) \nabla^2 \phi_i(\mathbf{r}), \quad (2.76)$$

with

$$\rho(\mathbf{r}) = \sum_i^N |\phi_i(\mathbf{r})|^2. \quad (2.77)$$

The functional $T_s[\rho]$ can then be expressed in terms of the full set of occupied orbitals $\{\phi_i[\rho]\}$ as $T_s[\{\phi_i[\rho]\}]$. The exact total energy of the interacting system becomes

$$\begin{aligned} E_v[\rho] &= T_e[\rho] + V_{\text{ee}}[\rho] + V_{\text{eN}}[\rho] \\ &= T_s[\{\phi_i[\rho]\}] + V_H[\rho] + E_{\text{xc}}[\rho] + V_{\text{eN}}[\rho] \end{aligned} \quad (2.78)$$

where we have introduced the Hartree potential $V_H[\rho]$ representing the electrostatic interaction of the charge distribution $\rho(\mathbf{r})$,

$$V_H[\rho] = \frac{1}{2} \int d\mathbf{r} \int d\mathbf{r}' \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \quad (2.79)$$

and the exchange-correlation energy functional $E_{xc}[\rho]$ defined by the difference between the true kinetic energy T_e and the non-interacting kinetic energy T_s , as well as the difference between the true electron-electron interaction potential V_{ee} and the Hartree potential in Eq. (2.79),

$$E_{xc}[\rho] = T_e - T_s + V_{ee} - V_H. \quad (2.80)$$

We now want to find the density ρ_0 that minimises the functional $E_v[\rho]$. The solution proposed by Kohn and Sham is to perform the minimisation indirectly. First they show that

$$0 = \frac{\delta E_v[\rho]}{\delta \rho(\mathbf{r})} = \frac{\delta T_s[\rho]}{\delta \rho(\mathbf{r})} + v_{\text{ext}}(\mathbf{r}) + v_H(\mathbf{r}) + v_{xc}(\mathbf{r}) \quad (2.81)$$

where $v_{\text{ext}}(\mathbf{r}) = \delta V_{eN}[\rho]/\delta \rho(\mathbf{r})$.

If we now consider a second system of non-interacting particles moving in a new potential $v_s(\mathbf{r})$, the minimisation is

$$0 = \frac{\delta E_v[\rho]}{\delta \rho(\mathbf{r})} = \frac{\delta T_s[\rho]}{\delta \rho(\mathbf{r})} + \frac{\delta V_s[\rho]}{\delta \rho(\mathbf{r})}. \quad (2.82)$$

Since the interactions are omitted in Eq. (2.82), the Hartree potential and the exchange-correlation terms are equal to zero. The solution of these two systems are identical if

$$v_s(\mathbf{r}) = v_{\text{ext}}(\mathbf{r}) + v_H(\mathbf{r}) + v_{xc}(\mathbf{r}). \quad (2.83)$$

This means that it is possible to compute the density of the fully interacting system in potential $v_{\text{ext}}(\mathbf{r})$ described by the complete Schrödinger equation by solving the equations of a non-interacting, single-particle systems in potential $v_s(\mathbf{r})$. The Schrödinger equation of the i -th system is

$$\left[\frac{\nabla^2}{2} + v_s(\mathbf{r}) \right] \phi_i(\mathbf{r}) = \varepsilon_i \phi_i(\mathbf{r}) \quad (2.84)$$

and the orbitals ϕ_i reproduce the density $\rho(\mathbf{r})$ of the original system,

$$\rho(\mathbf{r}) = \sum_i^N f_i |\phi_i(\mathbf{r})|^2 \quad (2.85)$$

and f_i is the occupation of the i -th orbital. The Eqs. (2.83) to (2.85) are known as the Kohn

Sham (KS) equations. They allow to use the solution of the Schrödinger equation of non-interacting systems to find the density ρ_0 that minimises the energy functional $E_v[\rho]$. Furthermore, these equations outline the procedure known as "self consistent field" (SCF). Starting from an initial guess for ρ , the potential $v_s(\mathbf{r})$ can be computed with Eq. (2.83), that can be then be used in Eq. (2.84) to find the KS orbitals $\phi_i(\mathbf{r})$. Finally, a new density ρ' is computed using Eq. (2.85) and can be fed back into Eq. (2.83). This loop can be repeated until the density ρ converges to the density of the ground state ρ_0 .

When ρ_0 is known, the total energy can be computed by

$$E_0 = \sum_{i=1}^N \varepsilon_i - \frac{1}{2} \int d\mathbf{r} \int d\mathbf{r}' \frac{\rho_0(\mathbf{r})\rho_0(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} - \int d\mathbf{r} v_{xc}(\mathbf{r})\rho_0(\mathbf{r}) + E_{xc}[\rho_0]. \quad (2.86)$$

The issue with Eqs. (2.83) and (2.86) is that there is no exact analytical form for $E_{xc}[\rho]$, nor for its functional derivative v_{xc} and this functional has to be approximated.

2.5 Approximations of the exchange-correlation functional

Despite the lack of an exact expression for the exchange correlation functional $E_{xc}[\rho]$, multiple attempts have been made using a broad range of approximations. In this chapter, we present a few approximations that will be used in the next chapters of this thesis.

2.5.1 Local Density Approximation (LDA)

One of the simplest approximations to the exchange correlation functional is the local density approximation (LDA). First, we separate the exchange correlation functional into an exchange term E_x and a correlation term E_c . The idea is to use results from another well-known system, the homogeneous electron gas, and apply it directly to our system. Thus, the exchange term is

$$E_x^{\text{LDA}}[\rho] = -\frac{3q_e^2}{4} \left(\frac{3}{\pi}\right) \int \rho^{4/3}(\mathbf{r}) d\mathbf{r}. \quad (2.87)$$

The correlation term E_c cannot be computed analytically for the homogeneous electron gas and it is usually approximated by estimates obtained with quantum Monte Carlo calculations.

2.5.2 Generalized Gradient Approximation (GGA)

In the LDA, only the information of the density at location \mathbf{r} is used. Real systems are obviously inhomogeneous, and information on the variation of the density could greatly help the description of the system. Therefore, different approaches referred to as generalized gradient approximations (GGA) incorporate the gradient of ρ in the form

2.5 Approximations of the exchange-correlation functional

$$E_{xc}^{GGA} = \int d\mathbf{r} f(\rho(\mathbf{r}), \nabla\rho(\mathbf{r})). \quad (2.88)$$

Multiple functions $f(\rho(\mathbf{r}), \nabla\rho(\mathbf{r}))$ have been suggested over the years and among the most popular ones are PBE [26, 27] and BLYP [28], both named after their authors (Perdew, Burke, Ernzerhof and Becke, Lee, Yang, Parr, respectively).

2.5.3 Hybrid functionals

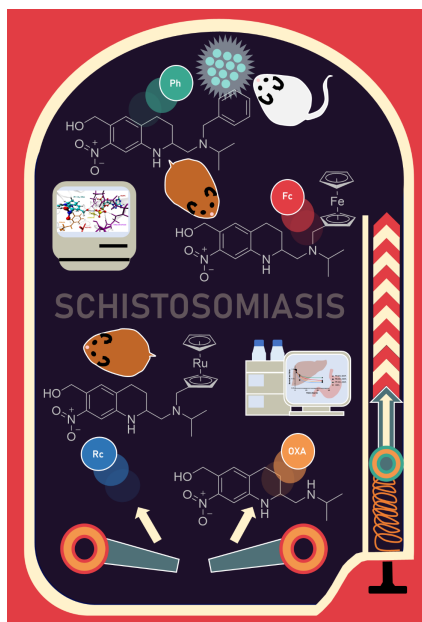
When the GGA functionals are not able to capture the essence of the desired quantum property, there is a broad range of methods that mix GGA exchange-correlation descriptions with approaches from wavefunction-based quantum chemical methods such as the direct calculation of the exact exchange integrals, resulting in the so-called hybrid functionals.

Popular functionals of this type are PBE0 [29] and B3LYP [30], from the same authors that created the PBE and BLYP functionals. However, the gain in accuracy comes at a cost. Hybrid functionals can be two orders of magnitude more computationally-expensive than LDA or GGA, especially in combination with plane wave basis sets.

3 Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis

Chapter 3 is a post-print version of an article published as:

Valentin Buchter, Yih Ching Ong, **François Mouvet**, Abdallah Ladaycia, Elise Lepeltier, Ursula Rothlisberger, Jennifer Keiser and Gilles Gasser. Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis. *Chemistry - A European Journal*, **2020**, 26, 15232



My contributions: all computational studies.

3.1 Abstract

Schistosomiasis is a disease of poverty affecting millions of people. Praziquantel (PZQ), with its strengths and weaknesses, is the only treatment available. We previously reported findings on three lead compounds derived from oxamniquine (OXA), an old antischistosomal drug: ferrocene-containing (Fc-CH₂-OXA), ruthenocene-containing (Rc-CH₂-OXA) and benzene containing (Ph-CH₂-OXA) OXA derivatives. These derivatives showed excellent in vitro activity against both *Schistosoma mansoni* larvae and adult worms and *S. haematobium* adult worms, and were also active in vivo against adult *S. mansoni*. Encouraged by these promising results, we conducted additional in-depth preclinical studies and report in this investigation on metabolic stability studies, in vivo studies on *S. haematobium* and juvenile *S. mansoni*, computational simulations, and formulation development. Molecular dynamics simulations supported the in vitro results on the target protein. Though all three compounds were poorly stable within an acidic environment, they were only slightly cleared in the in vitro liver model. This is likely the reason why the promising in vitro activity did not translate into in vivo activity on *S. haematobium*. This limitation could not be overcome by the formulation of lipid nanocapsules as a way to improve the in vivo activity. Further studies should focus on increasing the compounds' bioavailability, to reach an active concentration in the microenvironment of the parasite.

3.2 Introduction

Schistosoma mansoni, *S. haematobium*, and *S. japonicum* account for over 90% of the cases of schistosomiasis, an acute and chronic parasitic disease that affects over 200 million people worldwide [20, 21, 22] and threatens more than 700 million people who are at risk of infection. [31] In children, schistosomiasis stunts physical growth and the ability to learn, while in adults, the disease affects the ability to work and can cause organ failure and, ultimately, death; a situation that causes an enormous socioeconomic burden for developing communities. [32] Praziquantel (PZQ) is the only drug being used for periodic mass drug administration to control the disease. Considering the imminent threat of resistance [33] and considering other drawbacks that PZQ presents, our efforts are oriented towards identifying and developing a new molecule with the potential to become an alternative therapeutic option in the treatment of this disease.

Oxamniquine (OXA, Fig. 3.1) is an anthelmintic drug developed in the 1960s [34] that showed high activity and a very convenient drug profile in terms of safety and ease of administration. It became the cornerstone of the schistosomiasis eradication program in Brazil in the past and at the beginning of the 21st century, but fell into disuse for two main reasons: it was only active against adult *S. mansoni* [34, 35, 36] and resistance was clinically confirmed. The drug was therefore no longer commercialized after 2010 and replaced by PZQ. OXA is a prodrug that needs to be activated by the sulfotransferase of *S. mansoni* (SmULT) to an alkylating molecule that binds proteins and DNA, consequently killing the parasite. [37]

Different enzyme orthologues are present in all *Schistosoma* species, but only the active site of the sulfotransferase of *S. mansoni* can activate OXA.[38]

OXA was commercialized as a 1:1 racemic mixture, with both enantiomers having antiparasitic activity, although the (S)-OXA contributes the most. Crystal structures of (R)-OXA and (S)-OXA complexes with SmSULT target show similarities in the modes of OXA binding, but only the (S)-OXA enantiomer is observed in the structure of the enzyme exposed to racemic OXA.[39]

The mechanism of resistance and lack of activity against some schistosome species has been well studied. Resistance is based in one or more point mutations in the enzyme's active site that prevent the molecule from being sulfonated.[40] Taking into account that there is a 70% homology between the amino acid sequences of the sulfotransferases of *S. mansoni* and *S. haematobium*,[38] we derivatized OXA based on the hypothesis that modification of OXA could overcome the species and stage specificity. [41]

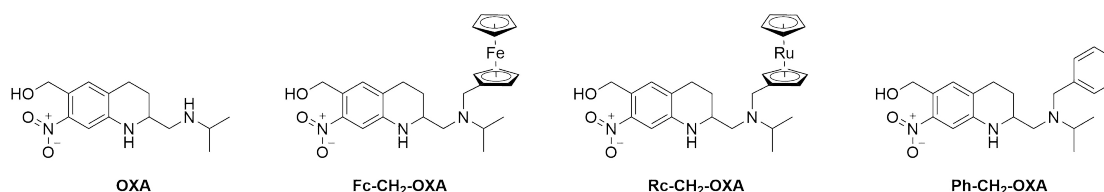


Figure 3.1: Structures of the compounds investigated in this study.

Previous studies by Jaouen and co-workers on the anticancer drug candidate ferrocifen [42, 43, 44] and Brocard, Biot and co-workers on the anti-malarial drug candidate ferroquine showed that the ferrocenyl analogues of tamoxifen and chloroquine, respectively, have improved bioactivity compared to the original organic drug compounds.[45, 46] This was due to several factors: the ferrocenyl component acted as a producer of reactive oxygen species (ROS), increased the lipophilic character of the molecule, and provided a mechanism of action different to that of the original drug.[47, 48, 49] With this concept in mind, we developed several metal-containing derivatives of OXA that were studied *in vitro* and *in vivo* against *Schistosoma* spp.[23, 50, 51] Among others, we demonstrated that the three derivatives of OXA, namely a ferrocene- (Fc-CH₂-OXA), ruthenocene- (Rc-CH₂-OXA), and benzene-containing (Ph-CH₂-OXA) derivative (Fig. 3.1), showed promising *in vitro* results, where all three OXA derivatives caused death of *S. mansoni* and *S. haematobium* adult worms [51] and worm burden reductions of 76 to 93% against adult *S. mansoni* *in vivo*. [23] Encouraged by our promising preliminary results, we decided to go further in the development and to fully characterize these three OXA analogues. *In vitro* studies were conducted against *S. mansoni* juvenile worms, and *S. japonicum* and *S. haematobium* adult worms. *In vivo* studies were carried out against adult *S. haematobium* and juvenile *S. mansoni* including studies with Ph-CH₂-OXA encapsulated in Lipid NanoCapsules (LNC). Only racemic mixtures were tested, as we focused on determining the presence of activity for the OXA derivatives *in vivo* with *S. mansoni* and *S. haematobium*, rather than on each of the enantiomers. As these drugs are intended to be administered orally, we evaluated the stability of the derivatives under acidic

Chapter 3. Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis

conditions. Our work was complemented by computational models and molecular dynamics simulations as well as microsomal stability and albumin binding studies.

3.3 Results and Discussion

3.3.1 In vitro studies

Fc-CH₂-OXA, Rc-CH₂-OXA, and Ph-CH₂-OXA were previously demonstrated to have promising activity as drug candidates in vitro against adult *S. mansoni* and *S. haematobium* and first stage of the larval development (NTS: newly transformed schistosomula) and in vivo against adult *S. mansoni*. [23, 51] To test the full potential of our compounds, we first elucidated the in vitro activity of the OXA derivatives against 28-day-old juvenile *S. mansoni* worms, because one of the important drawbacks of PZQ is its low activity against this developmental stage. All three compounds killed all the worms within 24 hours of incubation at a concentration of 100 μm (Table 3.1). Against juvenile *S. mansoni*, Fc-CH₂-OXA had the lowest IC₅₀ (0.5 μm) while Ph-CH₂-OXA was the drug with the highest IC₅₀ value (26.7 μm), i.e., the compound with the lowest activity. Interestingly, in the case of Ph-CH₂-OXA, the effect of the molecules on the juveniles was faster than against the adults: on adult *S. mansoni*, Ph-CH₂-OXA needed 72 hours to exert its maximal activity, whereas against juvenile stages, we observed a total lethal effect within 24 hours of incubation at a dose of 100 μm. For comparison, in the same incubation time, juvenile *S. mansoni* exposed to OXA showed an IC₅₀ of >100 μm, confirming previous studies of OXA being only slightly active in vitro and against juvenile stages of the parasite. [36, 51] Only after 72 hours of incubation at 100 μm, where the derivatives had long exerted their activity, a 48% reduction of the viability of OXA with respect to the control worms was found (Table 3.10).

Table 3.1: In vitro activity of Fc-CH₂-OXA, Rc-CH₂-OXA, and Ph-CH₂-OXA versus OXA against *S. mansoni*, *S. haematobium*, and *S. japonicum*.

Compound	<i>S. mansoni</i>				<i>S. haematobium</i>		<i>S. japonicum</i>
	IC ₅₀ adults 72h (μM)	IC ₅₀ adults in medium 45 g/L albumin. 72h (μM)	Onset of action on juveniles 100 μM (h)	IC ₅₀ 28 day juveniles 72h (μM)	IC ₅₀ adults 72h (μM)	IC ₅₀ adults in medium with albumin. 72h (μM)	IC ₅₀ adults 72h (μM)
Fc-CH ₂ -OXA	9.0	28.1	< 24	0.5	52.3	55.7	22.7
Rc-CH ₂ -OXA	6.0	NC	< 24	1.3	15.5	25.0	25.0
Ph-CH ₂ -OXA	13.5	90.7	< 24	26.7	32.6	70.6	70.6
OXA	>100	>100	72	>100	>100 *	ND	ND

* Hess et al.[23] , NC: no correlation, ND: not done.

Against *S. japonicum* adult worms, we observed the same behavior: while OXA was not active even with 100 μm after 3 days (Table 3.1 and Table 3.10), our derivatives showed considerable activity. Of the three derivatives tested, Fc-CH₂-OXA proved to be the most active of all three derivatives, killing all the parasites within 24 h at a concentration of 100 μm and having the

lowest IC50 value (22.6 μm). On *S. haematobium* instead, the most active compound was Rc-CH₂-OXA, also killing all parasites at a concentration of 100 μm and revealing the lowest IC50 value (15.5 μm).

Moreover, we incubated adult *S. mansoni* in medium containing albumin and compared the activity determined to our standard assay. A lower activity of the three drugs was observed in the enriched medium, with Fc-CH₂-OXA showing the least loss of activity of the three derivatives (Table 3.1). The albumin binding experiment was also performed for adult *S. haematobium*. Also in this case, the three compounds showed a reduction of the activity. These results are comparable to those obtained by Pasche et al.[52] who also identified a significant decrease in drug activity incubating antischistosomal drug candidates in vitro in the presence of albumin. PZQ also presents a high percentage (ca. 80 %) of drug bound to protein [53] and this might be a reason for the high doses needed to reach a significant effect. Protein binding is a major issue in drug development, since only the free fraction of the drug is able to interact with the target.[54]

3.3.2 Studies on juvenile *S. mansoni* in the mouse model

In terms of activity against juvenile parasites in vivo, we identified a lower activity of all three compounds in respect to the results on adult parasites.[23] Although the drug showed moderate activity, as shown by their shift to the liver due to the loss of vein attachment (data not shown), worm burden reductions were low, ranging from 39 to 47 % (Table 3.2). When considering the gender of the surviving worms, we found that there was no gender difference in susceptibility (binomial test, $p > 0.77$).

Table 3.2: Reduction of the juvenile worm burden in *S. mansoni* infected mice after treatment with 200 mg/kg of the OXA derivatives and after treatment with the nano encapsulated Ph-CH₂-OXA. Uncertainties are given by to the standard deviation.

Compound	No. mice	Worm burden [%]		WBR pure drug [%]		WBR nanocapsules [%]	
		Females	Total	Females	Total	Females	Total
Control group	8	6.5 ± 1.60	12.3 ± 2.50	-	-	-	-
Fc-CH ₂ -OXA	4	3.8 ± 0.96	7.5 ± 2.65	42.3 ± 14.7	38.8 ± 21.6	ND	ND
Rc-CH ₂ -OXA	4	3.0 ± 2.45	6.5 ± 4.43	53.8 ± 7.7	46.9 ± 36.2	ND	ND
Ph-CH ₂ -OXA	4	4.3 ± 0.50	7.5 ± 0.58*	34.6 ± 7.7	38.8 ± 4.7*	0	0

ND: Not done; WBR: worm burden reduction. *Statistically different from control on $p < 0.05$

3.3.3 In vivo studies on adult *S. haematobium*

Table 3.3 shows the worm burden reduction of the three compounds against *S. haematobium*: none of the compounds affected *S. haematobium* in vivo, contradicting the findings observed in vitro.

Chapter 3. Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis

Table 3.3: Change in the worm burden of *S. haematobium* infected hamsters after treatment with 200 mg/kg of the OXA derivatives.

Compound	No. of mice	Worm burden (SD)		WBR %
		Females	Total	
Control group	4	18.5 ± 8.3	40.0 ± 13.6	-
Fc-CH ₂ -OXA	4	25.8 ± 7.5	54.0 ± 16.2	0
Rc-CH ₂ -OXA	3*	55.0 ± 16.1	108.0 ± 28.9	0
Ph-CH ₂ -OXA	4	19.0 ± 3.7	51.0 ± 18.9	0

*One animal died during the experiment.

3.3.4 Computational studies

To understand the interaction between OXA analogues and the sulfotransferase proteins from *S. mansoni* (SmSULT) and *S. haematobium* (ShSULT), we performed classical molecular dynamics simulations of OXA, Fc-CH₂-OXA, Ph-CH₂-OXA, and Fc-CO-OXA to determine their binding poses within the active site of the two sulfotransferases at body temperature. Fc-CO-OXA was added to the comparison to include the case of a derivative that was shown to be less active against *S. mansoni* in vitro than the other compounds considered in the present study.[23]

We did not consider Rc-CH₂-OXA explicitly because, both from a geometrical and electrostatic point of view, the force field models for ferrocenyl and ruthenocenyl compounds are very similar and would likely yield comparable behavior at this simplified level of theory.

The free-energy difference between bound and unbound states of a receptor-ligand complex is a direct measure of the binding affinity. To estimate this quantity from our trajectories, we used the MM/PBSA (Table 3.4) and MM/GBSA methods (Table 3.8). All binding free energies are negative, meaning that the bound state is energetically favorable for all compounds. No systematic difference can be noted between the two proteins and all modified OXA compounds show a higher binding affinity than OXA itself. This is probably due to their larger size, forcing a tighter fit inside the protein and increasing the number of interactions with the binding pocket. Consequently, all analogues are strongly bound to their target proteins.

Table 3.4: Estimated binding free energies computed by the MM/PBSA method (kcal/mol).

Compound	<i>S. mansoni</i>		<i>S. haematobium</i>	
	R	S	R	S
Fc-CH ₂ -OXA	-39.3	-41.9	-36.3	-39.9
Fc-CO-OXA	-36.3	-29.9	-26.4	-32.0
Ph-CH ₂ -OXA	-30.3	-35.0	-36.6	-18.3
OXA	-12.5	-8.0	-17.1	-22.8

Since the drugs are supposed to react with PAPS within the target protein, the distance between the closest oxygen of the sulfate group of PAPS and the hydrogen in the hydroxyl group of each drug was measured every 40 ps. These distances are represented as histograms in Fig. 3.2 and their average is shown in Table 3.5. The near-attack configurations (NAC), i.e., those with

shorter distances between the reactive groups, are more likely to result in the activation of the compounds.

Table 3.5: Average O-H distance with standard deviation

Compound	S. mansoni		S. haematobium	
	R	S	R	S
Fc-CH ₂ -OXA	5.1 ± 1.1	1.8 ± 0.1	11.8 ± 0.7	6.1 ± 0.8
Fc-CO-OXA	12.2 ± 1.2	10.0 ± 0.9	12.8 ± 1.9	6.2 ± 0.7
Ph-CH ₂ -OXA	2.2 ± 0.7	2.0 ± 0.7	14.5 ± 1.1	4.3 ± 0.5
OXA	3.3 ± 0.4	3.5 ± 0.3	9.5 ± 1.6	6.1 ± 0.8

For OXA in SmSULT, the sulfate-hydroxyl distance is short and stable for both enantiomers, with an average slightly above 3 Å. The relative orientation of the reactive groups is fixed through their mutual interaction with ASN230, thus promoting the interaction (Fig. 3.5). The chain of residues 19 to 23 also binds to PAPS and OXA at different places, stabilizing the configuration. For OXA, the difference between R- and S-enantiomers is very small, which is consistent with previous experimental works that reported that both enantiomers bind in a similar fashion and that the activity difference originates from their binding kinetics.[39]

Both enantiomers of Ph-CH₂-OXA and Fc-CH₂-OXA show a very strong interaction with PAPS within SmSULT. Residues 18 to 21 bind both ligands in several places, highly stabilizing their binding (Figs. 3.7 and 3.9). Moreover, the residue ASN230 also binds to PAPS and to the chain of residues 18–21, further increasing the stability. This may explain the high activity of these compounds in vitro. On the other hand, the R-enantiomer of Fc-CH₂-OXA seems to drift slowly away from PAPS and a proper equilibrium is never reached within our simulation time. The distance increases over time, which suggests that the final configuration will not be active. Strikingly, Ph-CH₂-OXA does not seem to be impacted by the enantiomer selectivity observed for all other derivatives.

Fc-CH₂-OXA, which was the most active against *S. mansoni* according to the in vitro experimental studies, shows short average distance to PAPS. In *S. haematobium* instead, the shortest distance to PAPS is observed for the S-enantiomer of Ph-CH₂-OXA, which also shows a lower IC₅₀ against this species when compared to Fc-CH₂-OXA. Furthermore, our simulations of Fc-CO-OXA in SmSULT show NAC distributions in which the reactants are about three times more distant than in OXA, with averages above 10 Å. These configurations are very unlikely to activate the drug, which is again consistent with experimental results.

In practice, OXA is not active on *S. haematobium*. In simulations in which OXA is docked into ShSULT, we observe that the distance between the reactive groups is much larger than in SmSULT, by more than 6 Å on average. Furthermore, we observe a large difference between enantiomers, where the R-enantiomer shows a broad distribution of NAC distances with an average of 9.5 Å. This strong enantiomer dependence is present for all compounds and is more pronounced than in SmSULT with all R-enantiomers adopting less reactive configurations. The residue ASP80 seems to have a negative impact, forming a stable bond with the drug's hydroxyl group (Figs. 3.6, 3.8 and 3.10). This bond is far from PAPS, leading to unreactive

Chapter 3. Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis

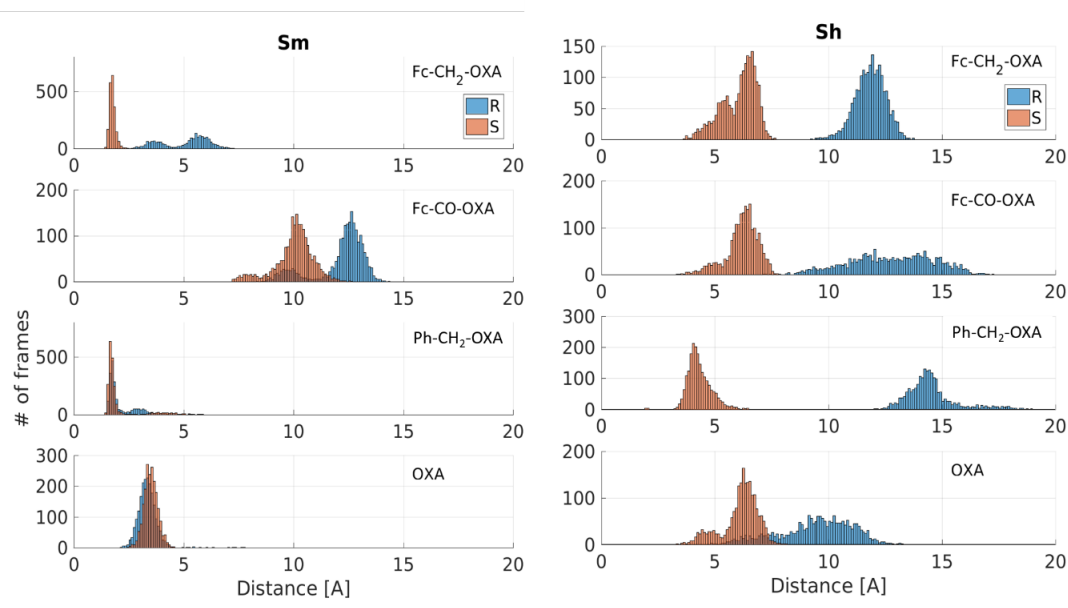


Figure 3.2: Distance between the closest oxygen of PAPS and the hydroxyl group of the analogues; (R)-enantiomers in blue and (S)-enantiomers in orange.

geometries shown by a widening of the distance distributions between the compounds and PAPS. This residue is conserved in SmSULT but stays at larger distances (more than 4.5 Å from the hydroxyl group of all considered compounds).

In ShSULT, the S-enantiomer of Ph-CH₂-OXA shows the most promising results. Our in vitro experiments showed that this compound was active against *S. haematobium* and our structures show significant improvements over OXA. The distances between reactant groups are significantly smaller, at about 4.3 Å on average and only slightly larger than for OXA in SmSULT, indicating that the compound could be activated. We also noted that the difference between enantiomers is much larger in ShSULT than in SmSULT, suggesting that using enantiopure samples of Ph-CH₂-OXA instead of a racemic mixture may significantly improve the drug's efficacy against *S. haematobium*. However, given the overall low activity of the racemate in vivo, this specific point was not further evaluated.

3.3.5 Stability of OXA Analogues in acidic environments and in the presence of microsomes

We further investigated whether the limited in vivo activity on juvenile *S. mansoni* and adult *S. haematobium* could be explained by physiological stability issues. We therefore evaluated two different conditions: an acidic environment and the co-incubation in the presence of liver microsomes, to simulate the environments within the stomach and the liver, respectively.

Simple HPLC methods with a short run time were used to visually check the elution of the

fragments before and after exposure to 1 M HCl (Fig. 3.11). Since all compounds after 24 h incubation with 1 M HCl produce completely different elution peaks after incubation, all three compounds have little stability in acidic media.

We conducted the metabolic stability assays using commercially available human liver microsomes, which are specific for Phase I processes catalyzed by cytochrome P450 (CYP) monooxygenases and flavin containing monooxygenases (FMO). We selected human instead of murine microsomes because the human is the final species of interest. Nonetheless, the results of these metabolic studies can give us an inference of the behavior in our mouse experiments.

The metabolic stability results are summarized in Table 3.6. The stability of the compounds decreased exponentially with >40 % compound remaining after 24 h, with similar half-life values ranging from 2.2 to 3.8 h. The intrinsic clearance of the compounds was low and intermediate, ranging from 7.5 to 13.3 $\mu\text{L min}^{-1} \text{mg}^{-1}$. [55]

From the values of intrinsic clearance in Table 3.6, according to McNaney et al., Fc-CH₂-OXA could be categorized as “low” clearance, whereas Rc-CH₂-OXA, Ph-CH₂-OXA, and OXA would be categorized as “intermediate”. [56, 57]

The elution peaks showed (Figs. 3.12 to 3.14) for Fc-CH₂-OXA, Rc-CH₂-OXA, and Ph-CH₂-OXA that only the Ph-CH₂-OXA derivative remained unchanged in an acceptable 56.

We previously reported [23] excellent in vivo efficacy in the *S. mansoni* murine model at 100 mg kg^{-1} . At that time, the stability of the compounds was unknown. Based on the stability results we obtained, it is entirely possible to attribute some activity of the derivatives to OXA itself, in addition to the derivatives' own bioactivity. Furthermore, we could infer that the factors of poor metabolic stability, solubility, and permeability of the compound to the membrane contributed to the poor in vivo results. To have a better idea of the permeability of Ph-CH₂-OXA (deemed the most stable candidate based on stability testing), we proceeded to calculate the MW, logP value, and the number of hydrogen-bond donors and hydrogen-bond acceptors, by using the software Molinspiration [58] to estimate its theoretical solubility and membrane permeability. These values gave us a rational basis for selecting a compound for formulation studies aimed at improving bioavailability. Traditionally, therapeutics have been small molecules that fall within Lipinski's rule of five [59] (i.e.; a molecule with a molecular mass ≤ 500 Da, ≤ 5 hydrogen bond donors, ≤ 10 hydrogen bond acceptors, and an octanol-water partition coefficient $\log P \leq 5$). Molecules violating more than one of these rules may exhibit limited bioavailability. Ph-CH₂-OXA has a molecular mass of 369 Da, 6 H-bond acceptors, 2 H-bond donors, and a calculated logP value of 4.26 with 7 rotatable bonds, which were well within the parameters of being an ideally permeable small molecule (Table 3.9).

It is important to note that it is solubility, permeability, and metabolic stability, collectively that have a bearing on the bioavailability of oral drugs. Ph-CH₂-OXA was active in vitro, and the in silico studies also predicted a very promising interaction with the parasite's active site on both evaluated species. Additionally, since Ph-CH₂-OXA was metabolically the most

Chapter 3. Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis

stable compound and with predicted acceptable permeability according to Lipinski's criteria, we proceeded with Ph-CH₂-OXA for further nanoencapsulation formulation studies on the hypothesis that the nanoencapsulated compound would have increased overall solubility and reduced degradation within the acidic environment of the stomach, a new panorama, which would allow better delivery of the compound to its parasitic target.

3.3.6 Lipid nanocapsules loaded with Ph-CH₂-OXA

Lipid nanocapsules (LNC) are the vector of choice to encapsulate lipophilic molecules.[60] LNC present an oily core formed of medium-chain triglycerides covered by a membrane made from a mixture of lecithin and a PEGylated surfactant[60] (Fig. 3.3): this core is able to encapsulate various lipophilic compounds and LNC have already been intensively used for in vivo administration of ferrocifens.[61, 62, 63, 64]

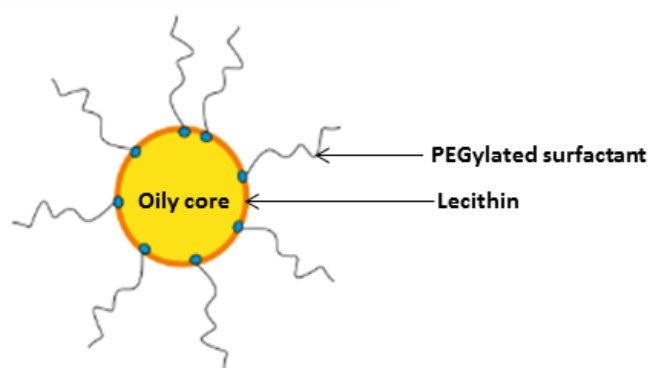


Figure 3.3: Schematic representation of LNC.

Table 3.6: Metabolic stability in human microsomes.

Compound	t _{1/2} [h]	k [min ⁻¹]	CL _{int} [μL min ⁻¹ mg ⁻¹]
Fc-CH ₂ -OXA	3.8	0.003	7.5
Rc-CH ₂ -OXA	2.2	0.0053	13.3
Ph-CH ₂ -OXA	2.4	0.0048	12
OXA	3.1	0.0037	9.3

Ph-CH₂-OXA was encapsulated at a concentration of 32.35 mg per mL of LNC, representing a drug loading of 4.35 % w/w. The physicochemical parameters (characterized by dynamic light scattering analysis), of blank LNC and Ph-CH₂-OXA loaded LNC were the same: diameters of approximately 50 nm, a PDI value below 0.2, which demonstrates a monodispersed population of nanoobjects, and a zeta potential close to neutrality as shown in Table 3.7.

The in vivo activity of the drug loaded nanocapsules was evaluated in mice harboring a 21-day *S. mansoni* infection, but this improvement in formulation was not translated into an increase in the activity, as summarized in Table 3.2. The low activity could be the result of a slow or even an absence of drug release, and/or a limited pathogen LNC internalization. In a study

Table 3.7: Physicochemical parameters characterized by DLS of empty LNC (blank) and Ph-CH₂-OXA loaded LNC.

Sample	Diameter [nm]	pdi*	Zeta potential [mV]
Blank LNC	57.4 ± 0.9	0.07 ± 0.02	3.9 ± 0.7
Ph-CH ₂ -OXA	53.0 ± 0.5	0.04 ± 0.02	3.1 ± 0.1

* pdi = polydispersity index

investigating the release of different dyes from LNC to a lipophilic compartment mimicking the cells' lipid membrane, lipophilic indocarbocyanine dyes were reported to stay entrapped in the surfactant shell of the LNC and no transfer was observed.[65] In a similar paper, studying the transfer in vivo of different dyes from LNC to different lipophilic acceptors, the absence of release of lipophilic indocarbocyanine from LNC was confirmed.[66] For Ph-CH₂-OXA, being similar in terms of hydrophobicity, the same behavior could be hypothesized and explain the lack of increase of activity of the LNC in comparison to the pure compound.

3.4 Conclusion

We followed up on three OXA analogues that had shown promising antischistosomal activity. The computational models forecast that the ferrocene- and benzene-containing analogues sample far more near attack configurations with the target sulfotransferase than the parent compound (OXA) for *S. mansoni*. In contrast, in *S. haematobium*, only the S-enantiomer of Ph-CH₂-OXA shows the most significant improvement over OXA and could be active against this species. In vitro, the derivatives showed improved activity compared to OXA, against adult worms of all three species evaluated (*S. mansoni*, *S. haematobium* and *S. japonicum*) and against juvenile *S. mansoni*. When considering the in vivo studies instead, we measured a lack of activity for all three derivatives against juvenile *S. mansoni* and adult *S. haematobium*. We found that all three compounds were only slightly cleared in the in vitro liver model but were poorly stable within an acidic environment. This is likely the reason why the promising in vitro results did not translate into in vivo activity. We further evaluated whether lipid nanoencapsulation of the lead compound (Ph-CH₂-OXA) could overcome this limitation, but unfortunately the formulated compound was also inactive. Since the stability and not the activity on the target seems to be the main limitation of these molecules, further steps should include additional strategies for improved drug formulation, to establish whether an enhanced bioavailability can overcome the loss of in vivo activity.

3.5 Experimental and computational details

¹H and ¹³C NMR spectra: All chemicals were either commercially available or were prepared by following standard procedures. Solvents were used as received or distilled using standard procedures. All preparations were carried out using standard Schlenk techniques.

Chapter 3. Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis

¹H and ¹³C NMR spectra were recorded in deuterated solvents with a Bruker 400 or 500 MHz spectrometer at RT. The chemical shifts, δ , are reported in ppm (parts per million). The residual solvent peaks have been used as internal references. The abbreviations for the peak multiplicities are as follows: s (singlet), d (doublet), t (triplet), m (multiplet). ESI mass spectrometry was performed with an LTQ-Orbitrap XL from Thermo Scientific. Elemental analysis was performed at Science Centre, London Metropolitan University with a Thermo Fisher (Carlo Erba) Flash 2000 Elemental Analyser, configured for % CHN. The scanned spectra of the compounds are available in the Supporting Information.

OXA-Derivatives preparation: The OXA derivatives were prepared starting from the parent compound oxamniquine (Pfizer) as described previously.[23] The analytical data matched that previously reported and is available in the Supporting Information for further reference. [23]

Animals and parasites: All animal experiments were conducted at the Swiss Tropical and Public Health Institute (Swiss TPH) and authorized by the animal welfare office Kanton Basel Stadt, Switzerland (Authorization no. 2070).

NMRI female mice were purchased from Charles River (Sulzfeld, Germany) at the age of three weeks and were left without intervention for one week of acclimatization. Mice were infected with a subcutaneous infection of around 100 cercariae in the back of the neck by following the procedure described by Lombardo et al. [67]

For the *S. haematobium* chronic infection, one-month old LVG hamsters (Charles River, NY) were provided by the National Institutes of Health (NIH)-National Institute of Allergy and Infectious Diseases (NIAID) Schistosomiasis Resource Center (SRC) for distribution by the Biomedical Research Institute in Rockville, USA, which were pre-infected with 350 *S. haematobium* cercariae. The animals were kept in the animal facility with humidity and light control (50% -12/12) for three months.

Swiss Webster mice infected with *S. japonicum* (Philippine strain) were also obtained from NIH NIAID SRC for our in vitro studies.

In vitro studies: Adult *S. mansoni*, *S. haematobium* and *S. japonicum* worms were collected by dissection from the mesenteric veins. Until use (within 24 h after dissection) and during the experiments, the worms were kept in an incubator at 37 °C and 5% CO₂ and the culture medium consisted of RPMI 1640 (Gibco-Thermo Fisher, Waltham, MA USA) supplemented with 1% penicillin/streptomycin (BioConcept, Allschwil, Switzerland) and 5% Fetal Calf Serum (FCS) (BioConcept). The control groups consisted of culture medium spiked with dimethyl sulfoxide (DMSO) at a concentration of 1% or 0.5%, equivalent to the content of DMSO present in the wells of worms treated with the highest drug concentration for that assay. The concentrations evaluated were 100, 50, 25, 12.5, and 6.25 μ m. The studies on *S. mansoni*

3.5 Experimental and computational details

and *S. japonicum* were performed as duplicates and repeated once, while the studies on *S. haematobium* were performed in duplicate. Every study condition included at least three worms.

We further evaluated the effect of the addition of albumin (AlbuMAX II, Gibco) to the culture medium to investigate if the activity of the derivatives was different. For this we set the same assay design as described before and added 45 g L⁻¹ albumin to the culture medium, corresponding to the content of albumin within the range of human plasma.[68] We performed the study with medium containing albumin on adult *S. mansoni* (duplicate and repeated once), and on adult *S. haematobium* (in duplicate). To assign a score to the viability, we used a previously described method [69] scoring motility, viability, and morphological alterations using a bright field inverted microscope (Carl Zeiss Oberkochen, Germany, magnification ×4 and ×10).

S. mansoni juvenile worms for in vitro studies: 28 days after infection, mice were euthanized and juvenile worms were obtained by blood perfusion. The perfusion solution consisted of 8.5 g L⁻¹ NaCl, 7.5 g L⁻¹ Na-citrate in distilled water. Parasites were in different stages of development, as reported elsewhere. [70] All juvenile worms were kept in culture medium as described before for adult *S. mansoni* until use within 24 h. To test the activity of the derivatives against the juvenile stages, we incubated the worms with a 100 μm concentration of each of the derivatives in duplicate and at least two worms per well. Duplicates of 100 μm OXA and 1% DMSO served as control conditions.

Calculation of IC₅₀ values: CompuSyn 1.0 (ComboSyn Inc, 2007) was used to calculate the IC₅₀ values of each of the derivatives after an incubation period of 72 h. Equation 1 was used to normalize the scores of the treated worms to the controls:

Statistical analysis: All statistics were performed using RStudio version 3.5.1.[71] For evaluating significance in the in vivo studies we applied the non-parametric Kruskal Wallis test for multiple comparisons, and the Dunn test with Bonferroni correction for individual differences. Significance was defined as an adjusted p value <0.05. To evaluate whether there was difference in gender susceptibility we applied the binomial test.

Drug suspension for oral administration: The derivatives were administered to the animals in the form of an oral suspension. The compounds were first dissolved in DMSO (Sigma-Aldrich, Buchs, Switzerland) corresponding to 10% of the total volume, and then a mixture of Tween 80 and ethanol in a proportion of 70:30 was added, also corresponding to 10% of the final volume. The remaining 80% of the volume consisted of distilled water, which was added under stirring in small aliquots.

Chapter 3. Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis

In vivo activity on juvenile *S. mansoni*: 21 days after infection, mice were treated orally with each of the derivatives and the nanocapsules loaded with Ph-CH₂-OXA at a concentration of 200 mg kg⁻¹. The control group consisted of four mice, which were infected on the same day and under the same conditions as the treatment arm, but were not treated, avoiding additional stress, in accordance to animal welfare regulations. Four weeks after treatment (seven weeks after infection), when the worms reached the adult stage of development, the mice were euthanized with CO₂, dissected, and the remaining live worms were picked out from the mesenteric veins and liver, sexed and counted.

In vivo activity on adult *S. haematobium*: The hamsters harboring an adult infection were treated with a dose of 200 mg kg⁻¹ of the compounds. Three weeks after treatment, the hamsters were euthanized with CO₂, dissected, and the worms were picked out from the mesenteric veins and liver, sexed and counted. The control group consisted of four untreated hamsters.

Computational studies: Classical molecular dynamics simulations were performed on the target proteins, that is, the sulfotransferase of *S. mansoni* (SmSULT) and of *S. haematobium* (ShSULT), in complex with OXA, Fc-CH₂-OXA and Ph-CH₂-OXA. Since the two enantiomers of OXA show different activities against schistosomes, [39, 72] we performed separate simulations for both enantiomeric forms of all molecules.

To develop force field parameters for the drugs, the geometry of all compounds was optimized in the gas phase with the Gaussian 16 software package, [73] using DFT with B3LYP functional and a 6-31+G* basis set for non-metallic atoms and LANL2DZ pseudopotential for the iron atom. The initial geometries were taken from crystallized (S)-OXA complexed to SmSULT (PDB: 4MUB11). Hess et al. crystallized in 2017 the (R)-enantiomer of Fc-CO-OXA [23], and we used this structure as a starting point for all (R)-isomers. The electrostatic potential was computed with the same functional and basis set to estimate the effective atomic point charges through RESP fitting.[74]

Classical molecular dynamics simulations were performed using Amber16.[75] The protein was modeled using the FF14SB force field and the Generalized Amber Force Field 2 (gaff2) was used as a base for the ligands. The ferrocenyl group was modeled using the force field published by Doman et al.[76] To estimate the missing dihedral parameters between the ferrocenyl group and the rest of the molecule, we scanned the angles of interest and performed DFT single-point energy calculations for all angles, using the same functional, basis set and pseudopotential as for the geometry optimization. We then used the software *paramfit* from AmberTools16 [75] to estimate the parameters. To prevent clashes between nuclei during the rotation, we performed these computations on subsystems containing solely atoms that are relevant for these parameters (e.g., Fc-CH₂-NH₂, Fc-CH₂-N-(CH₃)₂, Fc-CO-NH₂ and Fc-CO-N-(CH₃)₂). All parameters determined in this way accurately reproduce the corresponding ab

initio energy profiles (Fig. 3.8).

The crystallographic structures of SmSULT and ShSULT with OXA and the cofactor 3'-phospho-adenosine-5'-phosphate (PAP) served as a starting point for our simulations (PDB: 4MUB and 5TIY,[39, 72] respectively). OXA was replaced by its derivatives through alignment of their shared atomic groups. 3'-phosphoadenosine-5'-phosphosulfate (PAPS), the active version of PAP, was inserted in the same way by minimizing the distance between shared PAP atoms. Missing loops of the protein were added using the ModLoop web server.[77] Using *tleap*, the resulting system was put in an 84×84×84 Å periodic box filled with explicit TIP3P water molecules. Finally, the total charge was neutralized by adding Na⁺ counterions. The resulting systems consisted of about 50,000 atoms.

Classical trajectories were computed using Amber's CUDA version of the PMEMD program. After minimization, the SHAKE algorithm [78] was used to constrain covalent bonds involving hydrogen atoms and the system was heated to body temperature ($T = 310$ K) in two steps using a Langevin thermostat. First, the water molecules were heated to the target temperature while restraining the positions of the ligand and protein during 50 ps with a time step of 1 fs. Then, the restraints were released, and the whole system was thermalized in the NPT ensemble for 400 ps, with a time step of 1 fs and a pressure relaxation time of 3 ps. We then performed NPT simulations for 40 ns with a time step of 2 fs to reach an equilibrium state of the system. We finally performed 80 ns NPT production runs that were used for analysis.

The binding free-energy of the ligands inside the protein was estimated based on the MD trajectories using two methods available in Amber16: Generalized Born Surface Area (MM/GBSA) and Poisson Boltzmann Surface Area (MM/PBSA).[79]

pH and metabolic stability studies: The stability of the three OXA derivatives was studied by in vitro co-incubation in acidic environment. To 200 μ L of HPLC grade MeCN in a 1.5 mL Eppendorf, 2 μ L of 37% HCl were added to form a final 0.1 m HCl solution. Test compound (10 μ L; 5 mm in HPLC MeCN) was then added to this acidic solution. For the non-acidic control samples, test compound (10 μ L; 5 mm in HPLC MeCN) was added to 200 μ L of HPLC grade MeCN. The compounds were then incubated for 24 h at 37C and assessed at t=0 h and t=24 h.

Analytical HPLC measurement was performed with a 1260 Infinity HPLC System (Agilent Technology) comprising: 2 ×Agilent G1361 1260 Prep Pump system with Agilent G7115A 1260 DAD WR Detector equipped with an Agilent Pursuit XRs 5C18 (100 Å, C18 5 μ m 250×4.6 mm) Column. The solvents were acetonitrile (HPLC grade) and purified water (Pacific TII) with flow rate 1 mL min⁻¹. Detection was performed at 215, 250, 350, 450, 550, and 650 nm with a slit of 4 nm. The flow rate was 1 mL min⁻¹ and the max pressure was set 200 bar. Run parameters were as follows: 0 min 85% acetonitrile (MeCN) 15% H₂O; 3 min 85% MeCN 15% H₂O; 7 min 100% MeCN; 9 min 100% MeCN, 11 min 85% MeCN 15% H₂O.

The metabolic stability of the three OXA derivatives was studied by in vitro co-incubation with

Chapter 3. Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis

human liver microsomes. All three compounds were incubated in the presence of NADPH at 37 °C. The protocol was adapted from previous studies:[80, 81] 10 μL of 20 mg mL⁻¹ microsomes (GIBCO, 50 pooled), 463 μL of PBS (GIBCO, 1 \times PBS) and 2 μL of 40 mM NADPH (Sigma) were added to 1.5 mL Eppendorf tubes and incubated at 37 °C for 10 min to prime the microsomes. Following this, 10 μL of 50 mM test compound and an additional 15 μL of NADPH were added (1 mM final concentration of test compound in 500 μL total volume). The samples were incubated at 37 °C, and quenched at the desired time points of 1, 4 and 24 h by adding 2 mL of dichloromethane (DCM) or any other organic solvent. 2.5 μL of 5 mM caffeine (TCI Chemicals) in HPLC MeCN as internal standard were added during the quenching process. The mixture was shaken for 10 min to ensure complete extraction. The DCM layer was carefully removed and then evaporated to provide residues that were analyzed by LC-MS (HPLC Waters 2525/Mass Spectrometry Waters ZQ 2000) using a pure acetonitrile-water system with the same column as above. The residues were dissolved in 100 μL HPLC grade acetonitrile. 20 μL from each MeCN residue sample was injected manually by using the following run parameters: 3 min 5% MeCN 95% H₂O; 13 min 40 % MeCN 60% H₂O; 14 min 100% MeCN 0% H₂O; 20 min 100% MeCN 0% H₂O; 23 min 5% MeCN 95% H₂O. UV spectra were analyzed and compared at different time points.

By comparing the differences in respective m/z values in the MS spectra, m/z values for the parent compound and OXA could be identified. Semiquantitative analysis of the ratio of parent compound and different metabolites present in the mixture after incubation with human liver microsomes was achieved by comparing the areas under the respective peaks of different compounds visible in the UV traces of the LC analysis at 245 nm. To determine the in vitro half-life ($t_{1/2}$), the following process was derived from Tan et al.[82] The peak areas of the compounds at different time points are expressed first as a percentage of the internal standard, caffeine (Eq. (3.1)):

$$\text{Ratio at time point} = \frac{\text{Area under curve of compound}}{\text{Area under curve of internal standard}} \quad (3.1)$$

Following this, normalized ratios were calculated using the ratio of peak area of the test compounds to caffeine at t=0 h. Normalized ratios were calculated at each assessed time point of t=1 h, 4 h and 24 h (Eq. (3.2)):

$$\text{Normalized ratio} = \frac{\text{Ratio at } t \neq 0}{\text{Ratio at } t = 0} \quad (3.2)$$

The normalized ratio values are then plotted against incubation time. The $t_{1/2}$ values calculated via analyses methods in GraphPad Prism 8 (nonlinear regression, exponential one phase decay). The degradation rate constant, k was then calculated using the $t_{1/2}$ values (converted from hours into minutes). The predicted in vitro intrinsic clearance values (expressed in $\mu\text{L min}^{-1} \text{mg}^{-1}$ protein) were then calculated as a ratio of the degradation rate constant k (expressed in min^{-1}) and the microsomal protein concentration ($\text{mg } \mu\text{L}^{-1}$) (Eq. (3.3)):

3.5 Experimental and computational details

$$CL_{int} = \frac{k}{\text{microsomal protein content}} \quad (3.3)$$

Ph-CH₂-OXA loaded lipid nanocapsules: Lipid nanocapsules (LNC) were formulated by the phase inversion method.[83] Briefly, to prepare blank LNC, Labrafac® (Gattefossé SAS, France, 20.6% w/w), Lipoid® S 100 (Ludwigshafen, Germany 1.5% w/w), Kolliphor HS 15 (Florham Park, USA 17% w/w), NaCl (Sigma-Aldrich, USA 1.3% w/w) and water (59.6% w/w) were mixed and homogenized under magnetic stirring at 80 °C. Three cycles of progressive heating and cooling between 90 and 50 °C were then performed. During the last cooling cycle, the mixture was diluted by adding 2 °C purified water (28.7% v/v) in order to induce an irreversible shock and formulate LNC. To encapsulate Ph-CH₂-OXA inside the LNC, some slight changes to this protocol have been applied. The Ph-CH₂-OXA (4.35% w/w) was mixed with Labrafac®, Lipoid®S 100 and ethanol (Fisher, USA) to help the solubilization of the molecule in the lipid phase. This mixture was put under agitation at 50 °C until total solubilization of the Ph-CH₂-OXA. The ethanol was then evaporated under argon. Once the ethanol evaporated, Kolliphor HS 15, NaCl and water were added and three heating and cooling cycles were performed as prescribed for formulating blank LNC. During the last cooling cycle, the mixture was diluted by adding 2 °C purified water. Empty LNC and Ph-CH₂-OXA loaded LNC were characterized using dynamic light scattering (DLS) to determine their size, polydispersity index (PDI), and zeta potential.

NMRI female mice were purchased from Charles River (Sulzfeld, Germany) at the age of three weeks and were left without intervention for one week of acclimatization. Mice were infected with a subcutaneous infection of around 100 cercariae in the back of the neck, by following the procedure described by Lombardo et al.[42]

For the *S. haematobium* chronic infection, one-month old LVG hamsters (Charles River, NY) were provided by the National Institutes of Health (NIH)–National Institute of Allergy and Infectious Diseases (NIAID) Schistosomiasis Resource Center (SRC) for distribution by the Biomedical Research Institute in Rockville, USA, which were pre-infected with 350 *S. haematobium* cercariae. The animals were kept in the animal facility with humidity and light control (50% - 12/12) for three months.

Swiss Webster mice infected with *S. japonicum* (Philippine strain) were also obtained from NIH NIAID SRC for our in vitro studies.

All animal experiments were conducted at the Swiss Tropical and Public Health Institute (Swiss TPH) and authorized by the animal welfare office Kanton Basel Stadt, Switzerland (Authorization no. 2070).

3.6 Supplementary Figures

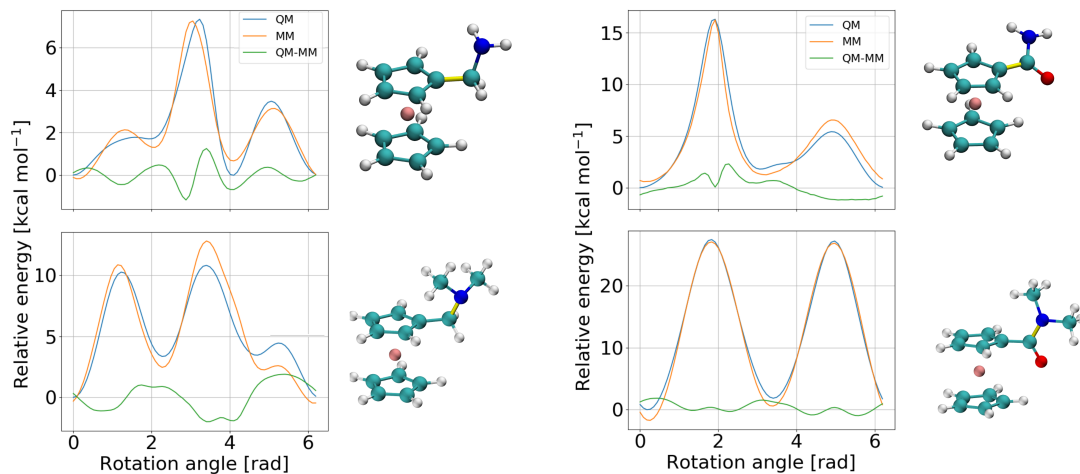


Figure 3.4: Rigid dihedral scan energy profiles for representative subsystems to determine missing force field parameters. The ab initio energy profile is plotted along with the fitted classical model. The rotation occurs around the bond represented in yellow and starts from the optimised geometry.

Table 3.8: Estimated binding free energies computed by the MM/PBSA method (kcal/mol).

Compound	<i>S. mansoni</i>		<i>S. haematobium</i>	
	R	S	R	S
Fc-CH ₂ -OXA	-52.04	-52.36	-46.57	-47.52
Fc-CO-OXA	-41.99	-40.85	-30.67	-45.81
Ph-CH ₂ -OXA	-43.48	-51.70	-45.49	-32.06
OXA	-29.40	-30.30	-22.09	-31.91

Table 3.9: Estimated binding free energies computed by the MM/PBSA method (kcal/mol).

Compound	miLogP	TPSA	nON	nOHNH	Nrotb	Nviolatn	Vol	MW
Fc-CH ₂ -OXA	3.22	81.32	6	2	9	0	435.69	477.39
Rc-CH ₂ -OXA	3.90	81.32	6	2	9	1	435.69	522.61
Ph-CH ₂ -OXA	4.26	81.32	6	2	7	0	351.94	369.46
OXA	2.62	90.11	6	3	5	0	263.34	279.34

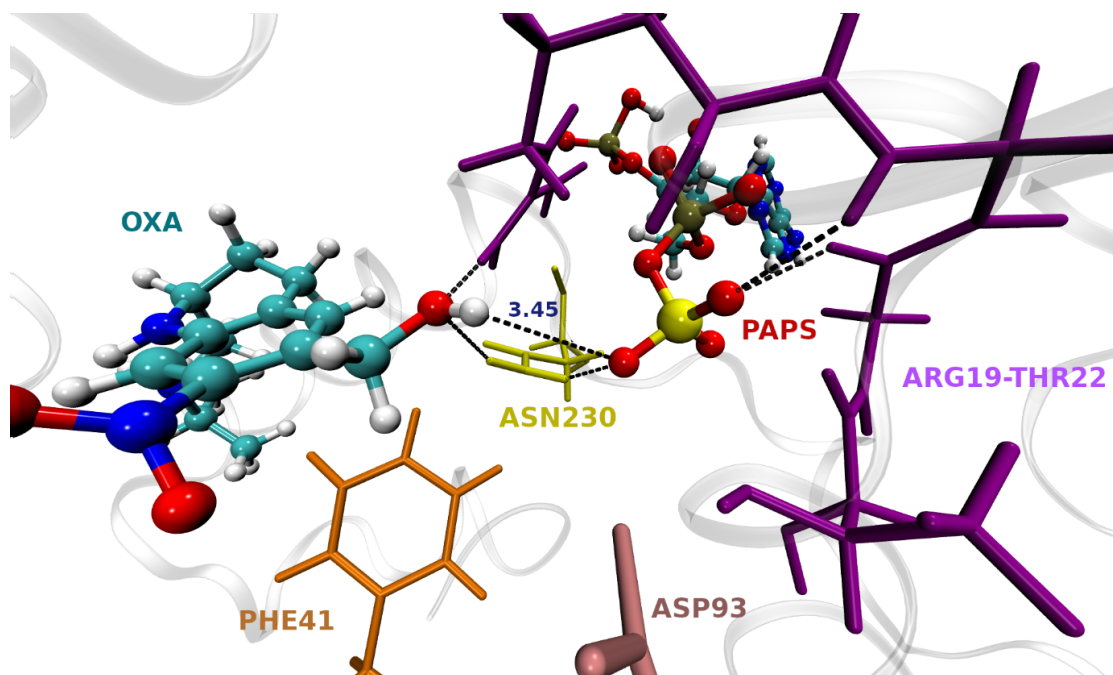


Figure 3.5: Snapshot of the simulation of S-OXA in SmSULT. The contacts represented are stable along the simulation.

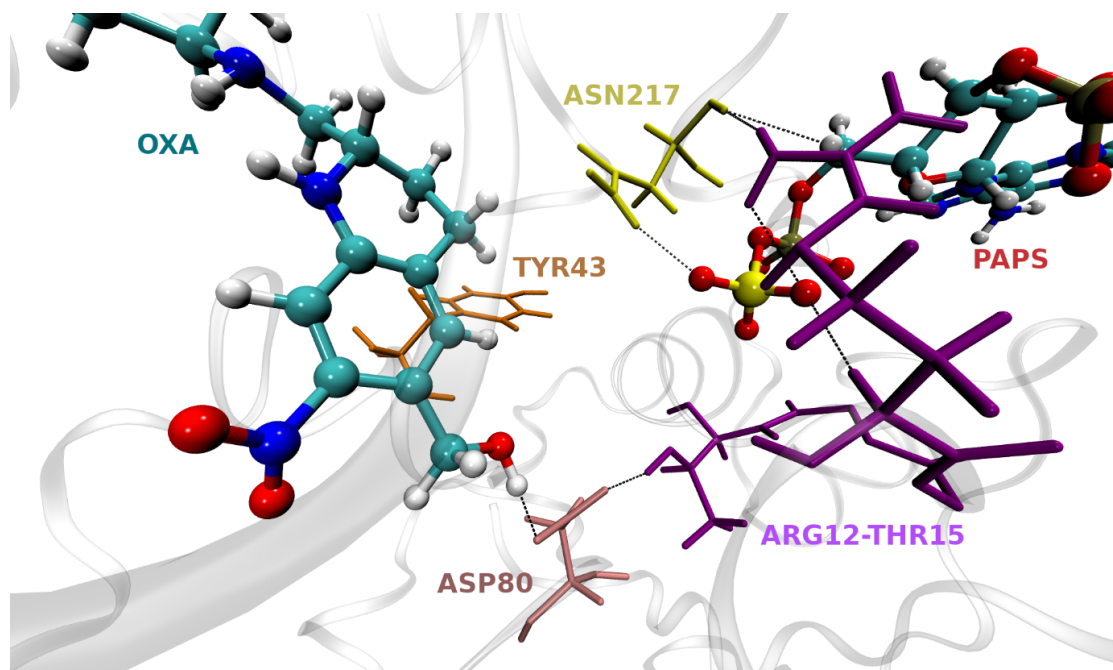


Figure 3.6: Snapshot of the simulation of S-OXA in ShSULT. The contacts represented are stable along the simulation.

Chapter 3. Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis

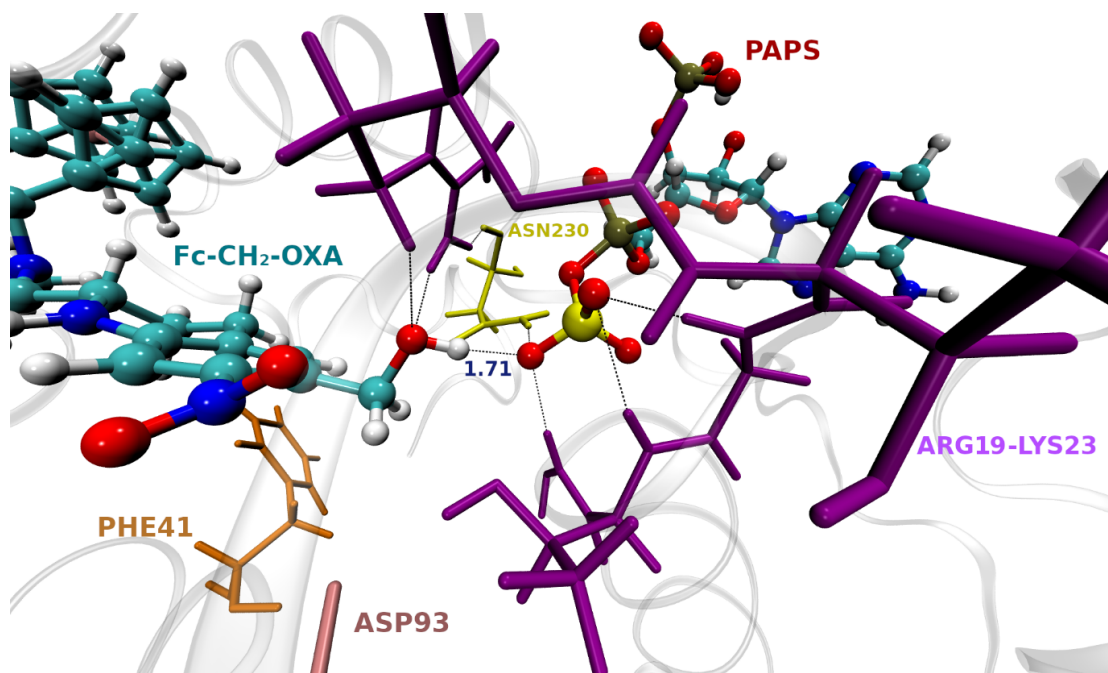


Figure 3.7: Snapshot of the simulation of S-Fc-CH₂-OXA in SmSULT. The contacts represented are stable along the simulation.

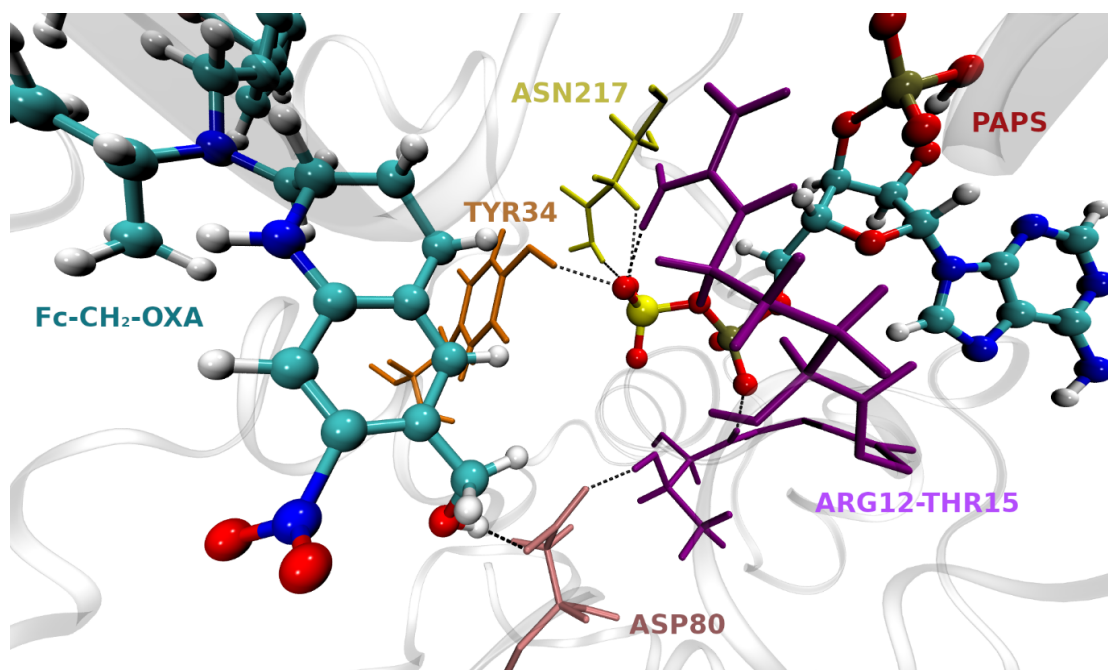


Figure 3.8: Snapshot of the simulation of S-Fc-CH₂-OXA in ShSULT. The contacts represented are stable along the simulation.

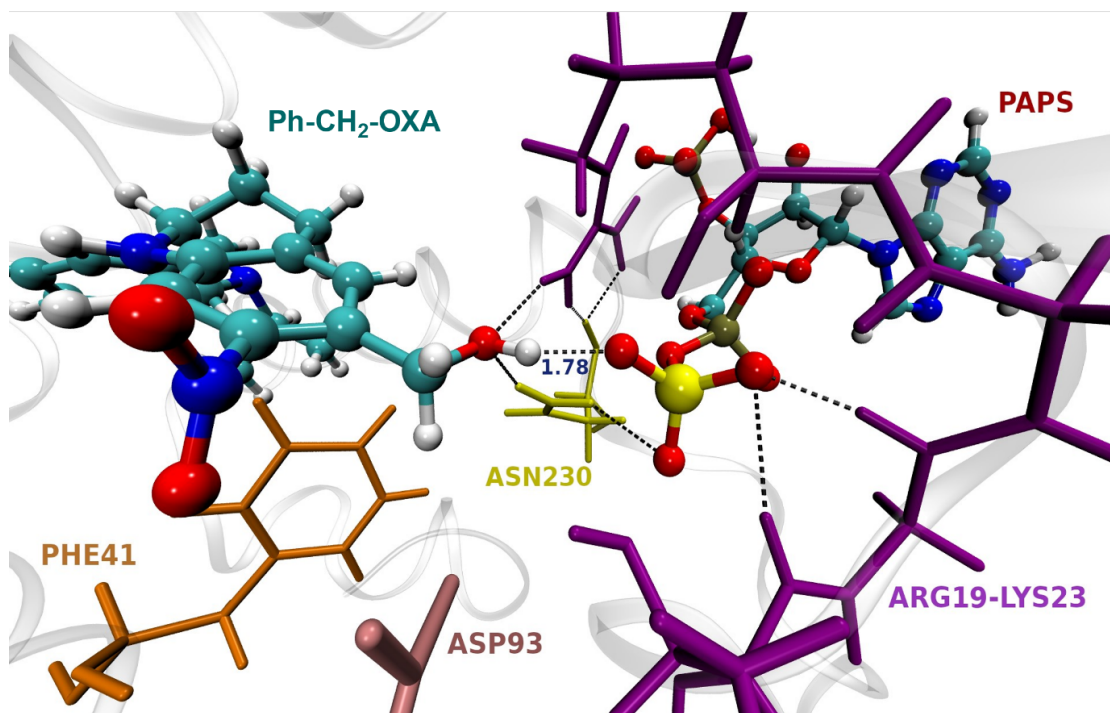


Figure 3.9: Snapshot of the simulation of S-Ph-CH₂-OXA in SmSULT. The contacts represented are stable along the simulation.

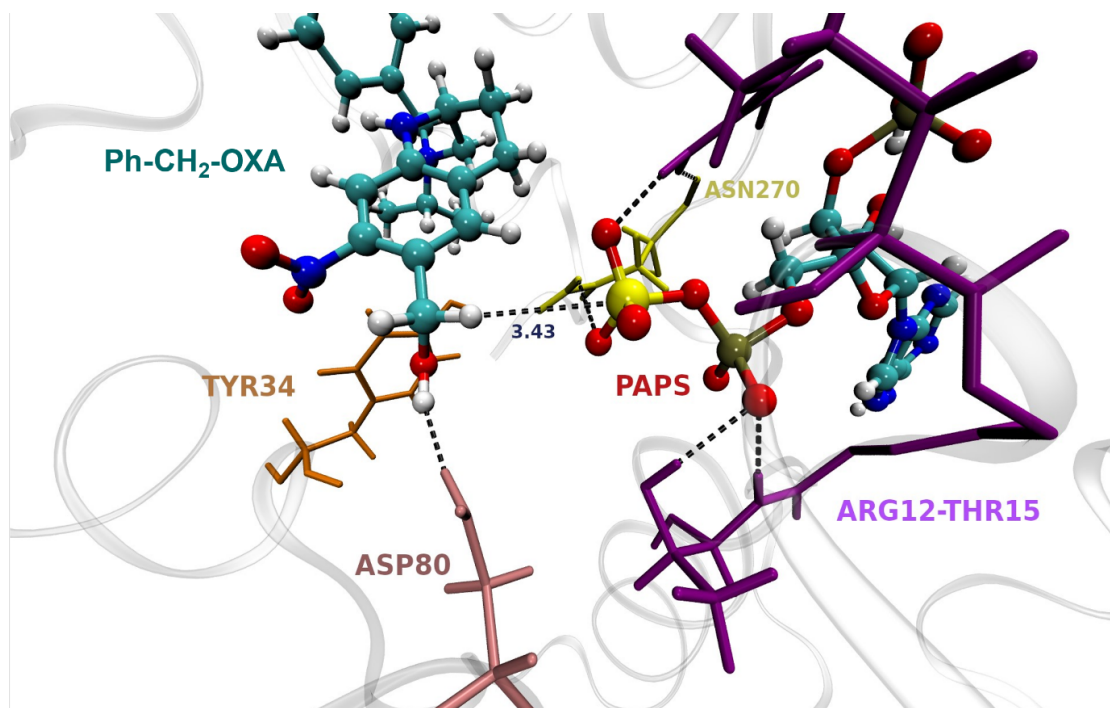


Figure 3.10: Snapshot of the simulation of S-Ph-CH₂-OXA in ShSULT. The contacts represented are stable along the simulation.

Chapter 3. Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis

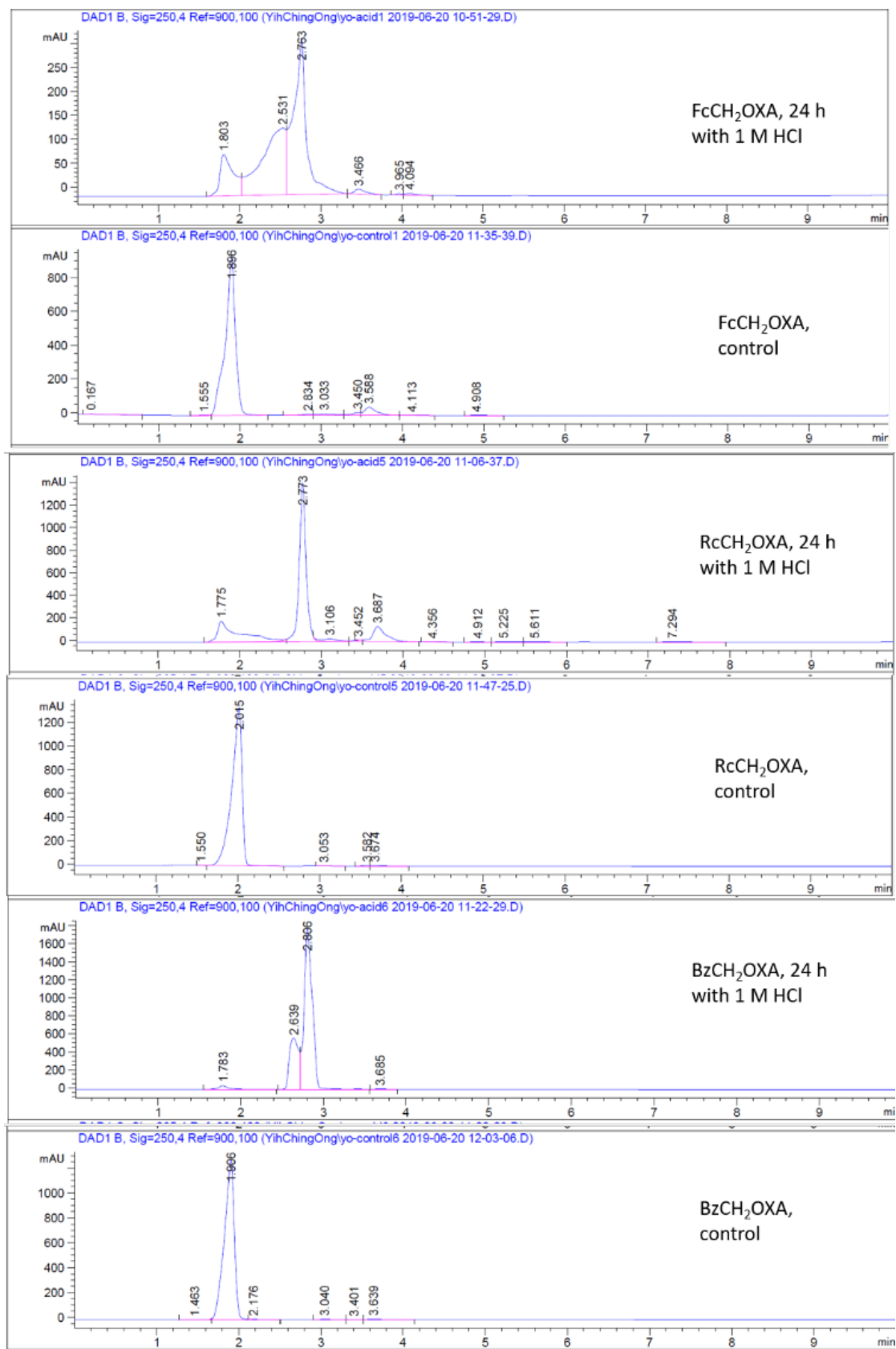


Figure 3.11: Comparison of compounds in the presence of 1 M HCl before and after 24 h incubation.

3.6 Supplementary Figures

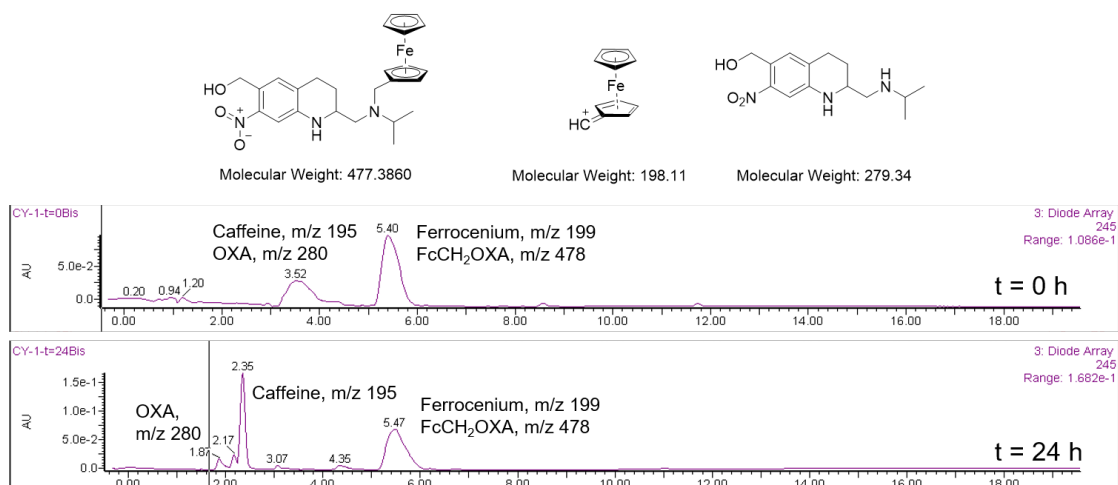


Figure 3.12: Comparison of LCMS trace for Fc-CH₂-OXA before and after incubation for 24 h.

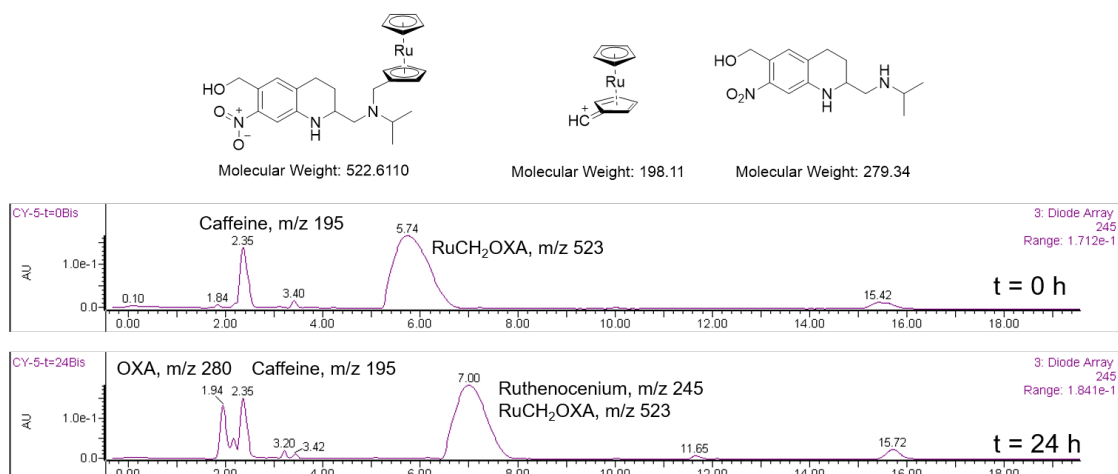


Figure 3.13: Comparison of LCMS trace for Rc-CH₂-OXA before and after incubation for 24 h.

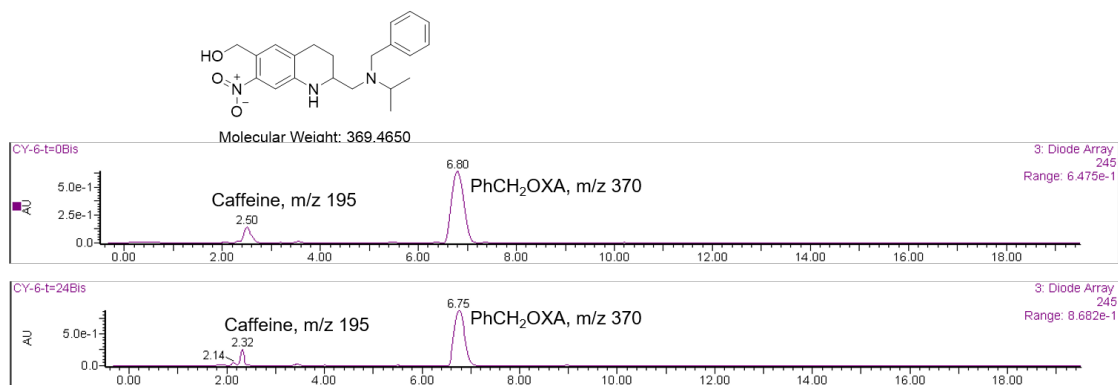


Figure 3.14: Comparison of LCMS trace for Ph-CH₂-OXA before and after incubation for 24 hours.

Chapter 3. Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis

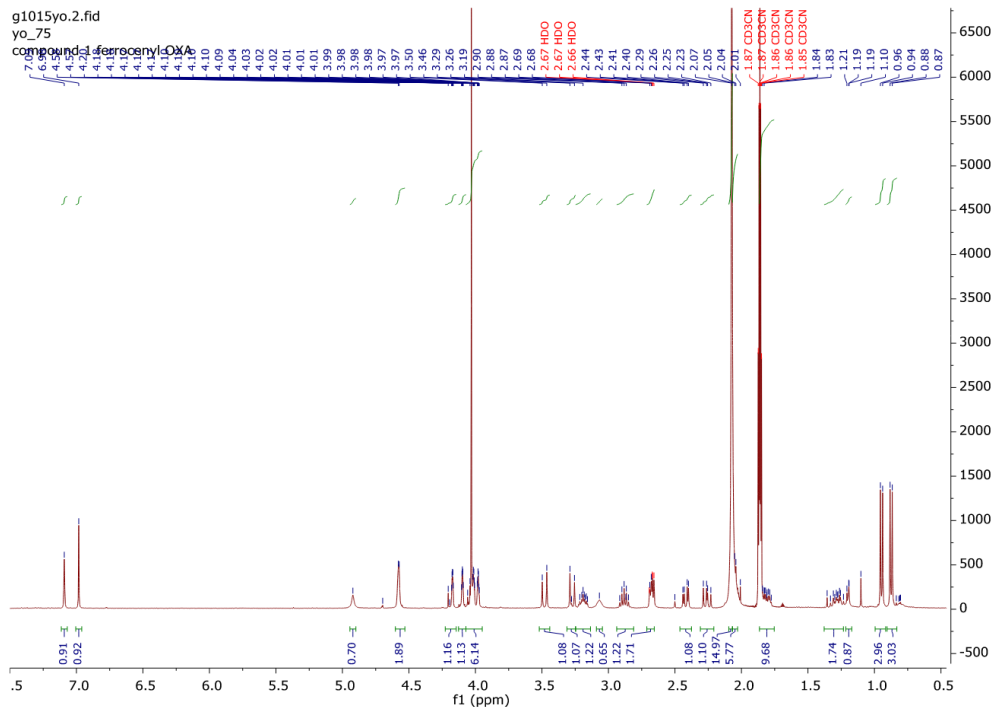


Figure 3.15: ¹H NMR spectrum for Fc-CH₂-OXA.

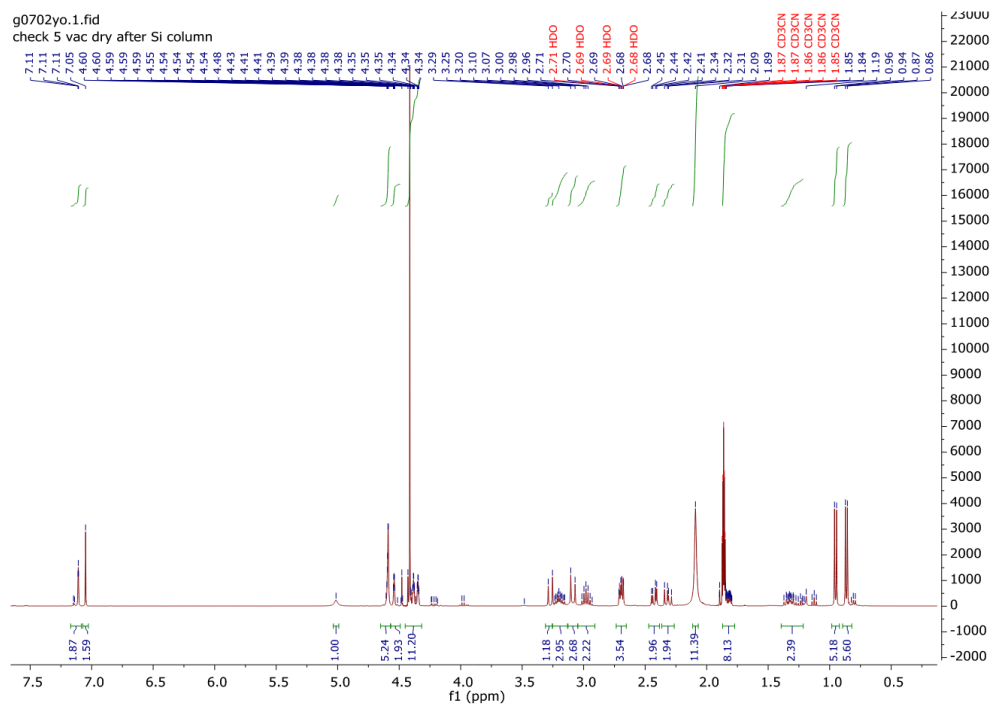


Figure 3.16: ¹H NMR spectrum for Rc-CH₂-OXA.

3.6 Supplementary Figures

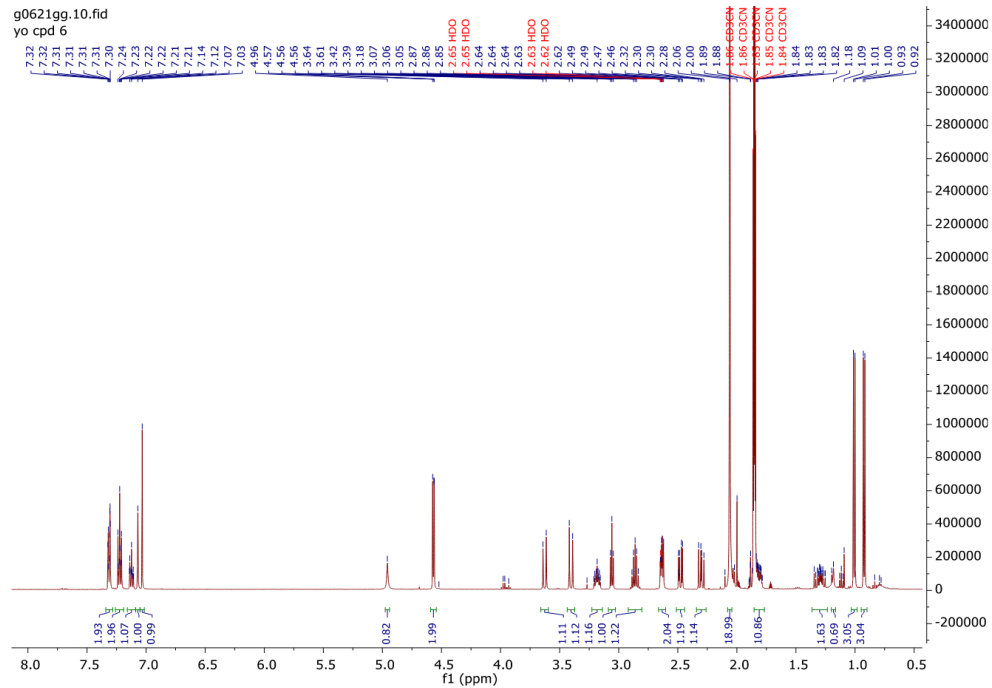


Figure 3.17: ¹H NMR spectrum for Ph-CH₂-OXA.

Table 3.10: Effect % ± SD of the different compounds against the species tested. 72 h.

Compound	Conc (µM)	S. mansoni			S. haematobium		S. japonicum
		Normal medium adults	Medium with 45 g/L albumin adults	juveniles	adults	Medium with 45 g/L albumin adults	adults
OXA	100			48.0±59.6			7.7 ± 10.0
	50			67.5±11.8			10.2 ± 17.3
	25			41.5±11.9			
	12.5			31.8±13.8			
Fc-CH ₂ -OXA	100	97.4 ± 4.4	64.4 ± 7.5	100 ± 0	71.9 ± 8.7	71.7 ± 17.2	100 ± 0
	50	87.0 ± 7.3	66.1 ± 10.8	100 ± 0	51.1 ± 12.3	46.3 ± 13.4	54.1 ± 12.3
	25	80.5 ± 6.7	61.0 ± 4.3	95.1 ± 4.9	15.6 ± 4.9	20.0 ± 0	16.9 ± 13.1
	12.5	69.9 ± 18.9	30.5 ± 34.6	79.7 ± 8.9	25.9 ± 20.0	12.0 ± 26.8	11.3 ± 4.3
	6.25	43.1 ± 8.3		64.5 ± 13.9	14.1 ± 19.2		
Rc-CH ₂ -OXA	100	94.6 ± 8.0	69.2 ± 12.3	100 ± 0	100 ± 0	80 ± 8.9	94.7 ± 2.6
	50	79.7 ± 15.0	64.0 ± 10.2	89.6 ± 12.3	66.2 ± 18.2	76.7 ± 10.3	64.1 ± 18.3
	25	86.1 ± 6.5	79.3 ± 10.5	91.5 ± 8.3	79.3 ± 4.6	62.5 ± 37.7	13.5 ± 0
	12.5	80.1 ± 14.0	78.4 ± 1.2	85.1 ± 15.5	36.3 ± 8.7	20.0 ± 24.5	11.7 ± 7.2
	6.25	69.2 ± 25.0		67.5 ± 19.4	5.2 ± 4.6		
Ph-CH ₂ -OXA	100	95.5 ± 6.3	52.3 ± 13.8	100 ± 0	90.3 ± 3.4	55 ± 24.39	97.1 ± 5.89
	50	53.6 ± 17.4	30.4 ± 18.8	31.8 ± 13.8	33.3 ± 0	13.3 ± 23.4	50.7 ± 19.2
	25	72.0 ± 13.0	21.4 ± 36.7	19.9 ± 3.6	26.7 ± 3.7	13.3 ± 10.3	17.8 ± 8.7
	12.5	50.8 ± 27.8	8.3 ± 29.6	20.4 ± 15.6		0 ± 8.2	0.5 ± 5.0
	6.25	30.9 ± 17.0		13.7 ± 15.3			

4 A multiple time step algorithm for trajectory surface hopping simulations

Chapter 4 is a post-print version of an article published as:

Pablo Baudin, **François Mouvet** and Ursula Rothlisberger. A multiple time step algorithm for trajectory surface hopping simulations. *Journal of Chemical Physics*, **2022**, 156, 034107

My contributions: Continued from work of Dr. Baudin. Finish the analysis, produce the figures, review the article for submission.

4.1 Abstract

A multiple time step (MTS) algorithm for trajectory surface hopping molecular dynamics has been developed, implemented, and tested. The MTS scheme is an extension of the *ab initio* implementation for Born–Oppenheimer molecular dynamics presented in [*J. Chem. Theory Comput.* **14**, 2834 (2018)]. In particular, the MTS algorithm has been modified to enable the simulation of non-adiabatic processes with the trajectory surface hopping (TSH) method and Tully’s fewest switches algorithm. The specificities of the implementation lie in the combination of Landau–Zener and Tully’s transition probabilities during the inner MTS time steps. The new MTS-TSH method is applied successfully to the photorelaxation of protonated formalimine, showing that the important characteristics of the process are recovered by the MTS algorithm. A computational speed-up between 1.5 and 3 has been obtained compared to standard TSH simulations which is close to the ideal values that could be obtained with the computational settings considered.

4.2 Introduction

Non-adiabatic phenomena such as photo-physical or photo-chemical processes are characterized by a failure of the Born–Oppenheimer (BO) approximation commonly invoked to describe molecular systems. The BO approximation, or the closely related adiabatic approximation, decouples the description of the nuclei and electrons. In non-adiabatic processes this coupling becomes important and the BO approximation breaks down.

Many different methods have been developed in order to describe non-adiabatic processes, ranging from fully quantum and formally exact models to mixed quantum/classical or semi-classical approaches (for a review see Refs.84, 85). Each method has its pros and cons, but in most cases it boils down to a compromise between computational efficiency and accuracy. In this work, we focus on one of the most popular methods, the trajectory surface hopping (TSH) approach.

In the TSH method (summarized in section 4.3.1), the evolution of the system is represented by a swarm of independent classical nuclear trajectories, which can hop from one electronic state to another in a stochastic way. The forces acting on the nuclei are calculated on-the-fly along each trajectory and transitions between electronic levels are considered simultaneously. In this way, the TSH method is thus able to describe non-adiabatic phenomena such as photo-chemical and photo-physical processes.

The TSH approach belongs to the mixed quantum/classical class of methods and is one of the computationally most expedient way to include non-adiabatic effects. Nonetheless, TSH simulations require the evaluation of the nuclear forces from first-principles simulations, *i.e.*, by solving the time-independent electronic Schrödinger equation, and those forces have to be evaluated for each nuclear geometry along the trajectory. Furthermore, due to its stochastic form (see section 4.3.1), the TSH approach requires to run a statistical ensemble of

trajectories, which means that, in practice, several hundred thousands of geometries have to be considered.[85, 86, 87] Solving the electronic Schrödinger equation is very computationally demanding and often have a very steep scaling with the number of electrons considered.[88, 89] These considerations limit considerably the applications of the TSH method. In particular, the three main limitations arise from: (i) the size of the systems that can be treated, (ii) the number of trajectories required to recover the proper statistical properties, and (iii) the duration of the physical processes that can be studied, *i.e.* the total length of the simulations. In recent years, several attempts to extend the application range of non-adiabatic MD have been proposed.[90, 91, 92] The present article is a further contribution to reduce the computational cost of non-adiabatic dynamics.

The strategy investigated in this work, is to reduce the computational requirements of the TSH method by relying on a multiple time step (MTS) algorithm. MTS techniques have first been introduced by Tuckerman *et al.* in the context of classical MD.[7] The MTS scheme relies on a decomposition of the atomic or nuclear forces into different components with different characteristic time scales. This decomposition enables to calculate the slow components of the forces less frequently than the fast one, while maintaining a fully time-reversible symplectic propagation. If the computational cost of the slow components is significant, large computational speed-ups can thus be obtained.

This article is organized as follows. After introducing an MTS algorithm for TSH simulations in section 4.3.2, the new method is applied to the photorelaxation of protonated formalimine (section 4.4). The MTS-TSH algorithm is compared to standard TSH simulations both in terms of accuracy and computational cost. Finally, some concluding remarks and perspectives are given in section 4.5.

4.3 Theory

In this first section, we briefly review the TSH formalism for non-adiabatic MD with particular emphasis on the version implemented in the CPMD plane-wave package.[93]

4.3.1 Trajectory surface hopping in CPMD

The TSH method can be seen as an attempt to introduce coupling between the electronic and nuclear degrees of freedom in Born–Oppenheimer molecular dynamics (BOMD).[94] Standard BOMD consists in a description of the nuclear coordinates based on classical mechanics,

$$M_\alpha \ddot{\mathbf{R}}_\alpha = \mathbf{F}_\alpha(\mathbf{R}), \quad (4.1)$$

where the index α denotes a given nucleus of mass M_α and classical coordinates \mathbf{R}_α and the two dots on top of the coordinates denote a second-order derivative with respect to time. When no index is specified, \mathbf{R} stands for all nuclear coordinates. The forces acting on the

Chapter 4. A multiple time step algorithm for trajectory surface hopping simulations

nuclei, $\mathbf{F}_\alpha(\mathbf{R})$, are usually calculated on-the-fly, from *ab initio* electronic structure calculations at fixed nuclear geometries,

$$\mathbf{F}_\alpha(\mathbf{R}) = -\nabla_\alpha H_{LL}, \quad (4.2)$$

where the diagonal matrix elements of the molecular electronic Hamiltonian, \hat{H} (under the BO approximation) are given by,

$$H_{LL} = \langle \psi_L(\mathbf{r}; \mathbf{R}) | \hat{H} | \psi_L(\mathbf{r}; \mathbf{R}) \rangle. \quad (4.3)$$

In Eq. (4.3) we have introduced the electronic wavefunction, ψ_L , for an arbitrary adiabatic state, L , which depends on the electronic coordinates \mathbf{r} . The parametric dependence of the electronic wavefunction on the nuclear geometry, \mathbf{R} , is also given. Generally, a single BOMD trajectory will thus propagate the classical nuclear coordinates on a single PES corresponding to a specific adiabatic electronic state.

However, when more than one electronic state is important to describe the dynamics of a system (for example in the description of photo-physical phenomena) it is important to go beyond the BO approximation and consider non-adiabatic algorithms.

Tully's fewest switches method

One of the most popular approaches used to describe such phenomena is the TSH method, in particular when combined with Tully's *fewest switches* algorithm.[95, 85, 96] In TSH a given trajectory can hop from one electronic state to another in a stochastic way, depending on the probability of the transition to occur. In order to get the transition probabilities one generally has to solve a time-dependent equation for the electrons of the system,

$$i\hbar \dot{C}_J(t) = C_J(t)\omega_J - i\hbar \sum_K^{N_{\text{states}}} C_K(t)\sigma_{JK}(\mathbf{R}) \quad (4.4)$$

Where the time-dependent coefficients $C_J(t)$ comes from an expansion of the time-dependent electronic wavefunction as a linear combination of time-independent adiabatic states. N_{states} is the total number of electronic states considered, ω_J is the excitation energy for state J and σ_{JK} is the non-adiabatic coupling (NAC) term,

$$\sigma_{JK} = \langle \psi_J | \partial_t \psi_K \rangle. \quad (4.5)$$

In CPMD, the NAC terms are computed by finite differences and using a CIS representation of the excited states.[97, 98, 99] Finally, in the fewest switches (FS) scheme, the probability of transition from adiabatic states J to K is evaluated as,

$$P_{J \rightarrow K}^{\text{FS}} = -\frac{2 \cdot \delta t}{|C_J|^2} \cdot \Re(C_K C_J^* \sigma_{JK}), \quad (4.6)$$

where δt is the classical time step used to integrate eq. (4.1), $\Re(z)$ denotes the real part of z and negative probabilities are set to zero. The decision to hop from state J to state K is then taken by generating a random number $r \in [0, 1]$ and evaluating the following condition,

$$\sum_{L=0}^{K-1} P_{J \rightarrow L}^{\text{FS}} < r < \sum_{L=0}^K P_{J \rightarrow L}^{\text{FS}}. \quad (4.7)$$

Landau–Zener transition probabilities

As a simpler alternative to the solution of eq. (4.4) for the calculations of the transition probabilities in eq. (4.6), it is possible to obtain approximate transition probabilities from Landau–Zener–Stückelberg (LZ) theory for non-adiabatic transitions.[100, 101] In the CPMD package, such probabilities are computed directly from the knowledge of the energy of the adiabatic electronic states as,

$$P_{J \rightarrow K}^{\text{LZ}} = \exp\left(-\frac{\pi^2}{h} \cdot \frac{|\Delta E_{JK}^{\text{adia}}|^2}{\max(d|\Delta E_{JK}^{\text{adia}}|/dt)}\right) \quad (4.8)$$

where $\Delta E_{JK}^{\text{adia}}$ is the gap between adiabatic states J and K directly obtained as a byproduct of DFT and TDDFT calculations.[100] In a TSH simulation relying on LZ theory, a hop from an electronic state to another is considered based on the following condition,

$$P_{J \rightarrow K}^{\text{LZ}} > r, \quad (4.9)$$

where r is again a random number chosen between zero and one.

For more details about the implementation of TSH in the CPMD package, see Refs.97, 102, 93.

4.3.2 Trajectory surface hopping with multiple time step scheme

MTS techniques have been introduced as a way to reduce the computational cost of molecular dynamics for systems in which the forces in action can be decomposed into different time scales. The success of MTS techniques is largely due to the development of the reversible reference system propagation algorithm (rRESPA) by Tuckerman *et al.* in Ref.7.

Recently, the rRESPA was implemented in the CPMD package for BOMD.[103] The important details of this implementation are summarized in section 4.3.3, while in section 4.3.4, we suggest an extension of the MTS algorithm to enable non-adiabatic dynamics in the context of TSH-MD.

4.3.3 Standard MTS algorithms

In Ref.7, Tuckerman *et al.* introduced the Trotter factorization of the Liouville operator as a convenient way to generate reversible MD integrators. This technique is summarized here. Let us first consider a phase space element $\Gamma(t=0)$ which describes the initial positions (x_j) and momenta (p_j) of all the nuclei of a system. The phase space element can be propagated in time using a classical propagator, $G(t)$,

$$\Gamma(t) = G(t)\Gamma(0) = e^{iL_t}\Gamma(0), \quad (4.10)$$

where L is the Liouville operator given by,

$$iL = \sum_j [\dot{x}_j \partial_{x_j} + F_j \partial_{p_j}] \quad (4.11)$$

and F_j is a single component of the nuclear forces (the index j is a collective index for Cartesian coordinates and nuclei). By assuming a time scale separation of the forces into fast (\mathbf{F}^{fast}), and slow (\mathbf{F}^{slow}) components, it is possible to rewrite the Liouville operator as,

$$iL = iL_x + iL_p^{\text{fast}} + iL_p^{\text{slow}} \quad (4.12)$$

$$iL_x = \sum_j \dot{x}_j \partial_{x_j} \quad (4.13)$$

$$iL_p^{\text{fast}} = \sum_j F_j^{\text{fast}} \partial_{p_j} \quad (4.14)$$

$$iL_p^{\text{slow}} = \sum_j F_j^{\text{slow}} \partial_{p_j} \quad (4.15)$$

Applying a Trotter factorization on the classical propagator and discarding terms of third order and higher in t , we can define a discrete time propagator, which can be translated into an MTS algorithm,[104, 7]

$$G^{\text{MTS}}(\Delta t) = e^{iL_p^{\text{slow}}(\Delta t/2)} \left[G^{\text{fast}}(\Delta t/N) \right]^N e^{iL_p^{\text{slow}}(\Delta t/2)}, \quad (4.16)$$

with

$$G^{\text{fast}}(\Delta t/N) = G^{\text{fast}}(\delta t) = e^{iL_p^{\text{fast}}(\delta t/2)} e^{iL_x \delta t} e^{iL_p^{\text{fast}}(\delta t/2)}. \quad (4.17)$$

In Eqs. (4.16) and (4.17) we have introduced two finite time steps; Δt , which reflects the time scale of the slow forces (\mathbf{F}^{slow}), and $\delta t = \Delta t/N$, which is adapted to the fast forces (\mathbf{F}^{fast}). If the slow forces are computationally expensive, this integrator can lead to significant computational savings without inducing any more approximation to the dynamics.

Algorithm 4.1 represents a pseudo-code that can be obtained by applying each term of the propagator in eq. (4.16) (one by one from the right to the left) onto an initial phase space

element $\Gamma(0) \equiv \{\mathbf{x}, \mathbf{p}\}$.

- 1: Initialize positions, velocities, fast and slow forces: $\mathbf{x}, \mathbf{v}, \mathbf{F}^{\text{fast}}, \mathbf{F}^{\text{slow}}$
 - 2: **for** $i = 1, \text{maxiter}$ **do** (MD loop for the slow component)
 - 3: *Slow* velocity update: $\mathbf{v} \leftarrow \mathbf{v} + \frac{\Delta t}{2m} \cdot \mathbf{F}^{\text{slow}}$
 - 4: **for** $j = 1, N$ **do** (MD loop for the fast component)
 - 5: *Fast* velocity update: $\mathbf{v} \leftarrow \mathbf{v} + \frac{\delta t}{2m} \cdot \mathbf{F}^{\text{fast}}$
 - 6: Position update: $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{v} \cdot \delta t$
 - 7: Get fast components of the forces: \mathbf{F}^{fast}
 - 8: Final *fast* velocity update: $\mathbf{v} \leftarrow \mathbf{v} + \frac{\delta t}{2m} \cdot \mathbf{F}^{\text{fast}}$
 - 9: **end for** (MD loop for the fast component)
 - 10: Get slow components of the forces: \mathbf{F}^{slow}
 - 11: Final *slow* velocity update: $\mathbf{v} \leftarrow \mathbf{v} + \frac{\Delta t}{2m} \cdot \mathbf{F}^{\text{slow}}$
 - 12: **end for** (MD loop for the slow component)
-

ALG. 4.1: Standard rRESPA MTS algorithm obtained as a direct translation of the discrete propagator in eq. (4.16). For comparison with the velocity Verlet algorithm, the steps concerned with momenta, \mathbf{p} , have been re-written in terms of velocities, \mathbf{v} .

In the context of first-principles BOMD, the separation of the forces into fast and slow components is not straightforward. In the CPMD package we have chosen to use different levels of electronic structure theory to decompose the forces. Typically, a “low” level density functional (*e.g.* GGA) is used to calculate the fast components of the nuclear forces, while the slow components are obtained as the difference between the forces obtained with a “higher” level functional (*e.g.* hybrid) and the “low” level forces,

$$\mathbf{F}^{\text{fast}} = \mathbf{F}^{\text{low}} \quad (4.18)$$

$$\mathbf{F}^{\text{slow}} = \mathbf{F}^{\text{high}} - \mathbf{F}^{\text{low}}. \quad (4.19)$$

The physical motivation for this separation comes from the fact that the chosen electronic structure levels differ only in their treatment of correlation (or exchange and correlation in the case of DFT). Since those contributions correspond to relatively weak interactions in terms of energy (with no explicit dependence on nuclear positions) they can be expected to represent relatively weak and slowly varying force contributions. Note that independent of the chosen low-level method, this MTS algorithm generates by construction trajectories at the level of the high-level method. The choice of the low-level method only determines the magnitude (and temporal oscillations) of the correction forces and thus the maximal possible time step ratio that can be applied to generate stable dynamics within a given energy conservation.

The implementation of the MTS algorithm in CPMD makes use of this separation as well as a slightly different but completely equivalent layout of the code.[103] This implementation presented in pseudo-code in algorithm 4.2, makes the MTS algorithm look like a velocity Verlet algorithm with effective forces, \mathbf{F}^{eff} , that are time step dependent.

Chapter 4. A multiple time step algorithm for trajectory surface hopping simulations

Even though, the force decomposition in terms of high and low electronic structure levels is done *ad hoc*, this kind of separation has already proven useful[105, 103] and the benefits in terms of computational cost are evident.

```
1: Initialize positions, velocities, high and low level forces:  $\mathbf{x}, \mathbf{v}, \mathbf{F}^{\text{high}}, \mathbf{F}^{\text{low}}$ 
2: Get effective force:  $\mathbf{F}^{\text{eff}} \leftarrow \mathbf{F}^{\text{low}} + (\mathbf{F}^{\text{high}} - \mathbf{F}^{\text{low}}) \cdot N$ 
3: for  $i = 1$ , maxiter do (MD loop)
4:   Velocity update:  $\mathbf{v} \leftarrow \mathbf{v} + \frac{\delta t}{2m} \cdot \mathbf{F}^{\text{eff}}$ 
5:   Position update:  $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{v} \cdot \delta t$ 
6:   if  $i \equiv 0 \pmod{N}$  then (outer step)
7:     Get both high and low level forces,  $\mathbf{F}^{\text{high}}, \mathbf{F}^{\text{low}}$ 
8:     Get effective force:  $\mathbf{F}^{\text{eff}} \leftarrow \mathbf{F}^{\text{low}} + (\mathbf{F}^{\text{high}} - \mathbf{F}^{\text{low}}) \cdot N$ 
9:   else (inner step)
10:    Effective forces are set to the low level forces:  $\mathbf{F}^{\text{eff}} \leftarrow \mathbf{F}^{\text{low}}$ 
11:   end if (outer/inner steps)
12:   Final velocity update:  $\mathbf{v} \leftarrow \mathbf{v} + \frac{\delta t}{2m} \cdot \mathbf{F}^{\text{eff}}$ 
13: end for (MD loop)
```

ALG. 4.2: MTS-BOMD algorithm as implemented in the CPMD package. This algorithm can be obtained straightforwardly from algorithm 4.1 by using the partitioning of the forces in Eqs. (4.18) and (4.19) and reshuffling a few steps.

4.3.4 MTS algorithm for trajectory surface hopping dynamics

When using the MTS implementation described in algorithm 4.2 in combination with a TSH algorithm, one has to decide how to hop from one electronic state to another. In order to get trajectories of high accuracy, it would be beneficial to consider electronic transitions based on Tully's FS criterion in eq. (4.7) calculated with the MTS high level functional. In the following, this type of calculation (with a velocity Verlet algorithm) will actually be used as a reference. However, when using an MTS algorithm, if the outer time step Δt becomes large, some parts of the PES where the transition probabilities are high might be treated only by the low level functional and the transitions would be missed at the high level.

In this work, we propose another strategy that can potentially solve that problem. This strategy consists in evaluating the transitions probabilities with the LZ formula in eq. (4.8) during the low level steps, while for high level steps, the electronic transition are evaluated according to the FS criterion in eq. (4.7) based on the high level quantities.

This is not yet completely satisfactory since it does not guarantee that a transition detected with the LZ probabilities during a low level step would have also been detected by the FS criterion calculated with the high level functional. To further improve on that issue, we suggest that if a transition is detected at the low level (using LZ theory), a high level calculation is triggered to confirm the transition using Tully's FS criterion. This strategy is described in

algorithm 4.3 and tested in the remaining sections.

Finally, it is important to note that in the case of high level calculations triggered by the low level LZ criterion, the random number used in the high level FS criterion in eq. (4.7), should be the same as in the low level LZ criterion in eq. (4.9). We also underline that the discrete time step δt in eq. (4.6) always correspond to the inner time step in the MTS scheme. This can be rationalized by realizing that at each inner time step, an electronic transition can occur if it is detected at the low level and confirmed at the high level, such that when eq. (4.6) is invoked, it is only to check for a transition in the last δt time window.

```

1: Initialize positions, velocities, and running electronic state:  $\mathbf{x}, \mathbf{v}, J$ 
2: Initialize high and low level forces:  $\mathbf{F}^{\text{high}}, \mathbf{F}^{\text{low}}$ 
3: Get effective force:  $\mathbf{F}^{\text{eff}} \leftarrow \mathbf{F}^{\text{low}} + (\mathbf{F}^{\text{high}} - \mathbf{F}^{\text{low}}) \cdot N$ 
4: for  $i = 1$ , maxiter do (MD loop)
5:     Velocity update:  $\mathbf{v} \leftarrow \mathbf{v} + \frac{\delta t}{2m} \cdot \mathbf{F}^{\text{eff}}$ 
6:     Position update:  $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{v} \cdot \delta t$ 
7:     if  $i \equiv 0 \pmod{N}$  then (outer step)
8:         Get both high and low level forces,  $\mathbf{F}^{\text{high}}, \mathbf{F}^{\text{low}}$ 
9:         Get effective force:  $\mathbf{F}^{\text{eff}} \leftarrow \mathbf{F}^{\text{low}} + (\mathbf{F}^{\text{high}} - \mathbf{F}^{\text{low}}) \cdot N$ 
10:        Get high level FS transition probabilities:  $P_{J \rightarrow K}^{\text{FS}}$ 
11:        if FS criterion in eq. (4.7) is met for any state  $K \neq J$  then
12:            Update index of adiabatic electronic state:  $J \leftarrow K$ 
13:        end if
14:    else (inner step)
15:        Effective forces are set to the low level forces:  $\mathbf{F}^{\text{eff}} \leftarrow \mathbf{F}^{\text{low}}$ 
16:        Get low level LZ transition probabilities:  $P_{J \rightarrow K}^{\text{LZ}}$ 
17:        if LZ criterion in eq. (4.9) is met for any state  $K \neq J$  then (check at high level)
18:            Effective forces are now set to the high level forces:  $\mathbf{F}^{\text{eff}} \leftarrow \mathbf{F}^{\text{high}}$ 
19:            Get high level FS transition probabilities:  $P_{J \rightarrow K}^{\text{FS}}$ 
20:            if FS criterion in eq. (4.7) is met for any state  $K \neq J$  then
21:                Update index of adiabatic electronic state:  $J \leftarrow K$ 
22:            end if
23:        end if
24:    end if (outer/inner steps)
25:    Final velocity update:  $\mathbf{v} \leftarrow \mathbf{v} + \frac{\delta t}{2m} \cdot \mathbf{F}^{\text{eff}}$ 
26: end for (MD loop)

```

ALG. 4.3: MTS-TSH algorithm as implemented in the CPMD package. This algorithm corresponds to a modified version of algorithm 4.2 that accounts for non-adiabatic transitions. See section 4.3.2 for details.

4.4 Results and Discussion

In this section, the protonated formalimine (CH_2NH_2^+ , denoted as system I) is used as a simple yet interesting example to investigate the capabilities of the new MTS-TSH method presented in section 4.3.2. In particular, we will investigate the possibility to use the MTS-TSH algorithm as a more efficient alternative to the standard FS-TSH algorithm. The physical process under investigation in this section is the photorelaxation of system I. This process is a typical example of photo-dynamics and is thus very handy to test new models for non-adiabatic molecular dynamics.[106, 97, 107, 108, 92]

All the calculations presented in this section have been performed with a local version of the CPMD plane-wave package.[93] For the reference TSH simulations, the nuclear forces and NACs are computed with the PBE0 hybrid functional.[29] The same functional is thus used for the high level forces in the MTS calculations, while the low level forces are obtained from the PBE functional.[26, 27] For a fair comparison, all remaining parameters are kept the same in the reference and the MTS simulations.

The five lowest singlet excited states obtained with the Tamm-Dancoff approximation of TDDFT are considered for all calculations. Norm-conserving Trouiller-Martins pseudopotentials are used with a plane-wave cutoff of 70 Ry and an isolated cubic box with an edge of 10 Å. Unless specified otherwise, the inner time step is set to $\delta t = 10$ a.u. See the supplementary materials for the raw data, the analysis scripts, and a full description of the computational details.

4.4.1 Investigating different approximations for the non-adiabatic couplings

First of all, it is important to rationalize the use of the LZ transition probabilities at the low level in the MTS-TSH method presented in section 4.3.2. For that purpose, we have run a reference trajectory for about 100 femtoseconds (fs) on the second excited state of system I. This trajectory was performed with the PBE0 functional and a standard velocity Verlet algorithm. The transition probabilities were calculated at each time step using the FS method (see section 4.3.1) but no electronic transitions were allowed such that the system stayed on the second excited states PES the whole time.

The exact same trajectory (velocities and coordinates) have then been repeated by calculating the transition probabilities using the LZ method (see section 4.3.1) together with the PBE functional (low level). For comparison, an additional run was performed at the PBE level and calculating transition probabilities using the FS method. No MTS algorithm was used in this section and the only difference between the three trajectories is the model used to calculate the transition probabilities (velocities and coordinates are the same for all three trajectories). Such conditions allows to compare the transition probabilities obtained with three different levels: PBE0-FS, PBE-LZ, and PBE-FS.

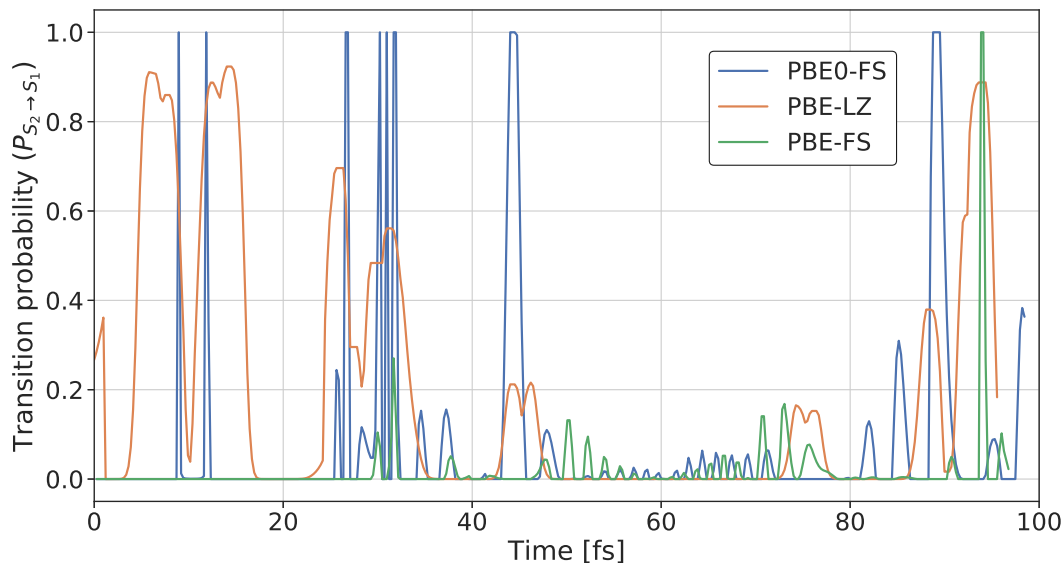


Figure 4.1: Evolution of the transition probabilities (from S_2 to S_1) along a single trajectory started in the second excited state of system I. Three different levels are compared; PBE0-FS, corresponding to eq. (4.6) together with the PBE0 functional, PBE-LZ, corresponding to eq. (4.8) together with the PBE functional, and PBE-FS, corresponding to eq. (4.6) together with the PBE functional. (Values of $P_{S_2 \rightarrow S_1}^{\text{FS}}$ larger than 1, have been set back to unity.)

In Fig. 4.1 we have represented the evolution of the transition probabilities ($P_{S_2 \rightarrow S_1}$) for the three different levels. From Fig. 4.1, it is clear that neither the PBE-LZ probabilities nor the PBE-FS probabilities represent a completely reliable approximation to the reference PBE0-FS transition probabilities. However, a relatively good correlation is observed. In particular, whenever the reference PBE0-FS transition probabilities are large, the PBE-LZ probabilities are also relatively large. For some reason this is less obvious for the PBE-FS curve.

The algorithm developed in section 4.3.2 relies on the low level calculations to detect potential electronic transitions. Whenever a low level transition is detected, it is double checked with a high level FS calculation such that the low level probabilities do not have to be quantitatively accurate. The transition probabilities in Fig. 4.1 indicates that using LZ theory at the low level for the calculation of the transition probabilities is enough. In other words, the LZ probabilities can be used as a proxy during the low level steps of the MTS-TSH method.

We note that the objective of the investigation performed in this section is to support the design of the algorithm presented in section 4.3.2 and that further tests could be performed to draw more general conclusions. Nonetheless, since the usage of LZ probabilities at the low level is only used to trigger high level calculations, we believe that a strong empirical support is not required at this stage.

4.4.2 Comparing single trajectories via deterministic surface hopping

As we have seen in section 4.3.1, non-adiabatic dynamics performed with a TSH algorithm are stochastic by nature due to the randomness used in the hopping procedure. This stochastic behaviour makes it difficult to compare individual trajectories obtained with a TSH algorithm. To properly compare different non-adiabatic models, one needs to look at a statistical ensemble of trajectories. Before we present such results in section 4.4.3, we first consider in this section a deterministic version of TSH in which the random number r used in Eqs. (4.7) and (4.9) has been fixed arbitrarily to $r = 0.3$. We note that, the only MTS algorithm tested here and in the next section is the one presented in section 4.3.2, i.e., low level (PBE) LZ transition probabilities are used during the inner steps to trigger a high level (PBE0) calculation which confirms or not the electronic transition.

Quality assessment

Six different runs have been produced, all starting in the second excited state and with the same nuclear configuration. For simplicity, the nuclear velocities are initialized to zero. The first run is a reference TSH trajectory at the PBE0 level, while the 5 remaining trajectories are obtained with the MTS-TSH algorithm, and an MTS time step factor of $N = \{2, 3, 4, 6, 8\}$.

In Fig. 4.2, the potential energy of the three lowest singlet states obtained from the reference PBE0 trajectory is represented. The trajectory starts in the second excited state and hops to the first excited state in less than 10 fs. The system stays in the first excited state for the next 70 fs until it intersects with and hops into the ground state. All trajectories are stopped whenever they collapse into the ground state.

In Fig. 4.3, the potential energies of all runs (reference and MTS) are represented. For the MTS runs, we only plot the energy from the PBE0 steps, which explains why for the larger MTS factors, the curves appear less smooth. One can see that all the MTS trajectories, except with MTS factor 8, successfully describe the first transition from S_2 to S_1 in the first 10 fs. Between 30 and 40 fs, the reference trajectory enters a new non-adiabatic region as the first and second excited states become close in energy for the third time. Until that point, the MTS trajectories with MTS factor 2, 3, 4, and 6 seem to describe the reference trajectory relatively well. Afterwards, the MTS trajectories with factor 2, 3, and 4 start to diverge from the reference run. The MTS trajectory with factor 6 is overall the closest from the reference run. However, an MTS factor of 8 seems to be too large to reproduce most of the features of the reference run. This analysis is confirmed by looking at the root-means-square-deviation (RMSD) of the nuclear positions of the MTS trajectories with respect to the reference run in Fig. 4.4.

The results presented in Figs. 4.3 and 4.4, could be interpreted as rather discouraging. Indeed, with only an MTS factor of 2, the MTS trajectory and the reference one start to diverge only after 35 fs. However, unlike with ground state MD, the presence of non-adiabatic events has a drastic impact on the chaotic behaviour of excited state dynamics. Very tiny differences in the

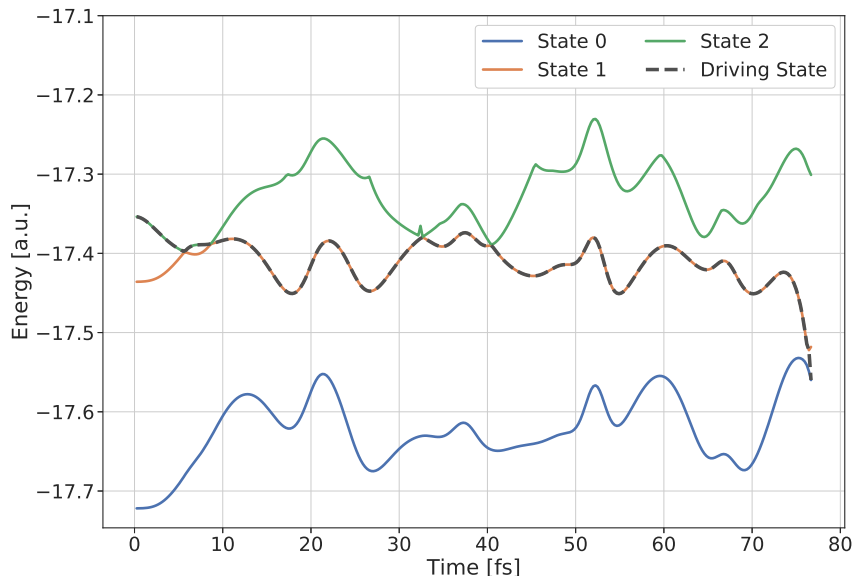


Figure 4.2: Potential energy surfaces of the three lowest singlet states of system I obtained from a 75 fs long PBE0 FS-TSH run started in the second excited state.

nuclear positions and velocities or wavefunction parameters can lead to completely different trajectories. The fact that some of the MTS trajectories presented in Fig. 4.3 differ significantly from the reference trajectory does not mean that those trajectories are not physical. Only a statistical analysis of the photorelaxation process can enlighten us on that matter.

The deterministic investigation of the MTS-TSH implementation presented here indicates that it is possible to recover the main characteristics of a reference calculation using the MTS-TSH algorithm presented in section 4.3.2, for example in the case of MTS factor 6. In section 4.4.3, we will investigate the possibility to recover statistically relevant quantities from the MTS-TSH algorithm, while in the next section we analyze the speed-ups obtained in the deterministic MTS-TSH simulations.

Efficiency assessment

Let us call $t^{\text{FS}} = t^{\text{high}}$ the average CPU time per step spent in a standard FS-TSH simulation with a “high” level functional. This time takes into account the SCF optimization, the solution of the TDDFT equations as well as the calculation of the FS probabilities. In the following we use t^{FS} as a reference CPU time. In order to provide fair comparisons, we also need to consider the average CPU time per step from a LZ-TSH simulation with a “low” level functional $t^{\text{LZ}} = t^{\text{low}}$.

The expected or ideal averaged CPU time per step in the MTS-TSH algorithm can then be

Chapter 4. A multiple time step algorithm for trajectory surface hopping simulations

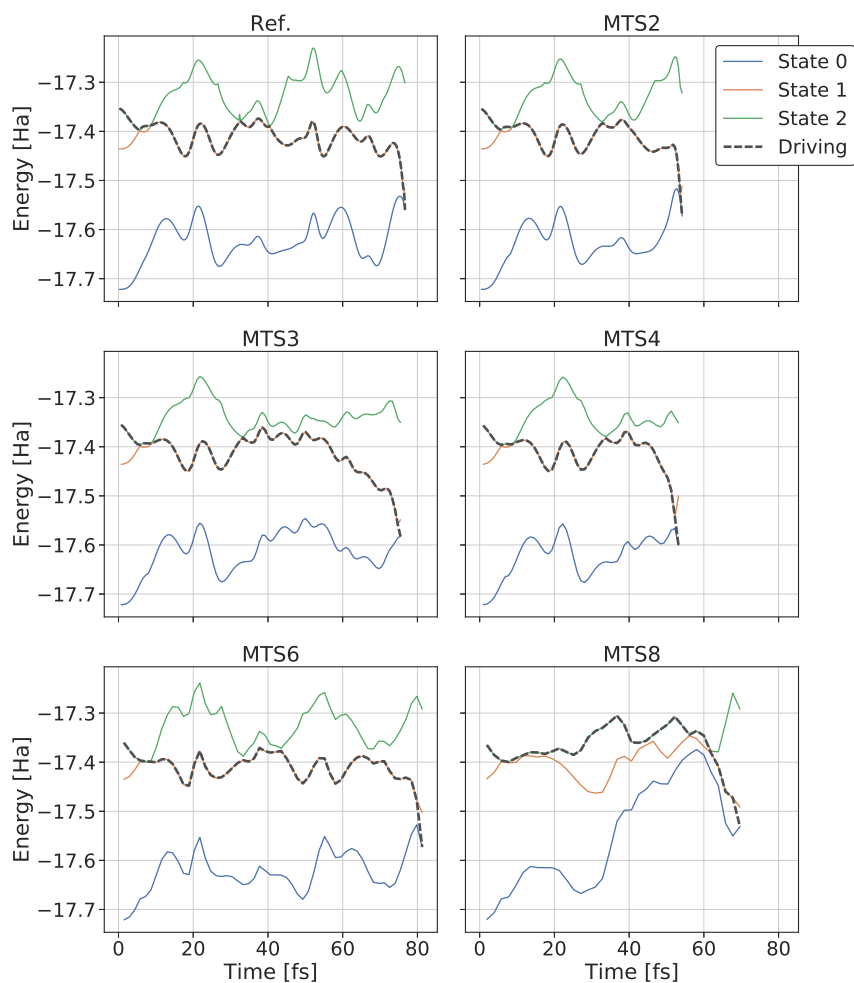


Figure 4.3: Potential energy surfaces of the three lowest singlet states of system I obtained from 6 different calculations. The upper-left panel constitutes the reference PBE0 FS-TSH run, while the other panels represent the MTS-TSH runs with different MTS factors ($N = \{2, 3, 4, 6, 8\}$). All simulations have been started from the same geometries and zero velocities. The driving state is represented with the dashed thick black line.

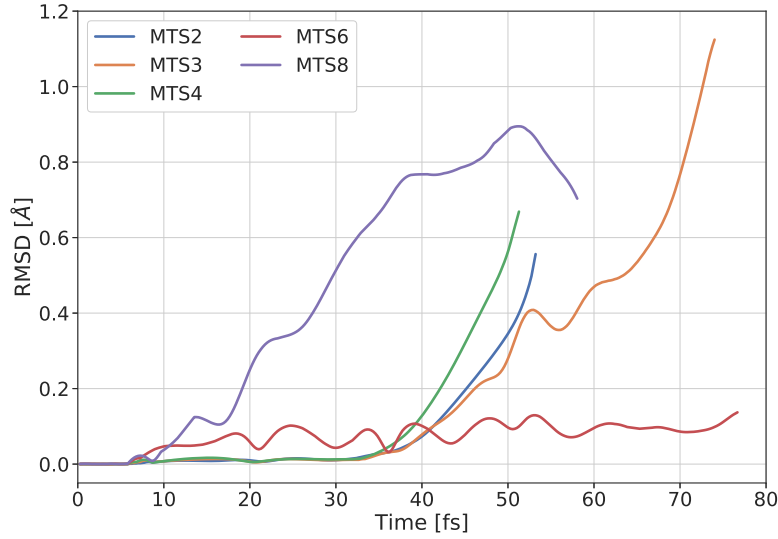


Figure 4.4: Root-means-square-deviations (RMSD) of nuclear positions of the MTS trajectories with different MTS factors ($N = \{2, 3, 4, 6, 8\}$) with respect to a reference PBE0 FS-TSH trajectory.

calculated as,

$$t^{\text{MTS}(N)} = t^{\text{low}} + \frac{1}{N} \cdot t^{\text{high}} + X \cdot t^{\text{high}}, \quad (4.20)$$

where N is the MTS factor and, t^{high} and t^{low} are CPU timings coming from standard (non MTS) simulations. X denotes the average frequency of triggered high level steps which is a quantity difficult to predict. To obtain the ideal MTS-TSH speed-up we set $X = 0$, and get,

$$S^{\text{ideal}} = \frac{t^{\text{FS}}}{t^{\text{MTS}(N)}} = \frac{t^{\text{high}}}{t^{\text{low}} + \frac{1}{N} \cdot t^{\text{high}}} = \frac{N}{N \cdot \frac{t^{\text{low}}}{t^{\text{high}}} + 1} \quad (4.21)$$

In the limit of a negligible cost of the low level steps (compared to the high level ones) we get,

$$S^{\text{limit}} = \lim_{t^{\text{high}} \gg t^{\text{low}}} S^{\text{ideal}} = N. \quad (4.22)$$

In practice, several things can impact the ideal and limit speed-ups. The most obvious one being the number of triggered high level steps. It is easy to realize that for large values of N , the number of triggered high level steps will tend towards a system dependent number, in most cases larger than zero. Such that one cannot achieve arbitrarily large speed-ups simply by increasing N . From a more practical point of view, the physics of the system (vibrational frequencies) will be the first parameter to consider as a limitation for the value of N and thus for the speed-up (too large values of N would lead to artifacts such as resonances[109]). A less straightforward impact on the effective (or real) speed-up is given by the overhead due to the convergence of the high level (TD)DFT parameters. Indeed, since with increasing values

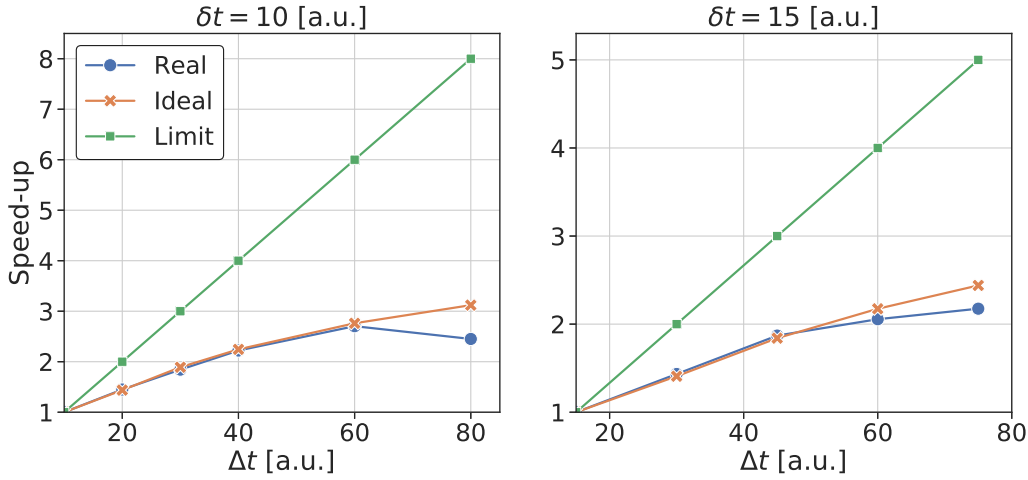


Figure 4.5: Real MTS speed-up obtained from deterministic MTS-TSH calculations with different inner and outer time steps (δt and Δt , respectively) compared to reference PBE0 FS-TSH timings. For comparison the ideal and limit speed-ups as defined in Eqs. (4.21) and (4.22), respectively, are also reported.

of N , larger nuclear displacements occur between high level steps, the initial guess for the electronic structure calculation becomes less appropriate, often resulting in more (*e.g.* SCF) iterations and thus higher computational requirements.

In Fig. 4.5, we have represented the ideal and limit MTS speed-ups as calculated from Eqs. (4.21) and (4.22), respectively, as well as the real MTS speed-up obtained for the calculations presented in section 4.4.2 (system I with the PBE0 and PBE functionals). The left panel of Fig. 4.5 corresponds to the MTS calculations with $\delta t = 10$ a.u. and MTS factors $N = \{2, 3, 4, 6, 8\}$, while, in the right panel a new set of calculations with $\delta t = 15$ a.u. and MTS factors $N = \{2, 3, 4, 5\}$ (all other parameters unchanged) are represented. The first points of both plots (with speed-up one) correspond to the reference FS-TSH calculations. The real MTS speed-ups are simply obtained from the ratio between $t^{\text{FS-PBE0}}$ and the averaged time per step in the actual MTS simulations,

$$S^{\text{real}} = \frac{t^{\text{MTS-total}}}{N^{\text{steps}}}. \quad (4.23)$$

From Fig. 4.5, we can see that both the real and ideal speed-ups are quite far from the limit speed-ups (up to a factor 3 of difference). This is simply a consequence of the fact that the condition, $t^{\text{high}} \gg t^{\text{low}}$ is not satisfied here. For $\delta t = 10$ a.u. $t^{\text{high}} \simeq 5.1 \cdot t^{\text{low}}$, while for $\delta t = 15$ a.u. $t^{\text{high}} \simeq 4.8 \cdot t^{\text{low}}$. It means that one way of improving the efficiency of MTS techniques is to consider cheaper “low” level models.

The real and ideal speed-ups are much closer from each-other and with an MTS factor between 3 and 6 it seems that reliable results could be obtained with a speed-up factor ranging from 1.5 to 2.5. The number of triggered high level calculations does not seem to affect the speed-up significantly since, in the considered calculations, the maximum averaged frequency of triggered high level calculations [X in eq. (4.20)] is equal to 0.026, which corresponds to a triggered call every 38 inner step. Most of the differences between the real and ideal speed-up in Fig. 4.5 can thus be attributed to the convergence overhead discussed above.

However, strong variations are observed for the individual time per step (not averaged) along the trajectories. Surprisingly, the average time per step in the low level steps of the MTS runs are often lower than in the standard low level runs ($t^{\text{LZ-PBE}}$). This explains why in Fig. 4.5 the ideal speed-up is sometimes lower than the real speed-up.

The different timings reported in this section are clearly subject to strong variations depending on the system considered as well as the computational parameters and the nuclear geometries. Therefore, a reliable comparison of the efficiency of the methods under investigation is a difficult task that will be further pursued in section 4.4.3.

4.4.3 Stochastic surface hopping

Starting from a PBE BOMD trajectory of 24 ps at 300 K in the ground state of system I, we have selected 100 equally spaced configurations. For each configuration, we have calculated the 5 lowest singlet excitation energies and oscillator strengths at the PBE0/TDDFT level. The starting state for the non-adiabatic dynamics is decided by randomly picking among the five lowest excited states with a distribution given by the normalized oscillator strengths at the corresponding configuration. From this random distribution, 94 runs started in the second excited state, one in the third, and five in the fourth excited state. The starting atomic positions and velocities were taken from the ground state MD. All the details concerning the preparation of the non-adiabatic simulations can be found in the SI.

With those initial conditions, 4 different types of simulations have been produced. A reference FS-TSH/PBE0 set of runs, and 3 different MTS-TSH batches with $N = \{4, 6, 8\}$ and using the PBE0 functional as high level and the PBE functional as low level. Fig. 4.9 shows two representative examples of the evolution of the high and low level forces during one of these simulations, as well as their difference. In total 400 trajectories were thus obtained. Each simulation is stopped when a transition to the ground state occurs or when it reaches a region of the PES where the calculation fails to converge. Most simulations reach the ground state in less than 100 fs. The calculations that failed to converge have not been considered in the statistical analysis below. In the reference calculations, 7 failed to converge, while for the MTS runs 15, 5, and 9 failed to converge for the MTS factor 4, 6, and 8, respectively. This seems to indicate that the MTS algorithm does not affect the convergence of the calculations significantly.

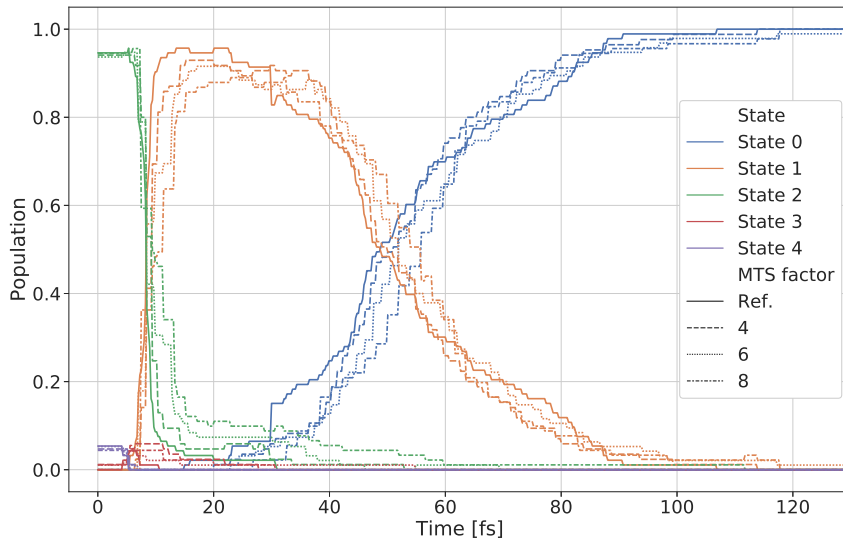


Figure 4.6: Collective evolution of the state populations along the FS-TSH/PBE0 and MTS-TSH molecular dynamics. For the MTS-TSH runs, the PBE0 functional was used as high level, while the PBE functional was used for the low level forces.

In Fig. 4.6, the collective evolution of the state populations along the dynamics are represented. The main characteristics of the photorelaxation process are well reproduced by all three MTS-TSH runs. In particular, the population is transferred from S_2 to S_1 in the first 10 to 20 fs and then slowly decays into the ground state. We note that the tail of the population of S_2 seem to become larger with the MTS factor, which shows the limit of the MTS scheme. This is reflected in the average lifetime of S_2 , for which we get 8.7 fs in the reference runs, while the MTS simulations lead to lifetimes of 10.1, 11.3, and 13.8 fs, for the MTS factors, 4, 6, and 8, respectively. The average lifetime of the first excited state is however well described with all simulations. The reference lifetime for S_1 is of 43.5 fs, while the MTS simulations lead to lifetimes of 42.5, 44.3, and 42.6 fs, for the MTS factors, 4, 6, and 8, respectively.

The populations of S_3 and S_4 are negligible. However, the trajectories starting in S_4 shows that different decay mechanisms are possible, hopping directly from S_4 to S_2 and then to S_1 or hopping first to S_3 and then directly to S_1 . Those mechanisms are rare due to the fact that the initial population of S_4 is much lower than the population of S_2 , but they are also part of the MTS-TSH swarm of trajectories, which indicates that the new MTS approach is reliable.

Following the work of Westermayr *et al.* in Ref.92, we have analyzed the geometries at which the S_2 to S_1 and S_1 to S_0 transitions occur. Fig. 4.7 represents the values of relevant geometrical parameters at the hopping geometries for the first and second transitions. The hopping geometries from the MTS-TSH runs spread basically over the same region as the hopping geometries from the reference TSH simulations.

As the MTS factor is increased, the average number of standard high level steps per trajectory

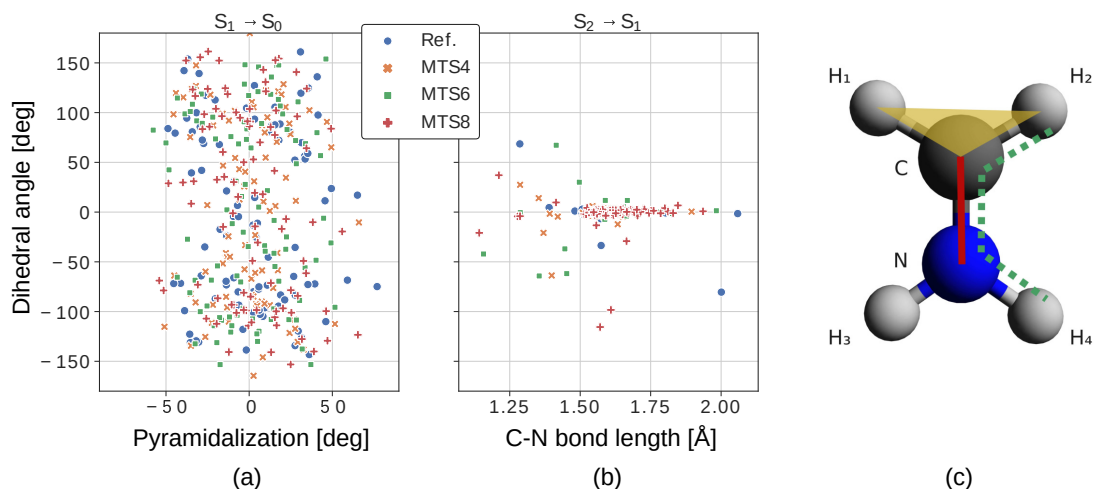


Figure 4.7: (a) Representations of the hopping geometries for the S_1 to S_0 transition and (b) for the S_2 to S_1 transition. (c) Representation of the geometric properties of the system. The C-N bond length is represented with the red solid line, the dihedral angle between atoms ($H_4, N, C,$ and H_2) is represented with the green dotted line, while the pyramidalization angle corresponds to the angle between the red solid line and the yellow triangle.

will decrease, which should decrease the computational cost of the MTS-TSH method. By standard high level steps, it is meant, high level steps which are not triggered by a low-level (LZ) transition. The number of triggered high level steps should increase with the MTS factor, since the average number of isolated low level steps (not linked to a high level calculation) will increase. Indeed, we obtain an average number of standard high level steps per trajectory of 55.6, 39.2, and 29.7 for the MTS factors 4, 6, and 8, respectively, while the average number of triggered high level steps per trajectory is 2.0, 2.5, and 2.6 for the MTS factors 4, 6, and 8 respectively.

Regarding the computational efficiency, the real speed-ups obtained over all the trajectories are 2.08, 2.56 and 2.86 for the MTS factors 4, 6, and 8, respectively. These speed-ups are in good agreement with the ideal speed-ups reporter in Fig. 4.5, which are 2.24, 2.76, and 3.12 for the MTS factors 4, 6, and 8, respectively, as can be seen from Fig. 4.8.

Overall, this statistical investigation indicates that the MTS-TSH algorithm introduced in section 4.3.2 allows to reproduce results from standard TSH simulations with a significant speed-up.

4.5 Conclusions and outlook

We have presented a new algorithm for non-adiabatic molecular dynamics simulations that is based on Tully's FS-TSH method combined with an MTS scheme for the integration of the nuclear classical equations of motion. The MTS scheme is an extension of the CPMD

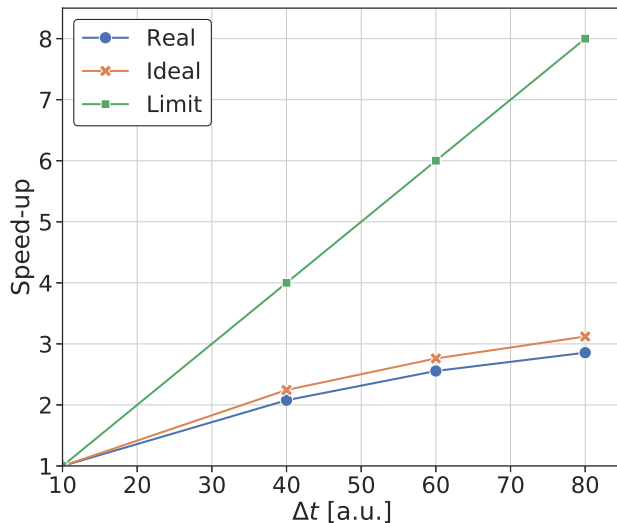


Figure 4.8: Real MTS speed-ups obtained from MTS-TSH calculations with different MTS factors compared to reference PBE0 FS-TSH timings. For comparison the ideal and limit speed-ups as defined in Eqs. (4.21) and (4.22), respectively, are also reported. The ideal speed-ups are identical to the ones in Fig. 4.5.

implementation introduced by Liberatore *et al.* in Ref. 103, in which the decomposition of the forces in terms of slow and fast components relies on the use of different electronic structure methods (*e.g.* different DFT functionals). In order to adapt the MTS scheme to the TSH method, it is important to enable electronic transitions in between outer steps. This is achieved by pre-evaluating the transition probabilities during inner steps with a low-level LZ criterion. If a transition is detected, a high level calculation is triggered, to confirm (or not) the electronic transition.

This new MTS-TSH algorithm has been tested successfully on the photorelaxation of protonated formalimine. We have shown that the MTS-TSH method is able to recover the correct state population along the reaction path as well as the correct geometries at the transitions. For this MTS scheme (combining PBE/PBE0 forces and time step factors between 2 and 8) a speed-up between 1.5 and 3 could be achieved compared to standard FS-TSH simulations. The obtained speed-ups are actually very close to the ideal speed-up that could be obtained with the computational settings considered, indicating that a better performance could be reached by considering cheaper models as low level in the MTS scheme.

This work constitutes a preliminary investigation and more tests should be performed on more complex systems to confirm the reliability of the presented results. Nonetheless, the presented MTS-TSH algorithm has shown promising results and this formulation opens the door to new developments such as combinations with other electronic structure models including machine learning techniques and QM/MM frameworks to target larger molecular systems and obtain computationally even less demanding algorithms for the description of non-adiabatic

phenomena.

4.6 Supplementary figures

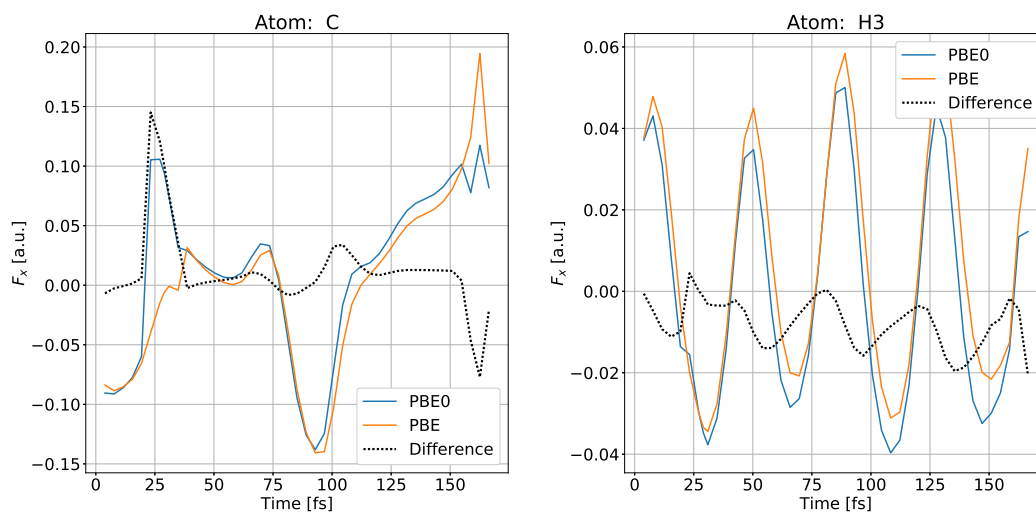


Figure 4.9: Time series of the x-Component of the total force computed using DFT with the PBE and PBE0 functionals, as well as their difference along one of the stochastic surface hopping MTS trajectories using a time step ratio of 4 (cf. Section III C of main text). The force is acting on (a) the carbon atom and (b) the H3 hydrogen of protonated formalimine (atom labels in Fig. 7c of main text).

5 Machine Learning-Enhanced Multiple Time-Step *Ab Initio* Molecular Dynamics

It's leviooosa, not leviosaaa!

Hermione Granger, Harry Potter and the Sorcerer's Stone Movie.

Chapter 5 is a pre-print version of the article:

François Mouvet, Nicholas J. Browning, Pablo Baudin, Elisa Liberatore and Ursula Rothlisberger. Machine Learning-Enhanced Multiple Time-Step *Ab Initio* Molecular Dynamics. *In preparation*.

My contributions: Continued from works of Dr. Browning. Implemented the ML software, ran the calculations, performed the analysis, finalised the article.

5.1 Abstract

The efficiency of molecular dynamics is limited by the time step that can be used to integrate the equations of motion, which is dictated by the highest frequency motion in the system. Multiple time step (MTS) integrators alleviate this issue by decomposing the forces acting on the particles into "fast" and "slow" components, which can then be integrated using different time steps. In *ab initio* MTS, an inexpensive, low level electronic structure method can be used to integrate the fast components, while its difference with an expensive but accurate high level method is used for the slow components. In this work, we present a machine learning-enhanced multiple time step (ML-MTS) method for performing accurate Born-Oppenheimer molecular dynamics at significantly reduced computational cost. We propose two alternative ML-MTS schemes which invoke different timescale separations and result in stable and accurate trajectories. In the first scheme, ML force estimates bypass the need for a high level calculation, resulting in speedups of 2 orders of magnitude over standard Velocity Verlet integration using a hybrid exchange-correlation functional. In the second scheme, we keep the high level calculation for the slow component and an ML correction is applied to the fast component, allowing a 4-fold increase in time step compared to modern *ab initio* MTS algorithms without any loss of stability, thus yielding speedups up to almost an order of magnitude over straightforward Velocity Verlet.

5.2 Introduction

The atomistic simulation of chemical systems using *ab initio* electronic structure methods is often required to calculate physical and chemical properties with a high level of accuracy. In particular, Born-Oppenheimer (BO) and Car-Parrinello [110] molecular dynamics (MD), in which the forces acting on classical nuclei are evaluated from *first principles* calculations (i.e. based on quantum mechanics), often provide an accurate description of the ground and even excited state properties of molecular systems [111, 6]. However, such simulations can quickly become computationally demanding. The computational cost of *ab initio* simulations crucially depends on three main factors: the electronic structure calculation of the nuclear forces (heavily dependent on the level of accuracy and the size of the system studied), the frequency of the fastest motion in the system, which determines the largest time step that can be used to integrate the classical equations of motion of the nuclei [112] and the timescale of the phenomena being studied (i.e. the required total duration of the simulation).

These factors put severe restrictions on the range of applications of *ab initio* MD. In order to widen this range, two main categories of techniques can be invoked. One can decide to (i) reduce the cost (level) of the electronic structure calculation or (ii) to reduce the total number of calls to the electronic structure method. The goal of the present work is to act on both points simultaneously by combining machine learning (ML) techniques and multiple time step (MTS) algorithms.

MTS algorithms [113, 9, 114] can be used to reduce the cost of MD simulations by exploiting the inherent timescale separation of the nuclear forces. Fully time-reversible MTS algorithms were first introduced in the context of force field-based MD for classes of systems where division of the forces into “slow” and “fast” components could be clearly identified [11, 115, 10]. A reduction of the computational cost can be obtained by realising that the “slower” components of the nuclear forces can be integrated less frequently, i.e. with a larger time step. In *ab initio* MD, the separation of the forces into “slow” and “fast” components is less straightforward. Nonetheless, MTS techniques applied to *ab initio* MD have been proposed by assuming different origins of the timescale separation. The earliest *ab initio* MTS implementations of Hartke *et al.* [116] and Tuckerman *et al.* [117] focused on splitting the electronic degrees of freedom in Car-Parrinello MD into rapidly and slowly varying components. Alternatively, the required separation of the nuclear forces can be imposed by including suitable splittings of the electronic Hamiltonian [118], by considering basis sets of different size [119], two-electron integral screening [120] or directly by calculating the nuclear forces with different levels of theory [121, 122, 8, 123]. Recently, we have implemented a highly efficient MTS scheme based on Density Functional Theory (DFT) in the plane-wave software CPMD [93]. Our implementation belongs to the latter kind of *ab initio* force separation schemes, where the “fast” force components are calculated from a “low” level DFT functional, while the “slow” components are defined as a correction provided by a higher level exchange-correlation functional or a wavefunction-based electronic structure method. This framework is flexible, in that the models chosen as low and high levels are not restricted. In this work we consider the benefits of this flexibility by investigating different combinations of models based on DFT and ML.

Over the last decade, great effort has been invested in the development of data-driven ML models of potential energy surfaces [124, 125, 126, 127, 128, 129, 130, 131, 132, 133], such that *ab initio* quality nuclear forces can be predicted with much reduced computational cost. These techniques replace expensive electronic structure calculations with a simpler regression model, fitted to a training database of geometries, potential energies and nuclear forces computed at some level of theory. Implicit prior knowledge can be introduced into these models via Δ -learning [134], where the difference between an affordable (and usually poor) and a very accurate (yet computationally expensive) level of theory is learned, which has shown to significantly reduce out-of-sample errors in many ML applications [135, 134]. Recently, the cost-efficient Operator Quantum Machine Learning (OQML)[136] method has been proposed, in which response properties, such as nuclear gradients, are obtained by applying an operator on the potential energy expressed in a basis of kernel functions. Here, we make use of both OQML and Δ -learning by learning the difference between an inexpensive local density approximation (LDA) functional and a more expensive functional, enabling us to perform accurate *ab initio* force evaluations at much reduced cost.

We first consider a case (“scheme I”) in which the “fast” components of the forces are computed from a “low” level functional, while the “slow” force components correspond to the ML correction targeting a “high” level functional. We also consider a another variant (“scheme II”)

Chapter 5. Machine Learning-Enhanced Multiple Time-Step *Ab Initio* Molecular Dynamics

in which the “low” level force components correspond to a low level functional plus an ML correction targeting a high level functional, while the “slow” force components are obtained from the difference between the low level (DFT+ML) and the true high level functional forces.

The paper is organized as follows. In Section 5.3, we summarize the ML method for calculating nuclear forces as well as the MTS strategy. Details regarding the different schemes and the implementation used in the present work are also included. In Section 5.4, we give information about the molecular systems and the computational settings used in Section 5.5, which illustrates the performance of both ML-MTS schemes in terms of accuracy and computational costs.

5.3 Theory

5.3.1 *Ab Initio* Nuclear Forces from Machine Learning

We make use of the Operator Quantum Machine Learning (OQML) kernel method [136], which in the following is introduced in the context of learning energies and nuclear forces. The extension for Δ -learning is also described.

For a given query system S , its potential energy E_S can be expressed as a sum of atomic energies [137], which can themselves be expressed as a linear combination of atomic kernel functions,

$$E_S = \sum_{I \in S} \epsilon_{\text{local}}(\mathbf{q}_I) = \sum_{I \in S} \sum_{J \in T} \alpha_J k(\bar{\mathbf{q}}_J, \mathbf{q}_I). \quad (5.1)$$

We introduced in this equation a training set of atomic environments T and $\bar{\mathbf{q}}_J$ is a vector representation of the environment of the J -th atom in this training set. α_J are the associated regression weights and \mathbf{q}_I is a vector description of the environment of the I -th atom in the query system S . Several contemporary atomic representations are available, which provide excellent accuracy on a variety of chemical properties [138, 139, 136, 125, 127]. In this work we use the descriptor FCHL19 because of its compactness and tested accuracy on a variety of chemical systems, including liquid water [140]. The kernel function $k(\mathbf{q}_I, \mathbf{q}_J)$ is a positive semi-definite function representing the similarity of the environment of two atoms I and J .

If n molecules are evaluated, Eq. (5.1) can be written in matrix form

$$\mathbf{E} = \mathbf{K}(T, S)\boldsymbol{\alpha} \quad (5.2)$$

where \mathbf{E} is a vector containing all potential energies and the elements of the matrix $\mathbf{K}(T, S)$ are given by

$$K_{ij}(T, S) = \sum_{I \in S_i} k(\bar{\mathbf{q}}_j, \mathbf{q}_I). \quad (5.3)$$

The force acting on the L -th atom in the query molecule can then be calculated by applying the nuclear position-derivative $\frac{\partial}{\partial \mathbf{R}_L}$ operator and is given by

$$\mathbf{F}_L = - \frac{\partial E_S}{\partial \mathbf{R}_L} \quad (5.4)$$

$$= - \sum_{I \in S} \sum_{J \in T} \alpha_J \frac{\partial k(\bar{\mathbf{q}}_J, \mathbf{q}_I)}{\partial \mathbf{q}_I} \frac{\partial \mathbf{q}_I}{\partial \mathbf{R}_L}. \quad (5.5)$$

Again, for n query systems, we can write this equation in matrix form as

$$\mathbf{F} = - \frac{\partial}{\partial \mathbf{R}} \mathbf{K}(T, S) \boldsymbol{\alpha} \quad (5.6)$$

The OQML method interpolates the energies and the forces simultaneously and the inference becomes

$$\mathbf{u}_S = \begin{bmatrix} \mathbf{E} \\ \mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{K} \\ \frac{\partial}{\partial \mathbf{R}} \mathbf{K} \end{bmatrix} \boldsymbol{\alpha} := \mathbf{K}^{\text{OQML}}(T, S) \boldsymbol{\alpha}. \quad (5.7)$$

The coefficients $\boldsymbol{\alpha}$ can be obtained by using the training set T as both the training and the query ensemble in Eq. (5.7) to infer the (known) training forces and energies and solving the resulting equations for $\boldsymbol{\alpha}$,

$$\boldsymbol{\alpha} = [\mathbf{K}^{\text{OQML}}(T, T)]^{-1} \mathbf{u}_T. \quad (5.8)$$

If T is constituted of M_T configurations containing n_T atoms each, the dimension of $\mathbf{K}^{\text{OQML}}(T, T)$ is $(M_T + 3n_T M_T) \times (n_T M_T)$. Since this kernel matrix is non-square, these equations cannot be solved by direct matrix inversion and instead a Singular Value Decomposition (SVD) factorisation is performed. To prevent overfitting, singular values σ_i smaller than a fraction λ of the largest singular value σ_{max} are discarded.

The benefit of this method over conventional kernel-ridge and Gaussian process regression [141, 132, 142] is that the kernel basis is not extended when derivative information is included. Thus, there is only one coefficient α_J per atom J in the kernel basis, regardless of the dimensionality of the derivative. This yields a kernel model with significantly reduced computational requirements for both training and prediction.

Chapter 5. Machine Learning-Enhanced Multiple Time-Step *Ab Initio* Molecular Dynamics

This formalism can be extended to the case of Δ -learning by replacing the property vector \mathbf{u}_S in Eq. (5.7) by another vector containing the energy and force differences between two methods,

$$\Delta \mathbf{u}_S = \mathbf{u}_S^1 - \mathbf{u}_S^2 = \begin{bmatrix} \Delta \mathbf{E} \\ \Delta \mathbf{F} \end{bmatrix} = \mathbf{K}^{\text{OQML}}(T, S) \boldsymbol{\alpha}. \quad (5.9)$$

Thus, the kernel \mathbf{K}^{OQML} will not change, but the learnt and inferred properties as well as the set of coefficient $\boldsymbol{\alpha}$ will differ in values.

5.3.2 Multiple Time Step Molecular Dynamics

The MTS algorithm used in this work is the microcanonical reversible reference system propagation algorithm (rRESPA) introduced by Tuckerman *et al* [9]. In this section we summarize the MTS rRESPA following the derivation and notation of the original authors.

If $\Gamma(t_0)$ is a phase space element of a system of N nuclei with positions \mathbf{q} and momenta \mathbf{p} at time t_0 , the temporal evolution can be computed by applying the classical time propagator

$$\Gamma(t_0 + t) = e^{iL t} \Gamma(t_0), \quad (5.10)$$

where L is the (classical) Liouville operator. For a set of forces \mathbf{F} acting on the nuclei, it is defined as

$$iL = \sum_{j=1}^{3N} \left(\dot{q}_j \frac{\partial}{\partial q_j} + F_j \frac{\partial}{\partial p_j} \right) := iL_q + iL_p. \quad (5.11)$$

These two operators unfortunately do not commute, making the application in Eq. (5.10) complicated. However, thanks to the symmetric Trotter theorem, it can be factorized for a small enough time increment Δt as

$$e^{iL \Delta t} \approx e^{(iL_q + iL_p) \Delta t} = e^{iL_p(\Delta t/2)} e^{iL_q \Delta t} e^{iL_p(\Delta t/2)}, \quad (5.12)$$

where third-order and higher terms have been discarded. These operators can now be applied successively on $\Gamma(0)$ and this propagator can be directly interpreted as a Velocity Verlet (VV) algorithm with an integration time step Δt .

This formalism is very useful to derive time-reversible algorithms by considering different separations of the Liouville operator. In particular if one can split the nuclear forces into “fast”

and “slow” components, \mathbf{F}^{fast} and \mathbf{F}^{slow} , the Liouville operator can be rewritten as

$$iL = \sum_{j=1}^{3N} \dot{q}_j \frac{\partial}{\partial q_j} + \sum_{j=1}^{3N} F_j^{\text{fast}} \frac{\partial}{\partial p_j} + \sum_{j=1}^{3N} F_j^{\text{slow}} \frac{\partial}{\partial p_j} \quad (5.13)$$

$$:= iL_q + iL_p^{\text{fast}} + iL_p^{\text{slow}}. \quad (5.14)$$

Using the same procedure, the operator can be factorised into

$$e^{iL\Delta t} = e^{iL_p^{\text{slow}}(\Delta t/2)} \left[e^{iL_p^{\text{fast}}(\Delta t/2n)} e^{iL_q(\Delta t/n)} e^{iL_p^{\text{fast}}(\Delta t/2n)} \right]^n e^{iL_p^{\text{slow}}(\Delta t/2)}. \quad (5.15)$$

The inner operator is equivalent to Eq. (5.12), corresponding to n VV propagation steps of the fast components of the force using a small time step $\delta t := \Delta t/n$. The outer terms correspond to a final velocity update with the slow force components, completing the propagation of the system by one full time step Δt . This MTS propagator reduces to the VV propagator when $n = 1$. In first-principles molecular dynamics, the separation of the forces into fast and slow components is not evident. The MTS implementation we use is the one described in [8], that features two levels of electronic structure methods that only differ in their (exchange and) correlation descriptions that contribute primarily to the slow force components.

In the CPMD software, this integrator is implemented as a standard VV algorithm using the small time step δt and an effective force is calculated as

$$\mathbf{F}(t_0 + t) = \begin{cases} \mathbf{F}^{\text{fast}}(t_0 + t) + n \cdot \mathbf{F}^{\text{slow}}(t_0 + t) & \text{if } (t - t_0) = k\Delta t, \forall k \in \mathbb{N} \\ \mathbf{F}^{\text{fast}}(t_0 + t) & \text{otherwise} \end{cases}, \quad (5.16)$$

where we have introduced a real number k to emphasize that the slow component contribution to the forces is calculated only every $n = \Delta t/\delta t$ iterations. We note that, by construction, a full integration step of MTS samples the phase space at the level of the high-level method.

5.3.3 Implementation

In this work, we use Δ -learning to compute an ML correction which is trained on the difference between a low level method and the target level, representing the fast and slow force contributions, respectively. This correction is calculated using equation Eq. (5.9). We then consider two different MTS schemes. In the first scheme, the slow forces are provided by the ML model directly,

$$\begin{aligned} \mathbf{F}^{\text{fast}} &= \mathbf{F}^{\text{low}} \\ \mathbf{F}^{\text{slow}} &= \Delta \mathbf{F}^{\text{ML}}. \end{aligned} \tag{Scheme I}$$

Here, no expensive high-level *ab initio* method is required and thus the computational cost of this scheme is simply the cost of the (inexpensive) low level method and the ML correction at the outer time steps. We note that the computation time of the ML correction, while significantly cheaper than a DFT calculation, can still become noticeable when a large training set containing a wide variety of atomic species is used. Thus, using MTS to avoid an ML estimation at every time step is beneficial compared to standard Δ -learning-driven dynamics. However, since in Scheme I, the slow components are not computed exactly, the ensuing phase space sampling corresponds to an ML-driven simulation mimicking the reference level of theory, not the reference level itself. Thus, the ML needs to be well-trained to have reliable results.

At contrast, in the second strategy, scheme II, the ML model is applied to the fast forces and a high-level method is computed at the outer time step,

$$\begin{aligned} \mathbf{F}^{\text{fast}} &= \mathbf{F}^{\text{low}} + \Delta \mathbf{F}^{\text{ML}} \\ \mathbf{F}^{\text{slow}} &= \mathbf{F}^{\text{high}} - \mathbf{F}^{\text{fast}}. \end{aligned} \tag{Scheme II}$$

Here, the slow forces \mathbf{F}^{slow} can be seen as a slow impulse correcting the error induced by the ML model. Because the slow components are computed explicitly, this scheme is more expensive than the first, but it samples by construction the same space as the high level method independent of the accuracy of the ML model.

5.4 Computational Details

The two ML-MTS schemes, scheme I and scheme II, have been implemented in a version of the CPMD [93] code in which other schemes can be implemented thanks to the flexibility of the MTS framework. All simulations were performed with norm-conserving Troullier-Martins pseudopotentials [143] with plane-wave cutoffs for wavefunctions and electron density of 80 Ry and 320 Ry, respectively. First, the LDA [144] and PBE functionals were used for calculating the low-level “fast” and corrective high-level “slow” force components to study the properties of these schemes. Then, the hybrid functional PBE0 will be used as the high level to showcase the possible computational gains of both schemes. The time step for VV BO molecular dynamics was $\delta t = 0.36$ fs. When running MTS, the same time was used for the inner step.

Our test system consists of 32 water molecules in a cubic periodic box ($L = 9.939$ Å). The

system was first equilibrated in the NVT ensemble at $T = 300$ K with a Nosé-Hoover chains thermostat using a chain length of 4 and a frequency of 1500 cm^{-1} for 7.2 ps, using the VV integrator. The forces were computed using DFT with the PBE functional. This trajectory was continued for another 7.2 ps using the MTS integrator with a time step ratio $n = 4$, still in the NVT ensemble at 300 K. This trajectory was used for the training set.

5.4.1 Training Data: Sample Selection

Training sets were generated using Farthest Point Sampling (FPS). The total data set for training consisted of a collection of m snapshots with atomic positions, potential energies and forces. If a given data point is denoted τ , the whole collection of available data is defined as $\mathbb{T} = \{\tau_1, \tau_2, \dots, \tau_m\}$ and the descriptor representation of atom I in τ is the vector \mathbf{q}_I , we can define a kernel-induced distance metric

$$D^2(\tau_i, \tau_j) = k_d(\tau_i, \tau_i) + k_d(\tau_j, \tau_j) - 2k_d(\tau_i, \tau_j), \quad (5.17)$$

where $k_d(\tau_i, \tau_j)$ is defined as

$$k_d(\tau_i, \tau_j) = \sum_{I \in \tau_i} \sum_{J \in \tau_j} k(\mathbf{q}_I, \mathbf{q}_J) \quad (5.18)$$

The goal is to obtain a subset $\mathbb{S} \subset \mathbb{T}$ containing N data points taken from \mathbb{T} such that elements in \mathbb{S} are maximally distant from each other, measured by this kernel-induced distance metric. First a set of 2 initial data points are randomly selected from \mathbb{T} . The next point $\tau_{next} \in (\mathbb{T} - \mathbb{S})$ to include in the training set is the one that has the largest distance to its closest neighbour in \mathbb{S} based on the distance metric Eq. (5.17),

$$\tau_{next} = \operatorname{argmax}_{\tau_j \in (\mathbb{T} - \mathbb{S})} \left[\min_{\tau_i \in \mathbb{S}} D^2(\tau_i, \tau_j) \right]. \quad (5.19)$$

The training set is then gradually increased with this rule until it contains N data points. Since the frames are ordered, it is possible to train ML models of different sizes smaller than N and measure their accuracy on test set independent of \mathbb{T} to establish the optimal size. Previous work has found that this selection scheme yields a reduction in the fraction of large errors [145, 146].

5.4.2 ML Model: Kernel Function and FCHL19 Parameters

We use the Gaussian kernel function

$$k(\mathbf{q}_I, \mathbf{q}_J) = \delta_{Z_I Z_J} \exp\left(\frac{-\|\mathbf{q}_I - \mathbf{q}_J\|^2}{2\sigma_k^2}\right) \quad (5.20)$$

Where Z_I and Z_J are the nuclear charges of atoms I and J , respectively $\delta_{Z_I Z_J}$ is the Kronecker delta function. We set the width σ_k to 32.5 throughout this work, which was previously identified to yield approximately minimal mean absolute errors (MAEs) and root-mean-square-errors (RMSEs) on an out-of-sample set. [136] We found the out-of-sample error to be relatively insensitive to the chosen width, and $\sigma_k \in [12.5, 50.0]$ all gave similar errors with minor fluctuations. The singular value threshold was set to $\sigma_i/\sigma_{\max} = 10^{-13}$.

For the FCHL19 descriptor, we used the parameters that were optimised in [140], but with a shorter radius cutoff $r_{cut} = 4.8 \text{ \AA}$ for both the two- and three-body terms instead of 8 \AA . We observed that reducing this distance yielded more accurate predictions for our system.

5.5 Results and Discussion

We have chosen liquid water as a benchmark system to investigate the capabilities of the ML-MTS procedure. Significant effort has been invested in the calculation of structural and dynamical properties of liquid water, due to its peculiar characteristics such as anomalously high polarisability, dipole moment and its self-dissociation capability. Kohn-Sham (KS) DFT has been widely applied to this system using a variety of exchange-correlation (XC) functionals, and their accuracy with respect to structural and dynamical properties has been investigated in several studies [147, 148, 149]. At ambient temperatures the local density approximation (LDA) strongly overbinds, which leads to glassy-like behaviour with over-structured radial distribution functions (RDFs) and diffusion coefficients that are an order of magnitude too small. Conversely, generalized gradient approximations (GGAs) perform reasonably well, however, they often produce RDFs that are still too structured with respect to experimental data [149]. In general, hybrid functionals provide an improved description of the structural and diffusive properties of liquid water [148, 122], however, the evaluation of exact-exchange contributions to the XC functional in plane-wave implementations typically increases the computational cost by approximately 2 orders of magnitude. Given that LDA and the GGA functional PBE already produce starkly different structural and dynamical properties, we use these as the low level and high level functionals respectively, as a highly nontrivial showcase for the performance of the ML-MTS method. Then, in Section 5.5.4, we will use PBE0 as the high level functional to evaluate the potential of this method to accelerate these simulations.

5.5.1 Model Performance

In general, the predictive power of an ML-based method heavily relies on the quantity and quality of training data fed into the algorithm. However, including redundant data into the training increases the computational cost without any gains. Thus, the first step is to test the

relation between the accuracy of the force predictions and the size of the training set. From a 7.2 ps long NVT trajectory, we selected $N = 25, 50, 75, 100$ training frames using FPS, resulting in models consisting of 2400, 4800, 7200 and 9600 atomic environments, respectively. After training, these models were tested on 20 atomic configurations extracted at random from another independent 12 ps liquid water simulation.

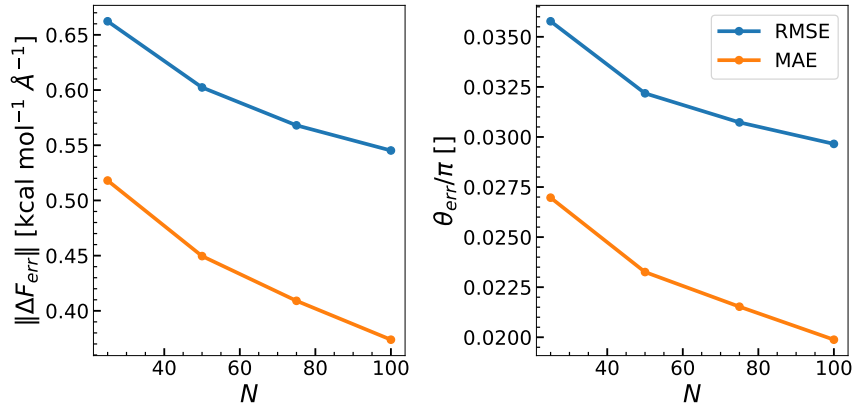


Figure 5.1: Learning curves for the ML-model trained on the liquid water database. N corresponds to the number of frames used to train the machine. The errors $\|\Delta\mathbf{F}\|_{\text{err}}$ and θ_{err} correspond to out-of-sample errors of the \mathbf{F}^* predictions with respect to target \mathbf{F}^{PBE} forces. Root-mean square errors (RMSE) are represented in blue and mean absolute errors (MAE) in orange.

For the force vector $\mathbf{F}^* = \mathbf{F}^{\text{LDA}} + \Delta\mathbf{F}^{\text{ML}}$, the prediction errors can be measured in terms of the force magnitude $\|\Delta\mathbf{F}\|_{\text{err}} = \|\mathbf{F}^{\text{PBE}} - \mathbf{F}^*\|$ and the angular error θ_{err} calculated as

$$\theta_{\text{err}}(\mathbf{F}^*, \mathbf{F}^{\text{PBE}}) = \arccos\left(\frac{\mathbf{F}^* \cdot \mathbf{F}^{\text{PBE}}}{\|\mathbf{F}^*\| \|\mathbf{F}^{\text{PBE}}\|}\right). \quad (5.21)$$

For clarity, θ_{err} was normalized by π to yield a value in $[0, 1]$, where 0 signifies that the two vectors are coincident and 1 signifies they point in opposite directions. Figure 5.1 displays the learning curves for the periodic water models. There is a healthy reduction in mean-absolute force magnitude and angular error with increasing training data. Models trained on $N = 100$ frames yield force prediction accuracies as low as $\sim 0.37 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$, with relative angular errors of $\sim 2.0\%$. Remarkably, good accuracy is already achieved even when only using 25 training frames, with a force magnitude and angular MAE of $\sim 0.52 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ and $\sim 2.7\%$ respectively. RMSEs, which provide more insight into outliers in the test set, improve at a similar rate for both measures.

Since the prediction accuracy is still improving going from 75 to 100 frames, the latter was used for the rest of this work. We first prolonged our reference trajectory for another 7.2 ps using ML-MTS Scheme **II** with a time step ratio $n = 4$. The resulting simulation contains information about both levels of theory, as well as the ML predictions at every outer time step. Figure 5.2 displays a parity plot of both $\|\mathbf{F}^{\text{LDA}}\|$ and $\|\mathbf{F}^*\|$ against the target $\|\mathbf{F}^{\text{PBE}}\|$

forces. It is evident that LDA both grossly under- and over-estimates atomic forces with larger angular errors present especially for smaller force magnitudes. Conversely, the ML correction markedly improves the force accuracy for both direction and magnitude. Relative to the extremal values of the PBE forces in the out-of-sample set, $\|F^{\text{PBE}}\|_{\text{max}} = 125.9 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ and $\|F^{\text{PBE}}\|_{\text{min}} = 1.1 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ respectively, the ML model produces excellent percentage errors given by $\|\Delta F\|_{\text{err}} / (\|F_{\text{max}}\| - \|F_{\text{min}}\|)$ with an average value of $\sim 0.14\%$. Figure 5.2 also shows a comparison $\|\Delta F\|$ for the ML (Δ -learning) prediction and target $\|\Delta F_{\text{LDA}}^{\text{PBE}}\| = \|F^{\text{PBE}} - F^{\text{LDA}}\|$ forces.

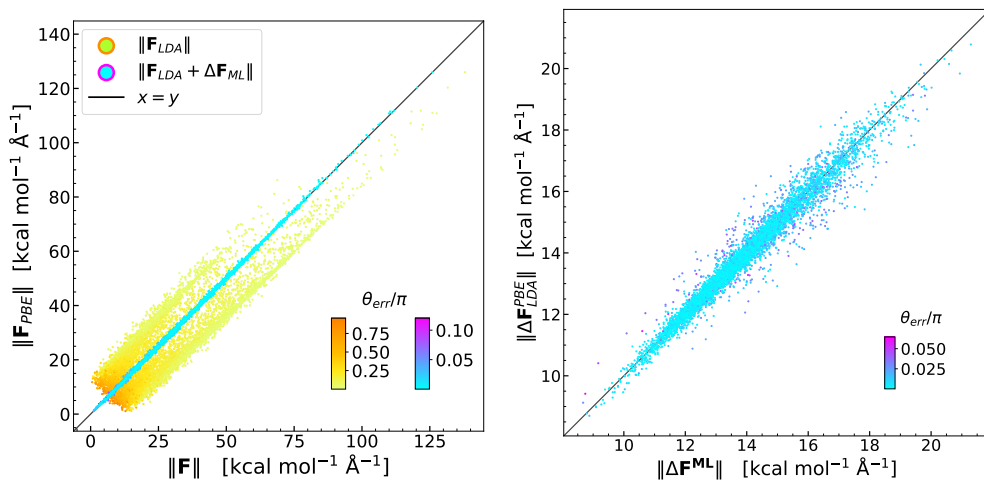


Figure 5.2: Left: parity plot of out-of-sample force norms $\|F\|$ with color-coded angular errors θ_{err} for LDA and LDA + ΔF^{ML} forces, with respect to the PBE reference. Right: parity plot displaying the accuracy of the ΔF^{ML} at predicting the corresponding target $\|\Delta F_{\text{LDA}}^{\text{PBE}}\|$ reference force differences. The ideal relationship is displayed in black in both figures.

5.5.2 Energy Conservation

Two primary factors constrain the maximum time step ratio in MTS simulations. First, the outer time step needs to accurately capture the characteristic frequency of the slow force component. Additionally, the presence of numerical resonances [150, 151, 152] occurs at specific values of the large time step when fast and slow motions resonate, causing instabilities and a degradation of energy conservation. Thus, the reliability of the ML-MTS method was first analysed by investigating the energy conservation of microcanonical (NVE) MD trajectories of our test system in which the outer time step Δt , i.e. the time step ratio n , is successively increased. We computed 7.2 ps long trajectories using regular MTS, Scheme I ML-MTS and Scheme II ML-MTS with an inner time step $\delta t = 0.36$ fs. On its own, BO MD simulations have a minor yet measurable energy buildup during long simulations. Thus, we included three VV trajectories computed with different time steps to have reference values for the typical energy conservation values, allowing us to estimate the relative impact of the MTS integration. A method to mitigate these resonances at larger time step is to use specific thermostats that were designed for this task [153, 154, 155]. Therefore, we used the same simulation setup but in the NVT ensemble, using a colored noise thermostat using the Generalised Langevin Equation (GLE).

To quantify the trend of the energy, we fitted the energy as function of time by a linear function $E = at + b$. Figure 5.3 displays the slope of this trend relative to the initial energy a/E_0 . For the standard MTS and Scheme I ML-MTS propagators, the outer time steps $\Delta t = n\delta t$ correspond to multiples of δt with $n = 4, 8, 10$, while for Scheme II ML-MTS $n = 10, 16, 24, 32$. For the VV trajectories, we used the time steps $\delta t = 0.36, 0.72$ and 1.44 fs.

As expected, the VV runs with appropriate time steps yield very good energy conservation, with a relative slope $a/E_0 = 3.1 \cdot 10^{-7} \text{ ps}^{-1}$ for $\delta t = 0.36$ fs and $a/E_0 = 6.0 \cdot 10^{-7} \text{ ps}^{-1}$ for $\delta t = 0.72$ fs. After this, the energy conservation degrades rapidly and time steps larger than 1.44 fs produce unstable trajectories.

In our test, the standard MTS scheme enables time step enlargement by $8\delta t$ without significant loss of accuracy over VV, with a relative slope a/E_0 of $1.22 \cdot 10^{-6} \text{ ps}^{-1}$. Scheme I enables very similar time step increases and our tests even showed slightly better energy conservation with respect to standard VV. We remind that the Scheme I does not explicitly compute the high level of theory and thus drastically reduces the cost of the simulation for similar results. In this case, the introduction of the GLE thermostat had a negative impact on energy conservation, with an increase of the relative slope to $4.2 \cdot 10^{-5} \text{ ps}^{-1}$. This issue may arise because the ML model was not trained using this specific thermostat and the precision of Scheme I depends entirely on the ML model's accuracy.

For Scheme II, the loss of energy conservation as a function of increasing time step is more gradual than for the other methods. Very stable trajectories are possible for as much as 16 times the inner time step without any compromise on the energy conservation ($a/E_0 = 3.4 \cdot 10^{-6} \text{ ps}^{-1}$) and up to 24 if slightly more deviation is tolerable ($a/E_0 = 1.5 \cdot 10^{-5} \text{ ps}^{-1}$). At $n = 10$,

the energy conservation is markedly better than standard MTS, with a drift that is one order of magnitude smaller than regular MTS or Scheme I. With Scheme II, using the stochastic thermostat GLE yields a good improvement over NVE simulations, systematically reducing the energy drift by a factor up to 2 at $\Delta t = 24\delta t$. Overall, the results show that the inclusion of Δ -learning models for the fast forces enables the use of significantly enlarged outer time steps while maintaining excellent energy conservation.

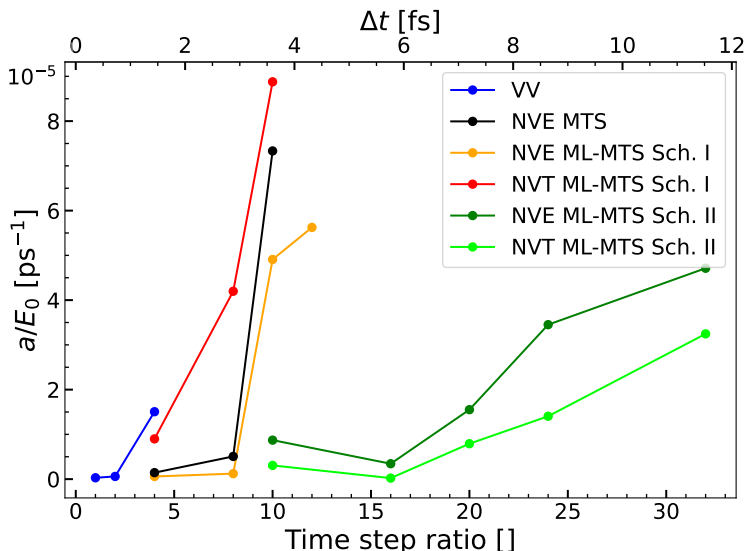


Figure 5.3: Energy conservation measured as the slope of the fitted linear function $E = at + b$, computed on trajectories using VV integration, regular MTS, Scheme I or Scheme II for different outer time steps, respectively ratios $n = \Delta t / \delta t$. The latter two are also represented in the NVT ensemble using a colored-noise stochastic thermostat (GLE). The corresponding outer time step is displayed on the top axis. All trajectories were started from the same initial conditions and with an inner time step of 0.36 fs.

5.5.3 Structural Properties

Figure 5.4 displays the oxygen-oxygen and oxygen-hydrogen radial distribution functions (RDFs) for liquid water, computed with VV-LDA, MTS LDA/PBE ($\Delta t = 4\delta t$), ML-MTS Scheme I ($\Delta t = 4\delta t$) and ML-MTS Scheme II ($\Delta t = 10\delta t$ and $\Delta t = 20\delta t$). Previous work[8] showed that standard MTS ($\Delta t = 4\delta t$) provides identical results to a VV trajectory with the high level functional. We will thus use a LDA/PBE MTS trajectory with this time step ratio as our reference. The first-shell maximum and minimum values $g(r_{\max})$ and $g(r_{\min})$ can be found in Table 5.1. Despite the somewhat higher temperature, the LDA RDFs are far too structured, with the first solvation shell oxygen-oxygen distance and the H-bonding distances being too small with values of $r_{\max}^{\text{OO}} = 2.62 \text{ \AA}$ and $r_{\max}^{\text{OH}} = 1.58 \text{ \AA}$, respectively. Reducing temperature to 300 K would make these results even more structured. The inclusion of the ML-PBE correction in scheme I significantly improves the first solvation sphere locations with all methods yielding values of $r_{\max}^{\text{OO}} = 2.75 \text{ \AA}$ and $r_{\max}^{\text{OH}} = 1.75 \text{ \AA}$, respectively, which are coincident with those measured for the standard MTS procedure. Maximum and minimum values g_{OO} and g_{OH} are also significantly

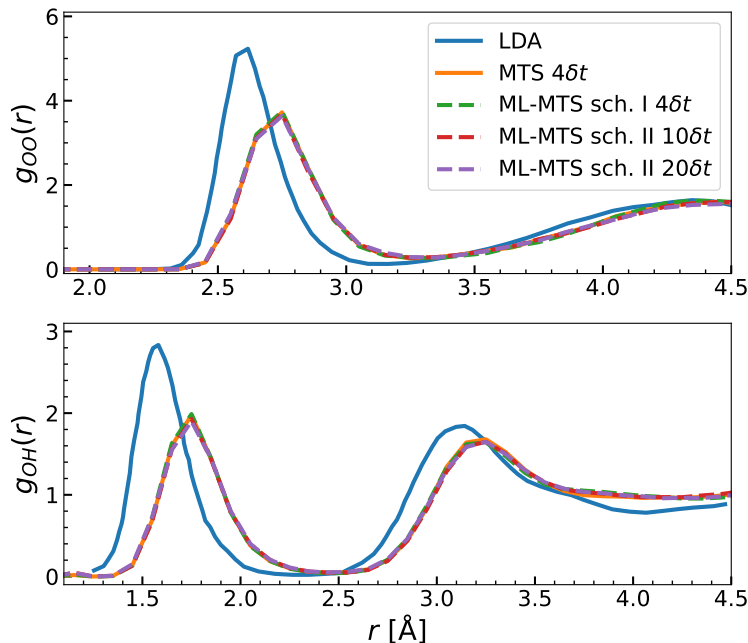


Figure 5.4: Inter-molecular oxygen-oxygen and oxygen-hydrogen radial distribution functions (RDFs) computed from NVE VV LDA MD [8] (350 K), LDA/PBE MTS ($\Delta t = 4\delta t$, 300 K), Scheme I ML-MTS ($\Delta t = 4\delta t$, 300 K) and Scheme II ML-MTS ($\Delta t = 10\delta t$ and $20\delta t$, 300 K).

improved.

The results of ML-MTS Scheme I are virtually identical to both ML-MTS Scheme II and the exact results provided by MTS ($\Delta t = 4\delta t$), thus the inclusion of the ML model indeed samples the phase space according to the high level of theory.

We would like to emphasize that of Scheme Scheme I solely relies on the ML inference for the outer integration of the MTS algorithm. The good reproduction of the structural properties suggest that the ML model accurately reproduces the correction to mimic the high-level PBE method. This level of accuracy is not guaranteed if the configurations explored during the simulations goes farther from the training set.

In contrast, Scheme Scheme II incorporates an actual high-level calculation for the final integration of the MTS, ensuring that the simulation properties align with the high-level method by construction. The ML component only serve as a proxy to increase the outer time step. This is observed in the significantly larger time step ratios, reaching up to 20, which are made possible without affecting the structural properties of water.

O-O	$r_{\max}[\text{\AA}]$	$g(r_{\max})$	$r_{\min}[\text{\AA}]$	$g(r_{\min})$	\bar{T} [K]
VV-LDA [8]	2.62	5.23	3.16	0.13	366.0
MTS $4\delta t$	2.75	3.72	3.25	0.26	289.0
ML-MTS sch. I $4\delta t$	2.75	3.74	3.35	0.25	295.2
ML-MTS sch. II $10\delta t$	2.75	3.68	3.25	0.28	306.3
O-H					
VV-LDA [8]	1.58	2.88	2.26	0.02	~
MTS $4\delta t$	1.75	1.98	2.45	0.05	~
ML-MTS sch. I $4\delta t$	1.75	1.99	2.45	0.04	~
ML-MTS sch. II $10\delta t$	1.75	1.93	2.45	0.05	~

Table 5.1: Position and height of the first maximum and minimum of the oxygen-oxygen and hydrogen-oxygen RDFs and the average temperature \bar{T} for each trajectory.

5.5.4 Efficiency assessment

Finally, we analyse the computational efficiency of the ML-MTS method by comparing it with the standard MTS integration scheme. Theoretically, if τ^H and τ^L are the times to compute the forces at the high and low level of theory and their ratio is $f = \tau^L/\tau^H$, the ideal speedup of the MTS over VV can be estimated as

$$s = \frac{n\tau^H}{\tau_n^{\text{MTS}}} = \frac{n\tau^H}{n\tau^L + \tau^H} = \frac{n}{nf + 1}. \tag{5.22}$$

Therefore, the speedup directly depends on the time step ratio and the time difference between the levels. In the extreme case where the time to compute the high level forces is significantly longer than the low level ($f \approx 0$), the speedup tends towards the time step ratio n .

Since the computational cost difference between LDA and PBE is not very large (factor 2), we used the more expensive hybrid functional PBE0 [156, 157] high level method in Scheme II to showcase the potential gains of the method. Figure 5.5 shows the speedups measured as the ratio between the total clock time required to compute a short (80 time steps = 36 fs) long simulation using a standard VV and ML-MTS Scheme II. Using only time step ratios that tightly maintain the conservation of the total energy, we obtain speedups up to 6.7 for an outer time step of $16\delta t$. The main reason for not reaching the ideal speedup of n is due to the increase of the number of self-consistent-field (SCF) iterations required to converge the wavefunction after displacement. The initial guess is based on recent history, which becomes less relevant as the time step increases. In our MTS procedure, the wavefunction is extrapolated based on the last 4 PBE0 time steps, which are increasingly poor initial guesses as Δt becomes large. Previous work showed that this limitation can be mitigated by using better extrapolation schemes [8].

The computational cost of the Δ -learning correction does not depend on the learnt level of theory, implying that the computational time of Scheme I does not depend on the chosen high

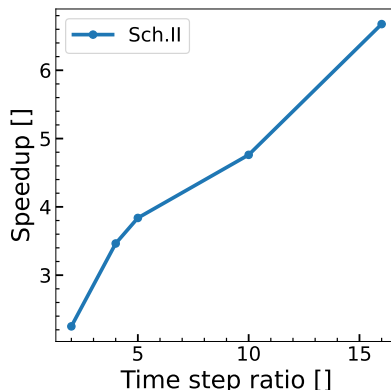


Figure 5.5: Speedup of MTS-ML Scheme **II** over standard VV using PBE0 functional, computed as the ratio of the total simulation times.

level functional. Therefore, we computed the speedup as the ratio between the computational cost of Scheme **I** at the PBE level and a VV using the PBE0 functional. Table 5.2 shows the average timings of the different steps involved in Scheme **I** and the associated speedups. Using only time steps that tightly conserve energy, we observe speedups of 2 orders magnitude. In this case, the worsening of the guess of Scheme **II** is not present since the high level method is completely bypassed. Furthermore, the speedup also increases with n because the average time for an ML force inference $\langle \tau^{\text{ML}} \rangle$ is not negligible compared to $\langle \tau^{\text{LDA}} \rangle$.

All calculations were made with 96 processors on Cray XC50 compute nodes.

	MTS2	MTS4	MTS8	VV
$\langle \tau^{\text{ML}} \rangle$ [s]	0.59	0.59	0.59	-
$\langle \tau^{\text{LDA}} \rangle$ [s]	2.00	2.02	1.75	-
$\langle \tau^{\text{PBE0}} \rangle$ [s]	-	-	-	297.55
Speedup	129.7	136.4	163.2	-

Table 5.2: Average computation times $\langle \tau \rangle$ for the different steps of ML-MTS Scheme **I** for a range of time step ratios. For comparison, the average time for a time step with PBE0 is computed. The average is done on a short 36 fs-long simulation. The speedup is computed with Eq. (5.22).

5.6 Conclusion

We have introduced a method which combines a multiple-time step scheme with machine learning for performing accurate *ab initio* molecular dynamics at significantly reduced computational cost. The numerical evidence presented here demonstrates that the addition of energy conserving machine learning corrections can yield significantly improved accuracies and speedups to standard MTS propagators. We have presented two schemes in which an ML correction is applied, through either a timescale separation argument analogous with other *ab initio* MTS applications[119, 121, 120, 122, 8] or a new scheme in which the error induced by

Chapter 5. Machine Learning-Enhanced Multiple Time-Step *Ab Initio* Molecular Dynamics

the ML model is corrected with an accurate high-level functional. In the first case, speedups of up to ~ 163 times relative to VV-PBE0 dynamics are possible, while in the latter speedups of ~ 6.7 can be reached using an outer-time step 4 times larger than what was obtained in previous work[8]. In addition, for Scheme **II**, ML-MTS poor initial guesses for the PBE0 SCF iterations result in an upper-bound to the maximum speedup possible for increasing outer time steps, which is also present in BO-MTS MD. More sophisticated extrapolation schemes would likely result in a vast improvement in computational efficiency. Nonetheless, the speedups provided by both ML-MTS schemes are encouraging and can be attached *ad hoc* to standard *ab initio* MTS procedures with minimal effort.

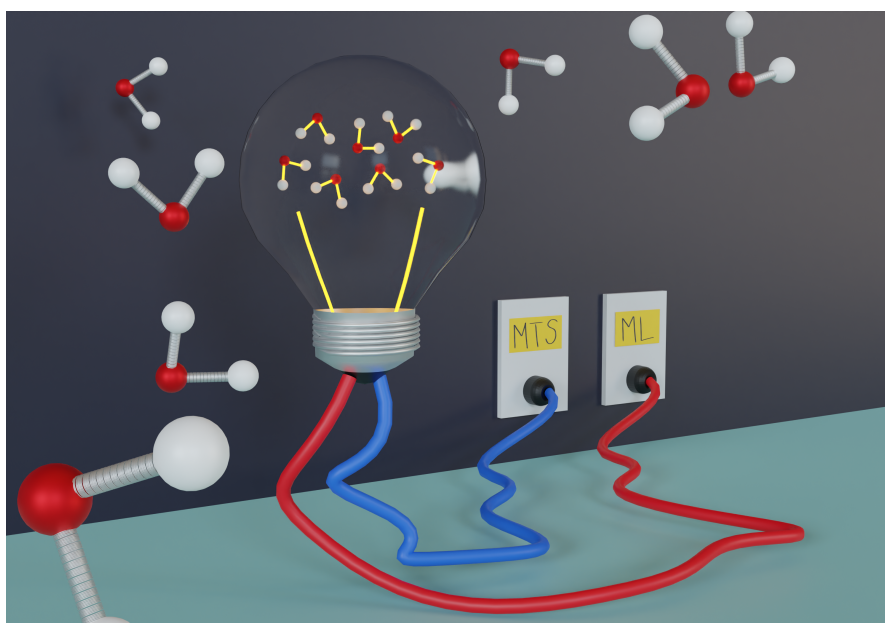
6 Multiple Time Step QM/MM Molecular Dynamics Enhanced With On-The-Fly Trained Machine Learning

A wizard is never late, Frodo Baggins, nor is he early. He arrives precisely when he means to.

Gandalf, The Fellowship of the Ring
Movie.

Chapter 6 is a pre-print article containing results of an ongoing project:

François Mouvet and Ursula Rothlisberger. Multiple Time Step QM/MM Molecular Dynamics Enhanced With On-The-Fly Trained Machine Learning. *In progress.*



6.1 Introduction

Molecular dynamics (MD) simulations are a very capable tool that can provide a detailed description of the complex mechanisms of biological macromolecules. However, the considerable number of atoms in these systems (often in the range of 10'000 to 100'000 atoms) is incompatible with a full quantum mechanical (QM) description of the atomic interactions. It is thus necessary to make some concessions to access relevant simulation times to model the desired phenomena.

A very popular choice is to rely on classical MD, where the interactions between atoms are computed using a classical potential mapped by a large set of parameters, the force field (FF). These parameters are usually estimated by fitting experimental data or energies computed using quantum-based equations. While these force fields generally have a good accuracy for certain applications, there are cases where the quantum nature of molecular systems is not reproducible by classical models. For example, the rearrangements of electrons during biochemical reactions or photoexcitations require a quantum description of the atomic interactions [6]. Moreover, most readily available FF do not contain parameters for molecules containing one or more (transition) metal centers and, in consequence, *ad hoc* FF parameters have to be computed. For these cases, hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) simulations [158], in which the system is divided into a QM region of interest and an MM environment, are a powerful alternative for studying complex chemical reactions involving both electronic and mechanical components.

The computational cost of QM/MM simulations is mostly dictated by the QM force calculation. The most direct way to extend the accessible timescale of QM/MM simulations is to minimize the time needed to compute each MD integration step by using clever software implementations that are capable of exploiting the full potential of modern computer architectures. Therefore, the parallelization of the force computation is critical to achieve good performance and most of the commonly used QM or MM software packages nowadays have implemented efficient parallelization schemes [159, 160, 161]. However, using two software packages together in a QM/MM workflow often imposes some constraints on the capabilities of each software component. To resolve this issue, the recently developed Multiscale Simulations in Computational Chemistry (MiMiC) framework is specifically designed to improve the communication between massively parallel programs [162, 163]. MiMiC loosely couples the QM and MM software packages, allowing multiple programs to run simultaneously with almost no impact on their parallelization.

Another way to increase the reachable simulation time is to reduce the number of computationally expensive force calculations by acting on the integration time step. Multiple time step algorithms (MTS), where force contributions are separated according to their characteristic time scales and integrated with a different time step, were developed for this purpose [164]. These techniques allow for frequent integrations of faster degrees of freedom while slower-evolving force components are integrated less frequently. Since the total force that

is used for propagation corresponds to the physical force, these MTS algorithms sample the same portion of phase space as if a more traditional integrator was used and by construction, these scheme can speed up calculations without any loss of accuracy. MTS algorithms were initially developed in the context of FF based (classical) MD simulations where a division into fast and slow force components emerges quite naturally but their application to QM or QM/MM MD is less straightforward, as the different force contributions are not as easily identified. However, different MTS approaches have been investigated for QM based MD[116, 117, 118, 121, 119, 120, 122, 12, 165] sometimes with varying success. One possible approach that proved successful in simulating quantum systems at a given high level of accuracy consists in introducing a second more affordable lower level of theory. The low level is used as the fast component of the forces and the slow component is then defined as the difference between the two levels, corresponding to a slowly evolving correction to the total force [8]. In the previous chapter, we have introduced a machine learning-aided multiple time step (ML-MTS) algorithm, that incorporates a Δ -learning correction to the fast, low-level forces to reduce their difference with respect to the high level description resulting in even smaller and slow-moving force corrections. This method made it possible to run stable simulations with intervals up to 7.2 fs between the expensive high-level QM calculations.

However, a caveat of any simulation that relies on a pre-trained machine learning model for the computation of the forces is that expensive training trajectories of the system still need to be computed in advance reducing the attractiveness of such methods. In addition, there may be unexplored portions of the configuration space where the ML model was not yet trained before and hence provides unreliable results. This is particularly important in the case of e.g. reactive events or phase changes in a crystal where the atomic environments change drastically.

In this chapter, we present a further evolution of this ML-MTS method destined to work in a highly efficient QM/MM environment using CPMD[93], GROMACS [166] and MiMiC [162, 163]. We also introduce the possibility to build up the ML model on-the-fly without the need of generating high-level energy and force data before hand. This is achieved by including retraining events that are triggered when the system explores atomic configurations that are too distant from the previously existing training set. This distance is measured by the kernel-induced distance metric. Because retraining events automatically trigger a calculation of the forces using the high level of theory, we also introduce a variation of the MTS algorithm where the intervals between high level corrections can be varied dynamically to match the retraining requirements. Since such simulations can be started with extremely poor ML models (trained on e.g. as few as 3 configurations) and the ML model is able to continuously adapt itself on-the-fly while the MTS framework ensures that the full accuracy of the high-level QM description of the dynamics is guaranteed at all times, this completely circumvents the need for prior generation of large amounts of costly high-level reference data. Instead in the true spirit of first-principles based MD, the ML model is trained where it is needed, i.e. on the actual configurations that are visited during the dynamics.

6.2 Theory

6.2.1 Machine learning-enhanced multiple time step MD

The ML-MTS algorithm used in this work is based on the microcanonical reversible system propagation algorithm (r-RESPA) [164] and the OQML [136] regression method. The formalism of the different elements of the ML-MTS algorithm have already been summarized in Section 5.3.

6.2.2 QM/MM MD

The QM/MM approach is a hybrid model where the atoms of a system are split into a section of interest that will be treated with QM methods and a classical MM environment. This method gives access to the simulation of very large (biological) systems that might exhibit phenomena that need to be described accurately by a QM method. The computation of the QM forces is typically the most expensive part of any QM/MM calculation and appropriate QM methods need to be selected to access relevant time scales. A popular choice is to use Density Functional Theory (DFT) with a functional that is capable of describing the underlying processes with sufficient accuracy. The MM calculations use a FF and thus only represent a small fraction of the total computational cost.

One of the main challenge of QM/MM methods is to appropriately describe the interaction between the QM and MM parts. The most straightforward way to compute the total energy of the system is to use a subtractive scheme [167]. With q labeling the quantum subsystem and c the classical environment, the total energy can be written as

$$E^{\text{tot}} = E_q^{\text{QM}} + E_{q+c}^{\text{MM}} - E_q^{\text{MM}}, \quad (6.1)$$

where the superscripts indicate the method used to compute the corresponding energy term. The main advantage of this method is its simplicity. By construction, there cannot be any double counting of interactions and subtractive QM/MM schemes can easily be set up for any pair of QM and MM software packages. However, this method has some drawbacks. For example, in their most straightforward implementation, the MM atoms cannot introduce a polarization of the electronic charge density.

The alternative is using an additive scheme, where the total energy is computed as

$$E^{\text{tot}} = E_q^{\text{QM}} + E_c^{\text{MM}} + E_{c+q}^{\text{QM/MM}}. \quad (6.2)$$

The last term $E_{c+q}^{\text{QM/MM}}$ describes all bonded and non-bonded interactions between the QM and MM regions. One complication of additive schemes is in the treatment of covalent

bonds between the QM and MM subsystems. In these cases, the bonded MM atoms are usually replaced in the QM calculation by a fictitious boundary atom such as a hydrogen atom [168, 169, 170, 171, 172] or a monovalent pseudopotential [173, 174, 175, 171]. The non-bonded terms is formed by the steric (van der Waals) interactions that are usually described by the same the force field used for the MM part, and the electrostatic interactions that can be treated at different levels of complexity [6]. In MiMiC, the MM atoms are split into two parts according their distance from the QM region. In the short-range shell, MiMiC uses an electrostatic embedding scheme calculated on the 3D mesh \mathbf{r} used by CPMD to map the total (electronic plus ionic) charge density $\rho(\mathbf{r})$. The short-range electrostatic interactions between the QM and MM subsystems are computed using a damped Coulomb potential to avoid electron spillover [176]. Specifically, if Q_i^{MM} is the partial charge of the i -th MM atom, $r_{c,i}$ its covalent radius and \mathbf{R}_i^{MM} its coordinates, the QM/MM electrostatic part of the energy is given by

$$E_{\text{ele}}^{\text{QM/MM}}(\rho(\mathbf{r}), \mathbf{R}) = \sum_{i=1}^{N_{\text{MM}}} Q_i^{MM} \int d\mathbf{r} \rho(\mathbf{r}) \frac{r_{c,i}^4 - |\mathbf{R}_i^{MM} - \mathbf{r}|^4}{r_{c,i}^5 - |\mathbf{R}_i^{MM} - \mathbf{r}|^5}. \quad (6.3)$$

The long-range interactions are computed by a multipolar expansion of the electrostatic potential of the QM region [176].

For more information on the implementation of the QM/MM calculations in MiMiC, we refer the reader to references [162] and [163].

6.3 Methods

6.3.1 ML-MTS in QM/MM MD

The MTS integration algorithm for QM/MM that we present here is based on the MTS Born-Oppenheimer (BO) MD formulation [103] that was summarized in Section 4.3.3 and a pseudo-code of its implementation in CPMD can be found in Algorithm 4.2. In this framework, the time scale decomposition of the total force is achieved by realising that the accuracy of a DFT calculation is determined by the description of exchange-correlation effects. These terms are expected to represent only a relatively small portion of the total energy. Furthermore these contributions do not have an explicit dependence on the ionic positions and can be expected to vary fairly smoothly along the atomic trajectory. It is thus possible to create two sets of force contributions described by two distinct levels of theory (differing only in the accuracy of the exchange-correlation part) that can be integrated with a different time step. Typically, a low level functional is used to compute the fast components of the forces \mathbf{F}_{fast} . Then, the slow components \mathbf{F}_{slow} are defined as the difference between the force computed at the target level of theory \mathbf{F}_{slow} and \mathbf{F}_{fast} .

$$\begin{aligned}\mathbf{F}^{\text{fast}} &= \mathbf{F}^{\text{low}} \\ \mathbf{F}^{\text{slow}} &= \mathbf{F}^{\text{high}} - \mathbf{F}^{\text{low}}.\end{aligned}\tag{6.4}$$

This scheme can easily be translated into a QM/MM workflow if the force acting on any atom is decomposed as

$$\begin{aligned}\mathbf{F}^{\text{low}} &= \mathbf{F}_{\text{QM}}^{\text{low}} + \mathbf{F}_{\text{QM/MM}}^{\text{low}} + \mathbf{F}_{\text{QM/MM}}^{\text{c}} + \mathbf{F}_{\text{MM}}^{\text{c}} \\ \mathbf{F}^{\text{high}} &= \mathbf{F}_{\text{QM}}^{\text{high}} + \mathbf{F}_{\text{QM/MM}}^{\text{high}} + \mathbf{F}_{\text{QM/MM}}^{\text{c}} + \mathbf{F}_{\text{MM}}^{\text{c}}.\end{aligned}\tag{6.5}$$

In this equation, $\mathbf{F}_{\text{MM}}^{\text{c}}$ is the classical force acting on the MM atoms, $\mathbf{F}_{\text{QM/MM}}^{\text{c}}$ is the classical part of the QM/MM forces, $\mathbf{F}_{\text{QM/MM}}^{\text{low}}$ and $\mathbf{F}_{\text{QM/MM}}^{\text{high}}$ are the parts of the QM/MM contribution treated at the QM level. Both $\mathbf{F}_{\text{MM}}^{\text{c}}$ and $\mathbf{F}_{\text{QM/MM}}^{\text{c}}$ force have no dependence on the level of theory and will cancel out in the calculation of \mathbf{F}_{slow} . This means that in practise, the MM atoms will be integrated with a standard velocity Verlet (VV) algorithm using the inner time step δt .

This method can be extended to exploit the ML-MTS method Scheme II exposed in Chapter 5, where we add a Δ -learning correction

$$\Delta\mathbf{F}_{\text{ML}} \approx \left(\mathbf{F}_{\text{QM}}^{\text{high}} + \mathbf{F}_{\text{QM/MM}}^{\text{high}}\right) - \left(\mathbf{F}_{\text{QM}}^{\text{low}} + \mathbf{F}_{\text{QM/MM}}^{\text{low}}\right)\tag{6.6}$$

to the low level of theory to reduce its difference with the high level and thus increasing the maximum attainable time step. The high and low forces of Eq. (6.5) become

$$\begin{aligned}\mathbf{F}^{\text{low}} &= \mathbf{F}_{\text{QM}}^{\text{low}} + \mathbf{F}_{\text{QM/MM}}^{\text{low}} + \mathbf{F}_{\text{QM/MM}}^{\text{c}} + \mathbf{F}_{\text{MM}}^{\text{c}} + \Delta\mathbf{F}_{\text{ML}} \\ \mathbf{F}^{\text{high}} &= \mathbf{F}_{\text{QM}}^{\text{high}} + \mathbf{F}_{\text{QM/MM}}^{\text{high}} + \mathbf{F}_{\text{QM/MM}}^{\text{c}} + \mathbf{F}_{\text{MM}}^{\text{c}}.\end{aligned}\tag{6.7}$$

6.3.2 Description of atomic environments

The atomic environments are described mathematically by the same FCHL19 [140] descriptor functions and parameters that were presented in Chapter 5.

In the context of force predictions for full QM based MD, the descriptor function of all atomic environments in a frame has to be computed and depends on the position of other atoms in the system within a cutoff distance. In the QM/MM case presented here, only the forces

acting on the QM atoms are needed. However, to compute the force acting on a given atom, the derivative with respect to the position of all atoms included in the descriptor needs to be computed. Overall, the computational cost of building the training kernel matrix that includes these derivatives (Eq. (5.7)) scales as $\mathcal{O}(6N^2M^3)$, where N is the number of frames containing M atoms each [136]. It is thus unrealistic to consider all atoms in a QM/MM simulation often constituted of thousands of atoms.

There are multiple possibilities to restrict the number of atoms considered in the descriptor. A schematic explanation of these possibilities is shown in Fig. 6.1. Let us define a subsystem labelled “ML section” that contains the atoms used to compute the descriptor of the QM atoms. First, all QM atoms need to be included in this ML section. Then, a fixed set of MM atoms of interest could be added to the ML section for all the simulation, as represented in the middle scheme of Fig. 6.1. This would make the calculations more affordable, but would be problematic when these atoms move away from the region of interest. This could be avoided by only including covalently-bonded atoms in this subsystem but this again could miss out on representing important environmental changes due to e.g. nearby mobile water molecules. Another possibility would be to include all atoms that are closer than a given distance threshold in the ML section (right scheme of Fig. 6.1). This would pose some technical challenges because the number of atoms in the descriptor would change constantly, changing the dimensions of the vectors of all descriptors.

Instead, the option we have chosen here was to begin with the absolutely minimal description consisting in limiting the ML descriptor to the isolated QM atoms, as represented by the left scheme of Fig. 6.1. This prevents all issues with redefining the ML subsection and greatly simplifies the implementation. This choice is especially reasonable for our application, since we always use the force estimation in tandem with a DFT calculation that computes all QM/MM interactions. Furthermore, the presence of close-by MM atoms can induce subtle deformations of the QM configuration that can be captured by the ML model, meaning the presence of MM atoms has still an impact on the estimated forces. If the force estimations provided by this zero order representations would prove to be too poor, more elaborate schemes could be implemented. However to anticipate our results, for the two prototypical test systems we assessed here, it turns out that this most expedient solution is already able to provide highly satisfactory results.

Another advantage of using Δ -learning is that the quantum charge distribution is computed using the lower level method, removing the need of an *ad hoc* charge estimation, which would be necessary for a direct learning scheme that can be investigated in later studies. In preparation of this, we have already implemented a kernel-based model trained on the dynamically generated restricted electrostatic potential (D-RESP) charges [177] that turns out to provide very accurate results.

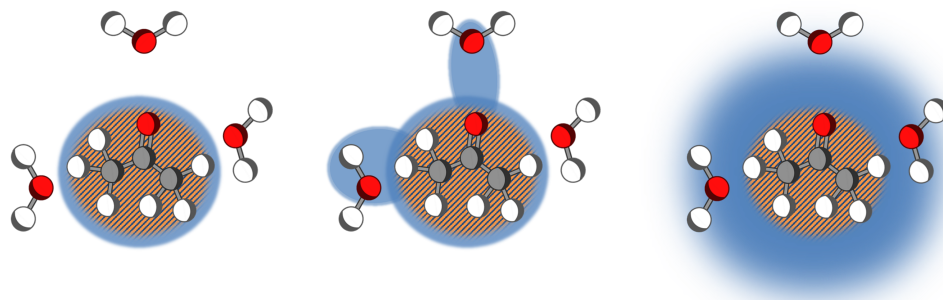


Figure 6.1: Different possible approaches for segmenting the QM, MM and ML subsystems. The QM region is shown in orange, the ML subsystem is shown in blue and the rest are MM atoms. Left: the ML and QM subsystem are identical. Middle: the ML subsystem contains QM atoms and a few selected MM atoms. Right: the ML subsystem contains the QM atoms and takes all MM atoms within a certain range into account.

6.3.3 Adaptive scheme

The adaptive ML-MTS scheme is a further evolution of the ML-MTS approach that incorporates an adaptive MTS integration time step ratio that depends on the accuracy of the ML estimates. Depending on this, the ML model can be retrained on-the-fly to incorporate unknown atomic configurations. A flowchart of this method is represented in Fig. 6.2.

The idea is to start the simulation with little to no prior data and to increase the training set as the trajectory progresses. When the simulation starts, the first step is to get both the low level forces \mathbf{F}_{low} and the correction ΔF_{ML} that are then combined to compute the fast components of the force \mathbf{F}_{fast} . Along with ΔF_{ML} , the ML program also computes a metric ξ (defined later in Eq. (6.10)) representing the confidence on the ML prediction. If ξ is smaller than a fixed threshold ξ_{max} , the MTS integration continues as intended with an inner update of the positions and velocities and the inner loop counter n is increased by one. If ξ is greater than the threshold, a retraining event is triggered, the inner loop is exited and the algorithm goes on with the computation of the high level force and finally the slow force contribution \mathbf{F}_{slow} . The outcome of the high level calculation is sent to the ML code to be incorporated into the training set and the ML model is immediately retrained. Finally, since the inner loop was interrupted after n iterations, the outer update of the positions and velocities is integrated using the time step $\Delta t = n\delta t$. The counter n is reset to 0 and the inner loop starts again. To prevent extravagant time step ratios when the ML model is sufficiently well trained, the number of inner loops n cannot exceed a maximum value, n_{max} . If this maximum outer time step is reached, the high level calculation is triggered without retraining the ML model.

6.3.4 Training Data Selection

The training set of all pre-trained ML models in this work is based on previous trajectories computed using the MTS algorithm with a small time step ratio of 4. The MTS implementation

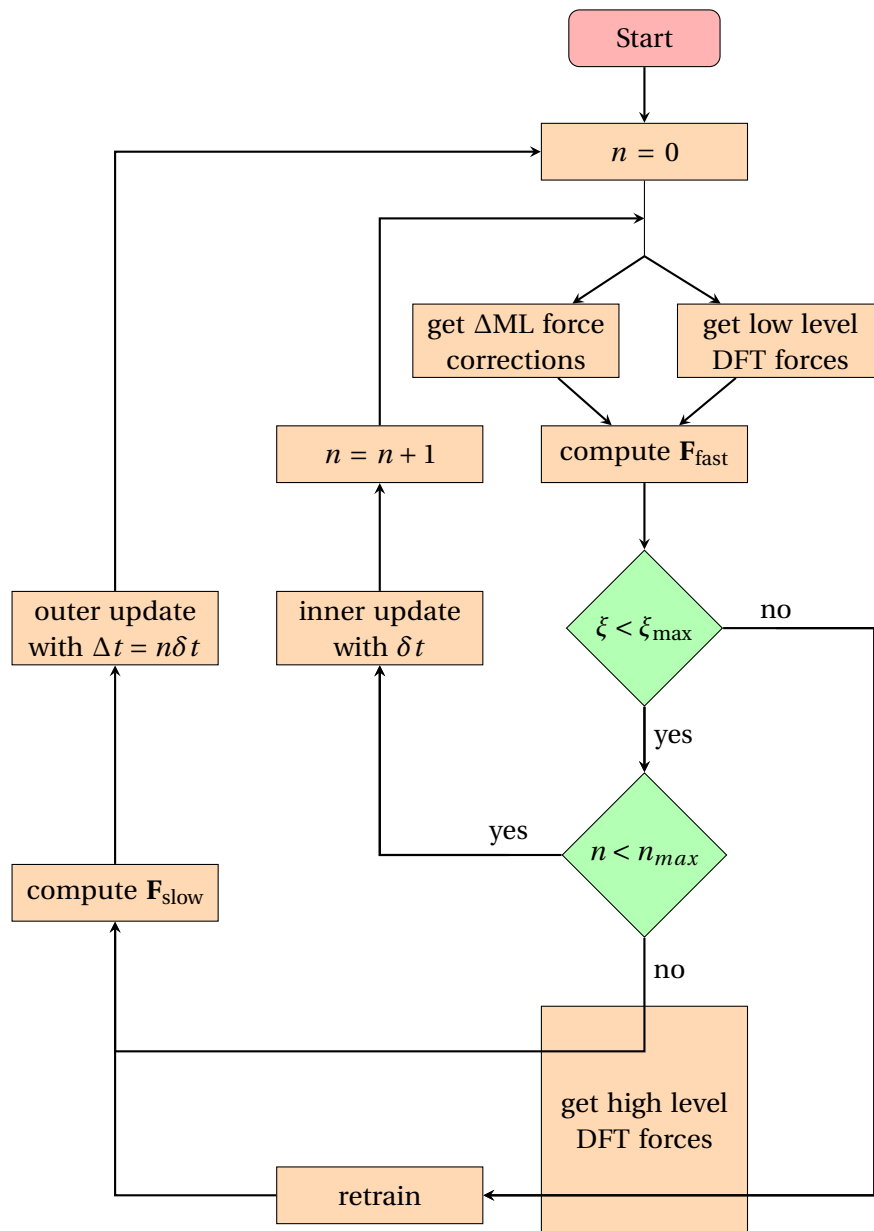


Figure 6.2: Flowchart of the adaptive MTS scheme, using an inner time step δt , an outer time step $\Delta t = n\delta t$ and a predefined maximum MTS ratio n_{\max} . The green boxes are conditions and the orange boxes are processes. The condition to arrive at the end of the simulation is not represented.

is used so that both levels of theory are computed during the outer time steps, thus providing the forces and energies of both levels and granting the possibility to compute their differences for a Δ -learning scheme. However, a trajectory can contain repeating atomic configurations that would bloat the ML model, increasing its computational cost and possibly induce errors. Therefore, all training trajectories were compressed into a smaller representative training set using the farthest point sampling (FPS) procedure presented in Section 5.4.1.

6.3.5 Confidence criterion

Several reliable uncertainty estimators have been developed during the last years both for kernel-based methods and neural networks [178, 179, 180, 181] but they usually involve a substantial increase of computational cost. We propose instead to use a metric based on the kernel-induced L_2 distance that we used previously in the context of the FPS resampling technique 5.4.1.

If \mathbf{q}_I is the descriptor of atom I in a frame τ_i , we can define the function

$$k_d(\tau_i, \tau_j) = \sum_{I \in \tau_i} \sum_{J \in \tau_j} k(\mathbf{q}_I, \mathbf{q}_J). \quad (6.8)$$

Since the kernel function $k(\mathbf{q}_I, \mathbf{q}_J)$ is a positive semidefinite function that measures the similarity between two atomic environments, this function is representative of the “total similarity” between the environments found in two frames τ_i and τ_j .

If the new configuration encountered during a simulation is τ_{new} , we can compute the kernel-induced distance with all frames in the training set. For a given frame j , it can be written as

$$D_j^2 = k_d(\tau_{new}, \tau_{new}) + k_d(\tau_j, \tau_j) - 2k_d(\tau_{new}, \tau_j). \quad (6.9)$$

Finally, if $\mathbf{D} = \{D_1^2, D_2^2, \dots, D_N^2\}$, its minimum value corresponds to the closest training frame and will be used as our confidence criterion,

$$\xi = \min \mathbf{D}. \quad (6.10)$$

Because the descriptor of a new frame is systematically computed for the Δ -learning correction and the training descriptors are stored, the cost of the computation of this metric is negligible compared to the rest of the ML machinery.

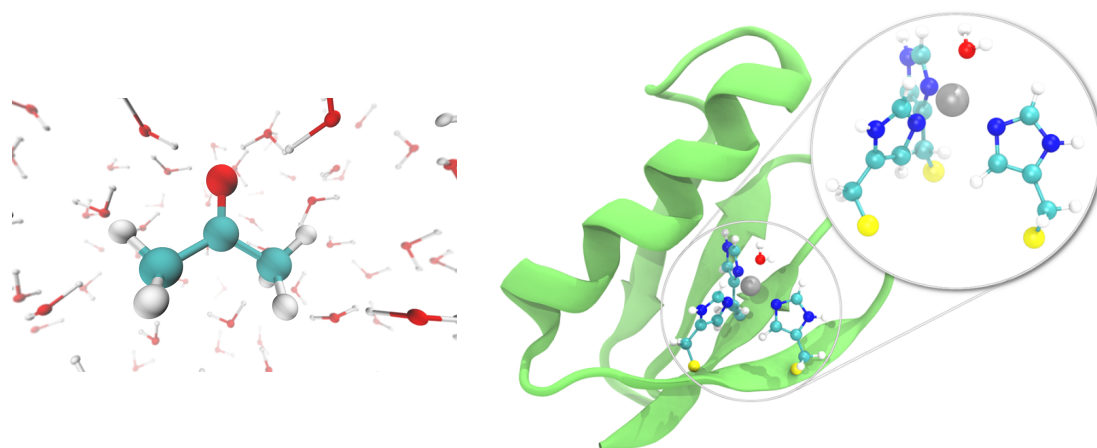


Figure 6.3: Representation of the two test systems. Left: solvated acetone (ACT). The QM region is restricted to the acetone molecule. Right: the Zinc-binding GB1 mutant. The QM region is the immediate zinc-binding site containing the zinc ion, three histidine residues and a water molecule represented as ball-and-sticks. The GB1 representation was taken from [162].

6.3.6 Software

All *ab initio* simulations are performed with the CPMD [93] software using the MiMiC [162, 163] high performance framework for the calculation of the QM/MM terms. The classical part is computed with the GROMACS software [166]. The machine learning method is implemented in an in-house code in C++, using the Eigen linear algebra library [182] and parallelized with MPI. Exploiting the flexibility of the MiMiC interface, the communications between CPMD and the ML software has been implemented in MiMiC. The calculations presented in this chapter are run on single-node workstations. Since the QM calculations are much more expensive than any other parts, a maximum of resources was given to CPMD, meaning that GROMACS and the ML software were run on single processor.

6.4 Results and Discussion

6.4.1 Simulation setup

We have selected two test systems of different complexity to showcase the properties and capabilities of the method. The first system is a single acetone molecule solvated in water (ACT). In this case, the QM region (and ML section) only contains the acetone molecule and all water is treated classically. The second, more complex test case is a biochemical system, namely a zinc-binding mutant of the GB1 protein [183, 184]. Both systems are shown in Fig. 6.3.

For all simulations, the integration time step is set to 0.24 fs. When MTS is used, the inner time step δt is also set to 0.24 fs and the outer time step Δt is specified by the time step ratio

Chapter 6. Multiple Time Step QM/MM Molecular Dynamics Enhanced With On-The-Fly Trained Machine Learning

$n = \Delta t / \delta t$. The convergence criterion for the optimization of the wavefunction is set to the CPMD default of 10^{-5} a.u..

The ACT system consists of 10 QM atoms immersed in 505 water molecules that are treated classically using a flexible TIP3P water model [185]. There are no covalent bonds between QM and MM atoms. The system was put in a periodically repeated cubic box of edge 24.74 Å. This system was created using the tools of the GROMACS software suite and first underwent a classical equilibration (energy minimization and gradual temperature ramping). Then, the system was reequilibrated at the QM/MM level and the input for this simulations was generated with the helper scripts of *mimicpy* [186]. The ACT simulations were performed using BO MD and DFT with the GGA functional BLYP [28, 187]. The plane wave basis cutoff for the wavefunctions was set to 70 Ry. To enforce the isolated system condition for the QM part, we used the Martyna-Tuckerman [188] method to solve the Poisson equation for an isolated system. The system was slowly annealed to 1 K and then heated back up to 300 K using a Berendsen thermostat with a time constant of $\tau = 5000$ a.u. The system was then equilibrated for 2.4 ps in the NVT ensemble at $T = 300$ K using a Nosé-Hoover chain (chain length: 4, frequency; 4000 cm^{-1}). We then extended this simulation in the NVT ensemble for another 2.4 ps with the same parameters but now using the MTS integrator with a time step ratio of 4. The functional describing the low level was LDA, while the high level remained BLYP. Because the forces are computed using both levels of theory at the outer time steps, this trajectory contains all the information necessary to create a training set for this system.

The GB1 system is more complex than ACT. The QM part contains a Zn ion, as well as three coordinating histidine side chains and an extra coordinated water molecule summing up to a total of 40 QM atoms. The rest of the protein (788 atoms) is treated classically using the FF14SB [189] force field and the surrounding water molecules use a rigid TIP3P water model (25149 atoms). Considering the multiple possible coordination pattern of the Zn ion, the movements of the coordinated water molecule, the possibility of water exchange and the covalent bonds between the QM and MM atoms, this system can serve as a good showcase of the capabilities and limitations of the method.

The initial structure of GB1 was taken from the SI of Ref. [162] and the equilibration step followed a similar procedure as for the ACT system. All simulations were performed using BO MD and DFT with the GGA functional BLYP. The energy cutoff for the single-particle wavefunctions was 85 Ry. For the QM/MM terms, the multipole expansion was terminated at $l=5$ and cutoff radius for short-range, respectively long-range electrostatic interactions between the QM and MM regions was set to 36 Å. The system was slowly annealed until a temperature close to 0 K was reached and was subsequently heated up to 300 K using a Berendsen thermostat with a time constant of $\tau = 5000$ a.u. Then, a longer BO MD equilibration simulation was performed using a Nosé-Hoover chain thermostat with a coupling frequency of 4000 cm^{-1} at a temperature of 300 K. All subsequent NVE trajectories of GB1 were computed from this point.

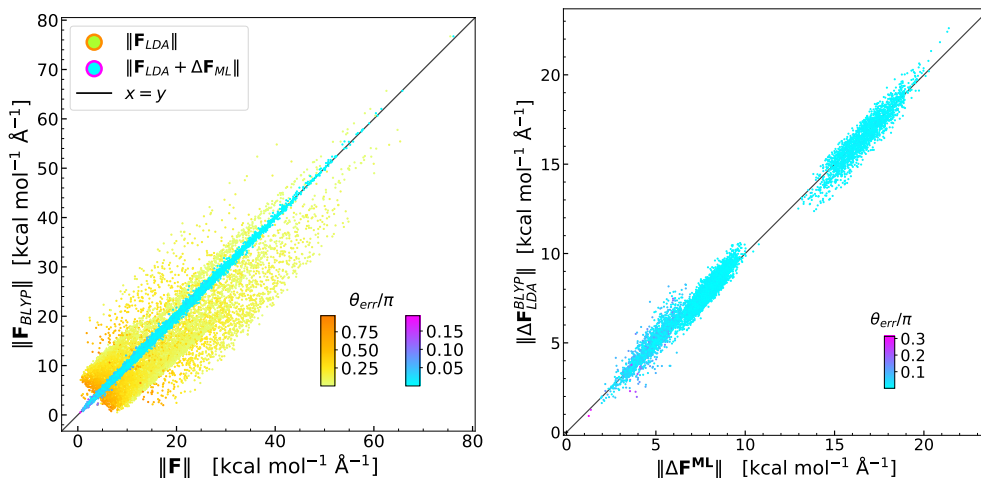


Figure 6.4: Left: Parity plot displaying the norms of the low level forces acting on the QM atoms of the solvated acetone system computed with ML-MTS ($n = 4$). The points are colored according to their relative angular error θ_{err}/π . Both LDA and LDA + $\Delta\mathbf{F}^{\text{ML}}$ are displayed. Right: parity plot of the norm of the ML force correction against its exact target. The ideal function $x = y$ is shown in black in both figures.

6.4.2 ML-MTS for QM/MM

Accuracy of the ML model

The accuracy of ML models strongly depends on the quality and amount of data that is fed into the training set. For the ACT system, the training set is based on a 2.4 ps long MTS ($n = 4$) trajectory in the NVT ensemble using LDA as the low level functional and the GGA functional BLYP as the high level. Given the time step ratio of 4, this provides 2500 data points containing positions, high and low level energies and forces. This training set is then compressed using the FPS procedure to extract the 100 maximally distant frames, providing a total of 1000 atomic environments. These frames form the training set for the ACT system.

This simulation was then extended for another 2.4 ps using the ML-MTS scheme for QM/MM with a time step ratio $n = 4$ to assess the accuracy of the ML predictions during a simulation. Fig. 6.4 shows a comparison of the high level force, the low level force and the ΔF_{ML} correction forces in a parity plot of the force norms colored with respect to their angular errors θ_{err} , computed as

$$\theta_{\text{err}}(\mathbf{F}_1, \mathbf{F}_2) = \arccos\left(\frac{\mathbf{F}_1 \cdot \mathbf{F}_2}{\|\mathbf{F}_1\| \|\mathbf{F}_2\|}\right). \quad (6.11)$$

In the following results, this angle is normalized by π to yield a value in $[0, 1]$ where 0 means that the vectors are aligned and 1 means that they are pointing in opposite directions.

In this simulation of ACT, the forces computed with LDA both under- and overestimate those

Chapter 6. Multiple Time Step QM/MM Molecular Dynamics Enhanced With On-The-Fly Trained Machine Learning

of the BLYP target. The mean average error (MAE), root mean squared error (RMSE) and the average angular error $\langle \theta_{\text{err}}/\pi \rangle$ are represented in Table 6.1. The addition of the $\Delta\mathbf{F}_{\text{ML}}$ correction greatly reduces the gap between the two levels of theory. The MAE is reduced by a factor 20.4, the RMSE by a factor 17.1 and the angular error by a factor 18.65. With the ML correction, the MAE of the force norm only represents 0.3% of the total range of values covered by the norms observed in this simulation.

For GB1, the training set is based on a very short (0.48 ps long) MTS ($n = 4$) trajectory computed using LDA as the low level functional and BLYP as the high level. With a time step of 0.24 fs, this gives a set of 500 data points. As for ACT, we extracted 100 representative frames, yielding a total of 4000 training environments. Using this ML model, this trajectory was extended for 2.4 ps using ML-MTS with a time step ratio of 4.

Fig. 6.5 shows the parity plots of the norms of the forces acting on the QM atoms of GB1. Again, adding the Δ -learning correction drastically reduces the gap between the two levels of theory, albeit a bit less than for ACT. This difference is probably due to the slightly less accurate ML model used for GB1 that was trained on very few configurations sampled over a narrow time interval. This is particularly noticeable in the right panel of Fig. 6.5, where some outliers are visible in the 0-10 kcal mol⁻¹Å⁻¹ region. Nevertheless, these outliers are few and the MAE is reduced by a factor 14.1, the RMSE by a factor 10.5 and the average angular error by a factor 12.08 compared to LDA.

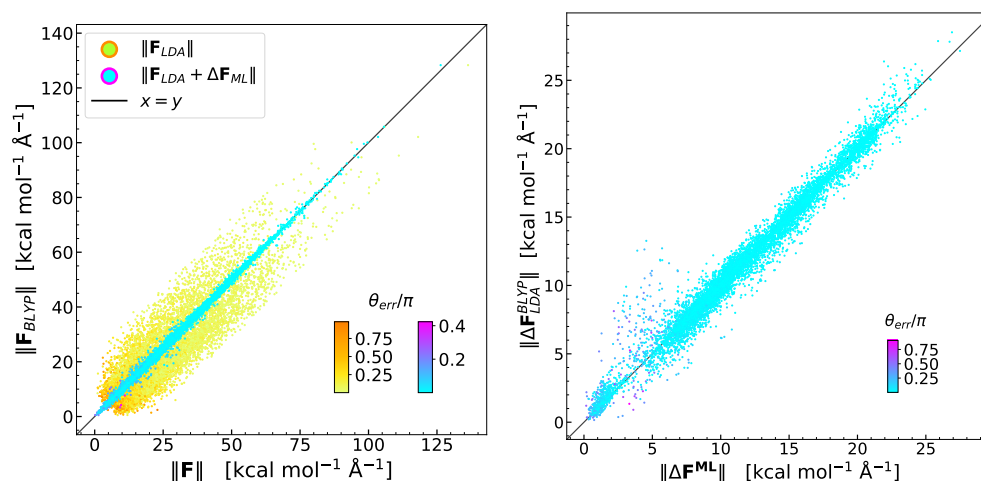


Figure 6.5: Left: Parity plot displaying the norms of the low level forces acting on the QM atoms of the GB1 system computed with ML-MTS ($n = 4$). The points are colored according to their relative angular error θ_{err}/π . Both LDA and LDA + $\Delta\mathbf{F}_{\text{ML}}$ are displayed. Right: parity plot of the norm of the ML force correction against its exact target. The ideal function $x = y$ is shown in black in both figures.

Energy conservation

To test how far the time step ratio n of the ML-MTS integration scheme can be pushed, we computed a series of microcanonical (NVE) trajectories with increasing time step ratios and

Table 6.1: Mean average error (MAE), root mean square error (RMSE) of the norm of the force acting on the QM atoms during QM/MM simulations of ACT and GB1, using ML-MTS. The average relative angular error of the forces $\langle\theta_{\text{err}}/\pi\rangle$ is also represented. The forces are expressed in $\text{kcal mol}^{-1}\text{\AA}^{-1}$.

	ACT			GB1		
	MAE	RMSE	$\langle\theta_{\text{err}}/\pi\rangle$	MAE	RMSE	$\langle\theta_{\text{err}}/\pi\rangle$
LDA	5.30	6.34	0.19	6.37	7.82	0.15
LDA+ $\Delta\mathbf{F}_{ML}$	0.26	0.37	0.01	0.45	0.74	0.01

monitored the conservation of the total energy. For both ACT and GB1, 2.4 ps long trajectories were computed with a constant inner time step $\delta t = 0.24$ fs.

Since BO MD can exhibit an energy drift (in addition to fluctuations, we considered a linear fit $E = at + b$ of the energy and used the slope as a metric describing the conservation of energy. We note that there is no physical reason why the energy drift should evolve linearly. In practice, the purpose of the fit is solely to give an indication of the trend. For ACT, we used time step ratios of $n = 4, 10, 20, 30, 40$ for the regular (without ML) MTS, corresponding to outer time steps between 0.96 fs and 9.6 fs while we added a trajectory with $n = 60$ for the ML-MTS simulations. The trends relative to the initial total energy are represented in Fig. 6.6. Strikingly, this system tolerates very large MTS ratios (likely due to the fact that the LDA forces are already quite a close approximation of their BLYP analogue) up to $n = 30$ even without the ML correction, with only minor consequence on the energy conservation ($a/E_0 = 3.88 \cdot 10^{-5} \text{ ps}^{-1}$). However, such large MTS ratios represent a limit for this system since the simulation with $n = 40$ shows very large instabilities and the energy increases rapidly at a relative rate of $a/E_0 = 2.21 \cdot 10^{-4} \text{ ps}^{-1}$, i.e. ten time faster than for the previous point. The use of the ML-MTS approach slightly increases the energy build up for time step ratios lower than $n = 30$, but drastically increases the maximal possible time step ratio. Very stable simulations are achieved even at $n = 40$ with a relative rate of $a/E_0 = 9.53 \cdot 10^{-6} \text{ ps}^{-1}$. In fact, we were able to push the time step ratio even further to $n = 60$ ($\Delta t = 14.4$ fs) and still have a good energy conservation during the simulation with $a/E_0 = 8.20 \cdot 10^{-5} \text{ ps}^{-1}$.

In the GB1 simulations, we used MTS ratios $n = 4, 8, 12, 16, 20$. The energy conservation for these trajectories is shown in Fig. 6.7. For $n = 4$, both MTS and ML-MTS simulations provide very accurate energy conservation with $a/E_0 = 2.82 \cdot 10^{-5} \text{ ps}^{-1}$ and $a/E_0 = 3.75 \cdot 10^{-5} \text{ ps}^{-1}$, respectively. Both curves reach a plateau around $\sim 1.1 \cdot 10^{-4} \text{ ps}^{-1}$, where no further degradation of the energy is observed. This plateau is marginally lower for the ML-MTS simulations. However, the curves dissociate at $n = 20$ where the regular MTS simulation gets unstable, while ML-MTS manages to keep the energy conservation at a reasonable level.

6.4.3 Adaptive ML-MTS

To define an optimal value of the a threshold that triggers a retraining event ξ_{max} , we ran two simulations using ML-MTS with a time step ratio of 12. The first simulation was run with the

Chapter 6. Multiple Time Step QM/MM Molecular Dynamics Enhanced With On-The-Fly Trained Machine Learning

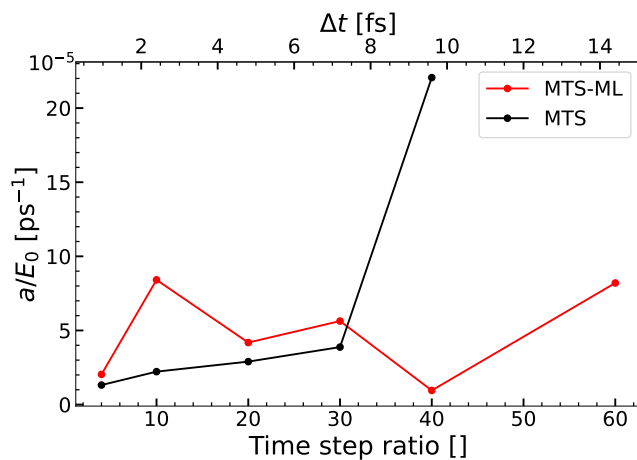


Figure 6.6: Trend of the total conserved energy during an NVE simulation of solvated acetone using both regular MTS and ML-MTS for different time step ratio n . The trend is represented relative to the initial energy $E_0 = E(t = 0)$. All trajectories are 2.4 ps long and start from the same initial positions and velocities. The top x -axis shows the corresponding outer time step Δt .

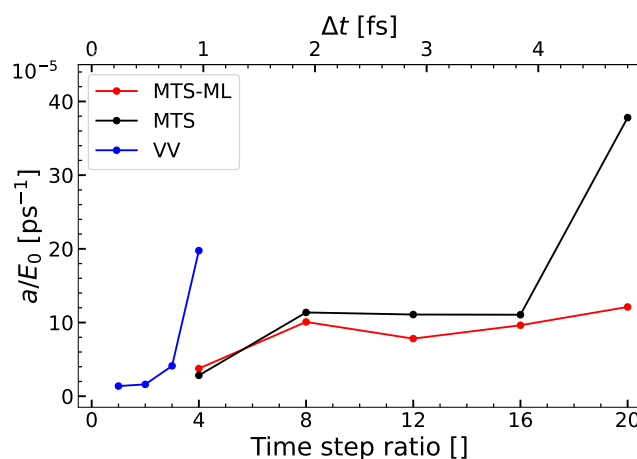


Figure 6.7: Slope of the total conserved energy during an NVE simulation of the GB1 system using both MTS and ML-MTS for different time step ratios n in comparison with regular velocity Verlet with the same time step (VV). All trajectories are 2.4 ps long and start from the same initial positions and velocities. The top x -axis shows the correspond outer time step Δt .

model described in Section 6.4.2. The second was run with a voluntarily poor training set, consisting of 3 consecutive frames taken from the beginning of the equilibration trajectory (N3). We measured the metric ξ (Eq. (6.10)) along the simulations and compare the results in Fig. 6.8. The distribution of ξ is significantly larger in the trajectory using the poor training set and takes a broad range of values up to $7.03 \cdot 10^{-3}$, with an average of $3.58 \cdot 10^{-3}$. Conversely, the appropriately trained model maintains much lower values during this test, with an average of $1.57 \cdot 10^{-3}$ and a maximum value of $3.57 \cdot 10^{-3}$. Given these distributions, we decided to use two different thresholds ξ_{\max} : a tight criterion on precision $\xi_{\max} = 3 \cdot 10^{-3}$ and a looser threshold of $\xi_{\max} = 5 \cdot 10^{-3}$.

We tested the adaptive ML-MTS procedure by running a set of 0.48 ps long trajectories of GB1. To challenge the method, we started the simulations using the the poor N3 training set. This will force the algorithm to retrain rapidly as simulation will quickly explore configurations that are out of the training set. We ran simulations with maximum time steps ratios of $n_{\max} = 12, 20, 40$, that correspond to cases where the energy was starting to deteriorate for the regular MTS algorithm (Fig. 6.7). To observe the potential gains of the method, another trajectory was computed using only the N3 training set without any further retraining (not adpative). Fig. 6.9 shows the conserved energy during these simulations along with markers indicating the moments where a retraining event was triggered. The initial and final size of the training sets is represented in Table 6.2, along with the relative energy trends.

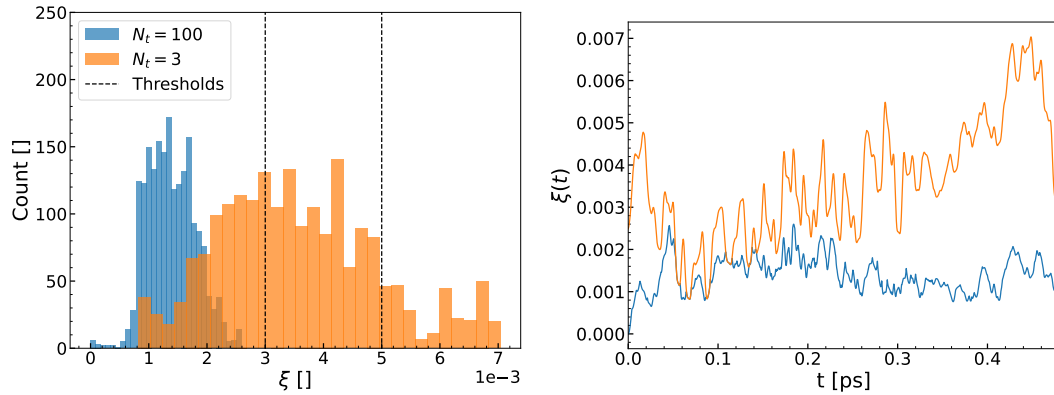


Figure 6.8: Distribution of the minimum kernel-induced distance ξ computed with Eq. (6.10) along two ML-MTS simulations of GB1 with a fixed pre-trained ML model containing N_t frames. The trajectory in blue uses a ML model described in Section 6.4.2. In orange, the model is trained on 3 consecutive frames, representing a very poor training set. The two chosen thresholds used for the adaptive scheme are shown with black dashed lines.

At $n_{\max} = 12$, all trajectories present good energy conservation even without retraining. In this case, the MTS integrator is capable of correcting the errors induced by the poor ML predictions and maintain energy conservation to a level similar to the simulations with a fixed ML model. The quality of the ML model has much more impact when the time step ratio is increased beyond the level where regular MTS is capable of maintaining energy conservation ($n > 20$). The inclusion of retraining sessions during the simulation vastly improves energy conservation,

Chapter 6. Multiple Time Step QM/MM Molecular Dynamics Enhanced With On-The-Fly Trained Machine Learning

decreasing the drifts by a factor of up to 6.2. Interestingly, the number of retraining events remains limited even with $n_{\max} = 40$ and the tight threshold, with only 16 retraining events in 10000 time steps, yielding a total of 19 frames in the final training set. This is much smaller than the training set of our pre-trained simulation in Section 6.4.2 consisting of 100 selected frames and still provides energy conservation of the same quality with a much larger time step ratio.

The maximal possible performance gains of this method might be limited by the time spent for retraining the ML model, especially since more data fed into the model translates into longer training times. To measure this time, we ran a very short simulation of GB1 where the model was retrained at every time step and we measured the timing of each subsequent retraining event. This evolution is presented in Fig. 6.10. In the simulation with the largest number of retraining events ($n_{\max} = 40$, $\xi_{\max} = 3 \cdot 10^{-3}$), the time spent for training sums up to a total of 52.46 seconds. For comparison, the total simulation time of this simulation was 24 hours and 37 minutes, meaning that the retraining phases only represent 0.059% of the total simulation time. However, the retraining time increases with the number of frames and the computational time could become a problem for very tight ξ_{\max} or very long simulations. We also note that these timings were obtained on a single-node machine and that only minimal resources were allocated to the ML part.

Table 6.2: Evolution of the size of the training set (N_{init} and N_{fin}) during the adaptive ML-MTS scheme. The relative slope of the energy a/E_0 is also shown in ps^{-1} .

	$n_{\max} = 12$			$n_{\max} = 20$			$n_{\max} = 40$		
	N_{init}	N_{fin}	a/E_0	N_{init}	N_{fin}	a/E_0	N_{init}	N_{fin}	a/E_0
$\xi_{\max} = 3 \cdot 10^{-3}$	3	12	$1.32 \cdot 10^{-4}$	3	12	$1.35 \cdot 10^{-4}$	3	19	$2.09 \cdot 10^{-4}$
$\xi_{\max} = 5 \cdot 10^{-3}$	3	6	$1.34 \cdot 10^{-4}$	3	8	$3.53 \cdot 10^{-4}$	3	10	$2.67 \cdot 10^{-4}$
Not adaptive	3	3	$1.32 \cdot 10^{-4}$	3	3	$8.33 \cdot 10^{-4}$	3	3	$1.27 \cdot 10^{-3}$

6.5 Conclusions and outlook

In this chapter, we presented two new algorithms. In the first part, we have implemented an adaption of the machine learning-aided multiple time step algorithm (ML-MTS) for the use in QM/MM simulations. This method was tested on two systems representing different complexities: a single acetone molecule in a water solution and a larger biological system containing a transition metal and covalent bonds crossing the QM/MM boundaries. For both systems, applying the Δ ML correction to improve the low level force estimate allowed to compute stable trajectories with an integration time step far beyond the capabilities of regular MTS. In particular for the first test case, acetone, the LDA/BLYP functional pair provided relatively close force descriptions allowing to increase the time step ratio up to 30 ($\Delta t = 7.2$ fs) without compromising the energy conservation in the NVE ensemble. Adding the ML correction allowed to double this MTS ratio to 60 ($\Delta t = 14.4$ fs) and still maintain a tight conservation of the total energy. For the more complex GB1 system, stable trajectories were achieved with ML-MTS with ratios up to at least 20 ($\Delta t = 4.8$ fs).

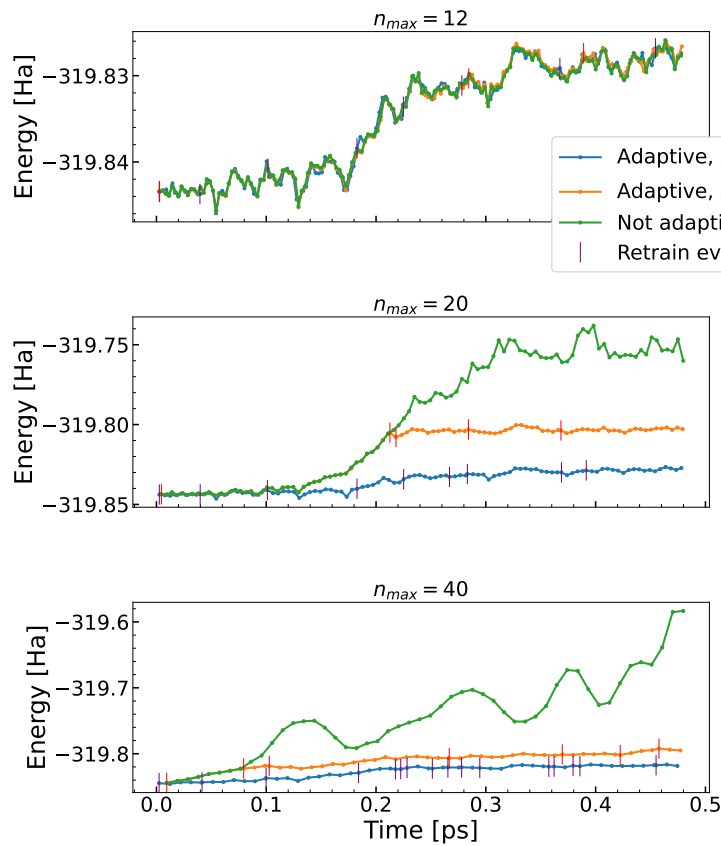


Figure 6.9: Energy evolution during 0.48 ps long simulations of GB1 with the adaptive ML-MTS trained on-the-fly for different retraining thresholds ξ_{max} and different maximum MTS ratios n_{max} . The initial training set is the same for all cases and consists of 3 consecutive frames randomly picked from the beginning of the equilibration trajectory. For all trajectories, retraining events are represented by vertical bars on the corresponding curve. We note that the scale of the y -axis is not the same for all graphs.

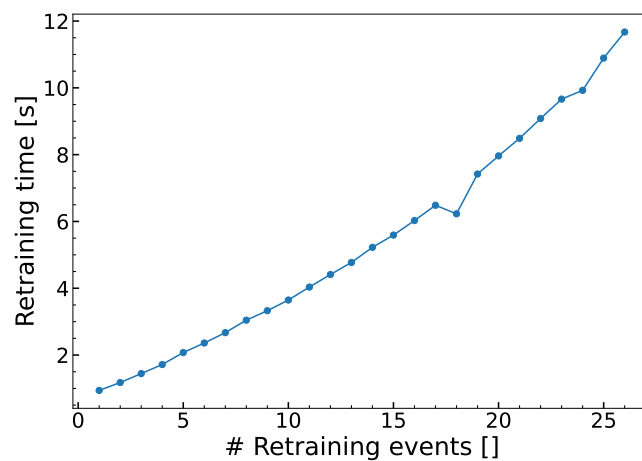


Figure 6.10: Evolution of the time required for retraining the GB1 machine learning model with the number of retraining events.

Chapter 6. Multiple Time Step QM/MM Molecular Dynamics Enhanced With On-The-Fly Trained Machine Learning

In the second part of this chapter we also introduced a new adaptive ML-MTS scheme where the time step ratio is dynamically modified to take the confidence of the machine learning prediction into account. This confidence is measured by a kernel-induced distance metric that compares the current atomic configuration with the configurations in the actual training set. This scheme also allows to incorporate the new configurations into the training set to improve the ML model on-the-fly.

We showed that the energy conservation in the ML-MTS scheme greatly depends on the accuracy of the ML model when large time step ratios are used. The adaptive ML scheme allowed to maintain energy conservation with time step ratios up to 40. Strikingly, this was achieved while only incorporating a very small number of atomic configurations thus reducing drastically the cost of both ML retraining and the inferences. In our tests, the total cost associated to retraining events was negligible compared to the overall simulation time.

Overall, the work constitutes a proof-of-concept investigation showing that these methods are capable of running stable trajectories with less frequent calls to expensive force computations. Several further analysis and developments could be investigated in future works to amend some of the limitations of the current study. First, the computational cost of the high level forces with BLYP is only twice longer as the low-level forces using LDA. For this reason, the computational gains of this method are not high in this case and in fact were not analyzed. Using a more expensive hybrid DFT method for the high level forces would be more appropriate to test the actual performance gains of this method. Furthermore, only NVE simulations were tested. Using specific stochastic thermostats for MTS [155, 153, 190, 154] often allows to further increase the time step. Finally, the other strategies suggested in Fig. 6.1 to express the atomic environments of the ML subsystem could further improve the accuracy of the ML corrections.

7 Conclusions and outlooks

7.1 Main results and outlooks

Throughout the chapters of this thesis, we have delved into different aspects of the field of computational chemistry. The upcoming sections provide a summary of the key results and outcomes obtained from the projects undertaken.

7.1.1 Computational study of drug candidates

Oxamniquine derivatives We investigated three organometallic analogues of oxamniquine, a drug against the neglected tropical disease schistosomiasis. This infection is caused by three different species of the flatworm “schistosome”. After creating a dedicated force field for each molecule, we conducted classical molecular dynamics simulations of the analogues docked into the binding pockets of their target proteins, the sulfotransferases of different worm species that activate the prodrugs. Despite the large overall homology of these enzymes, their binding pockets are rather different. We were able to identify two analogues (Fc-CH₂-OXA and Ph-CH₂-OXA) as better candidates against all species of schistosomes due to the higher numbers of sampled near-attack configurations favoring drug activation. An *in vitro* experiment further confirmed this result, showing good activities of the selected analogues against all three species. This unfortunately did not translate into the *in vivo* tests, where little to no activity was found in small animals inoculated with schistosomes of any species, despite an attempt to use lipid nanoencapsulation on Ph-CH₂-OXA to increase drug stability under *in vivo* conditions.

Given that drug stability in an acidic environment seemed to pose more problems than activity for the target proteins themselves, future work should prioritize research on improving the drugs formulation with the primary objective to improve bioavailability and hopefully mitigate the loss of *in vivo* activity.

7.1.2 Multiple time step algorithms

We have explored different ways to accelerate the computation of *ab initio* molecular dynamics through the development of multiple time step (MTS) algorithms. These new algorithms are all derived from the MTS for *ab initio* MD approach proposed by Liberatore *et al.* [8], where a more approximate but more expedient (exchange-) correlation description of a lower-level method is used to compute the fast force components $\mathbf{F}_{\text{fast}} = \mathbf{F}_L$ while a higher-level method supplies a correction term $\Delta\mathbf{F} = \mathbf{F}_H - \mathbf{F}_L$ that represents the slow force component \mathbf{F}_{slow} .

Trajectory surface hopping We presented a new approach for non-adiabatic molecular dynamics simulations based on linear response TDDFT implementation of Tully’s fewest switches trajectory surface hopping method combined with the MTS scheme for the integration of the nuclear classical equations of motion. The possibility of electronic transitions in between the large outer steps are enabled by pre-evaluating the transition probabilities during inner steps with a low-level Landau-Zener description. If a transition is detected, a high-level calculation including a full evaluation of the nonadiabatic couplings is invoked to confirm the transition. This method successfully reproduces the photorelaxation of protonated formalimine using a time step ratio up to 8, providing speedups up to 3 over standard fewest switches trajectory surface hopping simulations.

While this work served as a promising proof of concept, it is essential to conduct further tests on more intricate systems to rigorously evaluate the performance gains of method. Additionally, the flexibility of the implementation allows for different future possibilities, such as the exploration of combinations with other electronic structure methods, the incorporation of machine learning techniques, and the adaptation of this method for a QM/MM framework. These enhancements would expand the potential applications of and enable more efficient investigations of non-adiabatic effects in large biomolecular systems.

Machine learning enhanced-MTS We introduced two new algorithms to perform *ab initio* MD at a reduced cost by introducing a force correction inferred by machine learning (ML) to either the low or high level forces. In the first method (referred to as “Scheme I”), the ML correction serves as the slow component of the forces \mathbf{F}_{slow} , thus completely bypassing the need for any explicit high level calculation. The second method (“Scheme II”) introduces instead an ML correction to the low-level forces, with the objective of minimizing the difference between high and low level forces resulting in very small and smooth slow force components, \mathbf{F}_{slow} allowing a further increase in the outer time step and thus an overall efficiency enhancement. Both methods successfully reproduced the structural properties of liquid water. Scheme I provided speedups up to 2 orders of magnitude over a standard integration method with a hybrid functional. However, trajectories computed with this scheme may be hard to interpret, as they do not rely on physics but solely on the accuracy of the ML model, thus blurring the underlying hypotheses and limitations of the simulations. Furthermore, they might fail if

very different portions of phase space are sampled during e.g. the autodissociation of water. With Scheme II, the time step ratio could be pushed up to 20 without any loss of accuracy. When tested with a hybrid functional, this scheme accelerated the calculation by an order of magnitude compared to a standard velocity Verlet integrator.

The application of the colored-noise stochastics thermostat GLE during an NVT trajectory using Scheme II appropriately improved energy conservation by removing the remaining resonances. Conversely, the stochastic thermostat had a negative impact on Scheme I. This may originate from the fact that this thermostat was not used during the generation of the training set and this scheme relies entirely on the accuracy of the predicted forces.

In addition, it should be noted that the computational acceleration provided by Scheme II could be further improved by using a more adapted extrapolation scheme providing better initial guesses at the beginning of the wavefunction optimization. When the time step increases, the gap between high level calculations increases and a guess based on the previous steps becomes less relevant, leading to an increased number of self-consistency cycles required to achieve convergence at the high level. This limitation is present in both the original *ab initio* MTS and ML-MTS Scheme II. Therefore, it is essential for future research to explore novel extrapolation schemes that can significantly minimize this computational overhead and enhance overall computational efficiency. Such advancements would greatly benefit the performance and applicability of the methods.

Adaptive ML-MTS for QM/MM We presented two evolutions of the “Scheme II” ML-MTS algorithm. First, the algorithm was adapted for the highly efficient multiscale modeling framework MiMiC, thus opening the door to simulations of large biological systems. The method was tested on two systems of increasing complexity: a single acetone molecule solvated in water and a metallomutant of the GB1 protein. Including the ML correction, the MTS time step ratio could be pushed up to at least 60 ($\Delta t = 14.4$ fs) for acetone and up to 20 ($\Delta t = 4.8$ fs) for GB1, with very limited impact on energy conservation. We then presented a new adaptive ML-MTS scheme where the ML model is trained on-the-fly, thus removing the costly and tedious steps of generating a comprehensive training set. The high level force calculations imposed by the retraining events are also directly exploited in the dynamics by interrupting the inner loop of the MTS and the slow force components are integrated with an adapted time step. When tested on GB1, the adaptive ML-MTS scheme allows to maintain energy conservation with time step ratios up to 40. This was achieved while only incorporating a very small amount of atomic configurations to the training set (maximum 16 frames in 0.48 ps) and thus significantly reducing the cost of both ML retraining and inferences compared to a full pre-trained model. Furthermore, the total cost associated to retraining events was negligible compared to the overall simulation time.

There are several potential extensions for this work. First, it is important to note that the results presented in this work used a computationally affordable high-level functional, chosen

Conclusion

to efficiently demonstrate the capabilities of the new method. To measure significant computational accelerations, it is crucial to employ a more computationally intensive quantum mechanical (QM) method as the high-level reference, such as a hybrid-level functional. This would enable the measurement of notable speedups and validate the method's potential for practical applications. In addition, our tests were limited to one strategy for the ML description of the environments of the QM atoms. This strategy only considers the QM region in the descriptor. Including more information about the configuration of nearby MM atoms could improve the accuracy of the ML predictions, thus possibly leading to larger integration time steps with better energy conservation. Finally, this algorithm suffers from two limitations already observed in the previous research projects. The incorporation of MTS-specific stochastic thermostats could be highly beneficial in mitigating resonances between the different force components. By effectively minimizing these resonances, it becomes possible to further increase the integration time step, leading to even greater improvements in computational efficiency and overall speed. Moreover, this method is subject to the same issue encountered in the other MTS projects concerning the degradation of the initial guess for the wavefunction optimization when increasing the time step. An adapted extrapolation scheme for MTS applications could significantly improve the potential of this method.

7.2 Final word

This thesis has presented an exploration into the wide field of computational chemistry, with a large portion of the time dedicated to the acceleration of first-principles molecular dynamics simulations. By leveraging advanced computational techniques and novel algorithmic approaches, significant strides have been made in enhancing the efficiency and scalability of these simulations. The achievements outlined in this work not only pave the way for further advancements in the field, but also hold the promise of empowering researchers to tackle the simulation of intricate biomolecular processes with improved speed. By doing so, these advancements provide a compelling avenue for rapid screening and precise design of therapeutic compounds, ultimately enhancing the efficiency and effectiveness of drug discovery efforts. As the methodology continues to evolve and computational resources become more readily available, we can anticipate even more exciting breakthroughs in the future.

Bibliography

- [1] Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie, Colleen C. Persinger, Bernard H. Munos, Stacy R. Lindborg, and Aaron L. Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, 2010.
- [2] John P. Hughes and Sarah Rees. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6):1239–1249, 2011.
- [3] William L. Jorgensen. The Many Roles of Computation in Drug Discovery. *Science*, 303(5665):1813–1818, March 2004.
- [4] Giulia Palermo and Marco De Vivo. Computational Chemistry for Drug Discovery. In *Encyclopedia of Nanotechnology*, pages 1–15. Springer Netherlands, Dordrecht, 2014.
- [5] Douglas B. Kitchen, Herman Decornez, John R. Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949, 2004.
- [6] Elizabeth Brunk and Ursula Rothlisberger. Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chemical Reviews*, 115(12):6217–6263, 2015.
- [7] M. E. Tuckerman, B. J. Berne, and G. J. Martyna. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.*, 97(3):1990–2001, aug 1992.
- [8] Elisa Liberatore, Rocco Meli, and Ursula Rothlisberger. A versatile multiple time step scheme for efficient ab initio molecular dynamics simulations. *The Journal of Chemical Theory and Computation*, 14(6):2834–2842, 2018.
- [9] Mark E. Tuckerman, Glenn J. Martyna, and Bruce J. Berne. Molecular dynamics algorithm for condensed systems with multiple time scales. *The Journal of Chemical Physics*, 93:1287, 1990.
- [10] Mark E. Tuckerman, Bruce J. Berne, and Angelo Rossi. Molecular dynamics algorithm for multiple time scales: Systems with disparate masses. *The Journal of Chemical Physics*, 94:1465, 1991.

Bibliography

- [11] Mark E. Tuckerman, Bruce J. Berne, and Glenn J. Martyna. Molecular dynamics algorithm for multiple time scales: Systems with long range forces. *The Journal of Chemical Physics*, 94:6811, 1991.
- [12] Sagarmoy Mandal and Nisanth N. Nair. Speeding-up ab initio molecular dynamics with hybrid functionals using adaptively compressed exchange operator based multiple timestepping. *The Journal of Chemical Physics*, 151(15):151102, October 2019.
- [13] Volker L. Deringer, Albert P. Bartók, Noam Bernstein, David M. Wilkins, Michele Ceriotti, and Gábor Csányi. Gaussian Process Regression for Materials and Molecules. *Chemical Reviews*, 121(16):10073–10141, August 2021.
- [14] Bing Huang and O. Anatole von Lilienfeld. Ab Initio Machine Learning in Chemical Compound Space. *Chemical Reviews*, 121(16):10001–10036, August 2021.
- [15] Albert P. Bartók and Gabor Csányi. Gaussian approximation potentials: A brief tutorial introduction. *International Journal of Quantum Chemistry*, 115(16):1051–1057, 2015.
- [16] Ryosuke Jinnouchi, Kazutoshi Miwa, Ferenc Karsai, Georg Kresse, and Ryoji Asahi. On-the-fly active learning of interatomic potentials for large-scale atomistic simulations. *The Journal of Physical Chemistry Letters*, 11(17):6946–6955, 2020. PMID: 32787192.
- [17] Lennard Bösel, Moritz Thürlmann, and Sereina Riniker. Machine Learning in QM/MM Molecular Dynamics Simulations of Condensed-Phase Systems. *Journal of Chemical Theory and Computation*, 17(5):2641–2658, May 2021.
- [18] Julia Westermayr, Michael Gastegger, Maximilian Menger, Sebastian Mai, Leticia Gonzalez, and Philipp Marquetand. Machine learning enables long time scale molecular photodynamics simulations. *Chemical Science*, pages 1–25, 2019.
- [19] Zhenwei Li, James R. Kermode, and Alessandro De Vita. Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Physical Review Letters*, 114(9):096405, March 2015.
- [20] Daniel G Colley, Amaya L Bustinduy, W Evan Secor, and Charles H King. Human schistosomiasis. *The Lancet*, 383(9936):2253–2264, June 2014.
- [21] M. Rotger, T. Serra, M. González de Cárdenas, A. Morey, and M. A. Vicente. Increasing incidence of imported schistosomiasis in Mallorca, Spain. *European Journal of Clinical Microbiology and Infectious Diseases*, 23(11):855–856, November 2004.
- [22] Antoine Berry, Hélène Moné, Xavier Iriart, Gabriel Mouahid, Olivier Aboo, Jérôme Boissier, Judith Fillaux, Sophie Cassaing, Cécile Debuissou, Alexis Valentin, Guillaume Mitta, André Théron, and Jean-François Magnaval. Schistosomiasis Haematobium, Corsica, France. *Emerging Infectious Diseases*, 20(9):1595–1597, 2014.

- [23] Jeannine Hess, Gordana Panic, Malay Patra, Luciano Mastrobuoni, Bernhard Spingler, Saonli Roy, Jennifer Keiser, and Gilles Gasser. Ferrocenyl, Ruthenocenyl, and Benzyl Oxamniquine Derivatives with Cross-Species Activity against *Schistosoma mansoni* and *Schistosoma haematobium*. *ACS Infectious Diseases*, 3(9):645–652, 2017.
- [24] Joseph-Louis Lagrange. *Mécanique analytique*, 1788. Paris: Desaint.
- [25] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140(4A):A1133–A1138, November 1965.
- [26] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, 77(18):3865–3868, oct 1996.
- [27] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys. Rev. Lett.*, 78(7):1396–1396, feb 1997.
- [28] Axel D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6):3098–3100, September 1988.
- [29] John P Perdew, Matthias Ernzerhof, and Kieron Burke. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.*, 105(22):9982–9985, dec 1996.
- [30] Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, April 1993.
- [31] Peter Steinmann, Jennifer Keiser, Robert Bos, Marcel Tanner, and Jürg Utzinger. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. *The Lancet Infectious Diseases*, 6(7):411–425, July 2006.
- [32] Robert Bergquist, Xiao-Nong Zhou, David Rollinson, Jutta Reinhard-Rupp, and Katharina Klohe. Elimination of schistosomiasis: the tools required. *Infectious Diseases of Poverty*, 6(1):158, November 2017.
- [33] Sabine Geerts and Bruno Gryseels. Drug Resistance in Human Helminths: Current Situation and Lessons from Livestock. *Clinical Microbiology Reviews*, 13(2):207–222, April 2000. Publisher: American Society for Microbiology.
- [34] Eloi M Lago, Rogério P Xavier, Thaina R Teixeira, Lívia M Silva, Ademar A da Silva Filho, and Josué de Moraes. Antischistosomal agents: state of art and perspectives. *Future Medicinal Chemistry*, 10(1):89–120, January 2018. Publisher: Future Science.
- [35] Livia Pica-Mattocchia, A. Novi, and Donato Cioli. Enzymatic basis for the lack of oxamniquine activity in *Schistosoma haematobium* infections. *Parasitology Research*, 83(7):687–689, August 1997.

Bibliography

- [36] A. A. Sabah, Cathy Fletcher, G. Webbe, and M. J. Doenhoff. Schistosoma mansoni: Chemotherapy of infections of different ages. *Experimental Parasitology*, 61(3):294–303, June 1986.
- [37] Livia Pica-Mattocchia, Daniele Carlini, Alessandra Guidi, Velasco Cimica, Fabio Vigorosi, and Donato Cioli. The schistosome enzyme that activates oxamniquine has the characteristics of a sulfotransferase. *Memórias do Instituto Oswaldo Cruz*, 101:307–312, October 2006. Publisher: Instituto Oswaldo Cruz, Ministério da Saúde.
- [38] Claudia L. L. Valentim, Donato Cioli, Frédéric D. Chevalier, Xiaohang Cao, Alexander B. Taylor, Stephen P. Holloway, Livia Pica-Mattocchia, Alessandra Guidi, Annalisa Basso, Isheng J. Tsai, Matthew Berriman, Claudia Carvalho-Queiroz, Marcio Almeida, Hector Aguilar, Doug E. Frantz, P. John Hart, Philip T. LoVerde, and Timothy J. C. Anderson. Genetic and Molecular Basis of Drug Resistance and Species-Specific Drug Action in Schistosome Parasites. *Science*, 342(6164):1385–1389, 2013.
- [39] Alexander B. Taylor, Livia Pica-Mattocchia, Chiara M. Polcaro, Enrica Donati, Xiaohang Cao, Annalisa Basso, Alessandra Guidi, Anastasia R. Rugel, Stephen P. Holloway, Timothy J.C. Anderson, P. John Hart, Donato Cioli, and Philip T. LoVerde. Structural and Functional Characterization of the Enantiomers of the Antischistosomal Drug Oxamniquine. *PLoS Neglected Tropical Diseases*, 9(10):1–13, 2015.
- [40] Frédéric D. Chevalier, Winka Le Clec’h, Nina Eng, Anastasia R. Rugel, Rafael Ramiro de Assis, Guilherme Oliveira, Stephen P. Holloway, Xiaohang Cao, P. John Hart, Philip T. LoVerde, and Timothy J. C. Anderson. Independent origins of loss-of-function mutations conferring oxamniquine resistance in a Brazilian schistosome population. *International Journal for Parasitology*, 46(7):417–424, June 2016.
- [41] Anastasia Rugel, Reid S. Tarpley, Ambrosio Lopez, Travis Menard, Meghan A. Guzman, Alexander B. Taylor, Xiaohang Cao, Dmytro Kovalskyy, Frédéric D. Chevalier, Timothy J. C. Anderson, P. John Hart, Philip T. LoVerde, and Stanton F. McHardy. Design, Synthesis, and Characterization of Novel Small Molecules as Broad Range Antischistosomal Agents. *ACS Medicinal Chemistry Letters*, 9(10):967–973, October 2018. Publisher: American Chemical Society.
- [42] Elizabeth Hillard, Anne Vessières, Laurent Thouin, Gérard Jaouen, and Christian Amatore. Ferrocene-Mediated Proton-Coupled Electron Transfer in a Series of Ferrocifen-Type Breast-Cancer Drug Candidates. *Angewandte Chemie International Edition*, 45(2):285–290, 2006.
- [43] Christophe Biot and Daniel Dive. Bioorganometallic Chemistry and Malaria. *Medicinal Organometallic Chemistry*, 32:155–193, July 2010.
- [44] Gilles Gasser, Ingo Ott, and Nils Metzler-Nolte. Organometallic Anticancer Compounds. *Journal of Medicinal Chemistry*, 54(1):3–25, 2011.

- [45] Malay Patra and Gilles Gasser. The medicinal chemistry of ferrocene and its derivatives. *Nature Reviews Chemistry*, 1(9):1–12, September 2017. Number: 9 Publisher: Nature Publishing Group.
- [46] Yih Ching Ong, Saonli Roy, Philip C. Andrews, and Gilles Gasser. Metal Compounds against Neglected Tropical Diseases. *Chemical Reviews*, 119(2):730–796, January 2019.
- [47] Faustine Dubar, Timothy J. Egan, Bruno Pradines, David Kuter, Kanyile K. Ncokazi, Delphine Forge, Jean-François Paul, Christine Pierrot, Hadidjatou Kalamou, Jamal Khalife, Eric Buisine, Christophe Rogier, Hervé Vezin, Isabelle Forfar, Christian Slomianny, Xavier Trivelli, Sergey Kapishnikov, Leslie Leiserowitz, Daniel Dive, and Christophe . The Antimalarial Ferroquine: Role of the Metal and Intramolecular Hydrogen Bond in Activity and Resistance. *ACS Chemical Biology*, 6(3):275–287, March 2011. Publisher: American Chemical Society.
- [48] Jennifer Keiser, Mireille Vargas, Riccardo Rubbiani, Gilles Gasser, and Christophe Biot. In vitro and in vivo antischistosomal activity of ferroquine derivatives. *Parasites & Vectors*, 7(1):424, September 2014.
- [49] Christophe Biot, Donatella Taramelli, Isabelle Forfar-Bares, Lucien A. Maciejewski, Mlandzeni Boyce, Guy Nowogrocki, Jacques S. Brocard, Nicoletta Basilico, Piero Olliaro, and Timothy J. Egan. Insights into the Mechanism of Action of Ferroquine. Relationship between Physicochemical Properties and Antiplasmodial Activity. *Molecular Pharmacology*, 2(3):185–193, June 2005.
- [50] Jeannine Hess, Jennifer Keiser, and Gilles Gasser. Toward organometallic antischistosomal drug candidates. *Future Medicinal Chemistry*, 7(6):821–830, April 2015. Publisher: Future Science.
- [51] Valentin Buchter, Jeannine Hess, Gilles Gasser, and Jennifer Keiser. Assessment of tegumental damage to *Schistosoma mansoni* and *S. haematobium* after in vitro exposure to ferrocenyl, ruthenocenyl and benzyl derivatives of oxamniquine using scanning electron microscopy. *Parasites & Vectors*, 11(1):580, November 2018.
- [52] Valérien Pasche, Benoît Laleu, and Jennifer Keiser. Screening a repurposing library, the Medicines for Malaria Venture Stasis Box, against *Schistosoma mansoni*. *Parasites & Vectors*, 11(1):298, December 2018.
- [53] P. Olliaro, P. Delgado-Romero, and J. Keiser. The little we know about the pharmacokinetics and pharmacodynamics of praziquantel (racemate and R-enantiomer). *Journal of Antimicrobial Chemotherapy*, 69(4):863–870, April 2014.
- [54] P. Keen. Effect of Binding to Plasma Proteins on the Distribution, Activity and Elimination of Drugs. In Bernard B. Brodie, James R. Gillette, and Helen S. Ackerman, editors, *Concepts in Biochemical Pharmacology: Part 1*, Handbuch der experimentellen Pharmakologie/Handbook of Experimental Pharmacology, pages 213–233. Springer, Berlin, Heidelberg, 1971.

Bibliography

- [55] Measurement of in vitro intrinsic clearance using microsomes. available from: <https://www.cypotex.com/admepk/in-vitro-metabolism/microsomal-stability/>. Accessed: 04.06.2020.
- [56] Colleen A. McNaney, Dieter M. Drexler, Serhiy Y. Hnatyshyn, Tatyana A. Zvyaga, Jay O. Knipe, James V. Belcastro, and Mark Sanders. An Automated Liquid Chromatography-Mass Spectrometry Process to Determine Metabolic Stability Half-Life and Intrinsic Clearance of Drug Candidates by Substrate Depletion. *ASSAY and Drug Development Technologies*, 6(1):121–129, February 2008.
- [57] Karolina Słoczyńska, Agnieszka Gunia-Krzyżak, Paulina Koczurkiewicz, Katarzyna Wójcik-Pszczola, Dorota Żelaszczyk, Justyna Popiół, and Elżbieta Pękala. Metabolic stability and its role in the discovery of new chemical entities. *Acta Pharmaceutica*, 69(3):345–361, September 2019.
- [58] Molinspiration Cheminformatics, M.2001: <https://www.molinspiration.com/>. Accessed: 05.06.2020.
- [59] Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1-3):3–25, January 1997.
- [60] N.T. Huynh, C. Passirani, P. Saulnier, and J.P. Benoit. Lipid nanocapsules: A new platform for nanomedicine. *International Journal of Pharmaceutics*, 379(2):201–209, September 2009.
- [61] Anne-Laure Laine, Ngoc Trinh Huynh, Anne Clavreul, Julien Balzeau, Jérôme Béjaud, Anne Vessieres, Jean-Pierre Benoit, Joël Eyer, and Catherine Passirani. Brain tumour targeting strategies via coated ferrociphenol lipid nanocapsules. *European Journal of Pharmaceutics and Biopharmaceutics*, 81(3):690–693, August 2012.
- [62] Emilie Allard, Delphine Jarnet, Anne Vessières, Sandrine Vinchon-Petit, Gérard Jaouen, Jean-Pierre Benoit, and Catherine Passirani. Local Delivery of Ferrociphenol Lipid Nanocapsules Followed by External Radiotherapy as a Synergistic Treatment Against Intracranial 9L Glioma Xenograft. *Pharmaceutical Research*, 27(1):56–64, January 2010.
- [63] Reatul Karim, Elise Lepeltier, Lucille Esnault, Pascal Pigeon, Laurent Lemaire, Claire Lépinoux-Chambaud, Nicolas Clere, Gérard Jaouen, Joel Eyer, Géraldine Piel, and Catherine Passirani. Enhanced and preferential internalization of lipid nanocapsules into human glioblastoma cells: effect of a surface-functionalizing NFL peptide. *Nanoscale*, 10(28):13485–13501, 2018.
- [64] Anne-Laure Lainé, Eric Adriaenssens, Anne Vessières, Gérard Jaouen, Cyril Corbet, Emilie Desruelles, Pascal Pigeon, Robert-Alain Toillon, and Catherine Passirani. The in vivo performance of ferrocenyl tamoxifen lipid nanocapsules in xenografted triple negative breast cancer. *Biomaterials*, 34(28):6949–6956, September 2013.

- [65] Guillaume Bastiat, Christian Oliver Pritz, Clemens Roeder, Florian Fouchet, Erwann Lignières, Alexander Jesacher, Rudolf Glueckert, Monika Ritsch-Marte, Anneliese Schrott-Fischer, Patrick Saulnier, and Jean-Pierre Benoit. A new tool to ensure the fluorescent dye labeling stability of nanocarriers: A real challenge for fluorescence imaging. *Journal of Controlled Release*, 170(3):334–342, September 2013.
- [66] Carl Simonsson, Guillaume Bastiat, Marion Pitorre, Andrey S. Klymchenko, Jérôme Béjaud, Yves Mély, and Jean-Pierre Benoit. Inter-nanocarrier and nanocarrier-to-cell transfer assays demonstrate the risk of an immediate unloading of dye from labeled lipid nanocapsules. *European Journal of Pharmaceutics and Biopharmaceutics*, 98:47–56, January 2016.
- [67] Flavio C. Lombardo, Valérian Pasche, Gordana Panic, Yvette Endriss, and Jennifer Keiser. Life cycle maintenance and drug-sensitivity assays for early drug discovery in *Schistosoma mansoni*. *Nature Protocols*, 14(2):461–481, February 2019.
- [68] Executive Committee of the German Medical Association on the Recommendation of the Scientific Advisory Board. Cross-Sectional Guidelines for Therapy with Blood Components and Plasma Derivatives: Chapter 5 Human Albumin - Revised. *Transfusion Medicine and Hemotherapy*, 43(3):223–232, 2016.
- [69] J. Keiser. *In vitro* and *in vivo* trematode models for chemotherapeutic studies. *Parasitology*, 137(3):589–603, March 2010.
- [70] E. C. Faust and C. A. Jones. Life History of Manson's Blood Fluke (*Schistosoma mansoni*). III. The Blood Picture in Schistosomiasis Mansoni. *Experimental Biology and Medicine*, 31(4):478–479, January 1934.
- [71] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020.
- [72] Alexander B. Taylor, Kenneth M. Roberts, Xiaohang Cao, Nathaniel E. Clark, Stephen P. Holloway, Enrica Donati, Chiara M. Polcaro, Livia Pica-Mattoccia, Reid S. Tarpley, Stanton F. McHardy, Donato Cioli, Philip T. LoVerde, Paul F. Fitzpatrick, and P. John Hart. Structural and enzymatic insights into species-specific resistance to schistosome parasite drug therapy. *Journal of Biological Chemistry*, 292(27):11154–11164, July 2017.
- [73] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C.

Bibliography

- Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian-16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.
- [74] Christopher I. Bayly, Piotr Cieplak, Wendy Cornell, and Peter A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*, 97(40):10269–10280, October 1993.
- [75] D.A. Case and H.M. Aktulga and K. Belfon and I.Y. Ben-Shalom and J.T. Berryman and S.R. Brozell and D.S. Cerutti and T.E. Cheatham and III and G.A. Cisneros and V.W.D. Cruzeiro and T.A. Darden and N. Forouzesh and G. Giambasu and T. Giese and M.K. Gilson and H. Gohlke and A.W. Goetz and J. Harris and S. Izadi and S.A. Izmailov and K. Kasavajhala and M.C. Kaymak and E. King and A. Kovalenko and T. Kurtzman and T.S. Lee and P. Li and C. Lin and J. Liu and T. Luchko and R. Luo and M. Machado and V. Man and M. Manathunga and K.M. Merz and Y. Miao and O. Mikhailovskii and G. Monard and H. Nguyen and K.A. O’Hearn and A. Onufriev and F. Pan and S. Pantano and R. Qi and A. Rahnamoun and D.R. Roe and A. Roitberg and C. Sagui and S. Schott-Verdugo and A. Shajan and J. Shen and C.L. Simmerling and N.R. Skrynnikov and J. Smith and J. Swails and R.C. Walker and J. Wang and J. Wang and H. Wei and X. Wu and Y. Wu and Y. Xiong and Y. Xue and D.M. York and S. Zhao and Q. Zhu and P.A. Kollman. Amber 2016, 2016. University of California, San Francisco.
- [76] Thompson N. Doman, B. Bosnich, and Clark R. Landis. Molecular Mechanics Force Fields for Linear Metallocenes. *Journal of the American Chemical Society*, 114(18):7264–7272, 1992.
- [77] András Fiser, Richard Kinh Gian Do, and Andrej Šali. Modeling of loops in protein structures. *Protein Science*, 9(9):1753–1773, January 2000.
- [78] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, March 1977.
- [79] Bill R. Miller, T. Dwight McGee, Jason M. Swails, Nadine Homeyer, Holger Gohlke, and Adrian E. Roitberg. *MMPBSA.py* : An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation*, 8(9):3314–3321, September 2012.
- [80] Malay Patra, Katrin Ingram, Anna Leonidova, Vanessa Pierroz, Stefano Ferrari, Murray N. Robertson, Matthew H. Todd, Jennifer Keiser, and Gilles Gasser. In Vitro Metabolic Profile and in Vivo Antischistosomal Activity Studies of (η^6 -Praziquantel)Cr(CO)₃ Derivatives. *Journal of Medicinal Chemistry*, 56(22):9192–9198, November 2013.
- [81] Sarah Keller, Yih Ching Ong, Yan Lin, Kevin Cariou, and Gilles Gasser. A tutorial for the assessment of the stability of organometallic complexes in biological media. *Journal of Organometallic Chemistry*, 906:121059, January 2020.

- [82] Jackie Tan, Haresh Sivaram, and Han Vinh Huynh. Gold(I) bis(N-heterocyclic carbene) complexes: Metabolic stability, *in vitro* inhibition, and genotoxicity: Bioactivity of gold-NHC complexes. *Applied Organometallic Chemistry*, 32(8):e4441, August 2018.
- [83] Béatrice Heurtault, Patrick Saulnier, Brigitte Pech, Jacques-Emile Proust, and Jean-Pierre Benoit. A Novel Phase Inversion-Based Process for the Preparation of Lipid Nanocarriers. *Pharmaceutical Research*, 19(6):875–880, June 2002.
- [84] Basile F. E. Curchod and Todd J. Martínez. Ab Initio Nonadiabatic Quantum Molecular Dynamics. *Chem. Rev.*, page acs.chemrev.7b00423, 2018.
- [85] Rachel Crespo-otero and Mario Barbatti. Recent Advances and Perspectives on Nonadiabatic Mixed Quantum - Classical Dynamics. *Chem. Rev.*, 118:7026–7068, 2018.
- [86] Tammie Nelson, Sebastian Fernandez-Alberti, Vladimir Chernyak, Adrian E. Roitberg, and Sergei Tretiak. Nonadiabatic excited-state molecular dynamics: Numerical tests of convergence and parameters. *J. Chem. Phys.*, 136(5):054108, feb 2012.
- [87] Mario Barbatti. Nonadiabatic dynamics with trajectory surface hopping method. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1(4):620–633, jul 2011.
- [88] Trygve Helgaker, Poul Jørgensen, and Jeppe Olsen. *Molecular Electronic-Structure Theory*. John Wiley & Sons, Ltd, Chichester, UK, first edition, aug 2000.
- [89] Frank Jensen. *Introduction to Computational Chemistry*. John Wiley & Sons, Ltd, Chichester, UK, second edition, 2007.
- [90] Linjun Wang, Alexey Akimov, and Oleg V. Prezhdo. Recent Progress in Surface Hopping: 2011–2015. *J. Phys. Chem. Lett.*, 7(11):2100–2112, jun 2016.
- [91] Pavlo O. Dral, Mario Barbatti, and Walter Thiel. Nonadiabatic Excited-State Dynamics with Machine Learning. *J. Phys. Chem. Lett.*, 9(19):5660–5663, 2018.
- [92] Julia Westermayr, Michael Gastegger, Maximilian F. S. J. Menger, Sebastian Mai, Leticia González, and Philipp Marquetand. Machine learning enables long time scale molecular photodynamics simulations. *Chem. Sci.*, 2019.
- [93] CPMD. <https://www.cpmc.org/>, 2020. Copyright IBM Corp 1990-2019, Copyright MPI für Festkörperforschung Stuttgart 1997-2001.
- [94] John C. Tully and Richard K. Pkeston. Trajectory surface hopping approach to nonadiabatic molecular collisions: The reaction of H⁺ with D₂. *J. Chem. Phys.*, 55(2):562–572, 1971.
- [95] John C. Tully. Molecular dynamics with electronic transitions. *J. Chem. Phys.*, 93(2):1061–1071, jul 1990.

Bibliography

- [96] Mario Barbatti and Rachel Crespo-Otero. Surface hopping dynamics with DFT excited states. In N. Ferré, M. Filatov, and M. Huix-Rotllant, editors, *Density-Functional Methods Excit. States*, pages 415–44. Springer, Cham, 2015.
- [97] Enrico Tapavicza, Ivano Tavernelli, and Ursula Rothlisberger. Trajectory surface hopping within linear response time-dependent density-functional theory. *Phys. Rev. Lett.*, 98(2):1–4, 2007.
- [98] Mark E. Casida. Time-Dependent Density Functional Response Theory for Molecules. In Delano P. Chong, editor, *Recent Adv. Density Funct. Methods, Part I*, page 155. World Scientific, Singapore, 1995.
- [99] Sharon Hammes-Schiffer and John C. Tully. Proton transfer in solution: Molecular dynamics with quantum transitions. *J. Chem. Phys.*, 101(6):4657–4667, sep 1994.
- [100] Garth A. Jones, Barry K. Carpenter, and Michael N. Paddon-Row. Application of trajectory surface hopping to the study of intramolecular electron transfer in polyatomic organic systems. *J. Am. Chem. Soc.*, 120(22):5499–5508, 1998.
- [101] Hiroki Nakamura. *Nonadiabatic Transition*. World Scientific, Singapore, 2002.
- [102] Ivano Tavernelli, Enrico Tapavicza, and Ursula Rothlisberger. Nonadiabatic coupling vectors within linear response time-dependent density functional theory. *J. Chem. Phys.*, 130(12):124107, mar 2009.
- [103] Elisa Liberatore, Rocco Meli, and Ursula Rothlisberger. A Versatile Multiple Time Step Scheme for Efficient ab Initio Molecular Dynamics Simulations. *J. Chem. Theory Comput.*, 14(6):2834–2842, jun 2018.
- [104] H. F. Trotter. On the product of semi-groups of operators. *Proc. Am. Math. Soc.*, 10(4):545, apr 1959.
- [105] Ryan P. Steele. Communication: Multiple-timestep ab initio molecular dynamics with electron correlation. *J. Chem. Phys.*, 139(1):011102, 2013.
- [106] Mario Barbatti, Adélia J. A. Aquino, and Hans Lischka. Ultrafast two-step process in the non-adiabatic relaxation of the CH₂NH₂⁺ molecule. *Mol. Phys.*, 104(5-7):1053–1060, mar 2006.
- [107] Ivano Tavernelli, Enrico Tapavicza, and Ursula Rothlisberger. Non-adiabatic dynamics using time-dependent density functional theory: Assessing the coupling strengths. *J. Mol. Struct. THEOCHEM*, 914(1-3):22–29, nov 2009.
- [108] Sean A. Fischer, Bradley F. Habenicht, Angeline B. Madrid, Walter R. Duncan, and Oleg V. Prezhdo. Regarding the validity of the time-dependent Kohn–Sham approach for electron-nuclear dynamics via trajectory surface hopping. *J. Chem. Phys.*, 134(2):024102, jan 2011.

- [109] P. Minary, M. E. Tuckerman, and G. J. Martyna. Long Time Molecular Dynamics for Enhanced Conformational Sampling in Biomolecular Systems. *Phys. Rev. Lett.*, 93(15):150201, oct 2004.
- [110] Roberto Car and Michele Parrinello. Unified Approach for Molecular Dynamics and Density-Functional Theory. *Physical Review Letters*, 55(22):2471–2474, November 1985. Publisher: American Physical Society.
- [111] B. Kirchner, P. J. di Dio, and J. Hutter. *Real-world predictions from ab initio molecular dynamics simulations*, pages 109–153. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [112] Tamar Schlick, Margaret Mandziuk, Robert D. Skeel, and K. Srinivas. Nonlinear resonance artifacts in molecular dynamics simulations. *The Journal of Computational Physics*, 140(1):1 – 29, 1998.
- [113] William Streett, Dominic Tildesley, and G. Saville. Multiple time-step methods in molecular dynamics. *Molecular Physics*, 35:639, 1978.
- [114] Mark E. Tuckerman, Bruce J. Berne, and Glenn J. Martyna. Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics*, 97:1990, 1992.
- [115] Mark E. Tuckerman and Bruce J. Berne. Molecular dynamics in systems with multiple time scales: Systems with stiff and soft degrees of freedom and with short and long range forces. *The Journal of Chemical Physics*, 95:8362, 1991.
- [116] Bernd Hartke, Douglas A. Gibson, and Emily A. Carter. Multiple time scale hartree–fock molecular dynamics. *International Journal of Quantum Chemistry*, 45(1):59–70, 1993.
- [117] Mark E. Tuckerman and Michele Parrinello. Integrating the car–parrinello equations. ii. multiple time scale techniques. *The Journal of Chemical Physics*, 101(2):1316–1329, 1994.
- [118] Nathan Luehr, Thomas E. Markland, and Todd J. Martínez. Multiple time step integrators in ab initio molecular dynamics. *The Journal of Chemical Physics*, 140:084116, 2014.
- [119] Ryan P. Steele. Multiple-timestep ab initio molecular dynamics using an atomic basis set partitioning. *The Journal of Physical Chemistry A*, 119:12119, 2015.
- [120] Shervin Fatehi and Ryan P. Steele. Multiple-time step ab initio molecular dynamics based on two-electron integral screening. *The Journal of Chemical Theory and Computation*, 11:884, 2015.
- [121] Ryan P. Steele. Communication: Multiple-timestep ab initio molecular dynamics with electron correlation. *The Journal of Chemical Physics*, 139:011102, 2013.

Bibliography

- [122] Manuel Guidon, Florian Schiffmann, Jürg. Hutter, and Joost VandeVondele. Ab initio molecular dynamics using hybrid density functionals. *The Journal of Chemical Physics*, 128:214104, 2008.
- [123] Pablo Baudin, François Mouvet, and Ursula Rothlisberger. A multiple time step algorithm for trajectory surface hopping simulations. *The Journal of Chemical Physics*, 156(3):034107, January 2022.
- [124] Bingqing Cheng, Edgar A. Engel, Jörg Behler, Christoph Dellago, and Michele Ceriotti. Ab initio thermodynamics of liquid and solid water. *Proceedings of the National Academy of Sciences*, 116(4):1110–1115, 2019.
- [125] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98:146401, 2007.
- [126] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gabor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104:136403, 2010.
- [127] Albert P. Bartók, Risi Kondor, and Gabor Csányi. On representing chemical environments. *Physical Review B*, 87:184115, 2013.
- [128] Sergei Manzhos, Xiaogang Wang, Richard Dawes, and Tucker Carrington. A nested molecule-independent neural network approach for high-quality potential fits. *The Journal of Physical Chemistry A*, 110(16):5295–5304, 2006.
- [129] Venkatesh Botu and Rampi Ramprasad. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *International Journal of Quantum Chemistry*, 115(16):1074–1083, 2015.
- [130] Matthias Rupp, Raghunathan Ramakrishnan, and O. Anatole von Lilienfeld. Machine learning for quantum mechanical properties of atoms in molecules. *The Journal of Physical Chemistry Letters*, 6(16):3309–3313, 2015.
- [131] Huziel E. Saucedo, Stefan Chmiela, Igor Poltavsky, Klaus-Robert Müller, and Alexandre Tkatchenko. Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces. *The Journal of Chemical Physics*, 150(11):114102, 2019.
- [132] Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Saucedo, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5), 2017.
- [133] Andrea Grisafi, David M. Wilkins, Gabor Csányi, and Michele Ceriotti. Symmetry-adapted machine learning for tensorial properties of atomistic systems. *Physical Review Letters*, 120:036002, 2018.

- [134] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The δ -machine learning approach. *The Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015.
- [135] Nicholas J. Browning, Raghunathan Ramakrishnan, O. Anatole von Lilienfeld, and Ursula Roethlisberger. Genetic optimization of training sets for improved machine learning models of molecular properties. *The Journal of Physical Chemistry Letters*, 8(7):1351–1359, 2017.
- [136] Anders S. Christensen, Felix A. Faber, and O. Anatole von Lilienfeld. Operators in quantum machine learning: Response properties in chemical space. *The Journal of Chemical Physics*, 150(6):064105, 2019.
- [137] Richard F. W. Bader. Atoms in molecules. *Accounts of Chemical Research*, 18(1):9–15, 1985.
- [138] Felix A. Faber, Anders S. Christensen, Bing Huang, and O. Anatole von Lilienfeld. Alchemical and structural distribution based representation for universal quantum machine learning. *The Journal of Chemical Physics*, 148(24):241717, 2018.
- [139] Bing Huang and O. Anatole von Lilienfeld. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nature Chemistry*, 12(10):945–951, October 2020.
- [140] Anders S. Christensen, Lars A. Bratholm, Felix A. Faber, and O. Anatole von Lilienfeld. FCHL revisited: Faster and more accurate quantum machine learning. *The Journal of Chemical Physics*, 152(4):044107, January 2020.
- [141] Albert P. Bartók, Michael J. Gillan, Frederick R. Manby, and Gabor Csányi. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Physical Review B*, 88:054104, 2013.
- [142] Alexander Denzel and Johannes Kästner. Gaussian process regression for geometry optimization. *The Journal of Chemical Physics*, 148(9):094114, 2018.
- [143] N. Troullier and José L. Martins. Efficient pseudopotentials for plane-wave calculations. *Physical Review B*, 43:1993–2006, 1991.
- [144] David M. Ceperley and Berni J. Alder. Ground state of the electron gas by a stochastic method. *Physical Review Letters*, 45:566–569, 1980.
- [145] Albert P. Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gabor Csányi, and Michele Ceriotti. Machine learning unifies the modeling of materials and molecules. *Science Advances*, 3(12), 2017.
- [146] Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences*, 108(32):13023–13028, 2011.

Bibliography

- [147] I-Chun. Lin, Ari P. Seitsonen, Ivano Tavernelli, and Ursula Rothlisberger. Structure and dynamics of liquid water from ab initio molecular dynamics - comparison of BLYP, PBE, and revPBE density functionals with and without van der waals corrections. *The Journal of Chemical Theory and Computation*, 8(10):3902–3910, 2012.
- [148] Teodora. Todorova, Ari P. Seitsonen, Jürg Hutter, I.-Feng. W. Kuo, and Christopher J. Mundy. Molecular dynamics simulation of liquid water: hybrid density functionals. *The Journal of Physical Chemistry B*, 110(8):3685–3691, 2006.
- [149] Michael J. Gillan, Dario Alfè, and Angelos Michaelides. Perspective: How good is DFT for water? *The Journal of Chemical Physics*, 144(13):130901, 2016.
- [150] Tamar Schlick, Margaret Mandziuk, Robert D. Skeel, and K. Srinivas. Nonlinear Resonance Artifacts in Molecular Dynamics Simulations. *Journal of Computational Physics*, 140(1):1–29, February 1998.
- [151] Qun Ma, Jesús A. Izaguirre, and Robert D. Skeel. Verlet-I/R-RESPA/Impulse is Limited by Nonlinear Instabilities. *SIAM Journal on Scientific Computing*, 24(6):1951–1973, January 2003.
- [152] Jeffrey J. Biesiadecki and Robert D. Skeel. Dangers of Multiple Time Step Methods. *Journal of Computational Physics*, 109(2):318–328, December 1993.
- [153] Daniel T. Margul and Mark E. Tuckerman. A stochastic, resonance-free multiple time-step algorithm for polarizable models that permits very large time steps. *The Journal of Chemical Theory and Computation*, 12(5):2170–2180, 2016.
- [154] Charles R. A. Abreu and M. E. Tuckerman. Multiple timescale molecular dynamics with very large time steps: avoidance of resonances. *The European Physical Journal B*, 94(11):231, November 2021.
- [155] Joseph A. Morrone, Thomas E. Markland, Michele Ceriotti, and Bruce J. Berne. Efficient multiple time scale molecular dynamics: Using colored noise thermostats to stabilize resonances. *The Journal of Chemical Physics*, 134(1):014103, 2011.
- [156] John P. Perdew, Matthias Ernzerhof, and Kieron Burke. Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics*, 105(22):9982–9985, 1996.
- [157] Carlo Adamo and Vincenzo Barone. Toward reliable density functional methods without adjustable parameters: The pbe0 model. *The Journal of Chemical Physics*, 110(13):6158–6170, 1999.
- [158] Arieh Warshel and Michael Levitt. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, 103(2):227–249, May 1976.

- [159] Dimitrios G. Liakos and Frank Neese. Is It Possible To Obtain Coupled Cluster Quality Energies at near Density Functional Theory Cost? Domain-Based Local Pair Natural Orbital Coupled Cluster vs Modern Density Functional Theory. *Journal of Chemical Theory and Computation*, 11(9):4054–4063, September 2015.
- [160] Stefan Seritan, Christoph Bannwarth, B. Scott Fales, Edward G. Hohenstein, Sara I. L. Kokkila-Schumacher, Nathan Luehr, James W. Snyder, Jr., Chenchen Song, Alexey V. Titov, Ivan S. Ufimtsev, and Todd J. Martínez. TeraChem: Accelerating electronic structure and ab initio molecular dynamics with graphical processing units. *The Journal of Chemical Physics*, 152(22):224110, June 2020.
- [161] P. Bernát Szabó, József Csóka, Mihály Kállay, and Péter R. Nagy. Linear-Scaling Open-Shell MP2 Approach: Algorithm, Benchmarks, and Large-Scale Applications. *Journal of Chemical Theory and Computation*, 17(5):2886–2905, May 2021.
- [162] Jógvan Magnus Haugaard Olsen, Viacheslav Bolnykh, Simone Meloni, Emiliano Ippoliti, Martin P. Bircher, Paolo Carloni, and Ursula Rothlisberger. MiMiC: A Novel Framework for Multiscale Modeling in Computational Chemistry. *Journal of Chemical Theory and Computation*, 15(6):3810–3823, 2019.
- [163] Viacheslav Bolnykh, Jógvan Magnus Haugaard Olsen, Simone Meloni, Martin P. Bircher, Emiliano Ippoliti, Paolo Carloni, and Ursula Rothlisberger. Extreme Scalability of DFT-Based QM/MM MD Simulations Using MiMiC. *Journal of Chemical Theory and Computation*, 15(10):5601–5613, 2019.
- [164] Mark E. Tuckerman, Bruce J. Berne, and Glenn J. Martyna. Reversible multiple time scale molecular dynamics. *The Journal of Chemical Physics*, 97(3):1990–2001, 1992.
- [165] Sagarmoy Mandal, Vaishali Thakkur, and Nisanth N. Nair. Achieving an Order of Magnitude Speedup in Hybrid-Functional- and Plane-Wave-Based Ab Initio Molecular Dynamics: Applications to Proton-Transfer Reactions in Enzymes and in Solution. *Journal of Chemical Theory and Computation*, 17(4):2244–2255, April 2021.
- [166] Mark Abraham, Andrey Alekseenko, Cathrine Bergh, Christian Blau, Eliane Briand, Mahesh Doijade, Stefan Fleischmann, Vytautas Gapsys, Gaurav Garg, Sergey Gorelov, Gilles Gouaillardet, Alan Gray, M. Eric Irrgang, Farzaneh Jalalypour, Joe Jordan, Christoph Junghans, Prashanth Kanduri, Sebastian Keller, Carsten Kutzner, Justin A. Lemkul, Magnus Lundborg, Pascal Merz, Vedran Miletić, Dmitry Morozov, Szilárd Páll, Roland Schulz, Michael Shirts, Alexey Shvetsov, Bálint Soproni, David van der Spoel, Philip Turner, Carsten Uphoff, Alessandra Villa, Sebastian Wingbermühle, Artem Zhmurov, Paul Bauer, Berk Hess, and Erik Lindahl. GROMACS 2023.1 Manual. April 2023. Publisher: Zenodo.
- [167] Mats Svensson, Stéphane Humbel, Robert D. J. Froese, Toshiaki Matsubara, Stefan Sieber, and Keiji Morokuma. ONIOM: A Multilayered Integrated MO + MM Method for Geometry Optimizations and Single Point Energy Predictions. A Test for Diels-Alder

Bibliography

- Reactions and $\text{Pt}(\text{P}(\text{t-Bu})_3)_2 + \text{H}_2$ Oxidative Addition. *The Journal of Physical Chemistry*, 100(50):19357–19363, January 1996. Publisher: American Chemical Society.
- [168] U. Chandra Singh and Peter A. Kollman. A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the $\text{CH}_3\text{Cl} + \text{Cl}^-$ exchange reaction and gas phase protonation of polyethers. *Journal of Computational Chemistry*, 7(6):718–730, 1986.
- [169] Martin J. Field, Paul A. Bash, and Martin Karplus. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *Journal of Computational Chemistry*, 11(6):700–733, 1990.
- [170] Nathalie Reuter, Annick Dejaegere, Bernard Maigret, and Martin Karplus. Frontier Bonds in QM/MM Methods: A Comparison of Different Approaches. *The Journal of Physical Chemistry A*, 104(8):1720–1735, March 2000. Publisher: American Chemical Society.
- [171] Hans Martin Senn and Walter Thiel. QM/MM Methods for Biomolecular Systems. *Angewandte Chemie International Edition*, 48(7):1198–1229, 2009.
- [172] U. Ryde. Chapter Six - QM/MM Calculations on Proteins. In Gregory A. Voth, editor, *Methods in Enzymology*, volume 577 of *Computational Approaches for Studying Enzyme Mechanism Part A*, pages 119–158. Academic Press, January 2016.
- [173] Vincent Théry, Daniel Rinaldi, Jean-Louis Rivail, Bernard Maigret, and György G. Ferenczy. Quantum mechanical computations on very large molecular systems: The local self-consistent field method. *Journal of Computational Chemistry*, 15(3):269–282, 1994.
- [174] Jiali Gao, Patricia Amara, Cristobal Alhambra, and Martin J. Field. A Generalized Hybrid Orbital (GHO) Method for the Treatment of Boundary Atoms in Combined QM/MM Calculations. *The Journal of Physical Chemistry A*, 102(24):4714–4721, June 1998. Publisher: American Chemical Society.
- [175] R. B. Murphy, D. M. Philipp, and R. A. Friesner. A mixed quantum mechanics/molecular mechanics (QM/MM) method for large-scale modeling of chemistry in protein environments. *Journal of Computational Chemistry*, 21(16):1442–1457, 2000.
- [176] Alessandro Laio, Joost VandeVondele, and Ursula Rothlisberger. A Hamiltonian electrostatic coupling scheme for hybrid Car–Parrinello molecular dynamics simulations. *The Journal of Chemical Physics*, 116(16):6941–6947, April 2002.
- [177] Alessandro Laio, Joost VandeVondele, and Ursula Rothlisberger. D-RESP: Dynamically Generated Electrostatic Potential Derived Charges from Quantum Mechanics/Molecular Mechanics Simulations. *The Journal of Physical Chemistry B*, 106(29):7300–7307, July 2002.

- [178] Félix Musil, Michael J. Willatt, Mikhail A. Langovoy, and Michele Ceriotti. Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *Journal of Chemical Theory and Computation*, 15(2):906–915, 2019.
- [179] Justin S. Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E. Roitberg. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics*, 148(24):241733, May 2018.
- [180] Andrew A. Peterson, Rune Christensen, and Alireza Khorshidi. Addressing uncertainty in atomistic machine learning. *Physical Chemistry Chemical Physics*, 19(18):10978–10985, May 2017. Publisher: The Royal Society of Chemistry.
- [181] Jörg Behler. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angewandte Chemie International Edition*, 56(42):12828–12840, 2017.
- [182] Gaël Guennebaud, Benoît Jacob, et al. Eigen v3. <http://eigen.tuxfamily.org>, 2010.
- [183] Esra Bozkurt. *Reprogramming the B1 Domain of Streptococcal Protein G (GB1): A Joint Theoretical and Experimental Investigation*. PhD thesis, EPFL, Lausanne, 2018.
- [184] Esra Bozkurt, Marta A. S. Perez, Ruud Hovius, Nicholas J. Browning, and Ursula Rothlisberger. Genetic Algorithm Based Design and Experimental Characterization of a Highly Thermostable Metalloprotein. *Journal of the American Chemical Society*, 140(13):4517–4521, April 2018. Publisher: American Chemical Society.
- [185] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, July 1983.
- [186] Bharath Raghavan, Florian K. Schackert, Andrea Levy, Sophia K. Johnson, Emiliano Ippoliti, Davide Mandelli, Jógvan Magnus Haugaard Olsen, Ursula Rothlisberger, and Paolo Carloni. MiMiCPy: An Efficient Toolkit for MiMiC-Based QM/MM Simulations. *Journal of Chemical Information and Modeling*, 63(5):1406–1412, March 2023. Publisher: American Chemical Society.
- [187] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, January 1988.
- [188] 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. *The Journal of Chemical Physics*, 110(6):2810–2821, February 1999.
- [189] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, August 2015. Publisher: American Chemical Society.

Bibliography

- [190] Charles R. A. Abreu and Mark E. Tuckerman. Hamiltonian based resonance-free approach for enabling very large time steps in multiple time-scale molecular dynamics. *Molecular Physics*, 119(19-20):e1923848, October 2021.

François MOUVET

Doctoral research assistant in physics

👤 Swiss, born 28 January 1992

🏠 Avenue des Reneveyres 5, 1110 Morges, Vaud, Switzerland

☎ +41 76 443 08 33 | ✉ francois.mouvet@proton.me | **in** francois-mouvet



Education

- OCT 2017 - JULY 2023 PhD in PHYSICS, **EPFL**, Lausanne
Computational study of organometallic drug candidates
Development of new HPC methods to accelerate simulations in computational biochemistry using machine learning
- JAN 2016 - SEPT 2016 Master thesis in PHYSICS, **University of Edinburgh**, Scotland
- FEB 2014 - SEPT 2016 MSc in PHYSICS, **EPFL**, Lausanne
Emphasis on statistical physics and biophysics
Minor in computational neurosciences (neural networks and reinforcement learning)
- SEPT 2010 - FEB 2014 BSc in PHYSICS, **EPFL**, Lausanne

Skills

Software

Languages C++, C, Fortran, Python, MATLAB, Bash, awk, SQL
HPC OpenMP, MPI, SLURM
Development tools Vim, VSCode, Git, CMake
Data analysis NumPy, Numba, SciPy, Pandas, Eigen, Matplotlib, Seaborn
Operating systems GNU/Linux, macOS, Windows
Comp. Chem. CPMD, GROMACS, MiMiC, AMBER, Tinker, VMD, QML, LAMMPS, Gaussian

Scientific

Drug discovery Simulating protein-drug complexes using MM or QM/MM molecular dynamics
Free energy estimations (thermodynamic integration, metadynamics, PBSA)
Geometry and force field optimization of new compounds
Collaborate with chemists and biologists for preclinical studies

Method development Design new methods for quantum molecular dynamics (CP, DFT).
Strong experience in machine learning methods in computational chemistry
Building modules for large software meant to be used on supercomputers
Working in an international collaborative team

Other Strong knowledge in statistical physics
Knowledge in molecular biology, genetics and bioinformatics
Experience in collaborations with diverse backgrounds

Languages

FRENCH: First language GERMAN: Conversational (B1)
ENGLISH: Fluent (C1/C2) ITALIAN: Conversational (B1)

Work Experience

OCT 2017 - present Doctoral research assistant, **EPFL**, Switzerland
Laboratory of Computational Chemistry and Biochemistry (Prof. Röthlisberger)
Title: *Accelerating simulations of bioactive transition metal compounds using time scale decomposition and artificial intelligence.*

We started by studying a drug candidate against schistosomiasis through classical molecular dynamics. We then developed new methods using machine learning to speed up the simulation of biological systems at the quantum level. We implemented a parallel HPC ML application from scratch (C++) and implemented state-of-art dynamics integration schemes in the IBM software package CPMD (Fortran). The communication between software was implemented in the high performance library MiMiC (C++ and Fortran mix). All data analysis was performed in Python.

OCT 2016 - MAR 2017 Research Internship, **EPFL**, Switzerland
Laboratory of Computational and Systems Biology (Prof. Naef)
Analysis of experimental data from state of the art genomic experiments (Hi-C, ChIP-seq). We created statistical models to extract new insights on experimental data. The goal was to identify possible relations between the 3D structure of chromatin and circadian rhythms.

JAN 2016 - SEPT 2016 Master thesis in physics, **University of Edinburgh**, Scotland
Institute for Condensed Matter and Complex Systems (Prof. Marenduzzo)
Title: *Coarse-grained simulations of chromosome organisation and rearrangement.*
Study of a computational model reproducing the mesoscopic 3D configurations of chromatin using molecular dynamics. This model integrated several 1D genomic data to create a fitting free model able to accurately reproducing Hi-C experiments. We extended this method to model post-translational modifications.

OCT 2011 - OCT 2021 Teaching Assistant, **EPFL**, Switzerland
Received the 2021 SCGC **Teaching Excellence Award**.
Alongside my studies, I supervised the exercises of lectures *General Physics I-IV* (mechanics, thermodynamics, electromagnetism, quantum physics) both in English and in French, destined to students from 1st and 2nd year of most faculties.
During the PhD, I created and supervised exercises, gave lectures and graded exams for the courses *Introduction to Electronic Structure Methods* and *Molecular Dynamics and Monte Carlo* of Prof. Röthlisberger. I also supervised a student for a semester project: *Machine learning models towards assessing hybrid functionals accuracy on water properties*.





Interests and Activities

SPORTS Certified PADI Rescue Diver.
Won the Swiss Rowing Championship in eight in 2008.
Casually practice kayaking, volleyball, ski and snowboard.

TRAVEL Humanitarian trip to Tanguiga, Burkina Faso (Nouvelle planète, 2010).
Study trip in Japan, visiting research institutes in Osaka, Tokyo and Takayama (2014).
Many other trips to explore cultures, history and landscapes.

OTHERS Tabletop RPG, playing guitar, partner dances.

Publications

- 2023 Multiple Time Step *ab Initio* Molecular Dynamics With Machine Learning Trained On-The-Fly
F. Mouvet, U. Rothlisberger
In preparation
- 2023 Machine Learning Enhanced Multiple Time Step Ab Initio Molecular Dynamics
F. Mouvet, N. J. Browning, P. Baudin, E. Liberatore, U. Rothlisberger
In preparation
- 2022 Recent Advances in First-Principles Based Molecular Dynamics
F. Mouvet, J. Villard, V. Bolnykh, U. Rothlisberger
Accounts of Chemical Research 
- 2022 A multiple time step algorithm for trajectory surface hopping simulations
P. Baudin, **F. Mouvet**, U. Rothlisberger
Journal of Chemical Physics 
- 2020 Multidisciplinary Preclinical Investigations on Three Oxamniquine Analogues as New Drug Candidates for Schistosomiasis
V. Buchter, Y. C. Ong, **F. Mouvet**, A. Ladaycia, E. Lepeltier, U. Rothlisberger, J. Keiser, G. Gasser
Chemistry - A European Journal 
- 2017 Ephemeral Protein Binding to DNA Shapes Stable Nuclear Bodies and Chromatin Domains
C. A. Brackley, B. Liebchen, D. Michieletto, **F. Mouvet**, P. R. Cook and D. Marenduzzo
Biophysical Journal 
- 2016 Simulating topological domains in human chromosomes with a fitting-free model
C. A. Brackley, D. Michieletto, **F. Mouvet**, J. Johnson, S. Kelly, P. R. Cook, D. Marenduzzo
Nucleus 