

TOWARDS A MECHANISTIC UNDERSTANDING OF KINASE-CATALYSED N-PHOSPHORYLATIONS USING MULTISCALE SIMULATIONS

Zur Erlangung des akademischen Grades einer
DOKTORIN DER NATURWISSENSCHAFTEN
(Dr. rer. nat.)

von der KIT-Fakultät für Chemie und Biowissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte
DISSERTATION

von
M.Sc. Lena Eichinger
aus Worms

1. Referent: Priv.-Doz. Dr. Tomáš Kubař
2. Referent: Prof. Dr. Wolfgang Wenzel
Tag der mündlichen Prüfung: 20.04.2026

Abstract

Phosphorylation reactions are fundamental in biological systems. Although rarer than phosphorylation of oxygen atoms, nitrogen phosphorylations play important biological roles, including histidine autophosphorylation in signal transduction, and reversible phosphorylation of creatine to phosphocreatine, a rapidly mobilisable energy reserve. Despite their relevance, the detailed understanding of the molecular mechanisms of kinase-catalysed N-phosphorylation reactions remains incomplete.

Hybrid QM/MM methods enable the description of enzymatic processes at atomic resolution by employing quantum mechanical (QM) methods for the active site while treating the surrounding protein environment classically. However, biological processes often occur on micro- to millisecond timescales, whereas QM simulations are typically restricted to the nanosecond regime. Enhanced-sampling techniques are therefore employed to improve the exploration of the free-energy landscape. While semi-empirical methods like density-functional tight-binding (DFTB) provide the necessary computational efficiency, they do not always offer the accuracy needed.

To improve the accuracy of the QM description, a Δ -machine learning (Δ ML) correction based on high-dimensional neural network potentials (HDNNPs) is applied to DFTB, improving its accuracy towards density functional theory (DFT) while retaining the computational efficiency of DFTB. The Δ ML model is trained and evaluated in vacuum for a diverse set of N-phosphorylation model reactions. The resulting models substantially improve the accuracy of DFTB energies and forces. Applications to QM/MM metadynamics simulations reveal transferability challenges between vacuum and condensed environments, particularly for electrostatic interactions and Mg^{2+} coordination.

Building on these observations, the approach was applied to enzymatic systems. For histidine kinase, improved transferability was achieved by including the complete Mg^{2+} coordination sphere in the Δ ML model and by mitigating overpolarisation at the QM/MM interface. Enhanced sampling simulations provided both methodological and mechanistic insight into the autophosphorylation reaction, particularly regarding the influence of Mg^{2+} coordination and artificial restraints.

In addition, QM/MM metadynamics simulations were performed for mitochondrial creatine kinase to investigate conformational loop dynamics and the catalytic steps of proton and phosphoryl transfer. The simulations emphasise the importance of selecting flexible collective variables and highlight the influence of the electrostatic environment on the reaction pathway and product stability. Although the calculated barriers remain higher than experimental estimates, the simulations provide qualitative mechanistic insight into the catalytic cycle and identify important factors that influence the reaction energetics.

Overall, this work demonstrates how multiscale simulations combining QM/MM methods, enhanced sampling, and Δ ML corrections can provide mechanistic insight into kinase-catalysed N-phosphorylation reactions. Furthermore, the presented framework establishes a foundation for future Δ ML applications to large and biologically relevant enzymatic systems.

Zusammenfassung

Phosphorylierungsreaktionen sind zentrale Prozesse in biologischen Systemen. Obwohl seltener als Phosphorylierungen an Sauerstoffatomen, spielen Phosphorylierungen an Stickstoffzentren eine wichtige biologische Rolle. Beispiele hierfür sind die Autophosphorylierung von Histidin in Signaltransduktionskaskaden sowie die reversible Phosphorylierung von Kreatin zu Phosphokreatin, welches als kurzfristig verfügbare Energiereserve dient. Trotz ihrer biologischen Bedeutung sind die molekularen Mechanismen kinasekatalysierter N-Phosphorylierungen bislang nicht vollständig aufgeklärt.

Hybride QM/MM-Methoden erlauben die Beschreibung enzymatischer Reaktionen mit atomarer Auflösung. Hierzu wird das aktive Zentrum quantenmechanisch (QM) behandelt, während das umgebende Protein klassisch beschrieben wird. Viele biologische Prozesse finden jedoch auf Zeitskalen von Mikro- bis Millisekunden statt, während QM-Simulationen in der Regel auf Nanosekunden beschränkt sind. Daher werden Enhanced-Sampling-Methoden eingesetzt, um die Erkundung der freien Energielandschaft zu verbessern. Semiempirische Methoden wie die Density-Functional-Tight-Binding-Methode (DFTB) bieten zwar die erforderliche Recheneffizienz, liefern jedoch nicht immer die notwendige Genauigkeit.

Zur Verbesserung der Genauigkeit wird eine Δ -Machine-Learning (Δ ML) Korrektur auf Basis hochdimensionaler neuronaler Netzwerkpotenziale (HDNNPs) auf DFTB angewendet. Dadurch kann die Genauigkeit von DFTB in Richtung der Dichtefunktionaltheorie (DFT) erhöht werden, ohne dass die hohe rechnerische Effizienz beeinträchtigt wird. Das Δ ML-Modell wird zunächst im Vakuum mit einem umfangreichen Datensatz von N-Phosphorylierungs-Modellreaktionen trainiert und getestet. Die resultierenden Modelle verbessern die mittels DFTB berechneten Energien und Kräfte signifikant. Anwendungen in QM/MM-Metadynamik-Simulationen zeigen jedoch Herausforderungen bei der Übertragbarkeit zwischen Vakuum und kondensierten Umgebungen, insbesondere im Zusammenhang mit elektrostatischen Wechselwirkungen und der Koordination von Magnesium.

Aufbauend auf diesen Ergebnissen wurde der Ansatz auf enzymatische Systeme übertragen. Für Histidinkinase konnte die Übertragbarkeit verbessert werden, indem die vollständige Mg^{2+} -Koordinationssphäre in das Δ ML-Modell aufgenommen und die Überpolarisierung an der QM/MM-Grenzfläche reduziert wurde. Enhanced-Sampling-Simulationen lieferten sowohl methodische als auch mechanistische Einblicke in die Autophosphorylierungsreaktion, insbesondere hinsichtlich des Einflusses der Mg^{2+} -Koordination und kritischer Simulationsparameter.

Darüber hinaus wurden QM/MM-Metadynamik-Simulationen an einer mitochondrialen Kreatinkinase durchgeführt, um sowohl Konformationsänderungen als auch die katalytischen Schritte des Protonen- und Phosphoryltransfers zu untersuchen. Die Simulationen zeigen die Bedeutung flexibel gewählter kollektiver Variablen und verdeutlichen den Einfluss der elektrostatischen Umgebung auf Reaktionspfad und Produktstabilität. Obwohl die berechneten Barrieren über experimentellen

Abschätzungen liegen, liefern die Simulationen qualitative mechanistische Einblicke in den katalytischen Zyklus und identifizieren wichtige Faktoren, die die Reaktionsenergetik beeinflussen.

Insgesamt zeigt diese Arbeit, dass Multiskalen-Simulationen, welche QM/MM-Methoden, Enhanced Sampling und Δ ML-Korrekturen kombinieren, wertvolle mechanistische Einblicke in kinasekatalysierte N-Phosphorylierungsreaktionen ermöglichen. Gleichzeitig schafft der entwickelte Ansatz eine Grundlage für zukünftige Δ ML-Anwendungen auf große und biologisch relevante enzymatische Systeme.

Acknowledgements

First of all, I would like to thank Prof. Dr. Marcus Elstner for giving me the opportunity to pursue my PhD in his group. I am grateful for his scientific guidance, as well as for encouraging me to develop my presentation skills, which has greatly contributed to my personal and professional development. I would also like to sincerely thank PD Dr. Tomáš Kubař for always having an open ear, for his valuable scientific advice, and especially for his continuous support and encouragement throughout my time in this group. Furthermore, I would like to thank Sabine Holthoff for her help with all organisational matters and university bureaucracy.

I am grateful to the RTG 2450 for the financial support and to all its members for the shared activities, such as retreats and workshops. Furthermore, I would like to thank Prof. Qiang Cui for his supervision during my research stay at Boston University.

I would like to thank Monja Sokolov for being a constant throughout my time in the group – from supervising the MD experiment in the physical chemistry practical in the fourth semester, which first got me interested in simulations, to helping with scripts during my Master’s studies, and the time together in Boston. I am grateful to Mayukh Kansari, who supervised my Master’s thesis, initially got me interested in kinases and phosphorylation reactions, and continued to be available for questions even after leaving the group. Furthermore, I want to thank Van-Quan Vuong for his valuable insights on quantum chemistry and machine learning. A big thank you to Katharina Spies for the many scientific discussions, the countless chats in between that made everyday work more enjoyable, and overall for being a great office mate and friend to me over the last few years. I am grateful to Julian Böser for sharing an office since Campus West, for all the scientific discussions, and for his help whenever I needed it. A special thank you to Lukas Petersen for always helping with all my machine learning questions and for the fun AKK dance classes. Thank you to Eik Henigin for the many water and coffee breaks and the nice conversations that came with them. Thank you to Leonie Pfeffer for being a great student and giving me new perspectives on my research. I also want to thank both her and Maike Vogelbacher for the productive discussions in weekly subgroup meetings and beyond. A big thank you to everyone in the TCB group for the great time over the past years, whether through scientific discussions, help with proofreading this thesis, or through the many shared activities, such as barbecues at Campus West, hiking trips, Mensa lunches, visits to the Karlsruhe fair, and evenings at the pub.

I would also like to thank my friends for their support throughout the years. Lisa, thank you for being my best friend for more than 15 years and for always being there for me whenever I need you. Merten, thank you for all the hours-long phone calls that always brighten my day. Maarten, thank you for being a great roommate during my time in Boston and for many helpful discussions about scientific topics. All of you – Lisa, Merten, Maarten, Patrick, Christoph, and Sara – thank you for the many great times during vacations, our annual visits to the Christmas market, and school reunions.

Burçak and Ange, thank you for helping me survive the practical lab courses before I moved to the safety of theoretical studies, and for staying friends even after our bachelor's and master's studies. Ange, thank you in particular, for always having an open ear during the stressful final weeks of writing and for motivating me to do sports as a necessary and welcome balance.

I would also like to thank my family, especially my parents and my brother Jonas, for their constant support throughout the years.

Finally, my deepest thanks go to my husband. Benni, thank you for your endless love and constant encouragement and support over the last years – whether through the countless rounds of proofreading, providing delicious food to keep my spirits up, always having my back, or simply being there whenever I needed you. I could not have done this without you.

Contents

Abstract	I
Zusammenfassung	III
Acknowledgements	V
List of Abbreviations	XI
I. Introduction	1
1. Introduction	3
II. Theoretical Background	7
2. Computational Chemistry	9
3. Molecular Dynamics Simulations	11
3.1. Force Field	11
3.2. Energy Minimisation	14
3.3. Integration Method	15
3.4. Ensemble	16
3.5. Thermostat	16
3.6. Barostat	17
4. QM/MM Simulations	19
5. Quantum Chemistry Methods	23
5.1. <i>Ab Initio</i> Methods	23
5.1.1. Born–Oppenheimer Approximation	24
5.1.2. Variational Principle	25
5.1.3. Hartree–Fock Theory	25
5.1.4. Density Functional Theory (DFT)	27
5.2. Semi-Empirical Methods	31
5.2.1. Density-Functional Tight-Binding (DFTB)	32
6. Enhanced Sampling	35
6.1. Metadynamics	38
6.1.1. Well-Tempered Metadynamics	39
7. Machine Learning	41

III. Results	45
8. ΔML Model for N-Phosphoryl Transfer Reactions in Gas Phase	47
8.1. Introduction	47
8.1.1. Kinases Catalyse Phosphorylation Reactions	49
8.1.2. Quantum Chemical Description of Phosphorylation	50
8.1.3. Motivation and Aim of the Project	50
8.2. Methods	51
8.2.1. Pilot Simulation	53
8.2.2. Investigated Systems	55
8.3. Results and Discussion	57
8.3.1. Simulation-Based Generation of Training Data	57
8.3.2. DFTB- Δ ML – Training and Cross-Validations	66
8.4. Conclusions	72
9. Transferability of ΔML Models from Vacuum to Water	75
9.1. Introduction	75
9.1.1. Motivation and Aim of the Project	76
9.2. Methods	76
9.3. Results and Discussion	77
9.3.1. Model System for Histidine Kinase	77
9.3.2. Model System for Creatine Kinase	81
9.3.3. Computational Cost	83
9.4. Conclusions	84
10. Autophosphorylation in a <i>trans</i>-Histidine Kinase	85
10.1. Introduction	85
10.1.1. Motivation and Aim of the Project	87
10.2. Methods	88
10.3. Results and Discussion	91
10.3.1. Optimising Δ ML Corrections for Protein QM/MM Simulations	91
10.3.2. DFTB- Δ ML/MM Simulations for Mechanistic Modelling	93
10.3.3. Computational Cost	98
10.4. Conclusions	99
11. Understanding the Catalytic Mechanism of Creatine Kinase	103
11.1. Introduction	103
11.1.1. Motivation and Aim of the Project	105
11.2. Methods	106
11.3. Results and Discussion	109
11.3.1. Conformational Transition	109
11.3.2. Chemical Steps	110
11.3.3. Setting the Results into Context	114
11.3.4. Δ G Estimation of the Complete Process	115
11.4. Conclusions	117

IV. Summary and Outlook	119
12. Summary and Outlook	121
V. Appendix	123
A. Δ ML Model for N-Phosphoryl Transfer Reactions in Gas Phase	125
B. Transferability of Δ ML Models from Vacuum to Water	133
C. Autophosphorylation in a <i>trans</i> -Histidine Kinase	137
D. Understanding the Catalytic Mechanism of Creatine Kinase	141
Bibliography	151
List of Publications	167

List of Abbreviations

2PC	coordination by two ATP phosphoryl groups
3PC	coordination by three ATP phosphoryl groups
ACSF	atom-centred symmetry functions
ADP	adenosine diphosphate
AK	arginine kinase
aMD	accelerated molecular dynamics
Arg	arginine
Asn	asparagine
Asp	aspartate
ADP	adenosine diphosphate
ATP	adenosine triphosphate
CA	catalytic and ATP-binding domain
CK	creatine kinase
CPS	cysteine-proline-serine
cryo-EM	cryogenic electron microscopy
CV	collective variable
Cys	cysteine
DFT	density functional theory
DFTB	density-functional tight-binding
DHp	dimerisation histidine phosphotransfer domain
et al.	and others
FES	free energy surface
GGA	generalised gradient approximation
Glu	glutamate
His	histidine
HDNNP	high-dimensional neural network potential
HOMO	highest occupied molecular orbital
HF	Hartree–Fock
HK	histidine kinase
LCAO	linear combination of atomic orbitals
LDA	local density approximation
LUMO	lowest unoccupied molecular orbital
MD	molecular dynamics
ML	machine learning
MM	molecular mechanics
mCEC	modified centre of excess charge
NMR	nuclear magnetic resonance

NN	neural network
pCr	phosphocreatine
PDB	Protein Data Bank
PES	potential energy surface
pHis	phosphohistidine
PME	Particle Mesh Ewald
PMF	potential of mean force
PSFs	pair symmetry functions
PTM	post-translational protein modification
QM	quantum mechanics
QM/MM	quantum mechanical/molecular mechanical
REMD	replica exchange molecular dynamics
RMSE	root mean square error
RR	response regulator
SCC	self-consistent charge
SCF	self-consistent field
SRP	specific reaction parameterisation
TCS	two-component system
TSAC	transition-state analogue complexes
TST	transition state theory

Part I.

Introduction

1. Introduction

Phosphorylation reactions are among the most fundamental chemical processes in biological systems and play a central role in cellular energy metabolism, signal transduction, and metabolic regulation.¹ In particular, kinases catalyse the transfer of phosphoryl groups from ATP to a wide range of biomolecular substrates, thereby enabling the reversible regulation of protein activity and metabolic pathways.² While most phosphorylation reactions occur on oxygen atoms, phosphorylation at nitrogen centres also plays an important role in biological systems.³ Examples include histidine phosphorylation catalysed by histidine kinases in bacterial two-component signalling systems, and the reversible phosphorylation of creatine by creatine kinases, forming phosphocreatine as a rapidly mobilisable energy reserve (refer to Figure 1.1).^{4,5} Despite their biological importance, the detailed molecular mechanisms underlying kinase-catalysed N-phosphorylation reactions remain incompletely understood.⁶

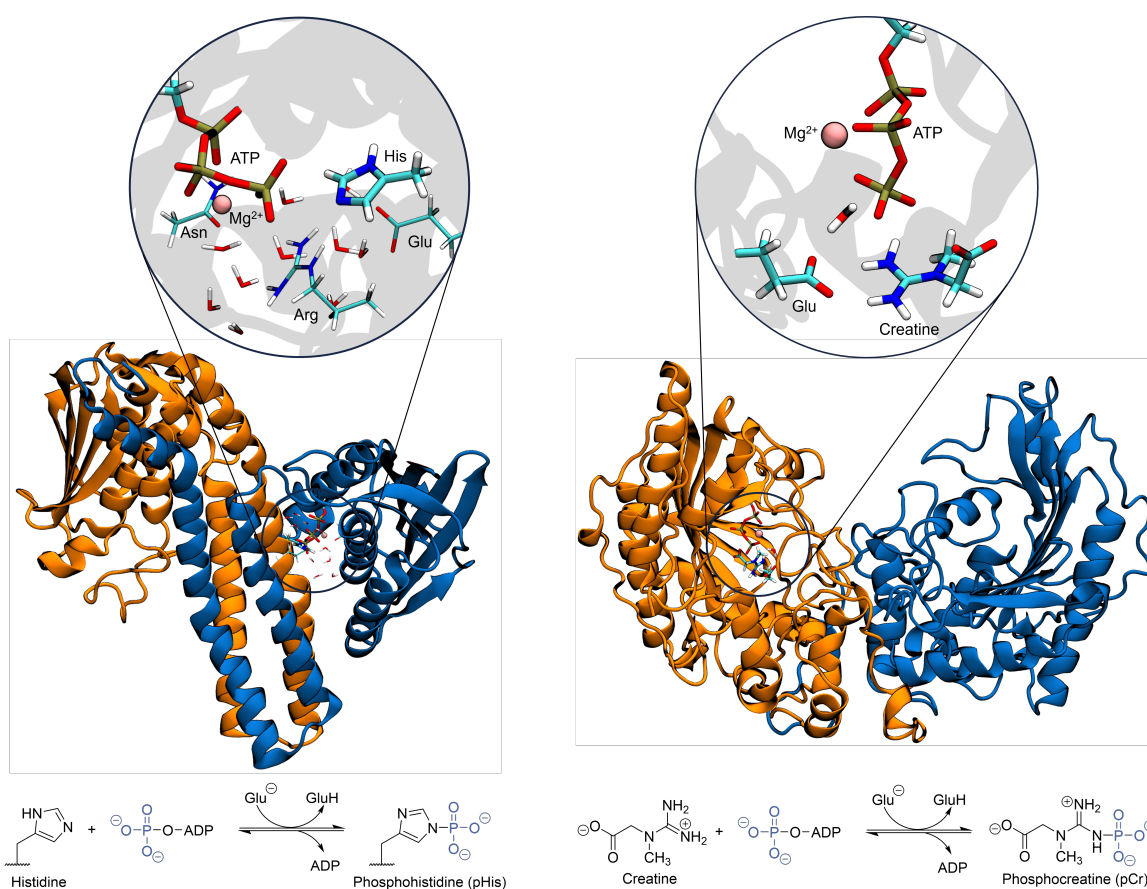


Figure 1.1.: Structures of histidine kinase and creatine kinase with enlarged active site and reaction equation. Histidine kinases (left, studied in chapter 10) are part of signalling systems in bacteria and archaea, and catalyse the autophosphorylation of a histidine residue. Creatine kinases (right, studied in chapter 11) catalyse the phosphorylation of creatine. The resulting phosphocreatine serves both as a rapid energy buffer and as a transport form of high-energy phosphate.

Experimental approaches have provided extensive structural and kinetic information for many kinase systems.^{7,8} High-resolution crystal structures reveal the organisation of active sites and the positioning of substrates, metal ions, and catalytic residues,^{9–12} while kinetic measurements provide information on catalytic rates.^{13,14} However, experimental techniques typically cannot directly resolve the atomic-scale details of the chemical reaction itself, such as the structure of the transition state or the precise sequence of proton- and phosphoryl-transfer events. Molecular simulations therefore play a crucial role in complementing experimental studies by providing atomistic insight into enzymatic reaction mechanisms.

The computational investigation of enzymatic phosphoryl transfer reactions presents several methodological challenges. The reaction involves the formation and breaking of bonds, which require a quantum mechanical (QM) description, while the surrounding protein environment and solvent can be represented on a classical level. Hybrid quantum-mechanical/molecular-mechanical (QM/MM) methods provide a suitable framework for this multiscale description. However, the accessible simulation times in QM/MM simulations are typically limited to a nanosecond timescale. At the same time, enzymatic catalysis often involves conformational rearrangements and rare events that occur on timescales of microseconds to milliseconds. Enhanced-sampling techniques, such as metadynamics, can address this challenge by accelerating the exploration of the free energy landscape, yet the computational cost of accurate quantum mechanical calculations remains a major limiting factor.

Semi-empirical electronic structure methods provide an attractive compromise between accuracy and computational efficiency for such simulations. In particular, the density-functional tight-binding (DFTB) method enables quantum mechanical simulations at a substantially reduced computational cost compared, e.g. to density functional theory (DFT). This efficiency makes it possible to perform extended QM/MM simulations in combination with enhanced-sampling methods for complex biomolecular systems. Nevertheless, semi-empirical methods typically exhibit systematic errors for certain classes of reactions, including phosphoryl transfer processes, which limits their predictive accuracy. This motivates the application of machine learning (ML) to bridge this gap.

The aim of this thesis is to advance the mechanistic understanding of kinase-catalysed N-phosphorylation reactions using multiscale molecular simulations. For histidine kinases, QM/MM simulations have already provided detailed insight into the autophosphorylation mechanism, revealing a penta-coordinated transition state and barrier in accordance with experiments.^{15,16} Building on these studies, the present work introduces a Δ ML correction to improve the accuracy of DFTB-based QM/MM simulations towards DFT-level accuracy while retaining their computational efficiency. To this end, a high-dimensional neural network potential (HDNNP) is employed. In chapter 8, the generation of the training data and the vacuum training procedure are described. In chapters 9 and 10, the vacuum-trained models are applied to simulations in aqueous solution and in histidine kinase, respectively, addressing challenges in transferability and demonstrating how structural and electrostatic aspects influence the protein simulations.

For creatine kinase, previous computational studies have investigated the catalytic mechanism, but relied on potential energies of only a limited number of structures and therefore did not explicitly account for the influence of protein dynamics on the reaction mechanism.¹³ To address these limitations, chapter 11 focuses on QM/MM simulations with DFTB3 as the QM method combined with metadynamics as an enhanced-sampling technique. This enables the exploration of the free energy

landscape of the catalytic cycle, providing insights into the influence of the electrostatic environment and the mechanistic steps of the catalytic mechanism.

This thesis highlights the importance of multiscale simulations for achieving a deeper mechanistic understanding of the fundamental chemical processes underlying kinase-catalysed N-phosphorylations.

Part II.

Theoretical Background

2. Computational Chemistry

Note on Prior Work Small parts of the Theoretical Background are based on the theory chapter of my Master’s thesis¹⁷. The material has been substantially revised, expanded and adapted to fit the broader scope and context of the present dissertation.

Understanding the mechanism of enzymatic catalysis at an atomic level is central to fully characterising an enzyme’s function. Experimental techniques offer a variety of approaches, including studying the structure of an enzyme with protein crystallography or nuclear magnetic resonance (NMR), investigating the influence of individual amino acids with site-directed mutagenesis, and conducting biochemical assays to measure the kinetics of catalysed reactions. However, these experimental studies are extremely complex and time-consuming, particularly for large enzymes.^{18,19}

Computational simulations enable atomic-level insights that would be difficult or impossible to obtain through experimentation alone. They allow the measurement of a wider range of observables with comparable time effort and act as a bridge between theory and experiment, permitting theories to be tested *in silico* first and confirmed with experimental results later. For example, an efficient computational evaluation of a wide variety of mutations can reduce the number of promising candidates that need to be investigated experimentally.¹⁹

Depending on the research question, different computational methods are employed (refer to Figure 2.1) to obtain information such as geometric arrangements, energies, or molecular interactions.

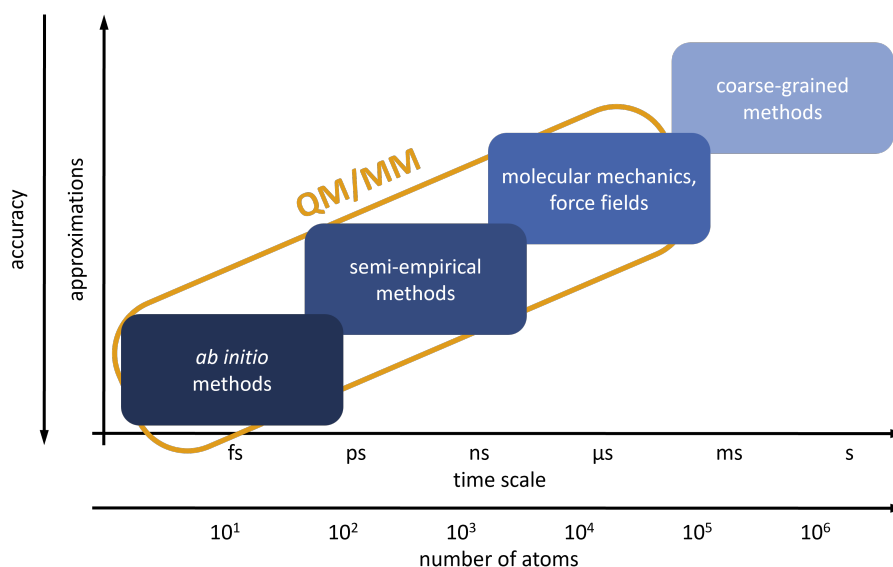


Figure 2.1.: Overview of different computational methods, adapted from Reference [20]. The multiscale quantum mechanical/molecular mechanical (QM/MM) method bridges different levels of theory by treating the chemically active region with more accurate methods and describing the surroundings with faster molecular mechanics.

Overall, the increasing use of computational methods has greatly expanded our knowledge of chemical and biological processes and led to numerous discoveries. In the field of materials and engineering, simulations of lithium ion batteries did not only help explain existing phenomena by revealing structural and electrochemical properties,²¹ but also assisted in the discovery of new materials that improve stability and extend battery life.^{22,23} In molecular biology, simulations provide essential insights into the functions of biomolecules by revealing their dynamics. They also help to refine experimentally determined structures, answer questions about the structure and energetics of enzyme reactions, and reveal how biomolecular systems respond to ligand binding or amino acid mutations.^{24,25} Based on computer-aided and structure-based drug design, structural and dynamical properties of ligand binding can be extracted, which are essential for the development of novel pharmacological therapeutics.²⁶ In neuroscience, simulations help to uncover mechanisms of protein aggregation associated with neurodegenerative diseases^{27,28} and to develop drugs targeting the nervous system.^{24,29–31} Similarly, in enzymology, simulations have become indispensable for understanding reaction mechanisms, intermediate states, and conformational changes that underlie catalysis.^{32,33}

The following sections of this chapter present a more detailed overview of the computational methods used in this work. Molecular dynamics (MD) simulations enable investigating the dynamics of large systems, such as enzymes in aqueous solution, by using Newtonian mechanics (refer to chapter 3). To study a chemical reaction inside an enzyme, hybrid quantum mechanical/molecular mechanical (QM/MM) methods (refer to chapter 4) are required since bond formation and breaking cannot be described with classical force fields, and treating the whole system quantum mechanically would be too expensive. In this approach, the active site is treated using quantum mechanics (QM) for an accurate description of a chemical reaction, while the remainder of the system can be modelled with faster molecular mechanics. Precise QM methods (refer to chapter 5) allow for the simulation and study of the formation and breaking of bonds, as well as their associated reaction energetics. However, these methods are computationally expensive, limiting their application to systems of up to about 100 atoms. Two methods are of particular focus: the *ab initio* method density functional theory (DFT, refer to Section 5.1.4), and the semi-empirical density-functional tight-binding (DFTB) method (refer to Section 5.2.1), which is two to three orders of magnitude faster while still keeping reasonable accuracy. To overcome the timescale limitations of conventional MD simulations, where rare but important events are difficult to observe due to high computational cost, enhanced sampling methods can be employed (refer to chapter 6) to accelerate the exploration of the free energy surface. Finally, to combine the computational efficiency of DFTB with the accuracy of DFT, machine learning (refer to chapter 7) can be used to predict the energy difference between the two methods and correct DFTB energies to DFT-level accuracy on the fly.

3. Molecular Dynamics Simulations

Molecular dynamics (MD) simulations enable the detailed study of the structure and dynamics of multi-million-atom complexes.³⁴ In order to keep the computational effort low, approximations and simplifications are applied, which, however, are accompanied by a loss of accuracy. This opens up the possibility to carry out simulations of large biomolecules in the microsecond range or longer. In 2008, for instance, Shaw et al. presented a specialised supercomputer, extending the limits of the then-accessible nanosecond regime to the biologically relevant millisecond timescale.³⁵

An essential approximation is the classical view of the atoms as charged spheres, in which their quantum mechanical properties are neglected. Their dynamics are described by Newton’s second law as a function of time, and the chemical bond is assumed to be a harmonic oscillator with force constant k and the equilibrium distance r_0 . For a diatomic molecule, this can be represented as

$$F = ma = m \frac{d^2 r}{dt^2} = -k(r - r_0). \quad (3.1)$$

To reduce the computational effort, not every single bond is subjected to an individual parametrisation of the variables k and r_0 . Instead, the bonds are classified by types described by combinations of k and r_0 , which depend on the chemical environment of the atoms. Thus, the parameters can be different for atoms of the same element, e.g. for aliphatic, carbonyl or aromatic carbon atoms.^{36,37}

3.1. Force Field

Molecular dynamics (MD) simulations determine how a system evolves over time by integrating Newton’s equations of motion according to the forces acting on each atom. These forces are obtained from force field methods, also known as molecular mechanics (MM) methods, which describe the potential energy of a system as a function of its atomic coordinates.

The term ‘force field’ may appear counter-intuitive, as it is defined in terms of potential energy rather than forces. However, this concept has become standard for conceptual and mathematical simplicity, since energy is a scalar quantity, whereas forces are vectors that can be calculated from it. The forces acting on the atoms are therefore derived from this potential energy. In the simplest one-dimensional case, the force F is related to the potential energy E by

$$F = -\frac{dE}{dr}. \quad (3.2)$$

The classical force field potential consists of bonding and non-bonding interactions and its general form can be expressed as

$$E_{\text{FF}} = E_{\text{cov}} + E_{\text{nb}} = \underbrace{E_{\text{str}} + E_{\text{bend}} + E_{\text{tors}}}_{\text{bonding}} + \underbrace{E_{\text{vdW}} + E_{\text{el}} + E_{\text{cross}}}_{\text{non-bonding}}. \quad (3.3)$$

The bonding energies are composed of covalent contributions (E_{cov}), i.e. the energies required to stretch a bond (E_{str}), to modify a bond angle (E_{bend}) or for rotations around a bond (E_{tors}). The non-bonding term includes van der Waals (E_{vdW}) and electrostatic interactions (E_{el}). Cross terms (E_{cross}) may also appear, representing couplings between covalent degrees of freedom.³⁸

The potential curve of a diatomic molecule is often described by the Morse potential

$$E_{\text{Morse}}(r) = D_e \left(1 - \exp(-\alpha(r - r^0)) \right)^2. \quad (3.4)$$

This potential depends on the internuclear distance r and captures both the attractive and repulsive interactions between the nuclei. Here, D_e is the dissociation energy and α is dependent on the force constant k by $\alpha = \sqrt{\frac{k}{2D_e}}$. A Taylor expansion of the Morse potential around the equilibrium distance r^0 up to the second order yields the harmonic potential

$$E_{\text{harmonic}}(r) = \frac{1}{2}k(r - r^0)^2. \quad (3.5)$$

This is often a good approximation for the energy to stretch a bond E_{str} , as illustrated in Figure 3.1, since bonds rarely deviate far from their equilibrium distance during an MD simulation.³⁶ However, the harmonic potential diverges for large bond extensions and therefore cannot describe bond dissociation. Furthermore, force fields are parametrised for fixed bond types, which are no longer valid if a bond is broken. Consequently, bond breaking and formation, together with the associated changes in the electronic structure, cannot be described within a force-field framework.

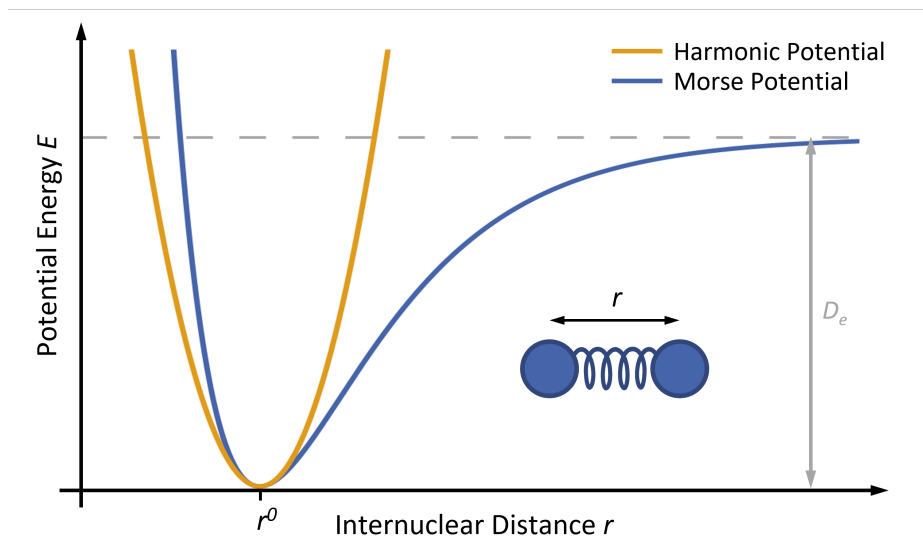


Figure 3.1.: Comparison of Morse and harmonic potentials, adapted from Reference [39]. The dissociation energy D_e is indicated. While the harmonic potential approximates the Morse potential near the equilibrium bond distance r^0 , it lacks the ability to describe bond dissociation at larger distances.

The harmonic potential described above represents the bond-stretching contribution E_{str} of the classical force field. Together with the remaining terms for angle bending, torsional rotation, and non-bonded interactions, the total force field potential can be expressed as

$$E_{\text{FF}} = \underbrace{\frac{1}{2} \sum_i^{N_{\text{bonds}}} k_i (r_i - r_i^0)^2}_{E_{\text{str}}} + \underbrace{\frac{1}{2} \sum_j^{N_{\text{angles}}} k_j^\vartheta (\vartheta_j - \vartheta_j^0)^2}_{E_{\text{bend}}} + \underbrace{\frac{1}{2} \sum_l^{N_{\text{dihedrals}}} \sum_n^{N_{\text{periodicities}}} V_{ln} [1 + \cos(n\omega_l - \gamma_{ln})]}_{E_{\text{tors}}} + \sum_i^{N_{\text{atoms}}} \sum_{j>i}^{N_{\text{atoms}}} \left[\underbrace{4\varepsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right)}_{E_{\text{vdW}}} + \underbrace{\frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r_{ij}}}_{E_{\text{el}}} \right] \quad (3.6)$$

A schematic representation of these contributions is given in Figure 3.2.

Analogously to the bond-stretching term, the angular-bending term (E_{bend} , Equation (3.6), Figure 3.2B) is also described via the harmonic approximation. Here, ϑ_j and ϑ_j^0 represent the angle and equilibrium angle, and k_j^ϑ indicates the force constant.

The torsional contribution (E_{tors} , Equation (3.6), Figure 3.2C), associated with rotation about a bond, is expressed as a sum of periodic cosine terms. Here, ω_l denotes the dihedral angle of the l -th dihedral, V_{ln} the relative rotational barrier height (amplitude), γ_{ln} the minimum of the rotation angle as a phase factor, and n the periodicity, i.e. the number of minima encountered during a full 360° rotation. Adding a constant of one to the cosine term is common practice to set the potential's minimum to zero.

The best known form of van der Waals interactions (E_{vdW} , Equation (3.6), Figure 3.2D) is the Lennard–Jones potential, which is composed of an attractive (r^{-6}) and a repulsive term (r^{-12}) and has two empirical parameters σ and ε . The first one represents the distance at which the energy is zero, the second one represents the depth of the energy minimum.

Electrostatic interactions (E_{el} , Equation (3.6), Figure 3.2E) between differently charged atoms or molecules are represented by Coulomb's law. The distance is given by r_{ij} and their charges by q_i and q_j , ε_0 is the electric field constant.^{37,38,40}

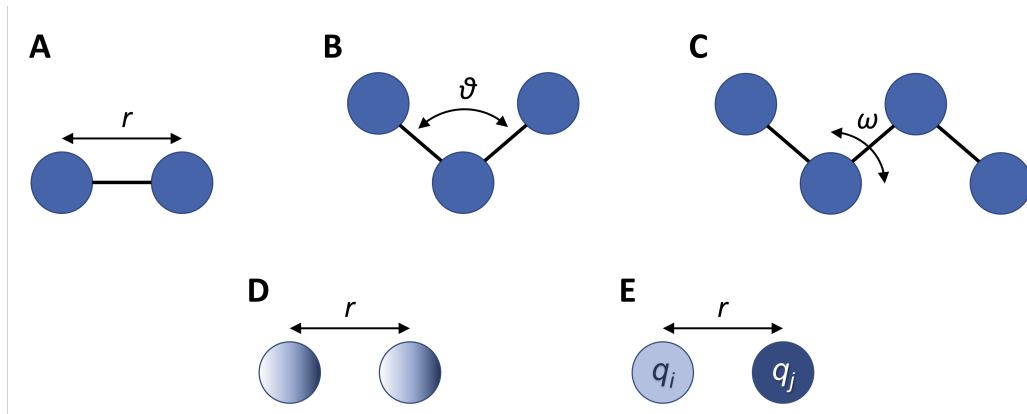


Figure 3.2.: Schematic representation of force field contributions, adapted from Reference [41]. A) Bond stretching, B) angle bending, C) dihedral rotation, D) van der Waals interaction, E) electrostatic interaction.

The force field can be used to determine the potential energy of molecular structures and predict their most stable and energetically favourable structures through energy minimisation (refer to Section 3.2). Furthermore, it forms the basis for molecular dynamics simulations, in which the time evolution of the system is obtained by integrating Newton's equations of motion (refer to Section 3.3).

3.2. Energy Minimisation

In order to simulate a protein, initial structures are typically obtained from the Protein Data Bank (PDB)^{42,43}. These structures are commonly determined by X-ray crystallography or cryogenic electron microscopy (cryo-EM), where hydrogen atoms are often not resolved. The computational addition of hydrogen atoms using tools such as the H++ server⁴⁴⁻⁴⁷ or PROPKA^{48,49} can introduce steric clashes or unrealistic geometries. Therefore, to obtain a physically meaningful starting structure for an MD simulation, an energy minimisation (also called geometry optimisation) must be performed first.^{38,50}

To find the optimal geometry of a molecule, its potential energy must be minimised, as this corresponds to the geometry where the net force on each atom is zero. In mathematical terms, a minimum is defined as a point where the first derivative of the energy with respect to the atomic positions is zero, while the second derivative is positive. Using the example of a diatomic molecule with the energy

$$E(r) = \frac{1}{2}k(r - r_0)^2, \quad (3.7)$$

the first derivative is given by

$$-F(r) = g(r) = \frac{\partial E(r)}{\partial r} = k(r - r_0), \quad (3.8)$$

where g is the first derivative of the energy with respect to the coordinates, generally called the gradient. The second derivative is obtained as the force constant

$$k = \frac{\partial^2 E(r)}{\partial r^2}. \quad (3.9)$$

In the case of a diatomic molecule, r represents a single interatomic distance. For larger systems, however, the potential energy depends on the three Cartesian coordinates x , y , and z of each of the N atoms. The complete set of atomic positions can therefore be written as a $3N$ -dimensional coordinate vector $\mathbf{r} = (x_1, y_1, z_1, x_2, \dots, z_N)$. Consequently, the gradient is no longer one-dimensional but a $3N$ -dimensional vector

$$\mathbf{g} = \nabla E(\mathbf{r}) = \left(\frac{\partial E}{\partial x_1}, \frac{\partial E}{\partial y_1}, \frac{\partial E}{\partial z_1}, \frac{\partial E}{\partial x_2}, \dots, \frac{\partial E}{\partial z_N} \right). \quad (3.10)$$

One common method to locate a minimum on the potential energy surface is the Steepest Descent algorithm. As one of the simplest optimisation methods, it determines the minimum iteratively by

taking steps in the direction opposite to the gradient, i.e. in the direction of the steepest energy decrease. Each step therefore lowers the energy within the local energy landscape, guiding the system towards a local minimum.³⁶

3.3. Integration Method

MD simulations describe the time-dependent motion of atoms. For this purpose, the temporal evolution of positions and momenta must be taken into account, which is also referred to as trajectory. It is calculated within the $6N$ -dimensional phase space, which results from each of the three position and momentum coordinates of the N atoms. The respective acceleration

$$\mathbf{a}(t) = \frac{\mathbf{F}(t)}{m} = -\frac{1}{m} \nabla E(\mathbf{r}(t)). \quad (3.11)$$

results from Newton's second law (refer to Equation (3.1)) or from the derivative of the energy with respect to the spatial coordinate.

Due to the complexity of a system containing more than two atoms, an analytical solution of the equations of motion is not possible, so a numerical integration over a finite number of small time steps must be applied. There are various integration methods, such as the Verlet⁵¹, Velocity Verlet⁵², or leap-frog⁵³ integrators, which are all based on Taylor series expansions of the velocities

$$\mathbf{v}\left(t + \frac{1}{2}\Delta t\right) = \mathbf{v}\left(t - \frac{1}{2}\Delta t\right) + \mathbf{a}(t) \Delta t \quad (3.12)$$

and atomic positions

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}\left(t + \frac{1}{2}\Delta t\right) \Delta t \quad (3.13)$$

up to the second order. Although mathematically equivalent, the latter two are mainly used nowadays, as they are numerically more accurate.⁵²

An important parameter in any integration method is the time step Δt . Since the Taylor series expansion is carried out up to the second order, the error scales with Δt^3 . It should therefore be kept as small as possible, which is achieved by using very small time steps. However, it must be noted that this results in a significantly longer simulation time – with a time step half as large, the computational cost doubles. An orientation is therefore provided by chemical aspects: Δt should be an order of magnitude smaller than the fastest motion of the system, which is often given by the oscillation period of the hydrogen atoms of about 10 fs. Therefore, 1 or 2 fs are usually chosen as time step in MD simulations.³⁶

3.4. Ensemble

In a simulation with the properties described so far, the system is isolated and neither exchanges energy nor matter with its surroundings. According to Newtonian mechanics, energy can be converted between kinetic and potential forms, but the total energy of the system remains constant. The number of particles stays fixed, as none enter or leave the system. The volume has not yet been explicitly addressed, but in the simplest case, it is assumed to be constant. Consequently, such a system corresponds to a microcanonical ensemble (*NVE* ensemble), where the number of particles N , the volume V , and the energy E are constant, while temperature T and pressure P fluctuate according to the system's dynamics.

Under experimental conditions, however, the temperature and the pressure are constant. Therefore, these two quantities must be adjustable in simulations to ensure the comparability of a simulation with experimentally obtained data. Temperature is controlled via a thermostat and pressure via a barostat.

3.5. Thermostat

A canonical ensemble (*NVT* ensemble) represents a closed system, where energy can be exchanged with the environment in the form of heat, while the temperature T remains constant.

The relation between temperature and the total kinetic energy of the system is given by

$$\langle E_{\text{kin}} \rangle = \left\langle \sum_i \frac{1}{2} m_i \mathbf{v}_i^2 \right\rangle = \left\langle \sum_i \frac{1}{2} m_i v_{x,i}^2 + \frac{1}{2} m_i v_{y,i}^2 + \frac{1}{2} m_i v_{z,i}^2 \right\rangle = \frac{3}{2} N k T, \quad (3.14)$$

which can be transformed into

$$T = \frac{2}{3} \frac{\langle E_{\text{kin}} \rangle}{N k}, \quad (3.15)$$

derived from the equipartition theorem and the fact that each atom has three degrees of freedom. A change in temperature is therefore always related to a change in velocity.

A canonical ensemble can be generated with the help of a thermostat that couples the system to a 'heat bath', which acts as a reference during the simulation and has the desired temperature T_{ref} . The Berendsen thermostat⁵⁴, for example, achieves temperature control by coupling the system to an external 'heat bath' and adjusting the temperature according to

$$T_{\text{new}} = T + \Delta T = T + \frac{\Delta t}{\tau} (T_{\text{ref}} - T). \quad (3.16)$$

The temperature change ΔT occurs between two time steps Δt and results from scaling the velocities. With a typical length of 0.1 to 10 ps, the relaxation time τ is about 100 to 10,000 times larger than the time step Δt , which reduces the temperature change by this factor. This allows a slower and more stable approach to the desired temperature.

As the Berendsen thermostat does not correctly reproduce the kinetic energy fluctuations expected for a canonical ensemble, it can lead to errors that prevent a correct canonical ensemble. For example, different parts of the system may exhibit different local temperatures even if the overall temperature matches the desired reference temperature. The velocity rescaling thermostat from Bussi et al.⁵⁵ therefore introduces a stochastic correction term to ensure proper sampling of the canonical temperature distribution, which is achieved by modifying the temperature according to

$$\Delta T = \frac{\Delta t}{\tau} (T_{\text{ref}} - T) + \underbrace{\frac{\Delta W}{\sqrt{\tau}} \cdot 2 \sqrt{\frac{T_{\text{ref}} T}{N_{\text{DoF}}}}}_{\text{correction}}, \quad (3.17)$$

where N_{DoF} is the number of degrees of freedom and ΔW represents a Wiener process (Brownian motion). This stochastic term introduces random fluctuations into the kinetic energy, ensuring proper temperature fluctuations and preventing artefacts such as the ‘flying ice cube’ problem⁵⁶ observed with the Berendsen thermostat.

Another algorithm to obtain a canonical ensemble is the Nosé–Hoover thermostat^{57,58}, which, unlike the thermostats described so far, does not use an external ‘heat bath’ but one integrated in the system. It is therefore also integrated in the equations of motion and thus an inseparable part of the integration algorithm.³⁶

3.6. Barostat

Since experiments are often carried out at constant atmospheric pressure, the pressure must also be controllable during a simulation. For this purpose, the number of particles N , the pressure P , and the temperature T are kept constant (isothermal–isobaric ensemble, NPT ensemble). Similar to the thermostat, where the velocities and thus the kinetic energy are adjusted to keep the temperature constant, a barostat scales the atomic coordinates $\mathbf{r}$ and box length L (and thus its volume $V = L^3$) to equalise the pressure.

Analogous to the Berendsen thermostat, there is also a Berendsen barostat⁵⁴, in which the system is coupled to an external ‘pressure bath’. The pressure is adjusted according to

$$P_{\text{new}} = P + \Delta P = P + \frac{\Delta t}{\tau} (P_{\text{ref}} - P), \quad (3.18)$$

where Δt is the time step and τ the pressure relaxation time. The corresponding scaling factor is given by

$$\lambda = \sqrt[3]{1 - \kappa \frac{\Delta t}{\tau} (P - P_{\text{ref}})}, \quad (3.19)$$

where κ is the isothermal compressibility of the system. The atomic coordinates and box length are then rescaled according to

$$\mathbf{r}_{i,\text{new}} = \lambda \mathbf{r}_i, \quad L_{\text{new}} = \lambda L, \quad (3.20)$$

with $i = 1, 2, \dots, N$.^{36,59}

The Parrinello–Rahman barostat⁶⁰, similar to the Nosé–Hoover thermostat, introduces an additional parameter to the Hamiltonian and also modifies the equations of motion accordingly.⁶¹

4. QM/MM Simulations

When investigating electronic processes such as chemical reactions, a purely classical description of molecules is insufficient, as this approach cannot describe the formation or breaking of covalent bonds. Quantum mechanical (QM) methods (refer to chapter 5) address such processes, but the computational effort required increases significantly with system size. This means they cannot be applied to very large molecules, such as proteins.

The quantum mechanical/molecular mechanical (QM/MM) approach is therefore used to simulate biomolecules, combining the strengths of both techniques in a hybrid approach. The significance of the hybrid QM/MM method was recognised with the 2013 Nobel Prize in Chemistry, awarded to Martin Karplus, Michael Levitt and Arieh Warshel for the development of multiscale models for complex chemical systems.^{62–64}

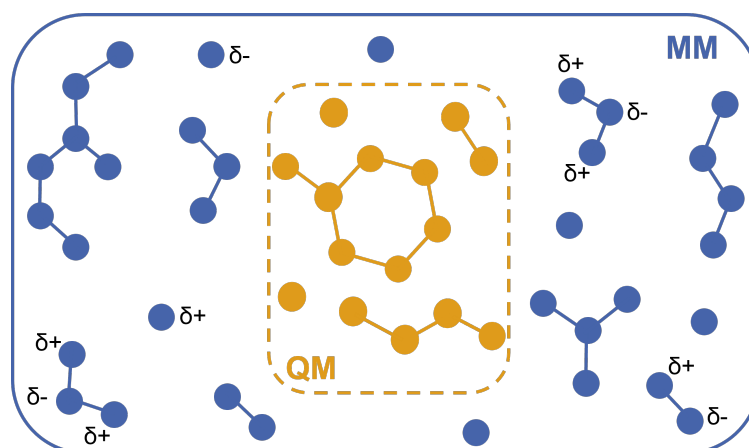


Figure 4.1.: Schematic representation of the QM/MM concept, adapted from Reference [65]. The QM region includes the part of the molecule where the chemical reaction occurs, such as the active site of an enzyme. The surrounding environment is part of the MM region.

As illustrated in Figure 4.1, the QM/MM approach involves the investigation of the reaction of interest (i.e. the active site in the case of an enzyme) using QM methods, while the surrounding components are treated at the MM level. Although the two regions are calculated separately, non-covalent interactions between them must be taken into account when calculating the total energy E_{total} . In the additive coupling scheme, the total energy is obtained as

$$E_{\text{total}} = E_{\text{QM}} + E_{\text{MM}} + E_{\text{QM-MM}}. \quad (4.1)$$

The energy of both the QM region E_{QM} and the MM region E_{MM} is calculated using the methods described in chapter 5 and 3, respectively. However, there exist conceptually different schemes to

describe the coupling term $E_{\text{QM-MM}}$ between the QM and MM region. They differ in whether the interactions between the QM and MM regions are treated or neglected. To be distinguished are the *mechanical*, *electronic*, and *polarised embedding* schemes.

In the *mechanical embedding* scheme, the QM region is treated as if in vacuum, meaning that the MM point charges do not affect the QM Hamiltonian. As a result, the QM electron density does not respond to the electrostatic field of the surrounding MM environment. Interactions between the QM and MM regions are computed at the force field level. The QM-MM coupling term is then obtained from Coulomb's law and the Lennard-Jones potential as

$$E_{\text{QM-MM}} = \sum_i^{\text{QM atoms}} \sum_m^{\text{MM atoms}} \left(\frac{q_i q_m}{r_{im}} + 4\epsilon_{im} \left(\frac{\sigma_{im}^{12}}{r_{im}^{12}} - \frac{\sigma_{im}^6}{r_{im}^6} \right) \right), \quad (4.2)$$

where q_i and q_m represent the partial atomic charges of the QM and MM atoms, respectively.

A drawback of this simplified procedure is that the two regions do not influence each other. For instance, the electronic structure of an enzyme's active site is unaffected by the electrostatic field of the charged amino acids in its vicinity. Consequently, the absence of mutual polarisation between the QM and MM regions can lead to noticeable discrepancies compared with experimental data.

A more accurate method is the *electronic embedding*, also known as *electrostatic embedding*, in which the MM region remains non-polarised, but the QM region is polarisable and thus influenced by the MM region. In the *mechanical embedding*, the QM point charges q_i in Equation (4.2) are classically fixed. In contrast, in the *electronic embedding*, the QM point charges q_i , which are determined by QM methods, also include the effect of the MM electronic potential. This leads to a polarisation of the QM subsystem by the MM environment and thus a more realistic description of electrostatic interactions. Although the computational effort increases only slightly compared to the *mechanical embedding* scheme, the *electronic embedding* scheme produces more reliable results with reasonable accuracy.^{38,65}

The most accurate results are obtained with the *polarised embedding* scheme, in which the QM and MM regions mutually polarise each other. Within the MM force field, each atom i is assigned a polarisability α , which, together with the electric field $\mathcal{E}_{\text{MM}}$ generated by the MM region, and the average charge density of the QM region $\langle \mathcal{E}_{\text{QM}} \rangle$ results in an induced dipole moment

$$\mu_i^{\text{ind}} = \alpha_i (\mathcal{E}_{i,\text{MM}} + \langle \mathcal{E}_{i,\text{QM}} \rangle), \quad (4.3)$$

that interacts with the surrounding point charges. The MM polarisation must be updated at every step of the self-consistent field (SCF) optimisation of the QM wave function. At the same time, the induced dipoles are determined self-consistently with the QM electron density, making the overall QM/MM calculation much more computationally demanding. Consequently, since the accuracy of the *polarised embedding* scheme comes at the expense of a very high computational cost, the *electronic embedding* scheme is usually employed.^{65,66}

The methods described so far deal exclusively with non-covalent interactions between the two regions. However, when a covalent bond crosses the QM-MM interface, each region contains one of the bonded atoms and thus effectively only half of the bond. This is particularly problematic for the QM region, which would be left with an unpaired electron. To avoid this, so-called link atoms are introduced (refer to Figure 4.2). These atoms are part of the QM region but are invisible to the MM calculation. When selecting a covalent bond to cross the QM-MM boundary, a bond that is as non-polar as possible is preferred, for instance one between two sp^3 -hybridised carbon atoms. Hydrogen atoms are commonly used as link atoms because their electronegativity is similar to that of carbon atoms. The link atoms are positioned along the covalent QM-MM bond at a typical C–H distance.^{25,65}

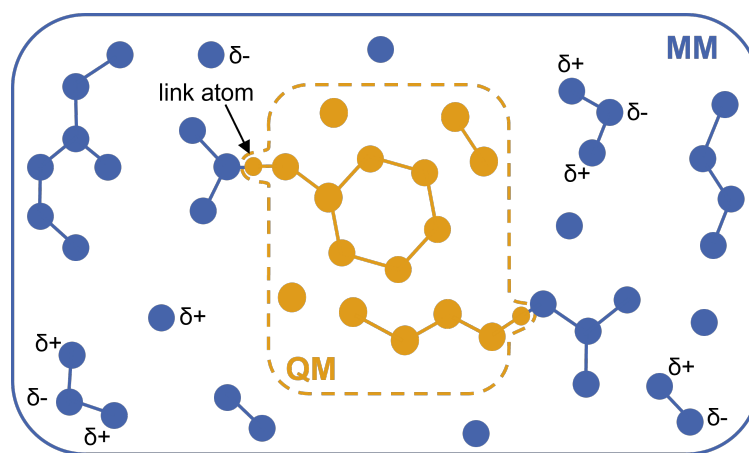


Figure 4.2.: Link atoms in the QM/MM concept, adapted from Reference [65]. To avoid unpaired electrons when cutting a covalent bond at the QM-MM interface, link atoms – often in the form of hydrogen atoms – are placed along the covalent QM-MM bond at a typical C–H distance.

5. Quantum Chemistry Methods

If a chemical reaction is to be investigated, conventional force-field MD simulations are not feasible, as they are typically parametrised for equilibrium configurations and provide limited insight when the system deviates significantly from these configurations.⁶⁷ Therefore, they do not allow for the breaking or formation of chemical bonds.

Quantum chemistry methods overcome these limitations by applying the laws of quantum mechanics (QM) to chemical systems.⁶⁸ QM methods are indispensable for describing chemical reactions and other electronic processes, for example, during catalysis in the active site of an enzyme, including electron transfer processes and excited-state isomerisation.⁶⁷ However, QM methods remain computationally demanding and are typically restricted to systems containing no more than a few hundred atoms.²⁵

QM methods can broadly be divided into two classes. *Ab initio* (Latin for ‘from the beginning’) or first-principles methods (refer to Section 5.1) aim to solve the Schrödinger equation without relying on empirical parameters. Examples include Hartree–Fock (HF) and density functional theory (DFT). Semi-empirical methods (refer to Section 5.2), on the other hand, are derived from HF or DFT by neglecting or approximating certain electronic integrals and introducing parameters obtained from reference calculations or experiments. These approximations reduce the computational cost by two to three orders of magnitude while maintaining reasonable accuracy, making semi-empirical methods suitable for a larger number of QM atoms and for longer simulation timescales.⁶⁹

The following section introduces the fundamental QM principles underlying these methods.

5.1. *Ab Initio* Methods

In order to obtain QM information on stationary states, such as the total energy of a molecule, the time-independent, non-relativistic Schrödinger equation

$$\hat{H}(r)\Psi(r) = E(r)\Psi(r) \quad \text{with} \quad \hat{H}(r) = \hat{T}(r) + \hat{V}(r) \quad (5.1)$$

must be solved. The wave function $\Psi(r)$ is an eigenfunction of the Hamiltonian $\hat{H}(r)$ with the corresponding energy eigenvalue(s) $E(r)$. For a system of N nuclei and n electrons, the molecular Hamiltonian $\hat{H}$ consists of five contributions

$$\begin{aligned} \hat{H} = & - \underbrace{\sum_A^N \frac{\hbar^2}{2M_A} \nabla_A^2}_{\hat{T}_N} - \underbrace{\sum_i^n \frac{\hbar^2}{2m_e} \nabla_i^2}_{\hat{T}_e} \\ & + \underbrace{\sum_A^N \sum_{B>A}^N \frac{Z_A Z_B e^2}{4\pi\epsilon_0 r_{AB}}}_{\hat{V}_{NN}} - \underbrace{\sum_A^N \sum_i^n \frac{Z_A e^2}{4\pi\epsilon_0 r_{iA}}}_{\hat{V}_{Ne}} + \underbrace{\sum_i^n \sum_{j>i}^n \frac{e^2}{4\pi\epsilon_0 r_{ij}}}_{\hat{V}_{ee}}, \end{aligned} \quad (5.2)$$

where $\hat{T}_N$ is the nuclear kinetic energy, $\hat{T}_e$ the electronic kinetic energy, $\hat{V}_{NN}$ the repulsive nucleus–nucleus interactions, $\hat{V}_{Ne}$ the attractive nucleus–electron interactions and $\hat{V}_{ee}$ the repulsive electron–electron interactions.

Using atomic units, where the reduced Planck constant $\hbar$, the electron mass m_e , the elementary charge e and the term $4\pi\epsilon_0$ are all defined to be equal to 1, Equation (5.2) simplifies to

$$\begin{aligned} \hat{H} = & - \underbrace{\sum_A^N \frac{1}{2M_A} \nabla_A^2}_{\hat{T}_N} - \underbrace{\sum_i^n \frac{1}{2} \nabla_i^2}_{\hat{T}_e} \\ & + \underbrace{\sum_A^N \sum_{B>A}^N \frac{Z_A Z_B}{r_{AB}}}_{\hat{V}_{NN}} - \underbrace{\sum_A^N \sum_i^n \frac{Z_A}{r_{iA}}}_{\hat{V}_{Ne}} + \underbrace{\sum_i^n \sum_{j>i}^n \frac{1}{r_{ij}}}_{\hat{V}_{ee}}. \end{aligned} \quad (5.3)$$

Atomic units will be used throughout the remainder of this chapter.

5.1.1. Born–Oppenheimer Approximation

The Born–Oppenheimer approximation⁷⁰ can be used to simplify Equation (5.2). It is based on the fact that protons have a mass about 2000 times greater than electrons. Owing to this large mass difference, nuclei move much more slowly, and the strong Coulomb attractions cause the electrons to adjust almost instantaneously to any change in nuclear positions. As a result, nuclear and electronic motion can be treated as decoupled, and the n electrons can be considered to move in the field of the N fixed nuclei. The electronic structure is thus obtained for fixed nuclear positions by solving the electronic Schrödinger equation

$$\hat{H}_{\text{elec}} \Psi_{\text{elec}} = E_{\text{elec}} \Psi_{\text{elec}} \quad (5.4)$$

with the electronic Hamiltonian

$$\hat{H}_{\text{elec}} = \hat{T}_e + \hat{V}_{Ne} + \hat{V}_{ee}. \quad (5.5)$$

With the electronic Hamiltonian defined, the central objective of QM methods is to find the best approximation for the wave function in order to approach the energy of a system as closely as possible.

5.1.2. Variational Principle

While the electronic Schrödinger equation can still be solved analytically for one-electron systems, such as H_2^+ , this is no longer possible for many-electron systems. In this case, the variational principle is used to obtain approximate wave functions: A trial wave function Ψ_{trial} is chosen and varied until the expectation value of the energy

$$\langle E_{\text{trial}} \rangle = \frac{\langle \Psi_{\text{trial}} | \hat{H} | \Psi_{\text{trial}} \rangle}{\langle \Psi_{\text{trial}} | \Psi_{\text{trial}} \rangle} \geq E_0 \quad (5.6)$$

reaches its minimum. As the energy expectation value $\langle E_{\text{trial}} \rangle$ is always above the lowest eigenvalue E_0 , it provides an upper bound to the exact energy of the system.

Two widely used QM methods based on the Born–Oppenheimer approximation and the variational principle are the Hartree–Fock method and density functional theory.^{71–73}

5.1.3. Hartree–Fock Theory

For many-electron systems, the electronic interactions make it impossible to solve the Schrödinger equation exactly, and approximations are therefore required. One such approximation, central to the Hartree–Fock (HF) theory^{74–76}, is to express the total wave function Ψ in terms of one-electron wave functions ϕ_q , also called *orbitals*.⁷⁷

Because electrons are fermions, the Pauli exclusion principle⁷⁸ must be respected, stating that no two electrons may possess the same set of quantum numbers at the same time. Therefore, each spatial orbital $\phi_q(\mathbf{r}_i)$, characterised by a set of spatial quantum numbers q , is combined with one of the two spin functions

$$\phi_q(\mathbf{r}_i)\alpha(i) \quad \text{and} \quad \phi_q(\mathbf{r}_i)\beta(i), \quad (5.7)$$

where spin-up α and spin-down β correspond to $m_s = +\frac{1}{2}$ and $m_s = -\frac{1}{2}$, respectively. Multiplying the spatial orbital with the corresponding spin function yields a one-electron wave function, or *spin orbital*,

$$\Phi_k(i) = \phi_q(\mathbf{r}_i)s_m(i) \quad \text{with} \quad k = 1, \dots, n \quad (5.8)$$

where the index m denotes the spin ($\pm\frac{1}{2}$), and k combines the quantum numbers q and m as $k = (q, m)$.⁷⁹

The HF theory uses a determinant to represent the total wave function Ψ with its one-electron wave functions Φ . This determinant, called the Slater determinant⁸⁰, inherently satisfies the required properties of a wave function. It reflects the indistinguishability of electrons, as interchanging two electrons by switching their coordinates corresponds to exchanging two rows of the determinant, which changes only its sign, fulfilling the antisymmetry requirement. Furthermore, the Pauli exclusion principle is obeyed since placing two electrons in the same spin orbital would generate two

identical columns, causing the determinant to vanish, meaning that such a configuration does not represent a physically allowed state. The Slater determinant can be written as

$$\Psi = \frac{1}{\sqrt{n!}} \begin{vmatrix} \Phi_1(1) & \Phi_2(1) & \dots & \Phi_n(1) \\ \Phi_1(2) & \Phi_2(2) & \dots & \Phi_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_1(n) & \Phi_2(n) & \dots & \Phi_n(n) \end{vmatrix}, \quad (5.9)$$

where n is the number of electrons and $\frac{1}{\sqrt{n!}}$ serves as a normalisation factor. Each row represents a different electron, while each column represents a different spin orbital.⁸¹

With the one-electron wave functions defined, the next step is to express the Hamiltonian in a one-electron form as well. Under the Born–Oppenheimer approximation (refer to Section 5.1.1), the electronic Hamiltonian (refer to Equation (5.5)) consists of the kinetic energy of the electrons $\hat{T}_e$, the attractive nucleus–electron interactions $\hat{V}_{Ne}$, and the repulsive electron–electron interactions $\hat{V}_{ee}$. The first two terms can easily be broken down into one-electron contributions, resulting in the one-electron Hamiltonian for the i -th electron as

$$\hat{h}_i = -\frac{1}{2} \nabla_i^2 - \sum_A \frac{Z_A}{r_{iA}}. \quad (5.10)$$

Since the remaining electron–electron repulsion term $\hat{V}_{ee}$ is formed as the sum of the repulsion between all electron pairs, it cannot be divided into one-electron terms. Instead, the electronic Hamiltonian is formed as

$$\hat{H}_{\text{elec}} = \sum_i^n \hat{h}_i + \sum_{j>i}^n \frac{1}{r_{ij}}. \quad (5.11)$$

The one-electron Hamiltonian $\hat{h}$ forms the basis of the Fock operator

$$\hat{f}_i = \hat{h}_i + v_i^{\text{HF}}, \quad (5.12)$$

where v_i^{HF} is the HF mean-field potential, i.e. the average potential experienced by the i -th electron in the presence of the other electrons. Replacing the electron–electron repulsion with an average term reduces the complexity of a many-electron system to a one-electron equation, which is the key principle of the HF method.^{81,82}

To clarify the role of the HF potential v_i^{HF} further, Equation (5.12) can be expressed as

$$\hat{f}_i = \hat{h}_i + \sum_k^n (\hat{J}_k - \hat{K}_k), \quad (5.13)$$

with the Coulomb operator $\hat{J}_k$, which represents the classical repulsion potential between two electrons in the k -th orbital, and the exchange-energy operator $\hat{K}_k$, which has no classical analogy and results from the antisymmetry of the electronic wave function.

By minimising the energy with the variational principle described in Equation (5.6), the HF equations

$$\hat{f}\Phi_k = \epsilon_k \Phi_k \quad (5.14)$$

can be derived. The HF equations have to be solved iteratively since $\hat{f}$ and $\hat{K}$ depend on the chosen spin orbitals Φ_k , i.e. the interaction experienced by each electron depends on the mean field of all the other electrons. Therefore, the HF method follows a self-consistent field (SCF) procedure, where first the spin orbitals Φ_k are guessed and iteratively improved until a convergence criterion is reached.

By approximating the electronic wave function with a single Slater determinant, the HF method is insufficient even for simple chemical reactions. This energy difference between the exact energy E_{exact} and the HF energy E_{HF} is called electron correlation^{83,84}

$$E_{\text{corr}} = E_{\text{exact}} - E_{\text{HF}}. \quad (5.15)$$

The neglect of the electron correlation arises because the HF theory treats the motion of opposite-spin electrons as independent, while in reality, electrons move in a correlated manner. Consequently, the resulting HF energies tend to be too high. This can be addressed using so-called post-HF methods such as the coupled cluster theory or the Møller–Plesset perturbation theory, which employ the HF solution as a starting point.^{36,38,82,84–86}

5.1.4. Density Functional Theory (DFT)

QM methods that use wave functions to describe a molecule as a function of $4N$ variables – three spatial coordinates and one spin coordinate for each of the N electrons – quickly become computationally challenging for larger systems. The density functional theory (DFT) offers an alternative approach. It is based on the theorems of Hohenberg and Kohn⁸⁷ and relies on the electron density $\rho(\mathbf{r})$, which depends on only three spatial coordinates, independent of the system size.

The first Hohenberg–Kohn theorem states that the external potential $V_{\text{ext}}(\mathbf{r})$, caused by the nuclei, can be determined by the ground-state electron density alone. The integral of the electron density

$$\int \rho(\mathbf{r}) d\mathbf{r} = N \quad (5.16)$$

defines the number of electrons N . Since the Hamiltonian is dependent on the number of electrons N and the external potential $V_{\text{ext}}(\mathbf{r})$, it follows from the Schrödinger equation (refer to Equation (5.1)) that the ground-state energy of the system

$$E = E[\rho(\mathbf{r})] \quad (5.17)$$

and all other ground-state electronic properties are also functionals of the electron density $\rho(\mathbf{r})$.

The second theorem proves that the variational principle (refer to Equation (5.6))

$$E[\rho(\mathbf{r})] \geq E[\rho_0(\mathbf{r})] \quad (5.18)$$

is valid for any electron density. Consequently, the trial density provides an upper bound to the exact ground-state energy $E[\rho_0(\mathbf{r})]$.^{36,84,88}

The energies shown in the equations above are expressed as functionals, recognisable by the square brackets. Unlike a function (e.g. $f(x) = x^2$), which assigns a value to another value, a functional (e.g. $E[\rho_0(\mathbf{r})]$) takes a function (e.g. the electron density $\rho(\mathbf{r})$) as input and returns a single value (e.g. the energy E).^{36,89}

Within the Born–Oppenheimer approximation (refer to Section 5.1.1), the electronic energy of a molecule

$$E[\rho(\mathbf{r})] = T[\rho(\mathbf{r})] + E_{\text{Ne}}[\rho(\mathbf{r})] + E_{\text{ee}}[\rho(\mathbf{r})] \quad (5.19)$$

is composed of the kinetic energy of the electrons $T[\rho(\mathbf{r})]$, the nucleus–electron attraction $E_{\text{Ne}}[\rho(\mathbf{r})]$ and the electron–electron repulsion $E_{\text{ee}}[\rho(\mathbf{r})]$. The latter can be divided into Coulomb and exchange–correlation contributions, $J[\rho(\mathbf{r})]$ and $E_{\text{xc}}[\rho(\mathbf{r})]$, respectively.

In the original DFT formulation by Hohenberg and Kohn, orbitals were not considered. However, since such a calculation usually lacks accuracy in the calculation of the kinetic energy, the Kohn–Sham theory⁹⁰ is typically applied, which (re-)introduces orbitals ϕ_i by modelling a fictitious system of non-interacting electrons designed to reproduce the exact ground-state electron density $\rho(\mathbf{r})$.^{36,88,91,92}

From this, the total DFT energy in the Kohn–Sham theory results as

$$E_{\text{DFT}}[\rho(\mathbf{r})] = T_{\text{S}}[\rho(\mathbf{r})] + E_{\text{Ne}}[\rho(\mathbf{r})] + J[\rho(\mathbf{r})] + E_{\text{xc}}[\rho(\mathbf{r})], \quad (5.20)$$

with the Kohn–Sham kinetic energy

$$T_{\text{S}}[\rho(\mathbf{r})] = \sum_i n_i \left\langle \phi_i \left| -\frac{1}{2} \nabla^2 \right| \phi_i \right\rangle, \quad (5.21)$$

calculated from a Slater determinant as a sum over the electrons i , and with the occupation number n_i of the i -th Kohn–Sham orbital, the classical nucleus–electron attraction

$$E_{\text{Ne}}[\rho(\mathbf{r})] = - \sum_A^{N_{\text{nuclei}}} \int \frac{Z_A \rho(\mathbf{r})}{|\mathbf{r}_A - \mathbf{r}|} d\mathbf{r}, \quad (5.22)$$

given as a sum over all nuclei A , and the energy of the classical Coulomb interaction of the electrons

$$J[\rho(\mathbf{r})] = \frac{1}{2} \int \int \frac{\rho(\mathbf{r}) \rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}', \quad (5.23)$$

where $\mathbf{r}$ and $\mathbf{r}'$ denote the coordinates of two independent electrons.

The exchange-correlation energy $E_{xc}[\rho(\mathbf{r})]$ contains the difference between the exact kinetic energy and T_S , the non-classical part of the electron–electron repulsion energy $E_{ee}[\rho(\mathbf{r})]$ and the correction for the self-interaction error inherent in $J[\rho(\mathbf{r})]$.^{36,88}

In Equation (5.20), the only unknown term is the exchange-correlation functional $E_{xc}[\rho(\mathbf{r})]$, which is a relatively small fraction of the total energy.^{36,93} It can be represented as

$$E_{xc}[\rho(\mathbf{r})] = E_x[\rho(\mathbf{r})] + E_c[\rho(\mathbf{r})], \quad (5.24)$$

with an exchange functional $E_x[\rho(\mathbf{r})]$ and a correlation functional $E_c[\rho(\mathbf{r})]$. To approximate the exchange-correlation functional $E_{xc}[\rho(\mathbf{r})]$, hundreds of non-empirical and semi-empirical functionals have been developed.⁹³ The different approximation approaches of $E_{xc}[\rho(\mathbf{r})]$ can be ordered by complexity and accuracy on ‘Jacob’s ladder’ (refer to Figure 5.1).

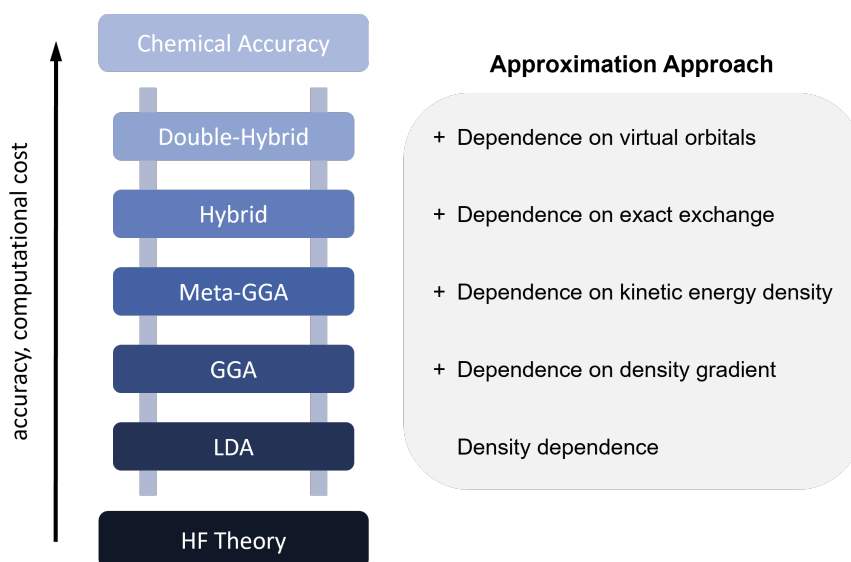


Figure 5.1.: Schematic representation of Jacob’s ladder of DFT, adapted from Reference [94]. Accuracy and computational cost increase from HF theory at the bottom to chemical accuracy (errors below 0.01 eV) at the top.

The local density approximation (LDA) assumes that the exchange-correlation energy $E_{xc}[\rho(\mathbf{r})]$ can be obtained solely from the local density $\rho(\mathbf{r})$ at the position $\mathbf{r}$.³⁸ However, since in reality there is no spatially homogeneous electron distribution, this method has clear limitations. Therefore, the generalised gradient approximation (GGA) additionally includes the density gradient to better capture the inhomogeneity of the charge distribution.³⁶ The first GGA functional was introduced by Perdew and Yang in 1986, introducing the second rung of Jacob’s ladder of DFT.⁹⁵ Further examples are PBE⁹⁶, LYP⁹⁷ and BLYP^{97,98}. The third rung consists of meta-GGA functionals, which rely on the kinetic energy density in order to detect electron delocalisation in molecules.

All of the exchange-correlation functionals on the first three rungs rely on local approximations, which makes them intrinsically limited in their ability to address self-interaction errors, long-range dynamic correlation (i.e. dispersion), and strong correlation. Hybrid GGA functionals on the fourth rung overcome these limitations by replacing the local exchange contribution with the exact HF

exchange functional, while retaining a local correlation functional that yields exactly zero correlation energy for any one-electron system (e.g. LYP). Most global hybrid GGA density functionals incorporate between 20% and 25% exact exchange. This range also includes B3LYP^{97–99}, one of the most frequently used functionals, which employs 20% exact exchange.^{93,100}

The fifth and highest rung on Jacob’s ladder comprises the double hybrid functionals, which describe electron correlation with wave-function-based expressions. These functionals are the most computationally demanding but also among the most accurate. Their usage of higher fractions of exact exchange reduces the self-interaction error and improves the description of thermochemistry and reaction barriers. One of the most promising high-accuracy functionals is the ω B97M-V¹⁰¹ functional.⁹³ It is a combinatorially optimised (i.e. developed through data-driven parameter optimisation), range-separated hybrid (i.e. different treatments of short-ranged and long-ranged Coulomb interactions), meta-GGA functional with a VV10 non-local correlation (i.e. explicitly accounting of the dispersion interactions).¹⁰¹

In addition to the exchange-correlation functional $E_{xc}[\rho(\mathbf{r})]$, so-called basis sets must be chosen in order to perform a DFT calculation. Basis sets consist of a set of functions (called basis functions), which represent the electronic wave function and approximate molecular orbitals. A molecular orbital can be expressed as a linear combination of atomic orbitals (LCAO) as

$$\phi = \sum_{i=1}^M c_i \chi_i, \quad (5.25)$$

where χ_i denotes the atomic orbitals or basis functions and c_i are the coefficients determined by the minimisation of the total energy, which indicate how much each atomic orbital χ_i contributes to the molecular orbital ϕ . The computational cost can be reduced by combining several functions into one contracted function. Because such contractions are applied mainly to core orbitals, the calculation is only slightly less accurate. For this, mostly Slater or Gaussian functions are used, although the latter ones are preferred due to their higher computational efficiency.⁴⁰

Commonly used basis set families are the Pople family with Gaussian type basis sets (e.g. STO-3G^{102,103} or 6-31G*^{104–106}) or the Karlsruhe def2 basis set family¹⁰⁷, also known as the Ahlrichs basis sets, with HF optimised polarisation functions (e.g. def2-SVP or def2-TZVP).¹⁰⁸

Basis sets can be extended by adding diffuse functions to selected elements to create a minimally augmented basis set (abbreviated as ‘ma’, e.g. ma-def2-SVP¹⁰⁹ or ma-def2-TZVP¹⁰⁹). Their use is recommended for calculations involving barrier heights, electron affinities, or for systems that are large or anionic. In the latter case, the inclusion of diffuse functions is often more important for achieving reliable results than increasing the underlying def2 basis set from def2-SVP to def2-TZVP.¹⁰⁹

The fundamental role of DFT in the field of computational chemistry was acknowledged when Walter Kohn and John A. Pople jointly received the 1998 Nobel Prize in Chemistry.^{110,111} Kohn’s formulation of DFT provided the theoretical basis for describing electronic structures in terms of electron density, while Pople’s development of efficient computational implementations, such as the Gaussian programme, made complex QM calculations accessible to a wider scientific community.

In their 2005 ‘Handbook of Materials Modeling: Methods’¹¹², Roberto Car et al. highlight that the good accuracy of DFT for ground-state properties, together with its reasonable scaling with system size, has made it the method of choice for many applications. Typical DFT implementations scale approximately as M^3 , whereas HF scales as M^4 and post-HF methods such as MP2 or CCSD(T) scale as M^5 or M^7 , respectively (with M as the basis set size).^{38,40} Over two decades later, the use of DFT calculations still continues to grow (refer to Figure 5.2).

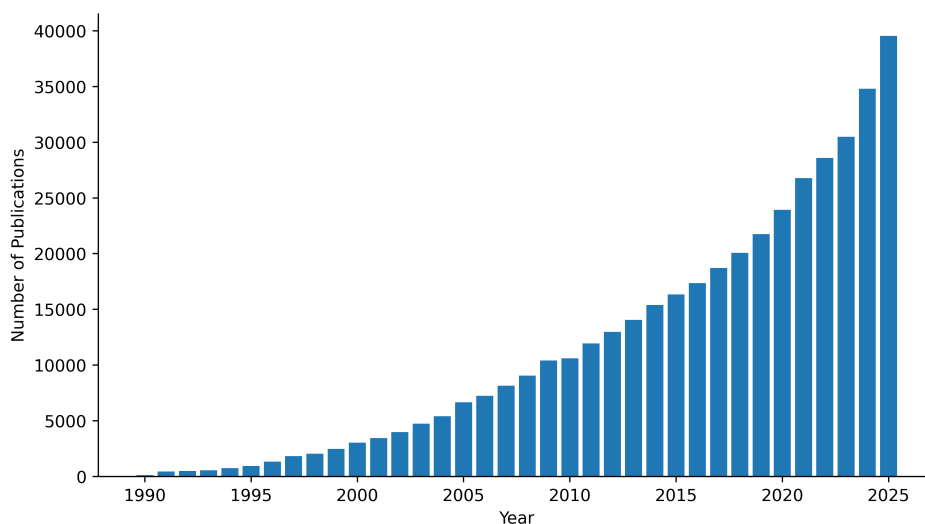


Figure 5.2.: Number of DFT publications per year (1990–2025), adapted from Ref. [73]. Publication data for the final publication year were obtained from Clarivate *Web of Science Core Collection* using the search terms ‘density functional theory’ or ‘DFT’ (accessed 2026-01-29). Copyright Clarivate 2025. All rights reserved.¹¹³

5.2. Semi-Empirical Methods

Ab initio methods provide accurate results, but their high computational cost makes them unfeasible for large systems. Their steep increase in computational effort largely originates from the number of two-electron integrals. Reducing the number of these integrals is therefore essential to lowering the overall complexity. This concept forms the basis of semi-empirical methods, which are derived from HF or DFT but avoid the explicit evaluation of many integrals by introducing approximations and empirical parameters.^{40,69}

A common strategy is to treat only the valence electrons explicitly while incorporating the effect of the core electrons implicitly, for example, by reducing the effective nuclear charge or introducing additional functions that mimic the combined repulsion of nuclei and core electrons. Restricting the treatment to valence electrons significantly reduces the number of electrons and required basis functions. As a result, semi-empirical methods can use minimal basis sets and thus achieve a substantially lower computational cost, making them two to three orders of magnitude faster than the HF and DFT methods from which they are derived. Despite this considerable speed-up, semi-empirical methods remain about three orders of magnitude slower than classical MD simulations. This makes them well-suited for bridging the gap between the *ab initio* and classical worlds.^{40,69} Even with medium-sized basis sets, DFT calculations are generally limited to the sub-nanosecond timescales, with an accuracy comparable to – or even lower than – that of semi-empirical methods.⁶⁹

An example of a semi-empirical method based on Kohn–Sham DFT is density-functional tight-binding (DFTB)^{114–117}, which can reach a level of accuracy comparable to DFT with GGA functionals and medium-sized basis sets, and is suitable for the application of large molecules like biomolecules^{118,119} use in biomolecular QM/MM simulations^{25,120}. Nevertheless, DFTB is an approximation of DFT and must be benchmarked for the specific system of interest.⁶⁹

5.2.1. Density-Functional Tight-Binding (DFTB)

The DFTB method is derived from a Taylor series expansion of the Kohn–Sham DFT total energy around a reference density ρ^0 with respect to charge density fluctuations $\delta\rho$. The reference density ρ^0 is obtained from a superposition of neutral atomic charge densities previously determined with DFT calculations. The Taylor expansion of the DFT total energy then results as

$$\begin{aligned}
 E_{\text{Taylor}}^{(0-3)}[\rho^0 + \delta\rho] = & \underbrace{\frac{1}{2} \sum_{AB} \frac{Z_A Z_B}{R_{AB}} - \frac{1}{2} \iint \frac{\rho^0(\mathbf{r}) \rho^0(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r} d\mathbf{r}' - \int V_{\text{xc}}[\rho^0] \rho^0(\mathbf{r}) d\mathbf{r} + E_{\text{xc}}[\rho^0]}_{E^{(0)}[\rho^0]} \\
 & + \underbrace{\sum_i n_i \langle \phi_i | \hat{H}[\rho^0] | \phi_i \rangle}_{E^{(1)}[\rho^0, \delta\rho]} + \underbrace{\frac{1}{2} \iint \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} + \frac{\delta^2 E_{\text{xc}}[\rho]}{\delta\rho(\mathbf{r}) \delta\rho(\mathbf{r}')} \bigg|_{\rho^0} \right) \delta\rho(\mathbf{r}) \delta\rho(\mathbf{r}') d\mathbf{r} d\mathbf{r}'}_{E^{(2)}[\rho^0, (\delta\rho)^2]} \\
 & + \underbrace{\frac{1}{6} \iiint \frac{\delta^3 E_{\text{xc}}[\rho]}{\delta\rho(\mathbf{r}) \delta\rho(\mathbf{r}') \delta\rho(\mathbf{r}'')} \bigg|_{\rho^0} \delta\rho(\mathbf{r}) \delta\rho(\mathbf{r}') \delta\rho(\mathbf{r}'') d\mathbf{r} d\mathbf{r}' d\mathbf{r}''}_{E^{(3)}[\rho^0, (\delta\rho)^3]},
 \end{aligned} \tag{5.26}$$

where, depending on which terms are included, different DFTB approaches result.¹¹⁹

5.2.1.1. DFTB1

If only the zero- and first-order terms are considered, the standard (non-self-consistent) DFTB method (DFTB1^{114,115}) is obtained. The zeroth-order term $E^{(0)}[\rho^0]$ (refer to Equation (5.26)) includes the nuclear–nuclear repulsion between the atoms A and B , the Coulomb interaction of the electrons at the positions $\mathbf{r}$ and $\mathbf{r}'$, and the exchange–correlation potential and energy, V_{xc} and E_{xc} , respectively. All of these depend solely on the reference density ρ^0 and can be summarised as a repulsive energy term, which is approximated as a sum of pairwise potentials

$$E^{(0)}[\rho^0] \approx E_{\text{rep}} = \frac{1}{2} \sum_{AB} V_{AB}^{\text{rep}} \tag{5.27}$$

between the atoms A and B .¹¹⁹ It can be obtained either from DFT calculations¹¹⁵ or by extrapolating from empirical data¹²¹.

The first-order term $E^{(1)}[\rho^0, \delta\rho]$ (refer to Equation (5.26)) is formed by an electronic Hamiltonian dependent on the reference density ρ^0 and the molecular orbitals ϕ_i which result from an LCAO ansatz of Kohn–Sham orbitals. The diagonal elements can be approximated as $H_{\mu\mu}^0 = \varepsilon_\mu$ to provide

the energy ε_μ of a neutral atom in vacuum, while the off-diagonal elements are calculated in a two-centre approximation. This results in the so-called band-structure energy or Hamiltonian energy

$$E^{(1)}[\rho^0, \delta\rho] \approx E_{H0} = \sum_{iAB} n_i \sum_{\mu \in A} \sum_{v \in B} c_\mu^i c_v^i H_{\mu v}^0 = \sum_i n_i \varepsilon_i, \quad (5.28)$$

where n_i is the occupation number of the i -th Kohn–Sham orbital, and μ and v are the valence atomic orbitals of the atoms A and B , respectively. The coefficients c_μ^i and c_v^i together with the Hamiltonian matrix elements $H_{\mu v}^0$ give the energy ε_i of the i -th orbital. The matrix elements are predetermined for specific interatomic distances and listed in tabular format to be accessed during a DFTB calculation, in order to reduce computational cost.^{117–119,122}

Together, the DFTB1 total energy is given as

$$E_{\text{DFTB1}} = \underbrace{\frac{1}{2} \sum_{AB} V_{AB}^{\text{rep}}}_{E_{\text{rep}}} + \underbrace{\sum_{iAB} n_i \sum_{\mu \in A} \sum_{v \in B} c_\mu^i c_v^i H_{\mu v}^0}_{E_{H0}}. \quad (5.29)$$

5.2.1.2. DFTB2

Including also the second-order term $E^{(2)}[\rho^0, (\delta\rho)^2]$ (refer to Equation (5.26)) leads to the self-consistent charge (SCC) DFTB method (DFTB2¹¹⁶). The second-order term comprises the charge fluctuation energy, and by approximating it as an exponentially decaying charge density, it can be expressed as

$$E^{(2)}[\rho^0, (\delta\rho)^2] \approx E_\gamma = \frac{1}{2} \sum_{AB} \Delta q_A \Delta q_B \gamma_{AB}, \quad (5.30)$$

where $\Delta q_A = q_A - q_A^0$ is the net charge on atom A . The γ function incorporates the electron–electron interaction so that at large atomic distances, long-range electrostatics are described as the Coulomb interaction of the net charge Δq_A and Δq_B . For the on-site term, Equation (5.30) describes the electron–electron interaction within one atom ($A = B$). In this case, it becomes

$$E_\gamma = \frac{1}{2} \sum_A \Delta q_A^2 \gamma_{AA}, \quad (5.31)$$

where the γ_{AA} function can be connected to the Hubbard parameter

$$U_A = \gamma_{AA}, \quad (5.32)$$

which equals twice the chemical hardness of an atom, and can be obtained from DFT calculation of a neutral atom.^{116,117,119}

Adding E_γ to Equation (5.29), the DFTB2 total energy results as

$$E_{\text{DFTB2}} = \underbrace{\frac{1}{2} \sum_{AB} V_{AB}^{\text{rep}}}_{E_{\text{rep}}} + \underbrace{\sum_{iAB} n_i \sum_{\mu \in A} \sum_{v \in B} c_\mu^i c_v^i H_{\mu v}^0}_{E_{H0}} + \underbrace{\frac{1}{2} \sum_{AB} \Delta q_A \Delta q_B \gamma_{AB}}_{E_\gamma}. \quad (5.33)$$

5.2.1.3. DFTB3

For the third-order term $E^{(3)}[\rho^0, (\delta\rho)^3]$ (refer to Equation (5.26)), the same approximations as for the second-order term are applied. It describes the change of the chemical hardness of an atom depending on its charge state and can be approximated as

$$E^{(3)}[\rho^0, (\delta\rho)^3] \approx E_\Gamma = \frac{1}{3} \sum_{AB} \Delta q_A^2 \Delta q_B \Gamma_{AB}, \quad (5.34)$$

with the function $\Gamma_{AB} = \frac{\partial \gamma_{AB}}{\partial q}$ as the derivative of the γ function introduced in Equation (5.30).¹¹⁹

Adding this new energy term to Equation (5.33) leads to the DFTB3¹¹⁷ formalism in which the total energy is given as

$$E_{\text{DFTB3}} = \underbrace{\frac{1}{2} \sum_{AB} V_{AB}^{\text{rep}}}_{E_{\text{rep}}} + \underbrace{\sum_{iAB} n_i \sum_{\mu \in A} \sum_{\nu \in B} c_\mu^i c_\nu^i H_{\mu\nu}^0}_{E_{\text{H0}}} + \underbrace{\frac{1}{2} \sum_{AB} \Delta q_A \Delta q_B \gamma_{AB}}_{E_\gamma} + \underbrace{\frac{1}{3} \sum_{AB} \Delta q_A^2 \Delta q_B \Gamma_{AB}}_{E_\Gamma}. \quad (5.35)$$

In DFTB2, a constant Hubbard parameter U_A is used, and thus no distinction is made between the chemical hardness of anions and cations, or of the neutral atom. This can lead to significant issues in highly charged systems. Furthermore, the shape of the atoms is fixed, as it is defined by the initial reference density. In DFTB3, these limitations are overcome by introducing E_Γ , as the Γ_{AB} function in this energy contribution contains the derivative of the Hubbard parameter (chemical hardness). This makes it possible to achieve the expected behaviour also for charged systems, e.g. correct scaling of the energy with charge, making DFTB3 especially well-suited to describe biomolecular systems.^{117,119}

In summary, DFTB is based on a Taylor expansion of the Kohn–Sham DFT energy around a reference charge density. Treating only the valence electrons enables the use of a minimal basis set. Together with predetermined values for the repulsive energy E_{rep} and Hamiltonian energy E_{H0} , this results in a significant reduction in computational cost of about two to three orders of magnitude compared to DFT, while still retaining reasonable accuracy. This makes the semi-empirical DFTB method attractive for investigating biomolecules and applications in QM/MM systems, as it provides access to long timescales and larger QM regions.

However, it should be noted that the approximations used reduce the accuracy for systems for which the method has not been benchmarked. For example, the element pair-wise parameterisation limits transferability to different systems or bonds, creating the necessity for specific reaction parameters (SRP).^{16,117,119,120,122,123}

Furthermore, even though the decrease in computational cost is impressive, other methods must be relied on if a high free-energy barrier has to be overcome during a chemical reaction, such as the enhanced sampling methods discussed in the next chapter.

6. Enhanced Sampling

Ergodicity is the fundamental concept that makes MD simulations meaningful.¹²⁴ While experiments yield the ensemble average, MD simulations obtain the time average over a single trajectory. The ergodic hypothesis states that the average over a large number of molecules (ensemble) is the same as the average over all states of a single molecule over a long period of time. This can be expressed as

$$\begin{aligned} \langle A \rangle_{\text{time}} &= \langle A \rangle_{\text{ensemble}} \\ \text{with } \langle A \rangle_{\text{time}} &= \frac{1}{t} \int_0^t A(x, p, t') dt' \\ \text{and } \langle A \rangle_{\text{ensemble}} &= \int \int \rho(x, p) A(x, p) dx dp, \end{aligned} \quad (6.1)$$

where t is the time, x and p denote the position and momentum variables that together span the $6N$ -dimensional phase space of an N -particle system, and $\rho(x, p)$ is the phase-space probability density.^{36,125} However, ergodicity can only be achieved if the simulation is carried out over a sufficiently long period of time to visit every point in phase space. As this is rarely the case, even for simple systems, quasi-ergodicity is used in practice instead of the ergodic theorem. It states that, while the system does not reach every point, it does come sufficiently close.^{125,126}

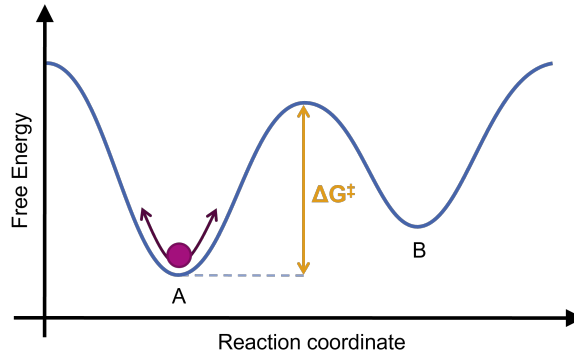


Figure 6.1.: Schematic representation of a free energy surface with two local minima A and B. If a high free-energy barrier $\Delta G^\ddagger$ separates the two minima, the barrier is unlikely to be overcome within the timescale accessible to conventional MD simulations. Consequently, sampling remains confined to configurations close to the system's initial state (represented by the purple circle). The purple arrows indicate the extent of phase-space fluctuations around the initial state.

Therefore, to achieve quasi-ergodicity, the phase space must be sampled properly. This becomes problematic when transitions between metastable states are separated by high free-energy barriers $\Delta G^\ddagger$ as highlighted by the transition state theory (TST)

$$k = \frac{k_B T}{h} \exp\left(-\frac{\Delta G^\ddagger}{RT}\right), \quad (6.2)$$

with the reaction rate k , the Boltzmann constant k_B , the temperature T , the Planck constant h , and the ideal gas constant R .¹²⁵ Equation (6.2) can also be expressed as the probability

$$P(A \rightarrow B) \propto \exp\left(-\frac{\Delta G^\ddagger}{RT}\right), \quad (6.3)$$

to go from one state A to another state B by crossing a barrier $\Delta G^\ddagger$. Consequently, high free-energy barriers $\Delta G^\ddagger$ lead to low transition rates k and low crossing probabilities $P(A \rightarrow B)$. For typical biomolecular barriers of several to tens of kilocalories per mol, simulation times required to observe a single transition can extend into the microsecond range or beyond. In contrast, the typical length of MD simulations is only up to 1 μ s, with QM/MM simulations reducing the accessible timescale to the nanosecond regime. As a result, simulations are confined to conformations that are energetically close to the initial structure, whereas high-energy regions and regions separated by high free-energy barriers are sampled very poorly or not at all, as illustrated in Figure 6.1.

This challenge becomes even more pronounced when not only barrier crossing events are of interest, but the free energy of the process has to be computed (e.g. barrier height or free-energy difference between two states). The Helmholtz free energy F is given as

$$F = -k_B T \ln Q, \quad (6.4)$$

with the Boltzmann constant k_B , the temperature T and the canonical partition function

$$Q = \sum \exp\left(\frac{-E(x, p)}{k_B T}\right). \quad (6.5)$$

When Equation (6.4) is expressed as

$$F = k_B T \ln \left\langle \exp\left(\frac{E(x, p)}{k_B T}\right) \right\rangle - \ln c, \quad (6.6)$$

with a constant c , the sampling problem becomes apparent. Since the energy enters an exponential term, high-energy regions contribute significantly to the free energy. If such regions are insufficiently sampled, the resulting free-energy estimates can be imprecise.¹²⁷

While the Helmholtz free energy F is computed in canonical NVT ensembles, the Gibbs free energy G is computed in isothermal–isobaric NPT ensembles and can be obtained from Equation (6.4) as

$$G = F + PV, \quad (6.7)$$

with the pressure P and the volume V . In the condensed phase, for example, biomolecules in aqueous solution, changes in the PV term are so small that they can be neglected, leading to $G \approx F$. Therefore, the term ‘free energy’ is often used without further specification.¹²⁸

To overcome these problems, a number of enhanced sampling methods have been developed in recent years, which can be classified as either collective variable-based or collective variable-free methods. The collective variable (CV), or reaction coordinate, is defined by a selected number of degrees of freedom that enable the reaction of interest to be described. The choice of CV depends on

the system and reaction being investigated. For example, in the case of a proton transfer between two molecules, it can be the distances between the atoms involved, while for visualising the conformational change in the secondary structure of a peptide, the backbone dihedral angles ψ and ϕ can be a good choice. Recently, machine learning-based methods have also emerged to identify suitable CVs.^{129–135} The free energy surface along a CV is called the potential of mean force (PMF).⁴⁰

Examples of CV-free and CV-based enhanced sampling methods are provided in Figure 6.2. CV-free methods aim to enhance sampling without explicitly defining CVs. They often focus on employing multiple replicas, modifying the thermodynamic ensemble, or introducing biasing potentials to decrease barriers on a potential energy surface (PES).¹³⁶ A prominent example of replica-based approaches is replica exchange molecular dynamics (REMD)¹³⁷, where multiple replicas of the same system are simulated in parallel at different temperatures. At predefined intervals, exchanges of configurations between neighbouring replicas are attempted according to a Metropolis criterion (stochastic criterion). This allows low-temperature simulations to overcome energy barriers by exchanging with high-temperature replicas.¹³⁸ Another CV-free method, accelerated molecular dynamics (aMD)¹³⁹, on the other hand, raises the PES in low-energy regions by adding biasing potentials, while high-energy regions are unaffected. Thereby, the energy barriers are effectively decreased.¹⁴⁰

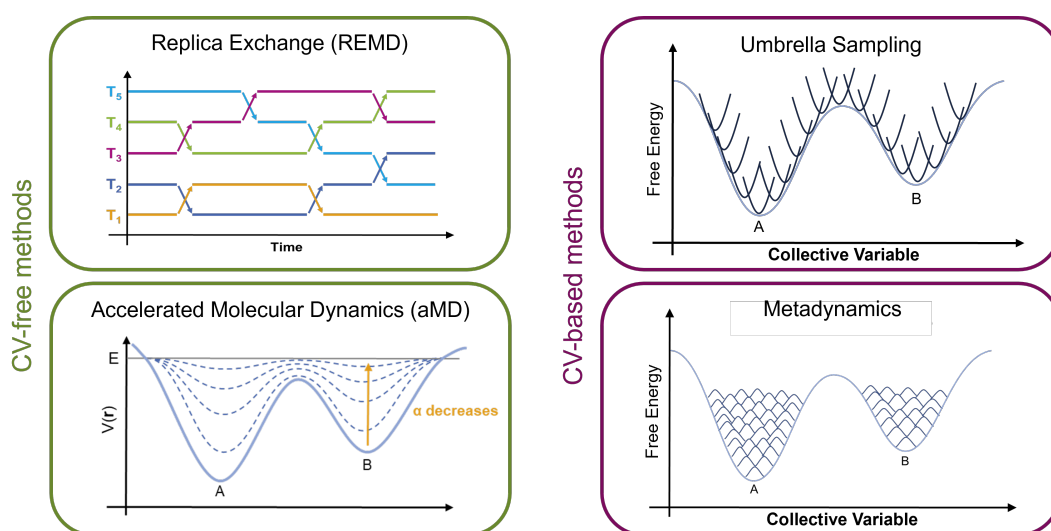


Figure 6.2.: Schematic representation of different enhanced sampling methods, adapted from References [141–144]. Enhanced sampling can be divided into CV-free (left) and CV-based approaches (right). In replica exchange (REMD) simulations, several replicas are simulated at different temperatures and periodically attempt configuration exchanges. Accelerated molecular dynamics (aMD) lowers free-energy barriers by adding a biasing potential below a chosen energy threshold E , scaled by an acceleration factor α . In umbrella sampling, a CV is defined *a priori* and sampled in overlapping windows, from which the free energy surface (FES) can be reconstructed. In metadynamics, the CV is likewise defined *a priori*, but a history-dependent biasing potential enables the system to explore multiple reaction pathways.

In CV-based methods, enhanced sampling is achieved by adding a biasing potential along one or more predefined CVs in order to facilitate barrier crossing.¹⁴⁵ In umbrella sampling^{146–148}, a series of overlapping biasing potentials are added along a chosen CV, often in the form of harmonic potentials due to their simplicity. These restraints confine the system to specific regions of CV space, ensuring efficient sampling along the whole CV. The individually sampled windows can afterwards be combined to reconstruct the full FES of the process.¹⁴⁸ Another example, metadynamics¹⁴⁹,

uses an opposite approach and does not restrain the system to fixed values of a CV. Instead, history-dependent bias potentials are deposited to discourage the system from revisiting regions of CV space that have already been explored. Metadynamics will be discussed in more detail in Section 6.1.

In summary, enhanced sampling methods help to overcome one of the main limitations of simulations: the timescale problem. This limitation arises from the need to choose a sufficiently small time step to ensure numerical stability, which in turn restricts simulations to relatively short timescales. Furthermore, CV-based methods can provide insight into the analysis and interpretation of the resulting high-dimensional trajectory by offering a projection onto the low-dimensional CV-space.¹²⁶

6.1. Metadynamics

One of the most popular CV-based approaches for investigating complex free energy surfaces is metadynamics.¹⁴⁵ It was first introduced by Laio and Parrinello in 2002¹⁴⁹ and is based on a history-dependent repulsive biasing potential, which is added to the potential energy in order to discourage the system from revisiting already sampled regions of the CV space (refer to Figure 6.3). The biasing potential $V_G(S, t)$ is constructed as a sum of Gaussian functions and can be written as

$$V_G(S, t) = \sum_{k\tau < t} W(k\tau) \exp \left(- \sum_{i=1}^{N_{CV}} \frac{(S_i - S_i(k\tau))^2}{2\sigma_i^2} \right), \quad (6.8)$$

where W and σ_i denote the height and width of the Gaussian potentials, respectively. These Gaussians are placed in a deposition stride τ along the i -th CV S_i . N_{CV} is the total number of biased CVs. In practice, metadynamics simulations are usually biased along one to three CVs.¹⁵⁰

Metadynamics simulations accelerate the sampling of rare events, such as the crossing of high free-energy barriers, by filling the initial minimum with biasing potentials. This effectively lowers the free-energy barrier and enables the system to reach other metastable states. In addition, metadynamics simulations enable the discovery of new reaction pathways as the system leaves this initial, local minimum across the saddle point with the lowest free energy. Once the simulation is complete, the sum of all biasing potentials

$$V_G(S, t \rightarrow \infty) = -F(S) + \text{const} \quad (6.9)$$

gives the negative of the underlying free energy surface (FES) as a function of the chosen CVs.¹²⁴ This highlights a key advantage of metadynamics: metastable states and reaction pathways can be discovered without requiring prior knowledge of the FES.¹²⁴

Choice of CVs and Parameters

The purpose of collective variables (CVs) is not only to accelerate transitions between metastable states, but also to provide a low-dimensional representation of a high-dimensional process. In practice, three prerequisites are important for selecting the CV(s). (1) The chosen CV(s) must be able to distinguish between the reactant, product and any relevant intermediate or transition states in order to accurately represent the process of interest. (2) The CV(s) must capture all slow modes of the system for the FES to converge within the simulation time. (3) The number of CVs should be limited, since a large number of CVs undermines the concept of a low-dimensional representation of

the process of interest. Furthermore, high dimensionality increases the simulation time and creates a more complex FES that is difficult to interpret.¹²⁴

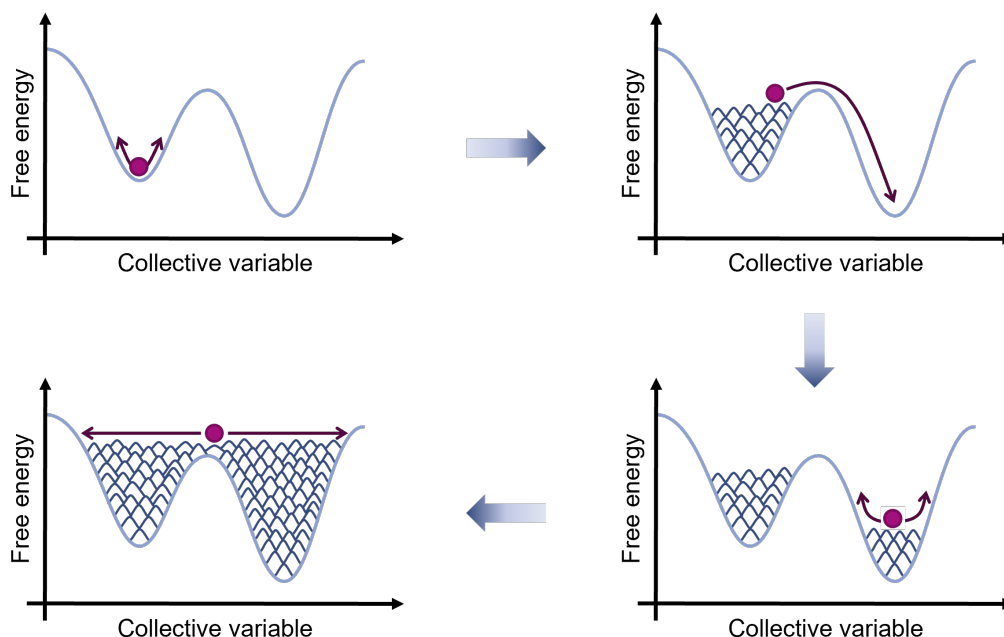


Figure 6.3.: Schematic representation of the metadynamics method, adapted from Reference [144]. History-dependent biasing potentials in the form of Gaussian functions are deposited along one or more predefined CVs. As the initial minimum is progressively filled, the effective free-energy barrier is reduced, enabling the system to escape the initial minimum and explore other metastable states. In the end, the system can explore the complete CV space freely. This enables not only barriers to be overcome, but also reaction pathways to be explored and the free energy surface (FES) to be obtained. The FES results from the negative sum of the Gaussian potentials.

For the deposition of Gaussian bias potentials along a CV, three parameters must be chosen: the height W , the width σ , and the deposition stride τ . For each parameter, a balance must be found between biasing the system too much, which would hide subtle changes in the FES (i.e. poor resolution and increased error), and biasing the system too little, which would lead to slow sampling and long convergence times. To determine an appropriate Gaussian width σ , it is recommended to perform an initial unbiased simulation and to choose a value of approximately one-half to one-third of the standard deviation of fluctuations along the CV. The height W and deposition stride τ can be combined into a single variable, the deposition rate $\omega = W/\tau$. In practice, the height W is typically chosen as a fraction of the thermal energy $k_B T$, and the deposition stride τ is adjusted to adapt the deposition rate ω .¹⁵¹

6.1.1. Well-Tempered Metadynamics

A drawback of standard metadynamics is that deciding when to stop the simulation is often non-trivial, as the FES does not converge in a single run. Instead, it oscillates around the exact free-energy profile, resulting in an average error that is proportional to the square root of the deposition rate. Furthermore, if a metadynamics simulation is continued for too long, the continuously growing bias may force the system into unphysical conformations.¹⁵²

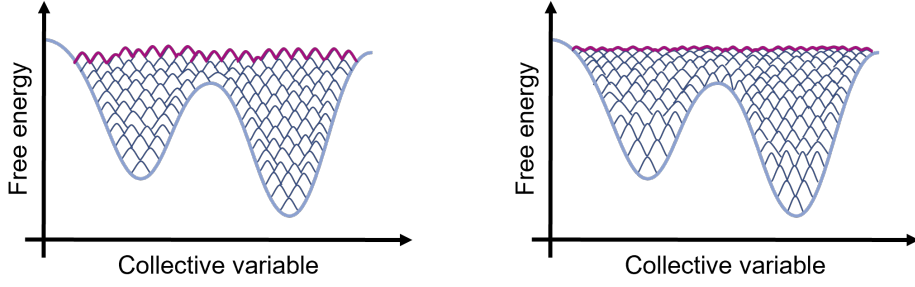


Figure 6.4.: Schematic representation of standard metadynamics versus well-tempered metadynamics. In standard metadynamics (left), Gaussian potentials of constant height are deposited along the chosen CV(s). While this fills the free-energy minima quickly, it can result in poor convergence and fluctuations when estimating the underlying FES. In well-tempered metadynamics (right), the height of the deposited Gaussian biases is reduced by a bias factor. This produces a smoother FES, avoids overfilling, and improves convergence.

Well-tempered metadynamics¹⁵² overcomes the aforementioned limitations by decreasing the bias heights as the simulation progresses, rather than using constant-height Gaussian functions (refer to Figure 6.4). The bias height W is reduced according to

$$W(k\tau) = W_0 \tau \exp\left(\frac{-V_G(S,t)}{\Delta T}\right), \quad (6.10)$$

with the initial Gaussian height W_0 and the deposition stride τ . The parameter ΔT has the dimension of temperature and represents the factor by which the Gaussian height is reduced. This way, the FES converges to

$$V_G(S, t \rightarrow \infty) = -\frac{\Delta T}{T + \Delta T} F(S) + \text{const}, \quad (6.11)$$

which allows the exact value to be obtained as the bias converges more smoothly. For $\Delta T = 0$, the bias is zero, resulting in an unbiased MD simulation. In contrast, when $\Delta T \rightarrow \infty$, Gaussians of constant height are deposited, thus yielding a standard metadynamics simulation. Between these two limits, the sample rate of the different conformations can be regulated by adjusting ΔT .¹²⁴

In practice, instead of specifying ΔT explicitly, a bias factor γ is commonly used, which is given as

$$\gamma = \frac{T + \Delta T}{T}. \quad (6.12)$$

T is the system temperature, while $T + \Delta T$ can be interpreted as the effective sampling temperature of the CVs. The bias factor γ controls how fast the bias heights decrease and must be chosen carefully. If it is too small, the bias decreases too quickly, causing the system to be trapped in the initial minimum. Conversely, an excessively large γ leads to slow bias decay, leading to problems similar to standard metadynamics, such as slow convergence and sampling of unphysical structures.¹⁵³

Multiple-Walker Metadynamics

One way to reduce wall-clock time is to parallelise several (well-tempered) metadynamics simulations, so-called walkers, to reconstruct the same FES. In this approach, known as multiple-walker metadynamics¹⁵⁴, the different walkers sample the CV space independently, but share a common biasing potential V_G , which is given by the sum of the Gaussians of all walkers. The accuracy of the metadynamics algorithm remains preserved.¹⁵⁴

7. Machine Learning

In recent years, machine learning (ML) methods have emerged as powerful tools in many areas of science. They enable the analysis of large datasets and the construction of predictive models for complex systems. This development was recognised by Nobel Prizes in 2024, both in Physics for fundamental contributions to machine learning using artificial neural networks,¹⁵⁵ as well as in Chemistry for advances in computational protein design and protein structure prediction.¹⁵⁶ Several ML approaches have been proposed in the context of atomistic simulations. In this work, the focus is on Δ -learning and high-dimensional neural network potentials (HDNNPs), specifically second-generation HDNNPs, which are therefore the focus in this chapter.

The accurate description of potential or free energy surfaces is essential for modelling chemical reactions, yet high-level QM methods often remain too computationally expensive for large systems or extensive sampling, motivating the development of ML based approaches. A prominent example is the Δ ML approach introduced by Ramakrishnan et al.¹⁵⁷, which aims to correct the errors of a faster and more approximate method (baseline method) by learning the difference to a more accurate reference method (target method), therefore combining the advantages of computational speed and accuracy. This strategy is effective because approximate methods already capture most of the underlying physics of the system. The remaining discrepancy, responsible for much of the computational cost of high-level quantum mechanical methods, can instead be learned using ML at a much lower cost.¹⁵⁷ This advantage has been explored by several groups to correct low-level methods, often semi-empirical approaches, towards high-level reference methods,^{158–163} but also to correct between two semi-empirical methods that differ in their underlying approximations.¹⁶⁴

To model the Δ ML correction for atomistic systems, HDNNPs have become a widely used approach. The first generation was restricted to smaller systems due to technical challenges such as poor scaling with system size, a fixed number of atoms and descriptors that did not satisfy translational, rotational and permutational invariance.¹⁶⁵ The second generation of HDNNPs, originally introduced by Behler and Parrinello¹⁶⁶, addresses these limitations by representing atomic interactions through local chemical environments, employing suitable atomic descriptors, and enabling efficient training-set generation through active learning.

Accordingly, the total energy of a system can be described as a sum of atomic energy contributions E_j^i , each determined by a neural network (NN) that depends on the local chemical environment of the atom. The short-range energy is given by

$$E_{\text{short}} = \sum_{i=1}^{N_{\text{elem}}} \sum_{j=1}^{N_{\text{atom}}^i} E_j^i, \quad (7.1)$$

with the number of chemical element N_{elem} , and N_{atom} as the number of atoms belonging to an element i . The short-range energy included all interaction within a cut-off radius R_c , capturing covalent interactions, as well as electrostatics and dispersion contributions.

In this framework, the atomic environments are described by atom-centred symmetry functions (ACSFs)^{166,167}, a type of descriptor that provides a representation of the local structure that is invariant with respect to translation, rotation, and permutation of identical atoms. When combined with the locality approximation, these descriptors enable the use of element-specific feed-forward NNs that predict atomic energy contributions E_j^i from the local chemical environment of each atom. The Cartesian coordinates $\mathbf{R}_j^i$ are transformed into symmetry function values G_j^i , which describe the local environment of the atom within the cut-off radius R_c . These symmetry function values serve as input to the NNs, while atoms of the same element share identical network parameters to ensure that chemically equivalent atoms are treated consistently. The sum of all atomic energy contributions yields the short-range energy E_{short} of the system (refer to Figure 7.1). In the case of Δ -learning, the energy difference ΔE is obtained. After training the network parameters, second-generation HDNNPs can be applied to systems with variable numbers of atoms, because the atomic NNs are evaluated independently for each atom and their energy contributions are summed to yield the total energy.

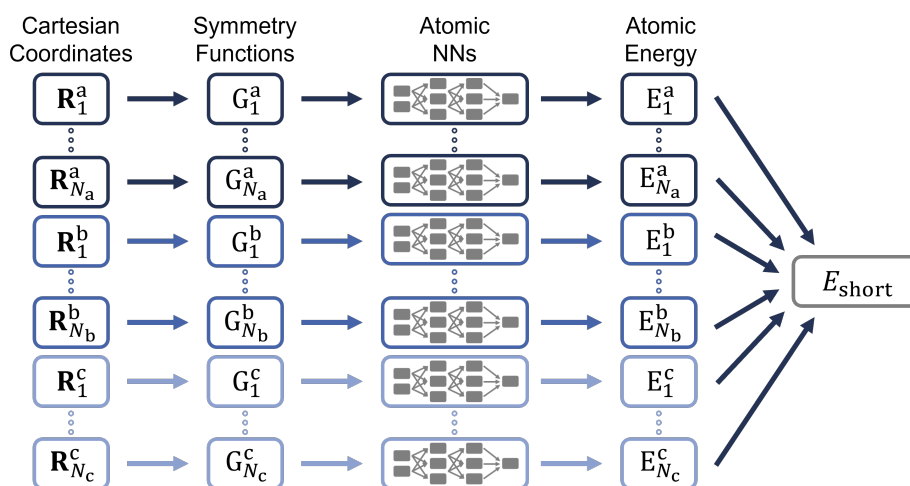


Figure 7.1.: Structure of a second-generation HDNNP as introduced by Behler and Parrinello¹⁶⁶, adapted from Reference [165]. The Cartesian coordinates $\mathbf{R}_i^\mu$ of each atom of the μ chemical elements a, b, and c, are transformed into symmetry function values G_i^μ , which describe the local atomic environment within a cut-off radius. They serve as input to element-specific NNs that predict atomic energy contributions E_i^μ . The sum of all atomic energies yields the short-range energy E_{short} of the system.

A central challenge in designing descriptors for atomistic ML potentials is ensuring invariance of the potential energy with respect to translation, rotation, and permutation of chemically equivalent atoms. ACSFs address this by providing a descriptor of the local atomic environment that is invariant by construction. Combined with the locality approximation, which restricts the description to neighbouring atoms within a cut-off radius R_c , this reduces the dimensionality of the representation and enables the atomic energy decomposition in Equation (7.1). Various functional forms can be used for the cut-off function $f_c(R_{ij})$, such as cosine,¹⁶⁶ hyperbolic tangent,¹⁶⁷ or exponential functions,¹⁶⁸ and these can be further refined by introducing an inner cut-off $R_{c_i} < R_c$ to focus the descriptor on chemically meaningful interatomic distances.

Radial and angular ACSFs provide a descriptor of the local atomic environment within the cut-off sphere of a central atom. The descriptors are constructed to decay smoothly at the cut-off radius, distinguish different local structures, and maintain fixed number of descriptor values independent

of the number of neighbouring atoms. In practice, a sum of Gaussian and cut-off functions is most commonly employed, which can be expressed as

$$G_{i,\mu}^{\text{rad}} = \sum_{j \neq i}^{N_{\text{atom}}} \exp(-\eta(R_{ij} - R_s)^2) f_c(R_{ij}), \quad (7.2)$$

with j as the number of atoms within the cut-off radius R_c . The parameter of η controls the spatial extent of the radial function, while the optional shift parameter R_s allows the Gaussian to be centred at different distances from the central atom, thereby focusing on neighbours within a specific distance interval (refer to Figure 7.2). Summing the contributions of all neighbouring atoms yields a single value that is independent of the number of atoms within R_c and can thus be interpreted as a continuous coordination number. As the radial symmetry function depends only on R_{ij} , separate sets of functions are defined for each element type, usually consisting of four to eight functions.

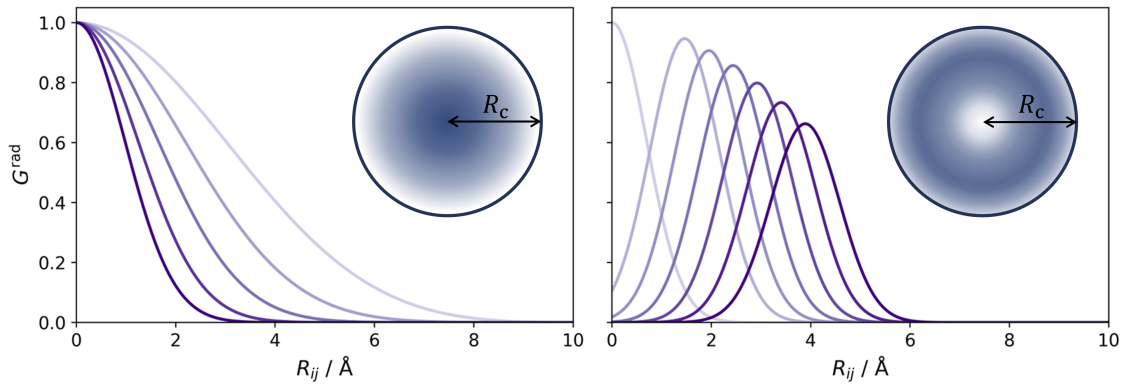


Figure 7.2.: Visualisation of radial symmetry functions G^{rad} , adapted from Reference [165]. Left: different η parameters illustrate the effect of the width parameter for a cut-off radius of $R_c = 10 \text{ Å}$ with $R_s = 0 \text{ Å}$. The values from light to dark purple are $\eta = 0.03, 0.08, 0.16, 0.3, 0.5 \text{ Å}^{-2}$. Right: $\eta = 1 \text{ Å}^{-2}$ is used while varying the shift parameter R_s with values $R_s = 0.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0 \text{ Å}$.

Radial functions alone cannot distinguish different coordination geometries, such as tetrahedral and square-planar arrangements, if the neighbour distances are identical. Therefore, angular ACSFs are employed, typically defined as

$$G_{i,\mu}^{\text{ang}} = 2^{1-\zeta} \sum_{j,k \neq i} (1 + \lambda \cos \theta_{ijk})^\zeta \exp(-\eta(R_{ij}^2 + R_{ik}^2 + R_{jk}^2)) f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}), \quad (7.3)$$

with the angle θ_{ijk} between the atoms i , j , and k . The parameter ζ controls the discrimination of different angular geometries, while $\lambda = \pm 1$ shifts the maximum of the cosine term to 0° or 180° (refer to Figure 7.3), and the prefactor $2^{1-\zeta}$ serves as a normalisation factor. Different values of η allow different radial regions to be captured. The cut-off functions $f_c(R_{ij})$, $f_c(R_{ik})$, and $f_c(R_{jk})$ ensure that only atom triplets within the cut-off radius are included.

An alternative to ACSFs are pair symmetry functions (PSFs), which describe the environment using pairs of atoms.^{169,170} While they can achieve a similar accuracy to ACSFs, their computational efficiency is lower because the number of atom pairs exceeds the number of atoms.¹⁶⁵

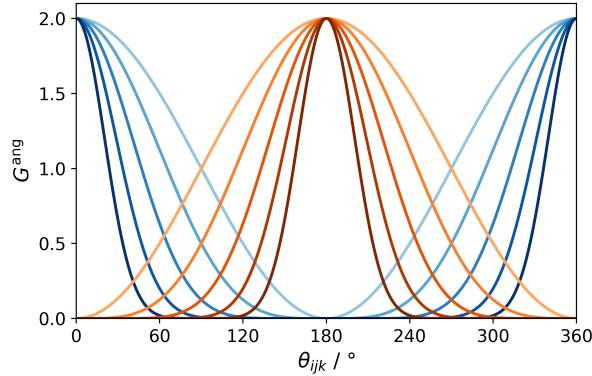


Figure 7.3.: Visualisation of angular symmetry functions G^{ang} , adapted from Reference [165]. The angular prefactor $2^{1-\zeta}(1 + \lambda \cos \theta_{ijk})^\zeta$ used in the angular ACSFs G^{ang} is shown in light to dark blue for $\lambda = 1$, and in light to dark orange for $\lambda = -1$. The employed exponents are $\zeta = 1, 2, 4, 8, 16$, with colours from light to dark corresponding to increasing ζ .

Despite the success of ACSFs, several challenges remain. For one, the parameters defining the spatial shape of the ACSFs must be chosen carefully. In practice, a typical set of starting parameters can often be chosen initially (e.g. as suggested in [165]) and can then be refined empirically, through fitting procedures,¹⁷¹ or by using optimisation approaches such as genetic algorithms.¹⁷² In addition, the number of symmetry functions scales unfavourably with the number of chemical elements, since all relevant element combinations must be represented explicitly. Atomic environments can also become highly dimensional, as e.g. 90 atoms within a cut-off radius of 6 Å already yield 270-dimensional atomic environments. To keep the descriptor size manageable, the cut-off radius is sometimes reduced, although this may reduce accuracy. One approach to address this issue is to use weighted ACSFs, which incorporate element information directly and reduce the number of required symmetry functions.¹⁷² Overall, ACSFs employ relatively few symmetry functions and therefore allow efficient HDNNP calculations, although achieving high accuracy often requires comparatively large training sets.¹⁷³

While the output of a second-generation HDNNP is energy, the atomic forces required for applications such as MD simulations can be obtained as the negative derivatives of the energy E_{short} with respect to the coordinate $R_{\mu\alpha}$ as

$$F_{\mu\alpha}^{\text{short}} = -\frac{\partial E_{\text{short}}}{\partial R_{\mu\alpha}} = -\sum_{i=1}^{N_{\text{elem}}} \sum_{j=1}^{N_{\text{atom}}^i} \frac{\partial E_j^i}{\partial R_{\mu\alpha}} = -\sum_{i=1}^{N_{\text{elem}}} \sum_{j=1}^{N_{\text{atom}}^i} \sum_{k=1}^{N_{\text{sym}}^i} \frac{\partial E_j^i}{\partial G_{jk}^i} \frac{\partial G_{jk}^i}{\partial R_{\mu\alpha}}. \quad (7.4)$$

This gives the force $F_{\mu\alpha}^{\text{short}}$ μ in direction $\alpha \in (x, y, z)$. The ACSFs are included through the chain rule. While atomic energies depend only on neighbours within the cut-off radius R_c , the forces can depend on atoms up to $2R_c$, since neighbouring atoms themselves interact with atoms in their own environments.¹⁶⁵

Part III.

Results

8. Δ ML Model for N-Phosphoryl Transfer Reactions in Gas Phase

Author Contributions The work in this chapter was carried out in collaboration with Christian Schmidt.¹⁷⁴ Christian Schmidt set up the neural network, coded the descriptors, training and inference workflows and end-to-end, versioned data pipeline. Lukas Petersen contributed to design choices, implementation details, and practical troubleshooting. Training data were generated jointly by Christian Schmidt and Lena Eichinger, covering metadynamics simulations, structure extraction, and single-point QM calculations. The division is as follows: datasets 1–3 (Schmidt and Eichinger), 4–8 (Eichinger), HK (Eichinger, metadynamics simulation created during her Master’s thesis¹⁷). An initial training of the network as well as the cross-validations were performed jointly by Lena Eichinger and Christian Schmidt. This chapter displays the results retrained by Lena Eichinger using an improved training set. Lena Eichinger performed the DFTB- Δ ML single-point calculations.

8.1. Introduction

Phosphates played a key role in the early evolution of life,¹ and phosphate-containing molecules are ubiquitous in all living cells today (refer to Figure 8.1).¹⁷⁶ In RNA and DNA, phosphate groups ensure the stable ligation of genetic information, while in phospholipids they constitute an essential component of cellular membranes.^{1,177} Additionally, phosphorylation is one of the most common and important types of post-translational protein modification (PTM). Reversible phosphorylation of amino acids can modify protein surfaces, allowing them to be recognised by phosphospecific binding domains in other proteins and thereby acting as an on–off switch for protein–protein interaction.^{2,176} It has been shown that approximately 30% of mammalian cell proteins are phosphorylated at any

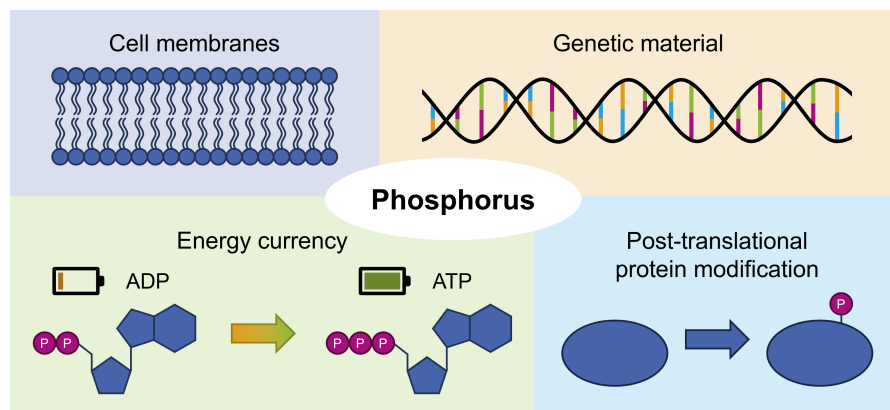


Figure 8.1.: Schematic representation of roles of phosphorus in biochemistry, adapted from [175]. Phosphorus is part of phospholipids, which are essential components of cellular membranes, and links nucleotides in the backbone of RNA and DNA. It is part of ATP, the biochemical energy currency, and is one of the most abundant PTMs.

given time, and that changes in the levels of protein phosphorylation coincide with diseases such as cancer, diabetes and rheumatoid arthritis.¹⁷⁸

Interestingly, phosphate esters are among the most inert compounds known, with half-lives for the breaking of phosphorus–oxygen bonds in the order of millions of years.¹⁷⁹ Nature turned this into a major advantage by using them as key components of adenosine triphosphate (ATP, refer to Figure 8.2), the biochemical energy currency. ATP is highly stable in aqueous solution at physiological pH, yet can be easily hydrolysed by the correct enzymes, creating an effective regulation.^{176,180,181} Through this, the energy-rich phosphoanhydride bonds in ATP can be used to drive energetically unfavourable metabolic reactions,^{1,180} making phosphates central to cellular energy metabolism and ATP the universal energy currency of all living organisms.

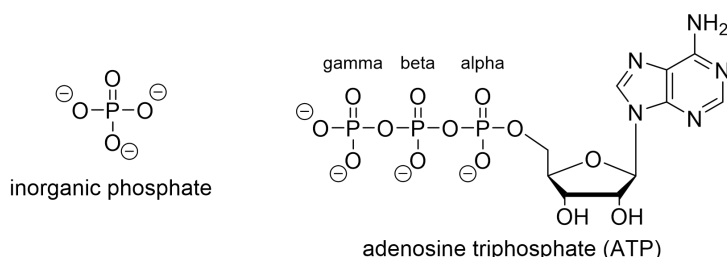


Figure 8.2.: Chemical structures of inorganic phosphate and organic adenosine triphosphate (ATP). In biological systems, phosphorus predominantly occurs in the form of inorganic phosphate. ATP is the universal energy currency of living systems and consists of a nitrogenous base (adenine), a sugar (ribose) and a triphosphate. Hydrolysis of the terminal γ -phosphate yields adenosine diphosphate (ADP), releasing free energy that can drive energy-intensive cellular processes.

In his 1987 review entitled ‘Why nature chose phosphates’¹⁸⁰, Westheimer identifies several reasons for the central role of phosphates in biochemistry. Besides the effective enzymatic regulation of phosphorylation reactions, phosphate groups are divalent and therefore well suited to act as linkers, for example in the backbone of DNA and RNA. Their negative charge furthermore helps to retain them within biological membranes.

Over 25 years later, Kamerlin et al. add in their review on ‘Why nature really chose phosphate’¹⁸², that electrostatic effects constitute the central factor in phosphate enzyme catalysis. To drastically reduce the exceptionally high barrier of phosphate hydrolysis, enzymes rely on charge redistribution upon passing from the reactant to the transition state. Electrostatic repulsion between negatively charged phosphate groups serves as a source of biological energy. This ionic character is also crucial for regulatory processes: by altering the magnitude and distribution of negative charge through phosphate hydrolysis (e.g. ATP to ADP), cells can exert precise allosteric control over proteins and drive complex signal transduction networks.¹⁸²

The topic of phosphorylation has been relevant for decades, as exemplified by the 1992 Nobel Prize in Physiology or Medicine, which was awarded to Edmond H. Fischer¹⁸³ and Edwin G. Krebs¹⁸⁴ for their research on reversible protein phosphorylation as a biological regulatory mechanism. Five years later, one half of the Nobel Prize in Chemistry was awarded to Paul D. Boyer¹⁸⁵ and John E. Walker¹⁸⁶ for their work on understanding the enzymatic mechanism underlying ATP synthesis. In 2013, Kamerlin et al.¹⁸² pointed out that the variety of reaction pathways that makes phosphate esters so versatile in catalysis also makes them exceptionally challenging to study. They call for the

need of computational investigations to achieve a detailed molecular understanding of phosphate biochemistry¹⁸², showing the significance of this topic to this day.

8.1.1. Kinases Catalyse Phosphorylation Reactions

Kinases and phosphatases are the enzymes that catalyse phosphorylation and dephosphorylation reactions, respectively. Kinases transfer a phosphoryl group from a high-energy species (usually ATP) to a specific substrate, while phosphatases remove phosphoryl groups through hydrolysis. Dysregulation of cellular signalling kinases through hyperactivity, malfunction, or overexpression is associated with numerous diseases, most notably cancer. This emphasises the importance of gaining a detailed mechanistic understanding of phosphorylation and kinase function, in order to develop more effective cancer therapies.²

During post-translational protein modifications (PTMs), usually the amino acids serine, threonine and tyrosine are phosphorylated (O-phosphorylation, whereby the phosphoryl group is attached to an oxygen atom to form acid-stable phosphomonoesters, refer to Figure 8.3). However, basic nitrogen-containing amino acids such as histidine, arginine and lysine can also be phosphorylated. This results in N-phosphorylation, where P–N bonds and acid-labile phosphoramidates are formed (refer to Figure 8.3). While most phosphorylation research has been focused on O-phosphorylation, it has become increasingly clear in the last years that N-phosphorylation PTMs play an important role in signal transduction and other regulatory processes.³

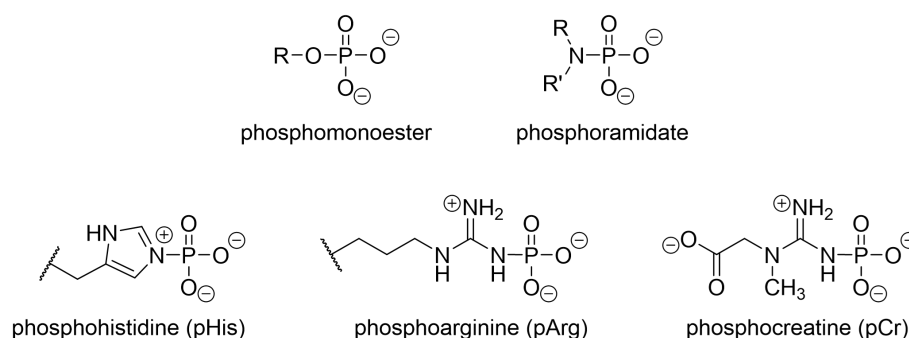


Figure 8.3.: Chemical structures of phosphomonoesters and phosphoramidates with examples. Phosphomonoesters are created through O-phosphorylation of amino acids such as serine, threonine and tyrosine. Phosphoramidates are formed through N-phosphorylation of a nitrogen-containing amino acid (such as histidine, lysine or arginine) or a phosphagen (such as creatine).

Prominent examples of kinases catalysing N-phosphorylation include histidine kinases (discussed in more detail in chapter 10), which play an important role in cellular signal transduction, and phosphagen kinases, which are involved in rapid buffering and redistribution of energy. The most well-known histidine kinases are part of two-component systems in bacteria, fungi and plants.⁴ These histidine kinases are located in the membrane and can detect environmental stimuli upon which a conformational change is induced. This, in turn, causes autophosphorylation, whereby a histidine residue on the kinase itself is phosphorylated. In a next step, this phosphoryl group is transferred to the second component, which can then regulate gene expression to adapt to the environmental signal.¹⁸⁷ Phosphagen kinases, on the other hand, catalyse the transfer of a phosphoryl group from ATP to guanidino compounds, thereby transforming biochemical energy from ATP into high-energy storage molecules known as phosphagens. This ensures temporal energy buffering and regulates

inorganic phosphate levels as well as intracellular energy transport.⁵ One of the most widely studied phosphagen kinases is creatine kinase (discussed in more detail in chapter 11), which catalyses the phosphorylation of creatine. Although it is also present in some invertebrate species, it is the sole phosphagen kinase found in vertebrates. It plays a key role in regenerating ATP in tissues with high and fluctuating energy demands, such as the brain and skeletal muscle.^{5,188} In contrast, arginine kinase catalyses the phosphorylation of arginine, and is the most widely distributed phosphagen kinase, occurring in all invertebrates and even several unicellular organisms, including bacteria.¹⁸⁸

8.1.2. Quantum Chemical Description of Phosphorylation

A quantitative understanding of phosphorylation reactions at the molecular level requires quantum chemical approaches to describe the changes in electronic structure that occur during the formation and breaking of bonds. Several computational studies have investigated the mechanisms of phosphoryl transfer, using different quantum mechanical (QM) and hybrid quantum mechanical/molecular mechanical (QM/MM) approaches to characterise reaction pathways, transition states and activation energetics. For example, QM/MM simulations at the density functional theory (DFT) level have been used to investigate GTP hydrolysis in the Ras–GAP complex¹⁸⁹, phosphoryl transfer in cyclin-dependent kinases¹⁹⁰, and phosphofructokinase from *Escherichia coli*¹⁹¹. Still, high-accuracy methods such as DFT remain computationally expensive, motivating the use of more efficient semi-empirical approaches such as density-functional tight-binding (DFTB). DFTB has been reparameterised to describe phosphoryl transfer reactions more accurately,^{16,192} and DFTB/MM simulations employing these parameters have successfully been applied to investigate the autophosphorylation reaction in a histidine kinase.¹⁵

Nevertheless, further refinements are required to achieve quantitative accuracy with DFTB. One approach is Δ -machine learning (Δ ML), which aims to combine the accuracy of DFT with the computational efficiency of DFTB by systematically correcting DFTB errors while retaining computational speed. This approach has already been shown to be effective, for example by Gómez-Flores et al. in the accurate correction of DFTB3 free energies to a B3LYP level of theory for thiol–disulphide exchange reactions.¹⁶⁰

8.1.3. Motivation and Aim of the Project

Motivated by these recent advances, this work aims to train and assess a Δ ML model for N-phosphorylation reactions. The goal is to improve the accuracy of DFTB simulations towards a DFT-level of accuracy on the fly, while retaining the computational efficiency of DFTB.

This approach relies on training a high-dimensional neural network potential (HDNNP), specifically a second-generation HDNNP¹⁶⁶, to predict energy differences (Δ -energies) between DFTB and DFT reference calculations based on chemical structures. By constructing training datasets that include multiple, chemically related systems, the transferability of the model can be assessed via leave-one-system-out cross-validation. Ultimately, the trained network should be able to predict Δ -energies, allowing it to correct DFTB energies even for systems not encountered during training.

To generate the required training data, metadynamics simulations are carried out in aqueous solution to sample relevant reaction pathways. Structures along these pathways are subsequently extracted and subjected to single-point energy calculations in vacuum, using DFT (ω B97M-V/ma-def2-TZVP) and DFTB3. The resulting Δ -energies serve as input for training the neural network, enabling accurate and efficient corrections of DFTB energies.

The workflow of this project comprises three main steps: generation of the training data, training and cross-validation of the Δ ML model in vacuum, and the evaluation of its transferability to aqueous and protein environments.

This chapter focuses on the generation of training data, as well as the training and cross-validation of the Δ ML model in the gas phase, while chapters 9 and 10 address the transferability of the trained model to aqueous and protein environments.

8.2. Methods

The chemical structures of the model systems were built using Avogadro 1.2.0^{193,194}. The systems were divided into a quantum mechanical (QM) region, where the chemical reaction takes place, and a molecular mechanics (MM) region, which accounts for the surrounding environment. The QM region was modelled using DFTB with the 3OB parameter set^{192,195,196}. Two specific reaction parameterisations (SRPs) were employed to describe the N-phosphorylation more accurately: OPhyd¹⁹², to be able to describe pentavalent phosphorus atoms, and PNmod¹⁶, to accurately represent biochemical phosphorylation and hydrolysis reactions involving P–N bonds. The QM region contained the model system consisting of a nitrogen and a phosphorus species and in some cases a magnesium cation (refer to Section 8.2.2). The MM region consisted of TIP3P water¹⁹⁷ and was described using the AMBER99SB-ILDN force field^{198,199}.

Energy minimisation, equilibration and production runs of the system were carried out using our local implementation of QM/MM in GROMACS 2020.2^{200–204} and DFTB+^{205,206}, which is available from a public repository^{207,208}. The energy minimisation was performed with the steepest descent algorithm. The subsequent equilibration was performed for 100 ps with a time step of 0.5 fs using the velocity rescaling thermostat⁵⁵ and Leap Frog integrator⁵³. Electrostatic interactions between the QM and MM regions were treated using the electronic embedding scheme. Periodic boundary conditions were applied within a simulation box of 3 nm and the electrostatic interactions were treated using the particle mesh Ewald method²⁰⁹. As a next step, an unbiased QM/MM simulation was performed for 100 ps with a time step of 0.5 fs to obtain the standard deviation of the two collective variables (CVs) in order to choose a suitable width for the bias potentials of the subsequent metadynamics simulation. For a more detailed discussion on the selection of the Gaussian width, refer to Section 8.2.1.

To enhance the sampling of the reaction pathway, two-dimensional well-tempered metadynamics simulations were performed, using the open-source, community-developed PLUMED library²¹⁰, version 2.5.1²¹¹. The distance between the atoms of the P–O and P–N bonds formed and broken in the reaction were used as CVs (refer to Figure 8.4). Bias potentials were introduced in the form of Gaussian functions with a height of 1.2 kJ mol^{-1} , a width of 0.004 nm, a deposition interval of

0.5 ps and a bias factor of 50, ensuring controlled exploration of the free energy surface. The time step was increased to 1 fs and each of the 16 walkers ran for 5–7.5 ns.

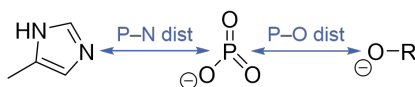


Figure 8.4.: Schematic representation of the collective variables (CVs) used for the well-tempered metadynamics simulations. The distances between the atoms of the P–O and P–N bonds are formed and broken during the phosphorylation reaction.

After the metadynamics simulations were completed, geometries were randomly extracted in a uniform distribution along the reaction pathways for subsequent quantum chemical single-point calculations. These single-point calculations were performed in vacuum using DFTB3/3OB with SRPs OPhyd and PNmod, as well as DFT with the ω B97M-V functional¹⁰¹ and the ma-def2-TZVP basis set^{107,109}. For better convergence in vacuum, Fermi temperatures of 300 K were chosen for the DFTB calculations, such that the resulting energies correspond to Mermin free energies rather than zero-temperature total energies. All DFT calculations were performed with ORCA²¹², version 5.0.5²¹³, while DFTB calculations were carried out using DFTB+^{205,206} in our local implementation²⁰⁸. For both DFT and DFTB energies, the atomisation energies

$$E_{\text{atomisation}} = E_{\text{total}} - \sum_i E_i^{\text{atomic}}, \quad (8.1)$$

were obtained by subtracting atomic energies (refer to Table A.1) from the total single-point energies. Atomic energies were initially obtained from B3LYP/aug-cc-pVTZ calculations and were updated in this work to ω B97M-V/ma-def2-TZVP in order to generate a more consistent dataset. Atomisation energies allow for a better comparison between different systems. The resulting Δ -energies, computed from the DFT and DFTB atomisation energies as

$$\Delta E = E_{\text{DFT}} - E_{\text{DFTB}}, \quad (8.2)$$

were used as training data for a second-generation HDNNP.

In addition to energies, forces were also calculated during the single-point calculations. Although the HDNNP was trained exclusively on Δ -energies, the quality of the correction can be assessed by analysing the corresponding forces, as they enter the equations of motion and ultimately determine the system’s dynamics.

Neural Network Specifications

A second-generation HDNNP, as introduced by Behler and Parrinello¹⁶⁶, was employed within a Δ -learning framework to correct DFTB energies towards DFT reference values. Atomic environments were represented using Behler–Parrinello symmetry functions, incorporating both radial and angular contributions. The radial symmetry functions used a cut-off radius of 10 Å with 7 R_s values ($R_s \in \{0.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ Å) and 5 η values ($\eta \in \{0.03, 0.08, 0.16, 0.3, 0.5\}$ Å⁻²), yielding 35 radial features per atom. The angular symmetry functions used a cut-off radius of 6 Å with $\lambda \in \{-1, 1\}$, $\zeta \in \{1, 2, 4, 8, 16\}$, and the same 5 η values, resulting in 50 angular features per atom. The dataset was randomly split into training, validation, and test sets in an 80:10:10 ratio. The validation set was used to monitor convergence and apply early stopping. Each network consists of

two hidden layers with 35 neurons each, with tanh activation functions. The output layer employed a linear activation function. The neural network was implemented in Python using the TensorFlow (version 2.15)^{214–216} and Keras²¹⁷ libraries and compiled using a mean-squared-error loss function. Training was carried out for up to 500 epochs using the Adam optimiser with a learning rate of $5 \cdot 10^{-5}$ and a batch size of 8, with early stopping applied if the validation loss did not improve for 100 consecutive epochs.

8.2.1. Pilot Simulation

To generate representative input structures for quantum chemical single-point calculations, geometries were extracted from the free energy surface (FES) generated by metadynamics simulations. The extraction was performed based on the positions of the deposited Gaussian hills, which were randomly selected to ensure a uniform sampling of the FES. The initial analysis of all Gaussians deposited during the first metadynamics simulations revealed a non-uniform scatter distribution. High-density regions were observed near the edges on all four sides, forming diffuse, band-like patterns that extend towards the centre as alternating regions of higher and lower point density (refer to Figure 8.5, bottom right). While these density accumulations are more pronounced near the boundaries, the interior patterns are less distinct. Notably, regions of lower density occur at the positions corresponding to the reactant and product states, implying that these configurations are underrepresented in the training data. This observation motivated a parameter study of the deposited Gaussians with the aim of lessening the band-like pattern and improving the uniformity of the sampling.

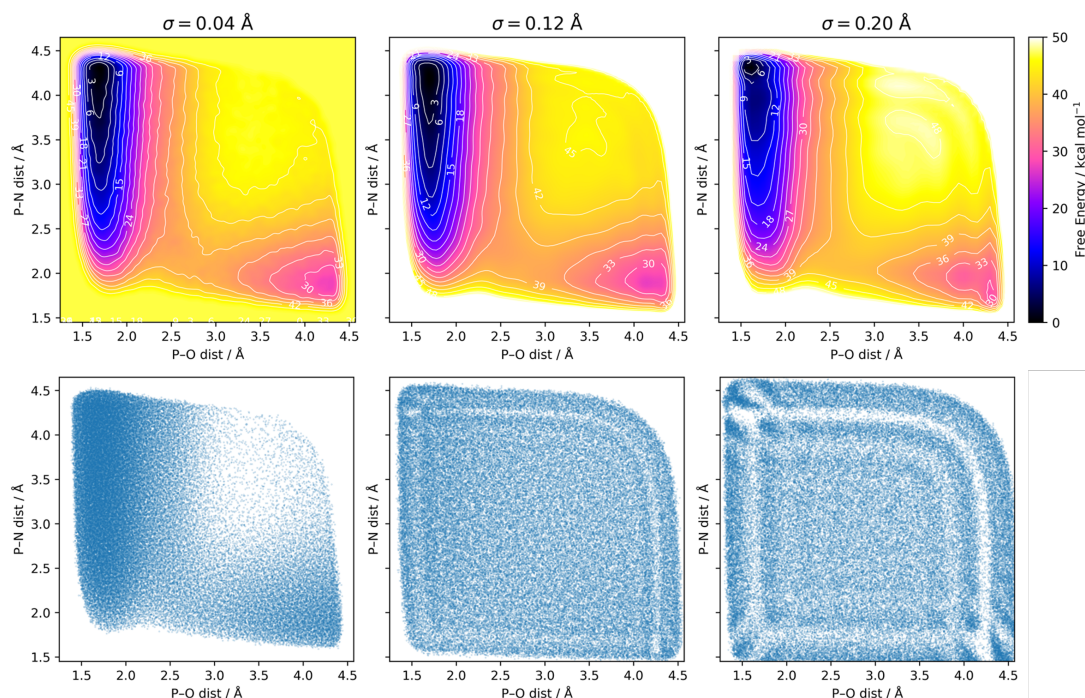


Figure 8.5.: Influence of the Gaussian width σ on the FES (top) and the Gaussian deposition distribution (bottom). Reducing the Gaussian width (right to left) suppresses the band-like pattern observed at bigger Gaussian widths. Energy values are displayed up to 50 kcal mol^{-1} , with areas exceeding this threshold and areas not explored during the simulation indicated in white. For $\sigma = 0.04 \text{ \AA}$, the surrounding area of the FES appears in yellow, as the FES could not be filled up to 50 kcal mol^{-1} within the same simulation time due to the smaller Gaussian width.

The parameters investigated were the Gaussian height and width. Varying the height affected the overall point density in the scatter plots. However, neither decreasing nor increasing the height eliminated the band-like pattern. Reducing the Gaussian width, on the other hand, progressively suppresses the pattern (refer to Figure 8.5, bottom). At the smallest width tested (0.04 Å, bottom left of Figure 8.5), no artefacts are visible, and the scatter distribution reproduces the shape of the FES. Furthermore, high point density in the scatter plot coincides with low free energy in the FES, and low density to high free energy regions, as expected for standard metadynamics simulations with constant Gaussian heights.

Regarding the origin of the pattern, it should be noted that the characteristic spacing between the bands aligns with the Gaussian width. Large-distance artefacts can be attributed to the artificial wall that keeps the two molecules in close proximity, while short-distance artefacts are the result of the natural P–N/P–O repulsion. As a side note on units, PLUMED-standard units (nm, kJ mol^{−1}) are used in the Methods section (refer to Section 8.2) for reproducibility, whereas Å and kcal mol^{−1} are used here for easier interpretation of the FES.

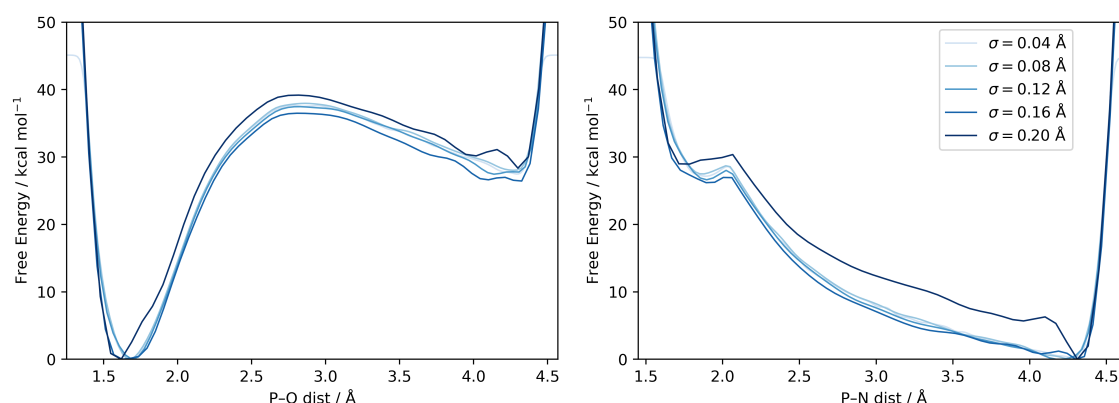


Figure 8.6.: Projection of the two-dimensional FESs onto the two CVs, simulated at different Gaussian widths σ . For the four smallest widths, the projections are qualitatively similar, with comparable positions and relative energies of the minima and maxima (corresponding to the transition states in the 2D FESs). In contrast, for the largest width tested ($\sigma = 0.20$ Å), the left minimum is shifted towards smaller distances in both projections. In addition, the energy of the maxima and ΔG (energy difference between the two minima) are increased, especially in the projection along the P–N distance. The most pronounced deviation is observed for the right minimum: the energy is increased on the left side of the minimum, and for the P–N projection, the minimum is noticeably steeper (~ 5 kcal mol^{−1}) compared to the other curves.

Analysing Figure 8.5 (top), it is evident that the FES is largely insensitive to the choice of Gaussian width. This is also observed in simulations with hill widths of 0.08 and 0.16 Å (refer to Figure A.1). Across all simulations, the free energy difference between reactant and product, ΔG , is approximately 28 kcal mol^{−1}, while the barrier height is consistently observed to be around 38 kcal mol^{−1}. The only exception is the largest width tested here (0.20 Å), where the pronounced band pattern in the scatter plot is also visible in the FES. Furthermore, an artificial stabilisation of the reactant minimum is observed, which is caused by an accumulation of Gaussians near the upper wall placed at large P–N and P–O distances. As a result, ΔG and the barrier height are increased by approximately 2 kcal mol^{−1}. This is also reflected in the projections of the FES onto the two CVs shown in Figure 8.6. The four smallest widths tested show a qualitatively similar behaviour, whereas for the largest width, the right minimum in the P–N projection (corresponding to the reactant state) is artificially steepened.

As demonstrated in the literature, this phenomenon arises in metadynamics simulations when Gaussians of finite width are employed, resulting in systematic errors at the boundaries of the CV space.^{218,219} These errors have been attributed to the inability of a sum of Gaussians to accurately reproduce discontinuities in the underlying free-energy profile. Although these boundary errors are initially small, they have been shown to increase progressively throughout the simulation. The increase of the artefact over time was also observed during this work and is visualised in Figure 8.7.

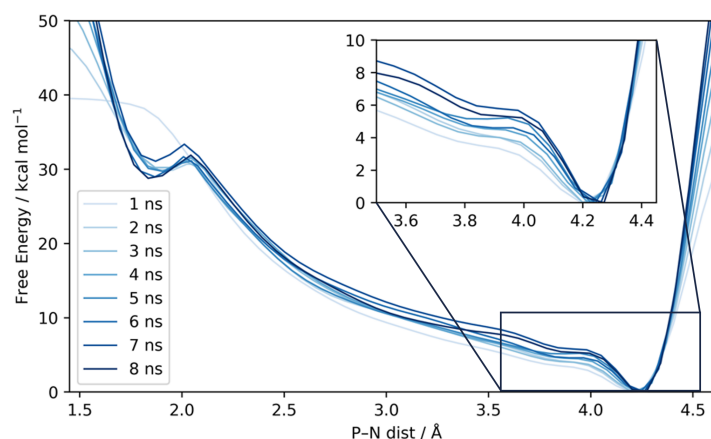


Figure 8.7.: Development of the 1D projection of the FES with $\sigma = 0.20 \text{ \AA}$ in the first 8 ns. The artificial steepening of the right minimum (corresponding to the reactant state) increases over time.

In summary, the FES is largely unaffected by the Gaussian width, with artefacts only emerging at comparatively large widths, substantially after appearing in the corresponding Gaussian deposition scatter plots. Consequently, the Gaussian width is not a critical parameter for the qualitative description of the FES in general simulations. However, for the generation of training data, a small enough Gaussian width is essential to ensure uniform representation of all structures. Therefore, the smallest Gaussian width tested ($0.04 \text{ \AA}$) was chosen for data generation simulations. This choice is consistent with literature recommendations, which suggest performing an initial unbiased simulation first and subsequently choosing a Gaussian width of approximately one-half to one-third of the standard deviation of the CV fluctuations.¹⁵¹ In the present case, the standard deviation obtained from unbiased simulations was $0.13 \text{ \AA}$, corresponding to a recommended Gaussian width of $0.04\text{--}0.07 \text{ \AA}$.

8.2.2. Investigated Systems

To generate the training set, several different model systems were selected, which vary both in the nitrogen nucleophile and in the leaving group of the phosphorus species (refer to Figure 8.8). Systems 1–4 were previously used in our group by Mayukh Kansari for the reparametrisation of the P–N repulsive potential for the SRP PNmod (system 4)¹⁶ and its benchmarking (systems 1–3)²²⁰. In systems 1–4, (methyl) imidazoles react with varying phosphate esters, each carrying a different leaving group to modulate electron-donating abilities. While these systems represent simplified model reactions, systems 5–8 are based on biologically relevant molecules. All of these systems contain methyl triphosphate coordinated by a magnesium cation to mimic the triphosphate moiety of ATP, while studying different nitrogen species acting as the phosphoryl acceptor. System 5 models the active site of a histidine kinase, in which a histidine residue is autophosphorylated by phosphoryl transfer from ATP. The arginine side chain in system 6 and creatine in system 8 model the active

sites of arginine and creatine kinases, respectively. While arginine serves as a phosphagen in invertebrates, the structurally related creatine fulfils the same role in vertebrates. System 7 features the side chain of δ -N-methyl-arginine, which, unlike arginine and creatine, is not a substrate for any known kinase but was included in the dataset to bridge the chemical space between these two systems. Similar to arginine, it lacks the carboxylate group present in creatine, while the additional methyl group at the guanidino moiety resembles the local chemical environment in creatine.

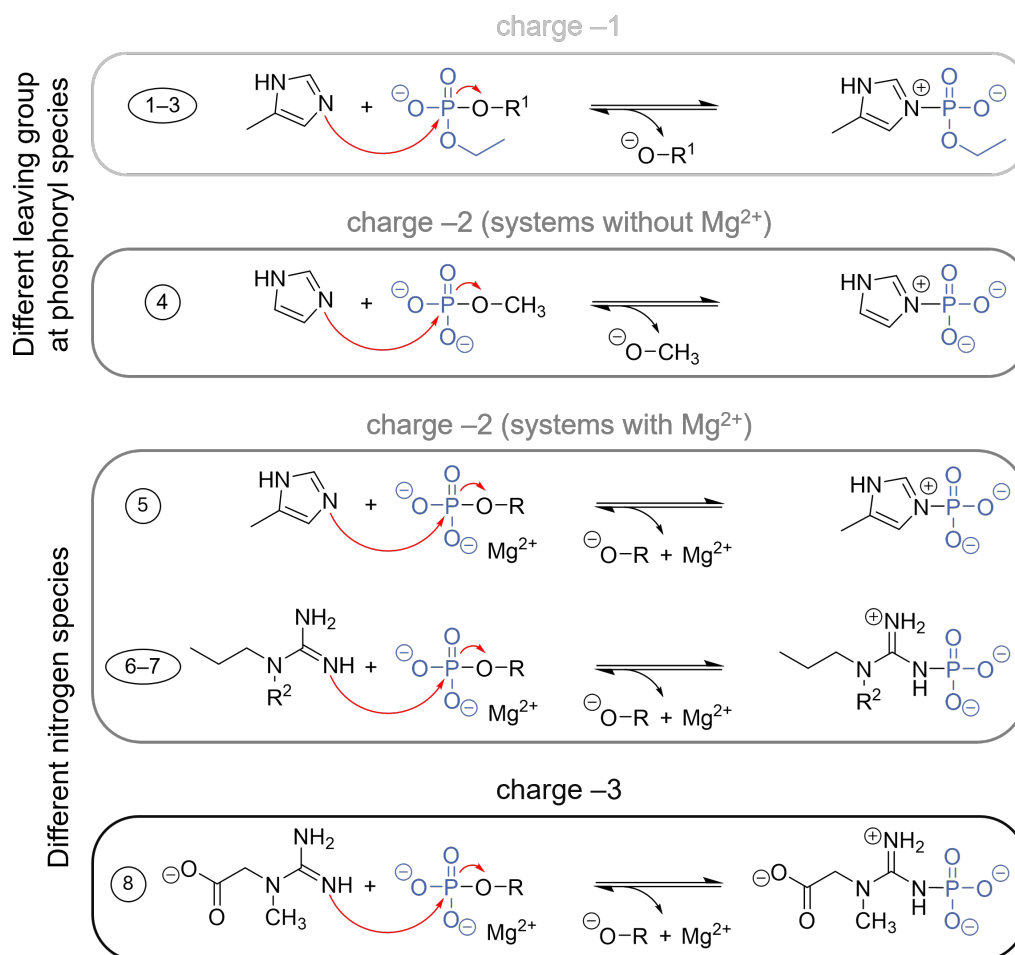


Figure 8.8.: Reaction equations of the simulated phosphorylation reactions with nitrogen as nucleophile, grouped by system charge. In systems 1–4, a (methyl) imidazole represents the nitrogen species while the phosphorus species (phosphate ester) varies in its leaving group (refer to Figure 8.9). In systems 5–8, the phosphorus species is represented by a methyl triphosphate (phosphoanhydride) and the nitrogen species is modelled to represent a 5) histidine, 6) arginine, 7) methylated arginine and 8) creatine.

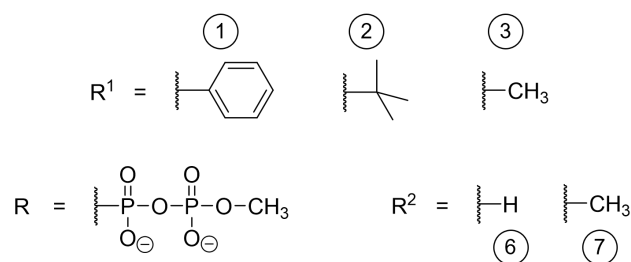


Figure 8.9.: Moieties for reaction equations abbreviated in Figure 8.8. R^1 is part of the leaving group in systems 1–3, varying their electron donating abilities. R in the phosphate species in systems 5–8 corresponds to a methyl triphosphate to represent ATP. Depending on R^2 , either an arginine side chain (system 6) or δ -N-methyl-arginine (system 7) fulfils the role of the nitrogen species.

8.3. Results and Discussion

In this section, the generation of the training data (refer to Section 8.3.1) as well as the training of the network and cross-validation (refer to Section 8.3.2) are discussed.

8.3.1. Simulation-Based Generation of Training Data

In this section, the simulation of phosphoryl transfer reactions and the generation of the corresponding training data are described. Metadynamics simulations were performed to generate trajectories from which representative structures could be extracted. The resulting free energy surfaces are analysed from a chemical perspective in Section 8.3.1.1. Along the sampled reaction pathways, structures were selected for subsequent single-point calculations in vacuum using DFT and DFTB. These single-point calculations are discussed in Section 8.3.1.2 and the resulting energy differences between DFT and DFTB are presented in Section 8.3.1.3. Forces were obtained as energy gradients during the single-point calculations, and the corresponding force differences between DFT and DFTB are analysed in Section 8.3.1.4.

8.3.1.1. Analysing Free Energy Surfaces

First, the free energy surface (FES) analysis is performed exemplarily for system 5 (refer to Figure 8.10), and then trends for systems 1–4 and 5–8 are discussed.

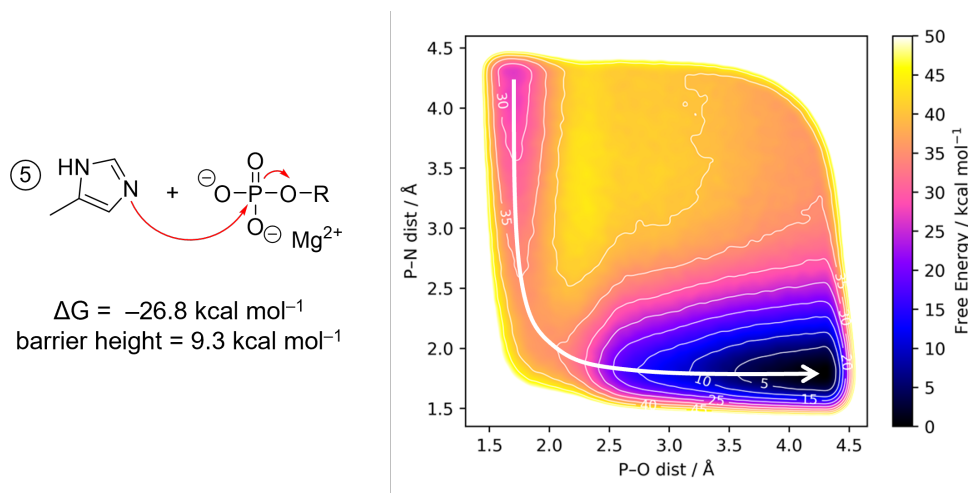


Figure 8.10.: Free energy surface (FES), free energy difference between reactant and products (ΔG), and barrier height for system 5 obtained from 2D well-tempered metadynamics. The collective variables (CVs) are defined by the distances between the phosphorus and oxygen atom, as well as the phosphorus atom and the attacking nitrogen atom (refer to Figure 8.4). Energy values are displayed up to 50 kcal mol^{-1} , with areas exceeding this threshold and areas not explored during the simulation indicated in white. The white arrow indicates the minimum energy path and the moiety R corresponds to a methyl diphosphate (refer to Figure 8.9).

Figure 8.10 shows the FES of system 5, which models the active site of a histidine kinase with a methyl imidazole (to represent the histidine residue) and methyl triphosphate (to represent the ATP cofactor) coordinated by a magnesium cation. The position of the minima of the reactant and product state (upper left and lower right corner, respectively) at approximately $4.3 \text{ \AA}$, are the result of artificial potentials introduced to keep the molecules in close proximity in the simulation box,

which is otherwise filled with water. The difference in free energy between reactant and product (ΔG) of $-26.8 \text{ kcal mol}^{-1}$ reveals the reaction to be strongly exergonic. The minimum energy path (indicated by the white arrow) illustrates that the reaction follows an S_N2 -like mechanism, which is indicated by the transition state being located at (1.88 | 2.11), corresponding to simultaneous short P–N and P–O distances. With $9.3 \text{ kcal mol}^{-1}$, the barrier height observed for system 5 is the lowest of all the model systems.

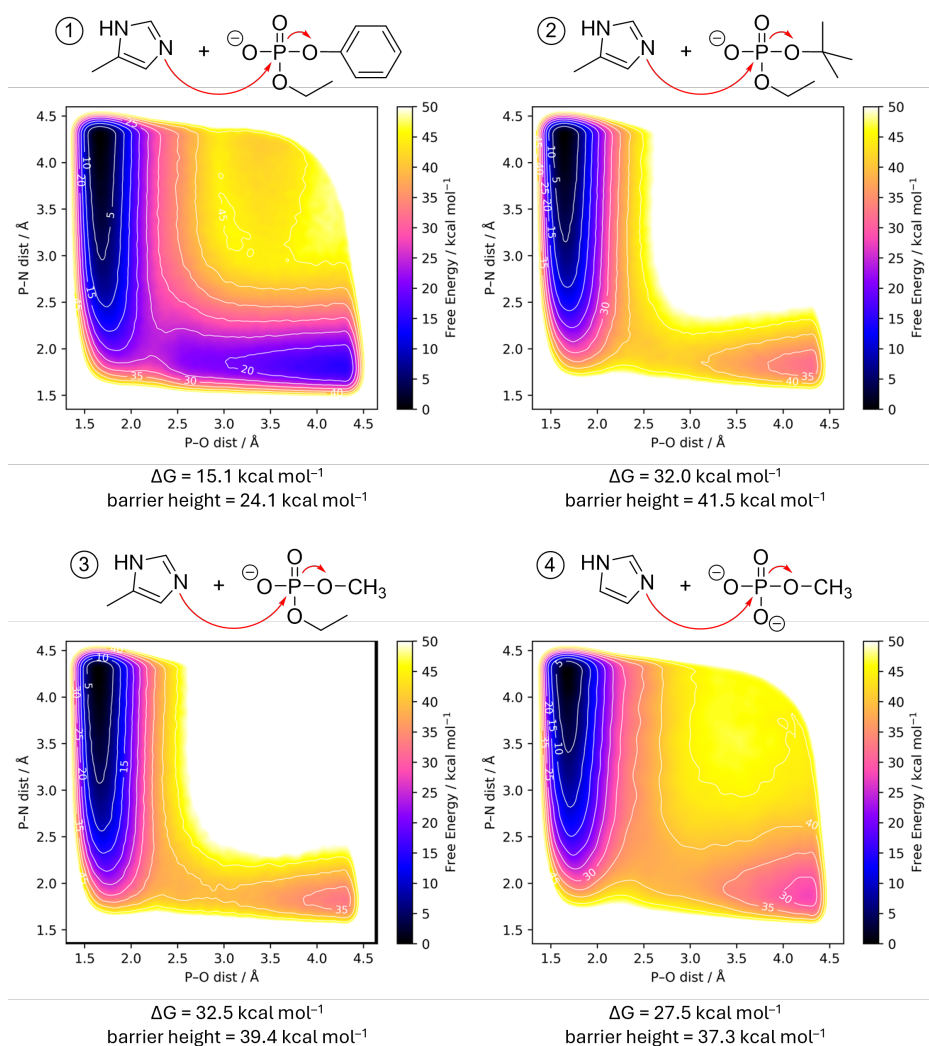


Figure 8.11.: Free energy surfaces (FES), free energy differences between reactant and products (ΔG), and barrier heights for systems 1–4 obtained from 2D well-tempered metadynamics. Energy values are displayed up to 50 kcal mol^{-1} , with areas exceeding this threshold and areas not explored during the simulation indicated in white.

Building on the detailed analysis of system 5, the FES of simplified model systems 1–4 are compared to identify systematic trends across different leaving groups (refer to Figure 8.11). System 1 with a phenoxide as leaving group exhibits the lowest energy barrier with $24.1 \text{ kcal mol}^{-1}$ and its product is identified as the most stable species in the context of these four reactions, with an energy difference of only $15.1 \text{ kcal mol}^{-1}$ between reactant and product state. Both the energies of the barrier and the product are approximately half as high compared to the FES of systems 2–4. This is to be expected as the phenoxide is resonance stabilised, and the negative charge is delocalised and stabilised by the polar TIP3P solvent. *tert*-Butoxide (system 2) and methoxide (system 3) on the other hand lack resonance stabilisation and are strong bases, making them poor leaving groups. This is visible in

the FES of systems 2 and 3 (refer to Figure 8.11), with barrier heights of 41.5 and 39.4 kcal mol⁻¹ and the product being 32.0 and 32.5 kcal mol⁻¹ higher in energy than the reactant, respectively. In system 4, the barrier is slightly lower at 37.3 kcal mol⁻¹, and the product lies 5 kcal mol⁻¹ lower in energy compared to system 3 – the only difference being the missing methyl group at the imidazole ring and the missing ethyl group at the phosphoryl group. A possible explanation is that the absence of the ethyl group (a weak electron donor via the inductive effect (+I)) makes the phosphoryl group more electron withdrawing, which could increase the leaving group ability. Similarly, the absence of the methyl group on the imidazole ring could reduce the electronic repulsion, thereby subtly influencing the reaction energetics.

It is particularly interesting to observe the thermodynamic contrast between system 4 and system 5, with the former undergoing an endergonic reaction and the latter exhibiting an exergonic reaction. System 4 involves an imidazole group and methyl monophosphate, while system 5 features methyl imidazole and methyl triphosphate-Mg²⁺. The product of system 5 is more stable by over

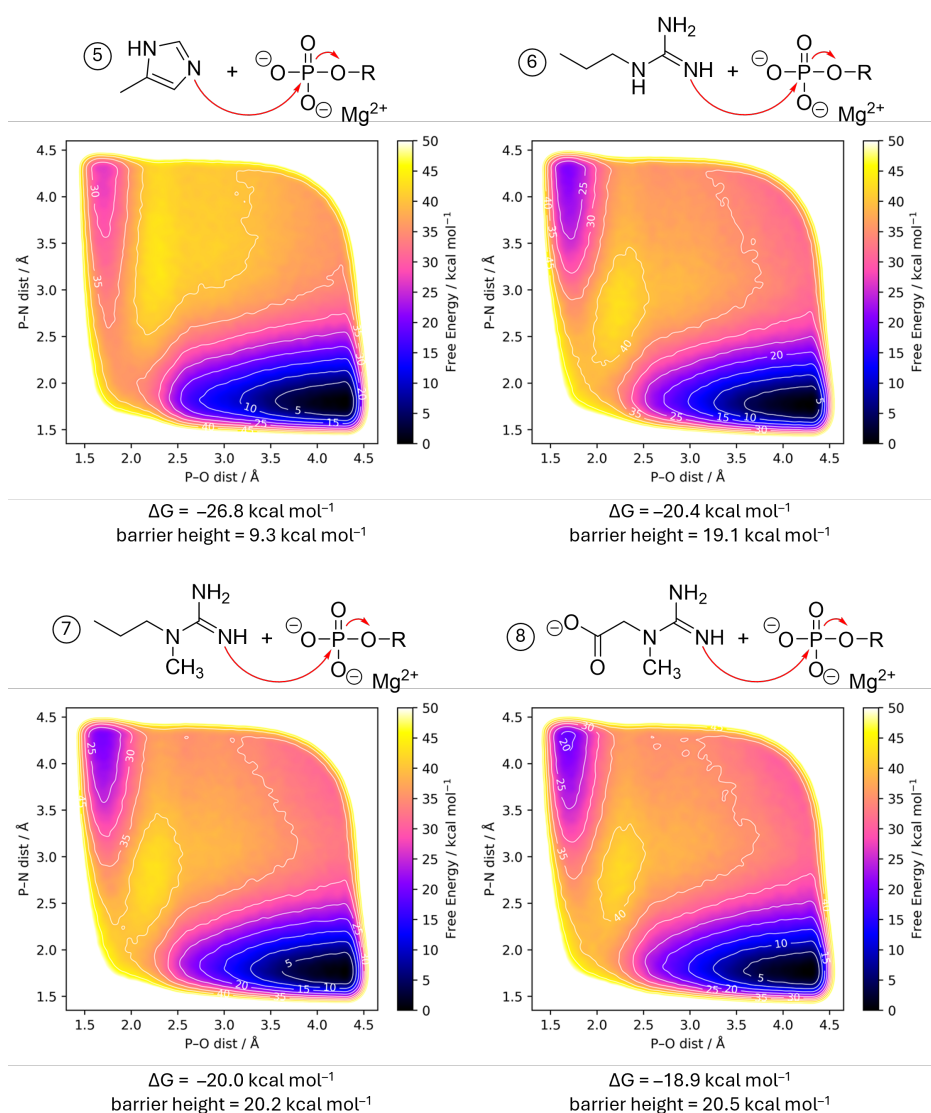


Figure 8.12.: Free energy surfaces (FES), free energy differences between reactant and products (ΔG), and barrier heights for systems 5–8 obtained from 2D well-tempered metadynamics. Energy values are displayed up to 50 kcal mol⁻¹, with areas exceeding this threshold and areas not explored during the simulation indicated in white. The moiety R is formed by a methyl diphosphate (refer to Figure 8.9).

50 kcal mol⁻¹ compared to the product of system 4, presumably because the diphosphate-Mg²⁺ in system 5 is highly stable, making it a good leaving group and resulting in a more thermodynamically favourable product. In contrast, the methoxide in system 4 is a poor leaving group, resulting in a higher ΔG and an energetically less favourable product.

Figure 8.12 shows the FES of systems 5–8 in comparison. Since systems 6–8 consist of a methyl triphosphate and a chemically similar nitrogen species (all guanidino-based), it is not surprising that their FES are also similar. The free energy differences (ΔG) between these systems vary by only 1.5 kcal mol⁻¹, and the barrier heights differ by 1.4 kcal mol⁻¹. Focusing on the overall thermodynamic nature, it is evident that the phosphorylation reactions in systems 5–8 are highly exergonic, with ΔG values ranging from -18.9 to -26.8 kcal mol⁻¹. As described in the comparison of systems 4 and 5, this can be attributed to the methyl-diphosphate-Mg²⁺-complex serving as an excellent leaving group, making the triphosphate reactions favourable and the product more stable. In contrast, the reaction in systems 1–4 are endergonic, with ΔG values ranging from 14.6 to 32.5 kcal mol⁻¹. Again, this difference can be explained by the poor leaving group abilities of these simplified model systems and the fact that a phosphate ester bond (P–O–C) is broken and not the phosphoanhydride bond (P–O–P) found in ATP.

Another interesting observation concerns the overall shape of the FES and its implications for the reaction mechanism. Systems with monophosphate consistently strictly follow an S_N2-like mechanism, characterised by a pentavalent phosphorus atom in the transition state. In contrast, systems involving triphosphate-Mg²⁺ show the possibility of an S_N1-like mechanism, with the formation of a trivalent metaphosphate intermediate stabilised by coordination to the magnesium cation. While the FES of system 5 shows a strong preference for an S_N2-like mechanism, the FESs of systems 6–8 indicate a similar probability for the two mechanisms, with system 7 even exhibiting a preference for the S_N1-like path (refer to Figure 8.12). However, the S_N1-like path is likely overstabilised due to the use of the SRP OPhyd, which was parametrised to reproduce the pentavalent transition state and its barrier height correctly but exhibits an error of approximately 10 kcal mol⁻¹ for the trivalent case.¹⁹² Consequently, the S_N1-like pathway is artificially stabilised, and the S_N2-like mechanism was accordingly considered as the minimum-energy path in this discussion. In addition, an S_N1-like mechanism would require the presence of a stabilised intermediate, which is not observed in the corresponding FESs (refer to Figure 8.13). Furthermore, experimental studies support a concerted reaction mechanism for phosphate monoesters.⁶ Nevertheless, the presence of the magnesium cation provides a chemical basis for an S_N1-like pathway, supporting it as a mechanistically relevant alternative.

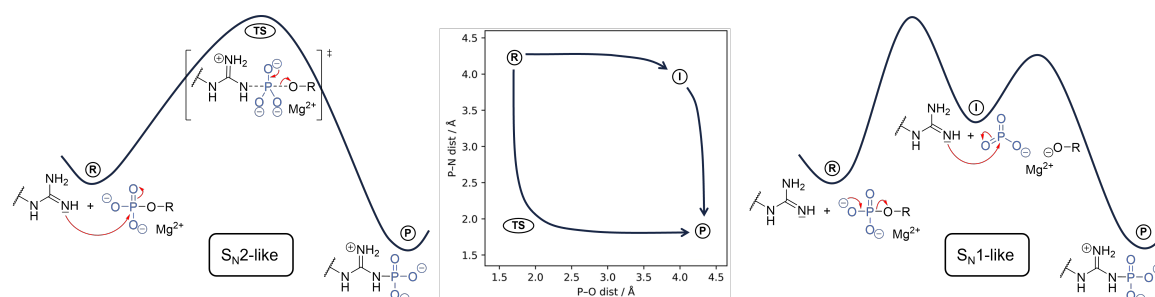


Figure 8.13.: Schematic representation of S_N2- and S_N1-like reaction mechanisms, illustrated for the phosphorylation of a guanidino group. S_N2-like reaction mechanisms are characterised by a concerted process with a pentavalent transition state at short P–N and P–O distances, whereas an S_N1-like pathway involves the formation of a stable intermediate, here represented by a metaphosphate species.

The detailed coordinates and relative free energies of the reactant, product, and transition state are summarised in Table 8.1. The table shows that the aforementioned minimum, which is introduced by artificial walls at a P–N distance of approximately 4.3 Å in the reactant state and a P–O distance of approximately 4.3 Å in the product state, is observed across all systems. The reactant coordinates indicate that the reactant state P–O bond length remains relatively consistent across all systems (ranging from 1.64 to 1.71 Å), which is expected since the immediate environment of the phosphorus atom is similar. In the product state of systems 1–5, which involve imidazole as the nitrogen species, the P–N distance is around 1.8 Å, with only system 4 showing a distance of nearly 1.9 Å. In contrast, systems 6–8, where the nitrogen species includes a guanidino group, show a slightly shorter P–N distance of 1.76 to 1.77 Å, suggesting a stronger bond. This is likely due to a combination of the guanidino group’s enhanced resonance stabilisation and stronger electrostatic attraction to the phosphoryl group.

Table 8.1.: Coordinates and energies of reactant, product and transition state. The first coordinate is the P–O distance, the second the P–N distance. The transition state energy is given relative to the reactant state energy. Coordinates are given in Å, energies in kcal mol^{−1}.

	Reactant state		Product state		Transition state	
	Coordinates	Energy	Coordinates	Energy	Coordinates	Energy
①	1.67 4.29	0.0	4.28 1.81	15.1	2.28 2.01	24.1
②	1.65 4.30	0.0	4.30 1.81	32.0	2.80 1.89	41.5
③	1.64 4.29	0.0	4.29 1.82	32.5	2.90 1.83	39.4
④	1.68 4.28	0.0	4.27 1.88	27.5	2.86 2.02	37.3
⑤	1.71 4.29	26.8	4.29 1.80	0.0	1.88 2.11	9.3
⑥	1.68 4.29	20.4	4.29 1.76	0.0	1.89 2.23	19.1
⑦	1.68 4.30	20.0	4.29 1.77	0.0	1.98 2.19	20.2
⑧	1.68 4.29	18.9	4.30 1.76	0.0	1.97 2.24	20.5

The barriers in systems 2–4 are the highest, as discussed previously, with system 1 having a barrier approximately half as high, at 24.1 kcal mol^{−1}. In systems 6–8, the barriers are consistent and lower, ranging from 19.1 to 20.5 kcal mol^{−1}, with system 5 showing the lowest barrier of only 9.3 kcal mol^{−1}. This trend aligns with expectations, especially given known values from protein systems, such as histidine kinase with 8 kcal mol^{−1} for the barrier of the chemical step¹⁵ and 16.3 kcal mol^{−1} for creatine kinase¹³. These values are only qualitatively comparable as the protein environment can affect the barrier heights in biological systems. The fact that the simulated barriers in water are in the same range as the ones in the protein might be due to the restraints applied during the simulation to keep the molecules in close proximity and at an optimal angle, thereby mimicking the role of an enzyme.

Looking at the coordinates of the transition state, it is evident that systems 2–4 exhibit longer P–O bond distances compared to the other systems. This is consistent with expectations that transition states with higher barriers (systems 2–4) show more significant distortion of the phosphorus atom and its surrounding bonds. Higher barriers often correlate with a more strained transition state, as the system requires more energy to overcome the transition state. However, this trend is not observed for the P–N bond distances, which do not display exceptional values. In fact, some of the systems

show even lower P–N bond distances in the transition state, suggesting that the P–N bond remains relatively stable throughout the reaction.

8.3.1.2. Single-Point Calculations in Vacuum

The structures for the single-point calculations were chosen to guarantee uniform sampling along the FES, giving equal representation to all regions rather than over-emphasising the minimum energy areas. Only structures along possible minimum energy paths were considered. This is why the full FES was sampled for systems 5–8, while for systems 1–4, the high-energy region in the upper right corner of the FES was excluded. Single-point calculations were performed in vacuum using DFT and DFTB.

DFT calculations generally converged well across all structures. In contrast, DFTB calculations showed convergence issues depending on the leaving group, particularly in the product state after dissociation. This is most pronounced in systems 1–4 since the leaving groups *tert*-butoxide, methoxide, and phenoxide are poorly stabilised in vacuum, leading to small HOMO–LUMO gaps or near-degenerate HOMO and LUMO energies. This, in turn, leads to self-consistent charge (SCC) convergence problems, which originate from the self-interaction error of the generalised gradient approximation (GGA)-based DFT on which DFTB is founded.^{93,221–224} Figure 8.14 (left) illustrates this for system 2 (*tert*-butoxide leaving group). At a Fermi temperature of 0 K, 3294 out of 12000 single-point calculation do not converge. Increasing the Fermi temperature allows fractional orbital occupations, which improves convergence in systems with small HOMO–LUMO gaps. It results in a smoothing of the electron distribution across orbitals, allowing fractional occupations of orbitals with similar energies. For system 2, increasing the Fermi temperature from 0 K to 300 K reduced the number of non-converged single-point calculations to 561 (refer to Figure 8.14, middle).

A downside of fractional orbital occupations is the risk of introducing a small artificial error (refer to Figure 8.14, right), as the resulting electron distribution does not reflect the true ground state. However, this error is minor, with deviations in single-point energies ranging from 0 to -0.8 kcal mol⁻¹ across all systems. Since this is within the uncertainty for P–O bonds calculated with the SRP OPhyd (~ 2.5 kcal mol⁻¹)¹⁹², the structures calculated at 300 K were included in the training set.

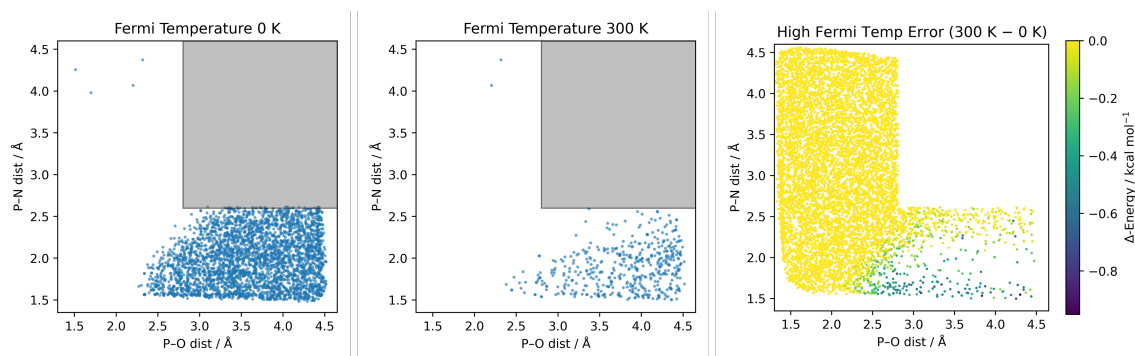


Figure 8.14.: Non-converged geometries of system 2 (*tert*-butoxide leaving group) at a Fermi temperature of 0 K (left) and 300 K (middle) and the error introduced by increasing the Fermi temperature (right). Grey areas indicate regions of the FES that were excluded from the single-point calculations as they are not part of the minimum energy path. Regardless of the chosen Fermi temperature, most convergence issues occur in the product state, where the leaving group is poorly stabilised. Increasing the Fermi temperature improves convergence due to fractional orbital occupations but also introduces a small energy error, which remains within the expected DFTB accuracy range.

For better-stabilised leaving groups, such as methyl diphosphate coordinated with a magnesium cation (systems 5–8), these convergence issues are less pronounced. For system 8 (creatine and methyl triphosphate–Mg²⁺), most DFTB calculations succeed even at a Fermi temperature of 0 K. An overview of single-point convergence and the error induced by increasing the Fermi temperature for all model systems can be found in Figures A.2 and A.3.

8.3.1.3. Δ -Energies Along the FES and Across Systems

The DFTB SRP OPhyd introduces an artificial overbinding of approximately 10 kcal mol^{−1} per P–O bond to lower the energy of the pentavalent S_N2-like transition state relative to the tetravalent reactant in order to improve the barrier height.¹⁹² In the original parameterisation, this correction was optimised using DFT with the B3LYP functional. In our case, we estimate the overbinding to be approximately 7 kcal mol^{−1} per P–O bond relative to ω B97M-V. Adding a magnesium cation increases the DFTB overbinding by an additional 21 kcal mol^{−1} ($\sim$ 7 kcal mol^{−1} per O–Mg²⁺ bond). The tendency of DFTB to overbind results in overstabilised interactions, leading to more negative and underestimated energies compared to DFT, and thus positive Δ -energy values.

This is visible in Figure 8.15 (left), where the pronounced difference between the box plots of systems 1–4 and 5–8 can be primarily attributed to the increase in the number of P–O bonds (eight additional P–O bonds in triphosphate compared to monophosphate) and the inclusion of a magnesium cation. Systems 4 and 8 exhibit the largest Δ -energies within their respective groups, which may be attributed to the higher system charge (systems 1–3: charge −1, systems 4–7: charge −2, system 8: charge −3). This stronger deviation suggests that DFTB describes highly charged systems less accurately. The larger standard deviation for systems 5–8 (refer to Figure 8.15, right) is caused by the inclusion of structures along the potential S_N1-like pathway, which exhibit larger deviations from the reference and increase the Δ -energy range (refer to Figure 8.16).

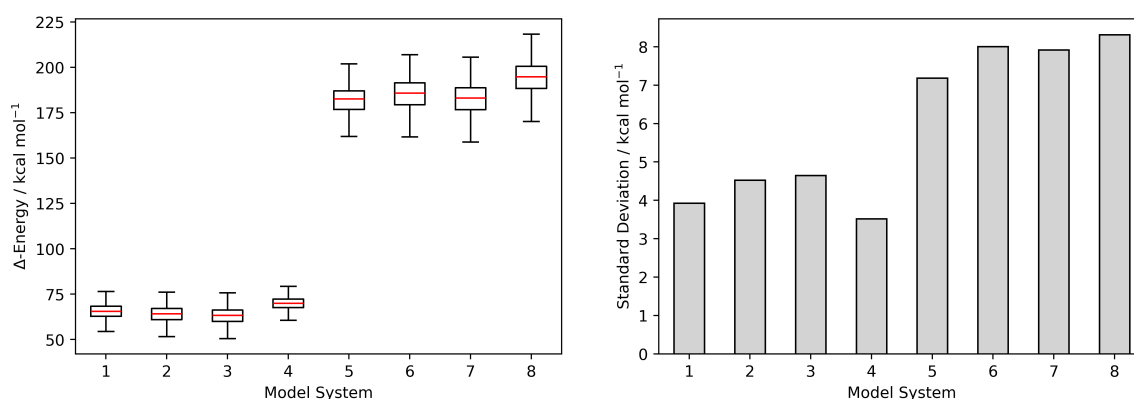


Figure 8.15.: Box plot representation of the Δ -energy distributions for the different model systems (left) and their standard deviations (right). The large separation between the box plots of systems 1–4 and 5–8 results from the inherent overbinding of the additional P–O bonds in DFTB. The increased standard deviation for systems 5–8 can be attributed to the inclusion of high free energy regions along a potential S_N1-like pathway, which results in a broader Δ -energy range.

Figure 8.16 shows an overview of Δ -energies across the FESs of the model systems. The largest values can be observed in the upper right corner of the FESs, where the phosphoryl group is neither bound to the leaving group nor the nitrogen nucleophile, but exists as a free metaphosphate. This region corresponds to an S_N1-like pathway with broken P–O bonds and not yet formed P–N bonds

(refer to Figure 8.13). The increased error in this area is presumably caused by DFTB using a minimal basis set, which limits its ability to capture the diffuse charge distribution of the free metaphosphate intermediate. Furthermore, although the SRP OPhyd reduces the error of the pentavalent transition state in the S_N2 -like pathway, a 10 kcal mol^{-1} error remains for the trivalent S_N1 -intermediate.¹⁹²

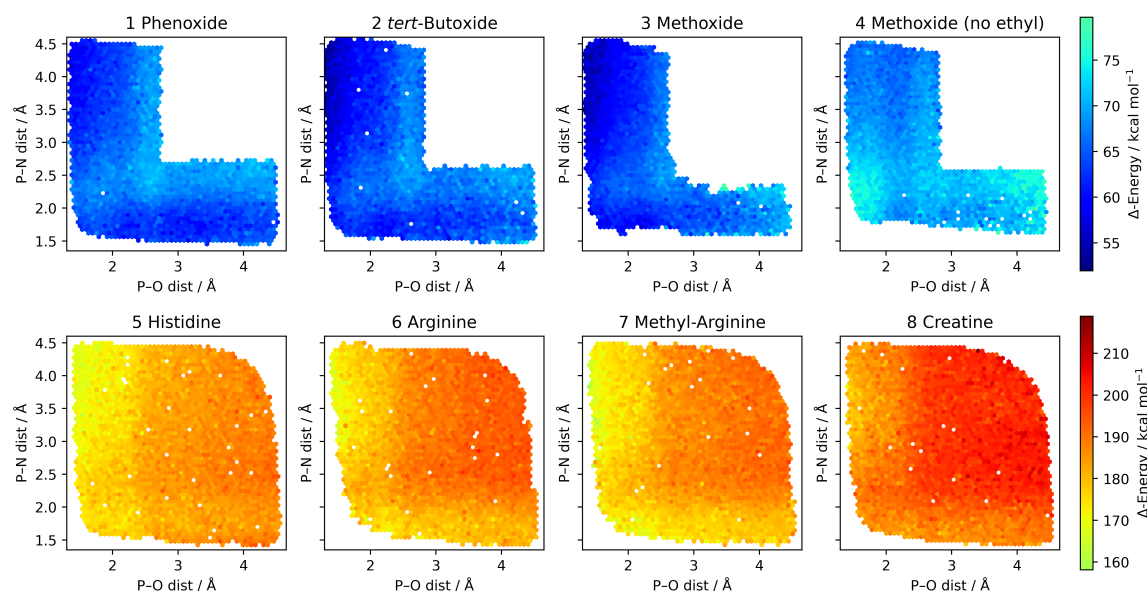


Figure 8.16.: Δ -Energies across the FESs of the model systems. The highest Δ -energy values are generally found in the upper right region of the FESs, where the phosphoryl group is not bound to the leaving group or nitrogen species but instead exists as a free metaphosphate. For systems 1–4, this region was excluded because only structures along the minimum energy path were considered. For systems 5–8, however, it was included in the training data because the respective FESs showed a potential S_N1 -alternative reaction pathway. Bins are coloured according to the mean value of the included Δ -energies.

For all systems, a slight increase in Δ -energy is observed in the area around $(2.0 \mid 2.5)$. In systems 5–8, this increase coincides with the position of the transition state, but in systems 1–4 the transition state is located at a different region. Notably, the same deviation is also visible for system 4, despite it being the same test system that was used during the PNmod reparametrisation. A related behaviour is already apparent in the original reparametrisation paper for the P–N repulsive potential.¹⁶ There, the DFTB FES is $10\text{--}20 \text{ kcal mol}^{-1}$ lower than the B3LYP FES in the area around $(2.0 \mid 2.5)$, indicating a stronger stabilisation at the DFTB level. This discrepancy was accepted as a trade-off when constructing the P–N repulsive potential in order to improve agreement at other relevant short P–N distances and obtain the correct structures and energetics for the transition and product states. As a result, an increased Δ -energy in this region is observed in all systems.

8.3.1.4. Force Differences Across Systems

While the Δ ML model was trained exclusively on energies, atomic force differences were evaluated to validate the quality of the correction (refer to Section 8.3.2.3). Hydrogen atoms consistently exhibit the smallest deviations among all atom types, indicating good agreement between DFT and DFTB for these atoms. While nitrogen atoms show slightly bigger deviations than oxygen and carbon atoms, phosphorus atoms display the greatest discrepancy between the two methods. Interestingly, the magnesium cation shows a deviation range comparable to that of hydrogen atoms,

but the entire distribution is shifted by approximately $170 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$, suggesting a systematic offset in force predictions for this atom type (refer to Figure 8.17).

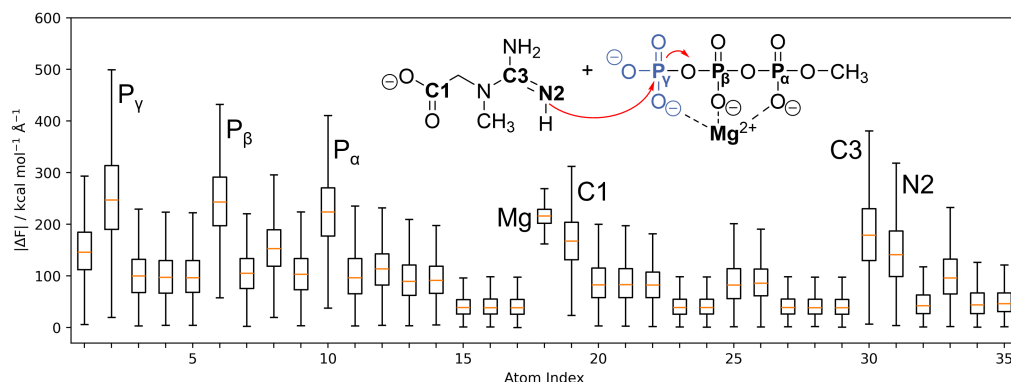


Figure 8.17.: Per-atom distribution of absolute force differences ($|\Delta F|$) for system 8. The deviations between DFT and DFTB are largest for the phosphorus atoms (especially the reacting P_γ). The force differences for magnesium show small deviations but a larger systematic offset. For creatine, the largest deviations are observed for the carboxylic carbon, the central guanidino carbon, and the reacting nitrogen.

For different systems, qualitatively similar trends can be observed, although the magnitude of the deviations varies. For the phosphorus atoms, the upper extreme of the box plot is similar for systems 1–3 (about $400 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$), slightly lower for system 4 (about $350 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$), and larger for systems 5–8, where deviations reach up to $500 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ (refer to Figure A.4). These larger deviations are particularly noticeable for the reacting phosphorus atom. In addition, in systems 6–8, the nitrogen atoms of the guanidino group also show higher deviations.

One potential biomolecular application is histidine kinase (HK). To generate input geometries from a protein environment, the relevant atoms, i.e. the histidine side chain, the triphosphate moiety of ATP, and the magnesium cation, were extracted from the QM region of QM/MM simulation of the HK autophosphorylation (*trans*-HK CpxA from *Escherichia coli* (PDB ID: 5LFK)²²⁵). These selected atoms are the same as in system 5, which was simulated in aqueous solution.

DFT and DFTB calculations of these geometries yielded a similar number of successfully converged structures, regardless of whether the structures originated from water or protein environments. Likewise, Δ -forces (refer to Figure 8.18, left and middle) reveal consistent qualitative trends across both systems: phosphorus atoms display the largest range of Δ -forces while those for the magnesium cation are shifted compared to other elements. Interestingly, the magnesium Δ -force distribution in the protein-derived structures is noticeably broader, suggesting that the protein environment may exert a particularly strong influence on the coordination dynamics of the magnesium cation.

This observation is consistent with the HK trajectory from which the structures were extracted. In this simulation, the active-site geometry did not permit full hexacoordination of the magnesium cation. Instead, an average of 4.5 coordinating oxygen atoms was observed, with the α - and γ -phosphoryl oxygens consistently involved in the coordination sphere. This behaviour is also reflected in the geometries used to generate input data for the network. The α -phosphoryl oxygen remains stably coordinated at around $2 \text{ \AA}$ to the magnesium cation, while the γ -phosphoryl oxygen is sometimes present at larger distances, particularly after phosphoryl transfer, when it is bound to the histidine and no longer coordinates the magnesium. The β -phosphoryl oxygen, on the other hand, is occasionally forced out of the coordination sphere and observed at distances of around $4 \text{ \AA}$

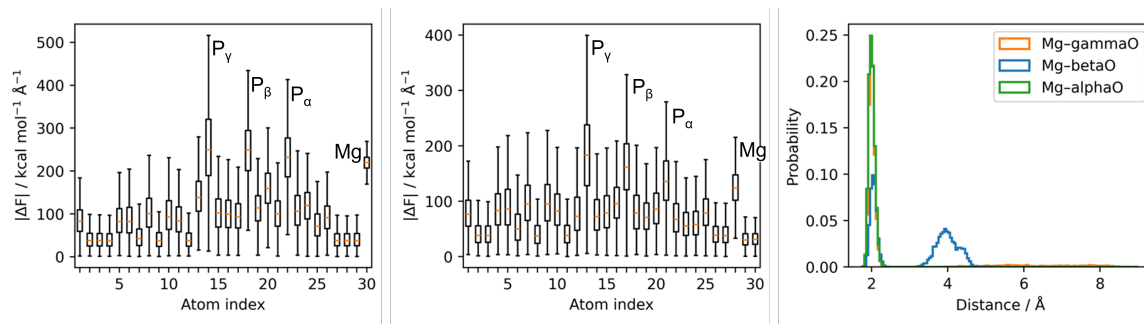


Figure 8.18.: Per-atom distribution of absolute force differences ($|\Delta F|$) for system 5 (left) and for the structures taken from the HK (middle). The distances between the magnesium cation and the coordinating ATP oxygen atoms from the HK structures are shown on the right.

from the magnesium cation (refer to Figure 8.18, right). In contrast, in aqueous solution, the magnesium cation remained hexacoordinated throughout the simulation by three ATP oxygen atoms and three water molecules. Unlike in the protein environment, where steric and geometric restrictions can prevent complete hexacoordination, the aqueous solution allows for greater structural flexibility.

8.3.2. DFTB- Δ ML – Training and Cross-Validations

In this section, the training of the network using the generated training data are discussed. To extend the dataset beyond sampling in aqueous solution, a metadynamics simulation of a histidine kinase (HK) was used to extract active-site geometries (histidine side chain, ATP triphosphate, and Mg^{2+}). Single-point calculations for this system were then carried out in vacuum, enabling a direct comparison to system 5, which was generated in aqueous solution. The HK system is therefore chemically identical to system 5, but differs in the environment in which it was simulated in (protein versus water). Furthermore, the HK was not biased along the P–N and P–O distance individually, but along their difference (P–O – P–N). This results in different structural distributions for the vacuum single-point calculations.

Training and testing on the same system is presented in Section 8.3.2.1 and subsequently applied to vacuum single-point calculations as a proof of concept, demonstrating that the network performs as intended. Cross-validation, in which the network is trained on a subset of systems and validated on an unknown system, is analysed in Section 8.3.2.2. The corresponding performance for vacuum single-point energies and forces is discussed in Section 8.3.2.3.

8.3.2.1. Single-System Case

In the single-system case, the network was trained and tested on the same reaction to illustrate that it performs as intended. The results are presented in Figure 8.19 (left), with the corresponding parity plots shown in Figure A.5. Across all systems, the coefficient of determination (R^2) shows a minimum of 0.969, indicating overall good predictive performance. For systems 1–4, the root mean square error (RMSE) does not exceed $0.699 \text{ kcal mol}^{-1}$. In contrast, systems 5–8 exhibit higher RMSE values of 0.851 – $1.039 \text{ kcal mol}^{-1}$, which can be attributed to the larger Δ -energy range and standard deviation in the corresponding training data (refer to Figure 8.15). This results from including the high free-energy regions at large P–N and P–O distances, as well as the larger

system size. Furthermore, since the second-generation HDNNP predicts the total Δ -energy as a sum of atom-wise contributions (refer to chapter 7), the RMSE increases with system size, even when the local per-atom accuracy is comparable. The correlation between RMSE and system size is visualised in Figure 8.19 (right). For HK, the R^2 value is comparable to that of system 5, while the RMSE is lower, presumably because the enzyme yields more optimal, lower-energy structures.

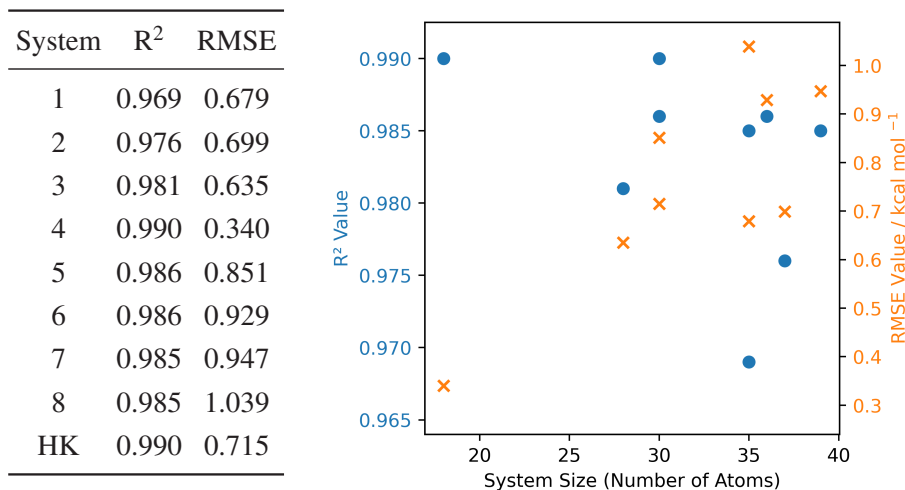


Figure 8.19.: Left: Overview of R^2 and RMSE values for single-system cases. RMSE values are given in kcal mol⁻¹. Right: R^2 and RMSE relative to the system size, showing that the RMSE increases with system size. For the respective parity plots, refer to Figure A.5.

As a proof of concept, Figure 8.20 exemplarily shows the effect of applying the single-system network trained on system 5 to vacuum single-point calculations of system 5. This demonstrates that the network is able to learn and reproduce Δ -energies for a given system. The mean Δ -energy is shifted from 50.39 kcal mol⁻¹ to -0.21 kcal mol⁻¹, showing that DFTB- Δ ML yields values that are very close to the DFT reference values, indicating improved accuracy. Furthermore, the standard deviation is reduced from 7.18 kcal mol⁻¹ to 0.47 kcal mol⁻¹, corresponding to a 93% reduction of the DFTB error and reflecting the improved precision.

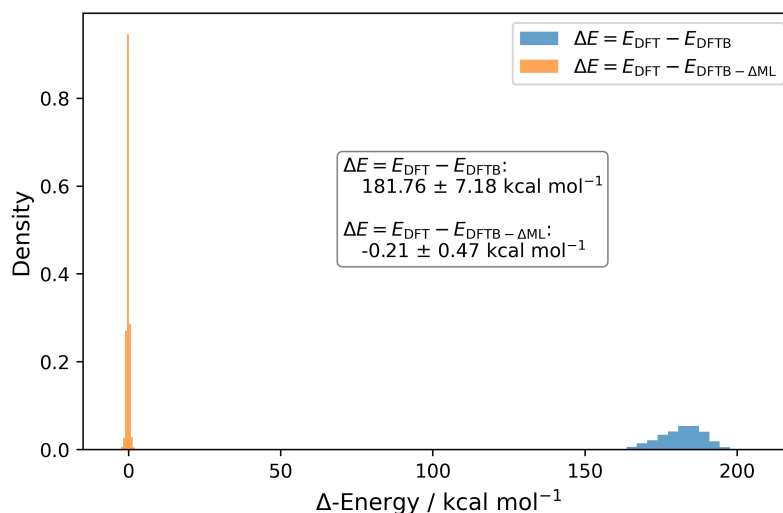


Figure 8.20.: Histogram of Δ -energies (DFT as reference) comparing DFTB and DFTB- Δ ML for the single-system case applied to vacuum single-point calculations of system 5. For DFTB- Δ ML, the mean Δ -energy is shifted towards zero, indicating that the corrected values closely reproduce DFT energies. The standard deviation is reduced by 93%. Together, this demonstrates improved accuracy and precision.

8.3.2.2. Cross-Validations

Cross-validations were performed using a leave-one-system-out scheme to evaluate the ability of the neural network to predict Δ -energies for systems that were not included in the training data. It was initially assumed that cross-validation performance would improve when the network was trained on systems with the same total charge as the validation system. Three different system charges were considered (-1 , -2 , -3), and systems with charge -2 were further divided into cases with and without Mg^{2+} (refer to Figure 8.8).

Systems with charge -1

For the three systems with a total charge of -1 (systems 1–3), parity plots are shown exemplarily for system 3 in Figure 8.21, while the corresponding plots for systems 1 and 2 are provided in Figure A.6. The comparison includes cross-validations trained either exclusively on the remaining systems with charge -1 or with system 4 (charge -2 and no Mg^{2+}) additionally included in the training set. Systems 5–8 were excluded because the presence of a magnesium ion introduces an additional chemical element, which would require an extended set of element-specific symmetry functions and thus prevents a direct comparison with models trained on Mg^{2+} -free systems.

Including system 4 in the training set does not lead to significant changes in the overall performance. A slight improvement in accuracy is observed when cross-validating against systems 1 and 3, whereas the accuracy decreases for system 2. However, the precision remains comparable across all cases. All cross-validation runs yield RMSE values of at least 2 kcal mol^{-1} , which is approximately three times larger than the errors observed in the single-system cases. Structurally, system 4 is closely related to system 3, differing only by the absence of the methyl group at the imidazole ring and the ethyl group at the phosphate, which results in an additional negative charge. Nevertheless, this modification does not substantially affect the overall performance, although an additional distribution in the predicted energies becomes visible when system 4 is included (refer to Figure 8.21), an effect that is also observed for system 2.

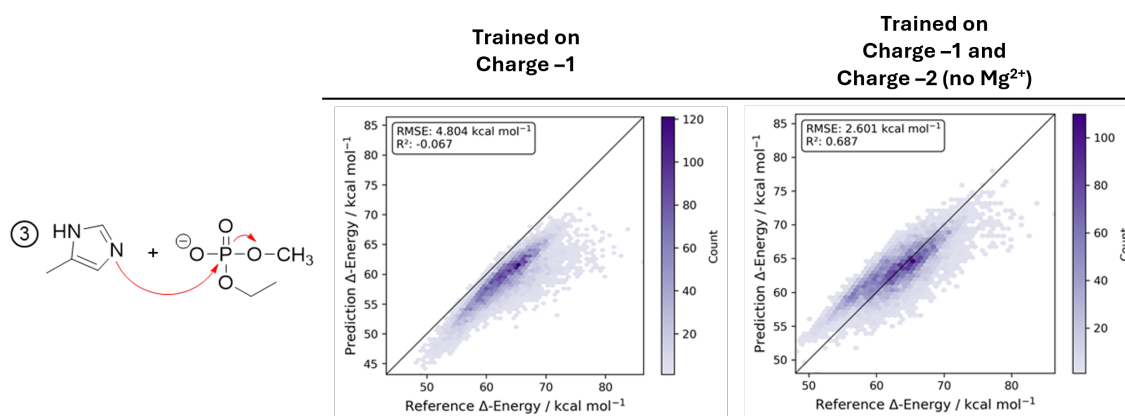


Figure 8.21.: Cross-validations for systems 3. For the training on charge -1 , systems 1 and 2 were included, while in the right plot, system 4 was included in the training set as well.

Systems with charge -2

System 4 is the only system with a charge of -2 and no magnesium cation, and can thus only be cross-validated against the three systems with charge -1 . The R^2 value is slightly negative and the RMSE is an order of magnitude larger compared to the single system case, which is also visible in the broad distribution in the parity plot (refer to Figure A.7).

Systems with charge -2 and Mg^{2+}

Systems 5–7 and HK all contain a triphosphate, a magnesium cation and a neutral nitrogen species, resulting in a net charge of -2 . In this case, the results from training on three of these systems, while cross-validating with the fourth system are compared to additionally including system 8 in the training set. System 8 has a net charge of -3 due to the presence of a carboxylate group in the nitrogen species. For systems 5, 6, and HK, no noticeable difference is observed between the two cross-validation setups (refer to Figure A.8). For these systems, the cross-validations yield R^2 values of 0.781–0.896 and RMSE values of 2.316–3.742, which are only 2.7–4.2 times larger than in the respective single system cases. The largest increases in RMSE values are mainly caused by systematic offsets in the Δ -energy predictions. These great cross-validation results can be explained with the strong structural similarity between the systems: systems 5 and HK contain identical molecular species extracted from simulations in different environments (water and protein), while systems 6 and 7 differ only in the presence of a methyl group at the central nitrogen atom N_C .

Notably, for system 7 (refer to Figure 8.22), the cross-validation trained on the other systems with charge -2 (systems 5, 6 and HK) performs exceptionally well, despite the network never having encountered an $\text{N}-\text{CH}_3$ bond, with an R^2 value of 0.993 and an RMSE of $2.051 \text{ kcal mol}^{-1}$. However, when system 8 is included in the training set, which explicitly contains the $\text{N}-\text{CH}_3$ bond, the corresponding cross-validation shows a systematic energy offset: the Δ -energies are predicted approximately 10 kcal mol^{-1} too high, which is presumably caused by the additional negative charge introduced by the carboxylate group in system 8.

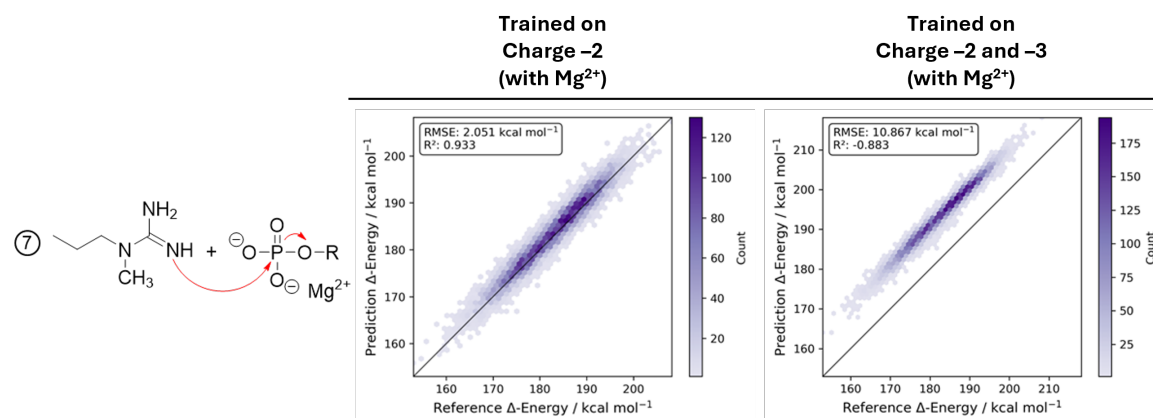


Figure 8.22.: Cross-validations for systems 7. For the training on charge -2 , systems 5, 6 and HK were considered, while in the second plot, system 8 was included in the training set as well.

Systems with charge -3 and Mg^{2+}

System 8 is the only system with a net charge of -3 and structurally closely related to system 7, differing only in the carboxylate group. Nevertheless, the cross-validation trained on systems 5–7 and HK (all with net charge -2) exhibits an offset with Δ -energies systematically predicted approximately 10 kcal mol^{-1} too low (refer to Figure 8.23). This systematic deviation can be attributed to the difference in Δ -energies between systems 7 and 8 of approximately 10 kcal mol^{-1} (refer to Figure 8.15), which is presumably caused by the additional negative charge introduced by the carboxylate group in system 8. Since the network is solely trained on Δ -energies without information on the system charge, it is unable to capture this energetic shift. Consequently, the similarity in chemical structures results in a systematic offset in Δ -energy predictions.

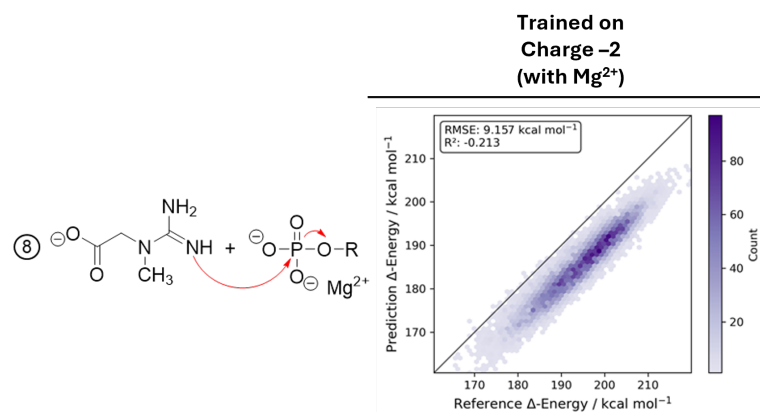


Figure 8.23.: Cross-validation for systems 8. The network was trained on systems 5–7 and HK, and cross-validated on system 8.

8.3.2.3. Overall Performances of DFTB- Δ ML: Energies and Forces

This section focuses on the influence of the Δ -energy prediction on energies and forces of DFTB- Δ ML calculations. The main focus is on systems 5 and 8, which represent the active sites of histidine and creatine kinase, respectively. For this, the single-point calculations of the training structures were recalculated with DFTB- Δ ML in vacuum.

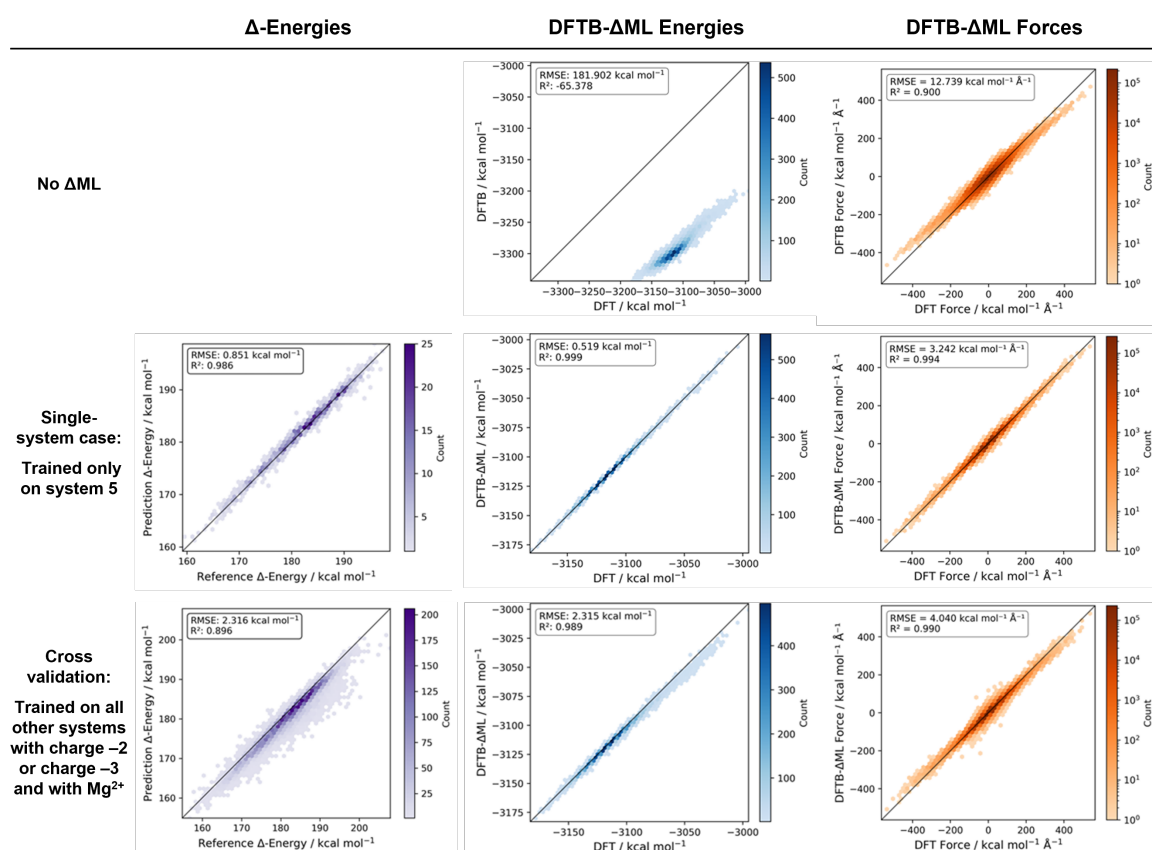


Figure 8.24.: Influence of single-system training and cross-validation on DFTB- Δ ML energies and forces of system 5. In the single system case, the error in DFTB- Δ ML forces is reduced by 75% compared to DFTB, in the cross-validation case, the error in forces is reduced by 68%.

Figure 8.24 shows the Δ -energies predicted by the network (left), as well as the energies and forces obtained from DFTB- Δ ML single-point calculations in vacuum (middle and right) for system 5. In the single-system case, the energies show near-perfect agreement, with an R^2 value of 0.999 and an RMSE of only $0.519 \text{ kcal mol}^{-1}$. In addition, the RMSE of the forces is reduced by 75% compared to DFTB. This behaviour is to be expected, as the network was trained on these exact structures. The cross-validation also shows very good performance. Although the Δ -energies exhibit a slight offset and a broader distribution, the DFTB- Δ ML energies deviate predominantly for a small number of high-energy structures. The error in forces is reduced by 68%, indicating substantial improvement. This result is also explicable, as the cross-validation training set includes the HK system, which contains the same molecular species although extracted from a protein environment instead of an aqueous solution as in system 5.

For system 8, the Δ -energy prediction shows a similar R^2 value in the single system case as observed for system 5 (refer to Figure 8.25) with near-perfect agreement between DFTB- Δ ML and the reference DFT energies ($R^2 = 0.999$) and a 72% reduction in force errors. Although the cross-validation results exhibit a systematic offset in both the Δ -energy predictions and DFTB- Δ ML energies, the RMSE in forces is considerably reduced by 48% compared to the uncorrected DFTB calculation. This demonstrates that the network can still substantially improve force predictions even when the Δ -energy predictions exhibit a systematic offset. Correcting forces is particularly important when focusing on free energy simulations, as they directly enter the equations of motion and ultimately determine the system's dynamics.

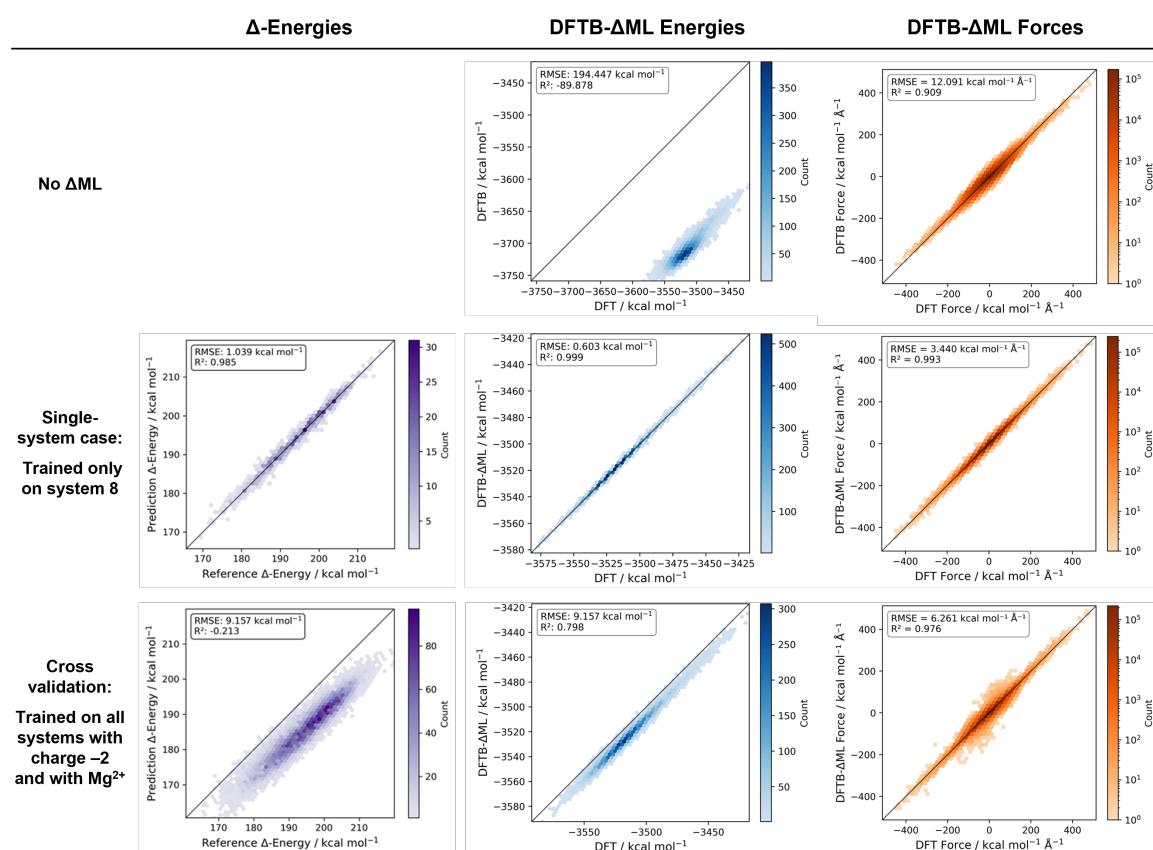


Figure 8.25.: Influence of single-system training and cross-validation on DFTB- Δ ML energies and forces of system 8. In the single system case, the error in DFTB- Δ ML forces is reduced by 72% compared to DFTB. In the cross-validation case, the Δ -energy predictions and DFTB- Δ ML energies exhibit a systematic offset to the reference. Nevertheless, the forces are considerably improved by 48% compared to uncorrected DFTB.

8.3.2.4. Computational Cost

When discussing ML corrections, it is important to consider their impact on the computational cost. Table 8.2 shows that the DFTB- Δ ML calculations are approximately 1.9–3.6 times slower than uncorrected DFTB, keeping them at a comparable cost and at least still three orders of magnitude faster than the respective DFT calculations. The relative computational cost increases with system size.

Table 8.2.: Mean CPU times per single-point calculation for the different model systems. The relative computational cost is defined as the ratio of the CPU times of DFTB- Δ ML to DFTB. Computation times are given in s. The system size corresponds to the number of QM atoms in each system.

System	Computation times			Rel. comp. cost	System size
	DFT	DFTB	DFTB- Δ ML		
1 Phenoxide	4682.62	0.258	0.934	3.62	35
4 Methoxide (no ethyl)	1307.58	0.291	0.556	1.91	18
5 Histidine	5469.51	0.329	0.867	2.63	30
6 Arginine	6270.79	0.321	1.000	3.12	36
8 Creatine	7149.78	0.329	1.009	3.07	35

8.4. Conclusions

Despite the central role of N-phosphorylation in many biological processes, its accurate simulation remains challenging. While the semi-empirical DFTB method provides a reasonable qualitative description, it is not always accurate enough for quantitative descriptions. DFT calculations, on the other hand, are too computationally demanding for extensive sampling. To overcome this gap, our research focuses on a Δ -learning approach based on a second-generation HDNNP trained on vacuum single-point Δ -energies, thereby correcting semi-empirical DFTB calculations towards DFT accuracy while retaining the computational efficiency of DFTB.

The training data were generated for a diverse set of chemical systems undergoing N-phosphorylation using two-dimensional well-tempered metadynamics simulations in aqueous solution, ensuring extensive sampling of the relevant phosphoryl-transfer reaction pathways. The dataset comprises four small test systems, including imidazole and monophosphate models with different leaving groups, as well as four biologically motivated systems featuring triphosphate- Mg^{2+} complexes with different imidazole and guanidino nitrogen species. For the biologically relevant systems, the FESs indicated the possible presence of an $\text{S}_{\text{N}}1$ -like reaction mechanism, which is considered to be over-stabilised at the DFTB level, yet still represents a chemically plausible alternative. These structures were therefore included in the training data.

From the metadynamics simulations, representative structures were extracted and subjected to single-point energy calculations in vacuum at both the DFTB3 and DFT (ω B97M-V/ma-def2-TZVP) levels of theory. In addition, structures extracted from the QM region of a histidine kinase were included alongside the systems sampled in aqueous solution. The resulting Δ -energies exhibit

the expected overbinding of approximately 10 kcal mol^{-1} associated with the OPhyd parameterisation. The largest Δ -energy differences are observed in regions corresponding to trivalent metaphosphate intermediates along an S_N1 -like pathway, consistent with the fact that the SRP OPhyd was parametrised with a main focus to improve the description of pentavalent S_N2 transition states.¹⁹²

These Δ -energies serve as input for the second-generation HDNNP and were analysed in two scenarios: interpolation, where training and testing were performed on the same system, and extrapolation, where the network was trained on a subset of systems and cross-validated on a system not included in the training set. In the interpolation case, the models achieve minimum R^2 values of 0.969 and RMSE values of 1 kcal mol^{-1} or lower, reaching chemical accuracy. The RMSE increases with system size, reflecting the fact that Δ -energies are predicted on a per-atom basis. When applied to vacuum single-point calculations, the Δ ML correction substantially improves both the accuracy and precision of DFTB energies, reducing their RMSE by 93%.

To assess the transferability of the trained networks, cross-validations were performed. In general, better cross-validation performance was observed for systems closer in chemical space, as this requires less extrapolation by the network. Systems with an additional negative charge exhibited a systematic offset in Δ -energies, indicating limitations in the underlying DFTB description of highly charged species. For structurally similar systems, this offset is reflected in the cross-validation results as a systematic offset in Δ -energy predictions. These observations highlight the importance of charge-sensitive training strategies, as an explicit treatment of different charge states would improve the transferability of Δ -energy predictions. For force calculations, however, systematic energy offsets are less critical, since they are obtained as gradients of the energy. Accordingly, despite systematic offsets in DFTB- Δ ML energies, the DFTB- Δ ML calculations reduce the initial DFTB force RMSE by 48%. This improvement in forces is particularly relevant for free-energy simulations based on metadynamics. Overall, the cross-validations demonstrate that the Δ ML approach can improve both energies and forces for systems not included in the training data. However, achieving good transferability would require a more accurate prediction of Δ -energies between different systems.

The relative computational cost increases with system size. For the considered systems, the DFTB- Δ ML single-point calculations are approximately 1.9–3.6 times slower than uncorrected DFTB calculations, still remaining at least three orders of magnitude faster than the corresponding DFT calculations. In summary, DFTB- Δ ML single-point calculations retain the computational efficiency of DFTB while bringing the accuracy closer to the DFT level.

So far, the Δ ML approach has been applied to single-point calculations in vacuum. As a next step, the method is extended to QM/MM metadynamics simulations in aqueous solution, with a particular focus on biologically relevant systems (refer to chapter 9). In particular, systems 5 and 8 are considered, as they represent reactions catalysed by two key N-phosphorylation enzymes, histidine kinase and creatine kinase. After this initial validation of the transferability from vacuum to water, the ultimate goal is to apply the vacuum-trained HDNNP to simulations in a protein environment, which is discussed in chapter 10.

9. Transferability of Δ ML Models from Vacuum to Water

Author Contributions The work in this chapter was carried out in collaboration with Christian Schmidt.¹⁷⁴ Christian Schmidt set up the neural network, coded the descriptors, training and inference workflows and end-to-end, versioned data pipeline. Lukas Petersen contributed to design choices, implementation details, and practical troubleshooting. The coordinates and partial charges of the atoms of the MM region are retrieved from QM/MM simulations based on a protocol from Katharina Spies.²²⁶ Lena Eichinger performed the single-point calculations in vacuum and water, as well as the metadynamics simulations and their analyses. The DFTB- Δ ML calculations were performed using the networks trained on the improved dataset from chapter 8.

9.1. Introduction

In chapter 8, a neural network was introduced that was trained and validated on single-point vacuum calculations, with the ultimate aim of application to protein environments. As an intermediate step, its transferability to aqueous solution is assessed in this chapter.

In recent years, several research groups have explored the combination of DFTB with ML techniques in order to improve its accuracy while retaining its computational efficiency. After the introduction of the Δ ML approach by Ramakrishnan et al.¹⁵⁷ in 2015, which learns the energy difference between two levels of quantum chemical calculations, the method has been applied in numerous subsequent studies. In an early work of neural network accelerated QM/MM free-energy simulations, Shen et al. developed a neural network extension of the HDNNP introduced by Behler and Parrinello¹⁶⁶, which enabled the prediction of high-level QM/MM potential energies along reaction paths determined with low-level QM/MM simulations.¹⁵⁸ This was followed by the work of Zhu et al.¹⁵⁹, Gómez-Flores et al.¹⁶⁰ and Bösel et al.¹⁶² who all applied the Δ ML approach with a Behler-Parrinello HDNNP in order to correct DFTB towards DFT. In conventional DFTB, the pairwise repulsive potential is fitted to reproduce DFT reference data, while remaining errors can be corrected using Δ -learning. A different strategy was recently proposed by Sandonas et al.¹⁶³, who replace the standard repulsive term entirely with a ML-based many-body potential.

While the aforementioned papers focus mostly on neutral or slightly polar test systems, the group of Darrin York concentrates on O-phosphorylation reactions. In their 2021 publication, they introduce a Δ ML model that corrects short-ranged QM/MM interactions to improve free energy profiles, including both internal QM contributions and QM/MM interaction energies and forces. Other than in previous studies, where a low-level method is corrected towards a high-level reference, they applied the Δ -learning approach between two semi-empirical methods, MNDO/d and DFTB2, which differ in their underlying approximations. Phosphoryl-transfer reactions relevant to RNA-cleaving enzymes

are employed as test systems. The reference system contains five oxygen atoms, while in five related systems one or two oxygen atoms are substituted by sulphur, enabling transferability between the test systems. While the CPU implementation results in a roughly fivefold increase in computational cost, a GPU implementation makes the ML overhead almost negligible.¹⁶⁴ A recent study by Giese et al. extends this framework by incorporating the graph neural network MACE to reproduce density functional theory reference data at the PBE0/6-31G* level for QM/MM simulations of transphosphorylation reactions in model systems designed to mimic RNA transphosphorylation.²²⁷ The objective was to enhance the accuracy of semi-empirical AM1/d calculations while retaining their computational efficiency. To this end, Δ ML corrections based on MACE were trained to recover hybrid functional QM/MM energies and forces. The resulting models successfully predict a stable pentacoordinated phosphorus intermediate on the free-energy surface, despite the absence of such bonding patterns in the training data. Free-energy surfaces are calculated using umbrella sampling with 200 ps of production sampling per window. In terms of performance, MACE-corrected QM/MM simulations on CPUs are approximately 30 times slower than standard AM1/d QM/MM, whereas GPU acceleration reduces this factor to 1.28.

9.1.1. Motivation and Aim of the Project

In chapter 8, a vacuum-trained Δ -learning approach was present with the aim of correcting semi-empirical DFTB towards DFT accuracy, with the ultimate aim of application in protein environments. As an intermediate step, its transferability to aqueous solution is assessed in this chapter.

First, single-point DFTB and DFTB- Δ ML energies and forces are compared to the DFT reference in vacuum and in an MM water electrostatic embedding, where the solvent is represented by point charges to mimic a QM/MM-like environment. Although the Δ ML model was trained only on Δ -energies, forces are obtained as the gradient of these energies and must therefore also be validated, since they directly govern the system's dynamics.

Finally, QM/MM metadynamics simulations are performed in aqueous solution, using either DFTB or DFTB- Δ ML as the QM method, to evaluate the impact of the correction in a dynamical setting.

9.2. Methods

In this chapter, systems 5 and 8 are investigated in water (refer to Figure 9.1), representing the active sites of histidine kinase and creatine kinase (refer to chapters 10 and 11), respectively.

Single-Point Calculations

The structures for the single-point calculations were extracted from metadynamics simulations from chapter 8 using MDanalysis (version 2.6.1)^{228,229}. The metadynamics parameters are provided in Section 8.2. For each system, 700 structures were extracted randomly and uniformly across the free energy surfaces (FESs) resulting from these metadynamics simulations.

Single-point calculations were performed using DFT (ω B97M-V/ma-def2-TZVP)^{101,107,109}, DFTB (DFTB3/3OB^{192,195,196} with OPhyd¹⁹² and PNmod¹⁶ SRPs), and the corresponding DFTB- Δ ML approach, in which the DFTB energies are corrected using the trained Δ ML model. As in chapter 8,

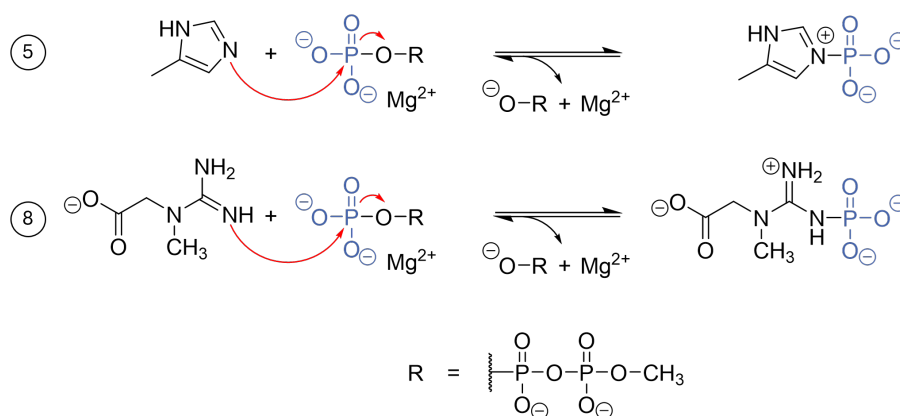


Figure 9.1.: Test systems considered for water simulations. Systems 5 and 8 were selected due to their biological relevance and their importance for this work.

DFT calculations were performed using ORCA²¹², version 5.0.5²¹³, while DFTB calculations were carried out using DFTB+^{205,206} in our local implementation²⁰⁸. For both DFT and DFTB energies, the atomisation energies (refer to Equation (8.1)) were calculated and used for comparison between the methods. Atomic force differences ΔF were computed as

$$\Delta F = F_{\text{DFT}} - F_{\text{DFTB}}. \quad (9.1)$$

In addition to vacuum calculations, a QM/MM electrostatic embedding was applied by placing point charges at the positions of the water atoms in the extracted structures to mimic the electrostatic environment of a QM/MM setup. For all calculations, energies and atomic forces were computed.

Metadynamics Simulations

The QM/MM metadynamics simulations were started from the equilibrated structures that were also used as starting points for the metadynamics simulations from which the single point structures were extracted. For the well-tempered metadynamics, the same parameters as in chapter 8 were used. However, the hill width was increased by a factor of two to 0.008 nm (compared to 0.004 nm used for data generation) for a computationally more efficient FES exploration.

9.3. Results and Discussion

In this section, the transferability of the vacuum-trained ΔML correction to aqueous solution is assessed by comparing energies and forces relative to DFT reference data, as well as FESs. The focus is on systems 5 and 8, which serve as model systems for histidine kinase and creatine kinase, respectively.

9.3.1. Model System for Histidine Kinase

System 5 (refer to Figure 9.1) represents a model system for histidine kinase, including the reacting molecules of the active site: the histidine side chain and triphosphate of ATP coordinated to a magnesium cation.

Single-Point Energies

For the total energies, the vacuum and aqueous calculations show different distributions prior to the application of the correction, as evident from the parity plots of DFTB against DFT (refer to Figure 9.2, top). After applying the Δ ML correction (refer to Figure 9.2, bottom), the vacuum energies show excellent agreement with DFT, yielding an R^2 value of 0.999 and an RMSE of $0.840 \text{ kcal mol}^{-1}$. The aqueous energies, on the other hand, exhibit a slight sigmoid-shaped deviation. Nevertheless, the overall agreement with DFT is improved significantly. The R^2 value increases to 0.945, while the RMSE, although remaining comparatively large at $11.276 \text{ kcal mol}^{-1}$, is reduced by 94% relative to uncorrected DFTB. The observed sigmoidal behaviour in water can be attributed to the different shapes of the vacuum and aqueous energy distributions prior to correction. In general, a stronger similarity between the uncorrected vacuum and water energy distributions facilitates more effective transferability of the correction.

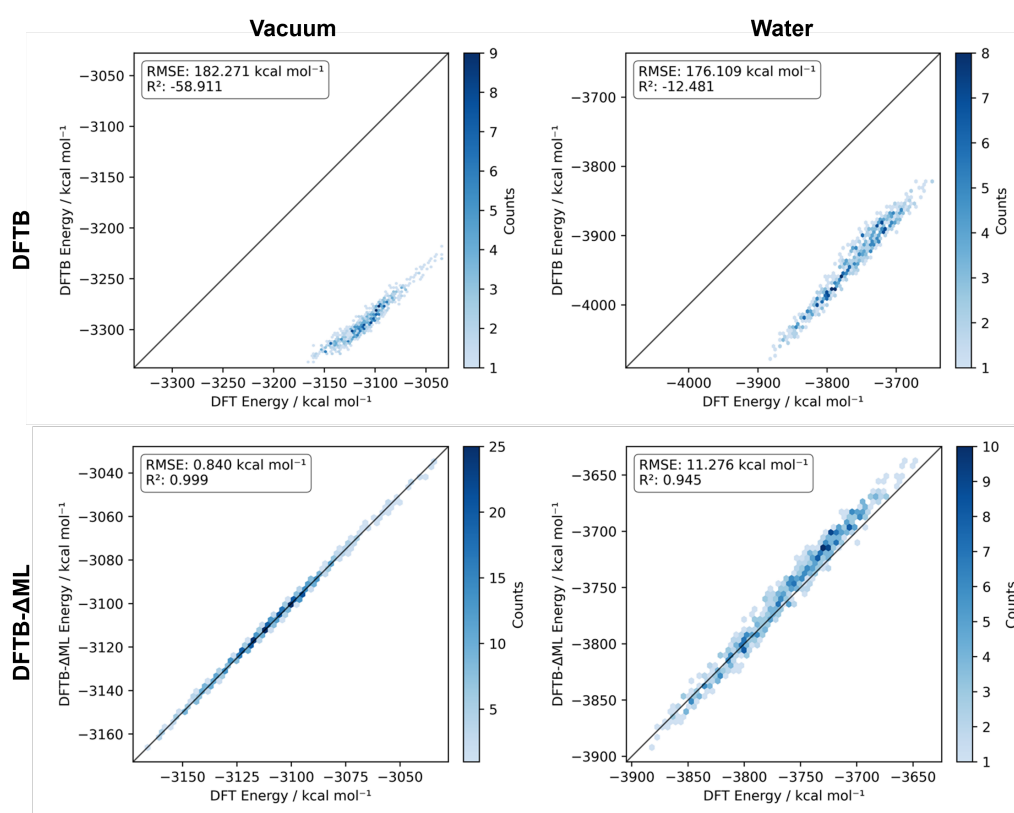


Figure 9.2.: Parity plots illustrating the transferability of the vacuum-trained Δ ML correction to water for the total energies of system 5. Depicted are total energies obtained with DFTB (top row) and DFTB- Δ ML (bottom row) compared against DFT reference energies in vacuum (left) and aqueous solution (right).

Single-Point Forces

Figure 9.3 shows how the network improves forces in vacuum and aqueous solution. In vacuum, the agreement between DFTB- Δ ML and DFT is strong, with an R^2 value of 0.993 and an RMSE of $3.472 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$, corresponding to a reduction of 74% relative to uncorrected DFTB. In aqueous solution, the agreement is weaker, yielding an R^2 of 0.828 and an RMSE of $10.665 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$. Although the error remains comparatively large, it is still reduced by approximately 40% relative to the uncorrected calculations. The improvement in the forces is particularly important, as they enter the equations of motion and thus directly govern the molecular dynamics simulations.

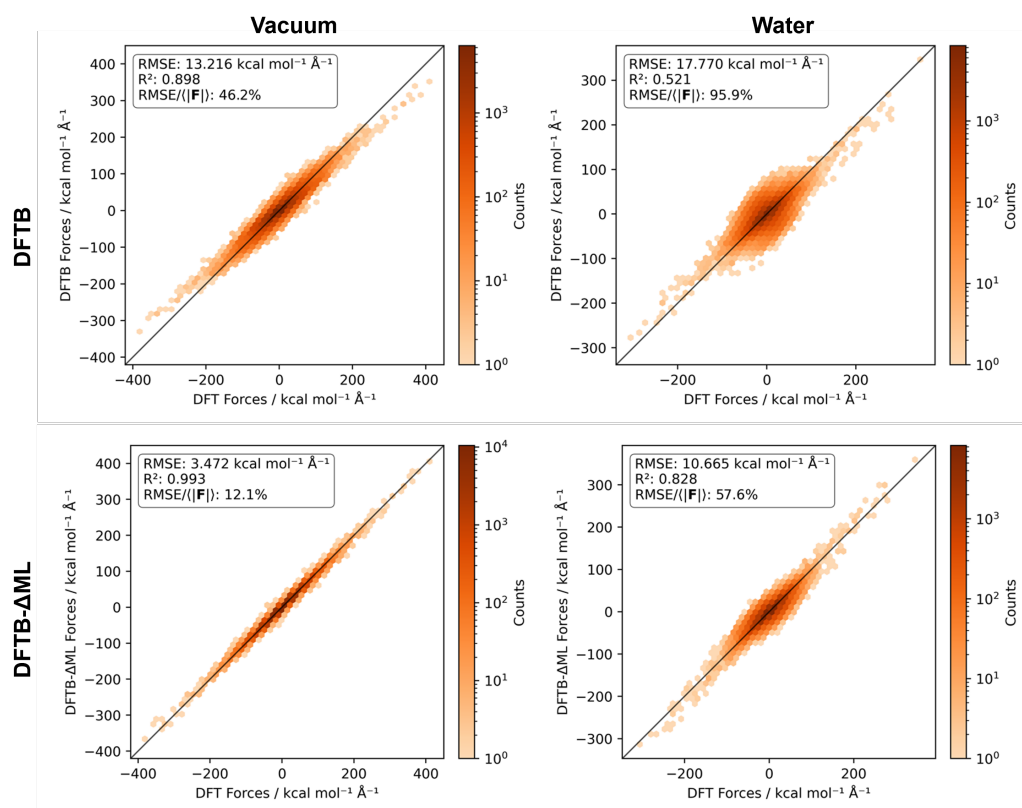


Figure 9.3.: Parity plots illustrating the transferability of the vacuum-trained Δ ML correction to water for the atomic forces of system 5. Shown are atomic forces obtained with DFTB (top row) and DFTB- Δ ML (bottom row) compared against DFT reference forces in vacuum (left) and aqueous solution (right). In vacuum, the Δ ML correction reduced the error by 74%, in water by 40%. The forces were not included in the training, but are obtained as the gradient of the respective total energies.

Analysing the forces by element (refer to Table 9.1) reveals that for uncorrected DFTB, the smallest RMSE values are obtained for hydrogen atoms, both in vacuum and in water. In vacuum, the largest RMSE values are observed for nitrogen and magnesium atoms. In aqueous solution, hydrogen, carbon and nitrogen atoms remain largely unaffected due to their small partial charge and consequently weak electrostatic interactions, such that the RMSE values in vacuum and water are similar. For phosphorus atoms, the RMSE is increased. The strongest increase is observed for oxygen, with a more than $10 \text{ kcal mol}^{-1} \text{ \AA}^{-1}$ larger RMSE, making it the element with the largest force RMSE in water. This behaviour can be explained by the negative partial charge of oxygen atoms, which leads to strong interactions with the surrounding point charges and increased force fluctuations. In contrast, magnesium exhibits a pronounced decrease in RMSE in water, reduced by almost a factor of two. This can be attributed to a systematic offset in DFTB vacuum forces relative to the DFT reference (refer to Figure B.1). In vacuum it is coordinated by the triphosphate, while in water it is hexacoordinated, as it has three additional oxygen point charges around it (refer to Figure 9.4). The incomplete coordination shell in the vacuum training data presumably causes the systematic offset observed for the calculated forces for magnesium in water. Considering the standard deviation of the Δ -forces, magnesium behaves similar in vacuum and in water (refer to Table B.1).

When applying the Δ ML correction in vacuum, the RMSE values for all elements decrease by at least 50%. The largest reductions are observed for oxygen and magnesium atoms, indicating that magnesium–oxygen interactions are most strongly affected by the Δ ML correction. In aqueous

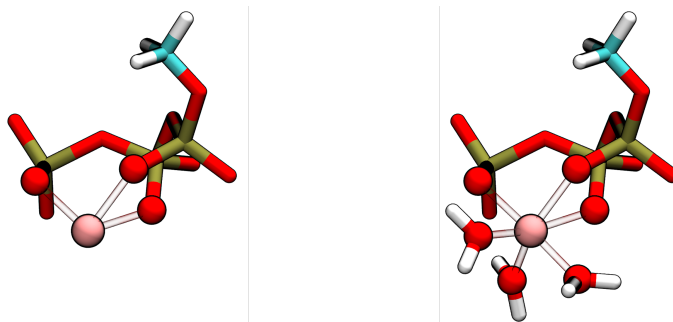


Figure 9.4.: Magnesium coordination in vacuum and water. In vacuum (left), the magnesium cation is only coordinated by the triphosphate. In water (right), water molecules coordinate as well, resulting in a hexacoordinated magnesium cation.

solution, the Δ ML correction also reduces the RMSE values of most atoms, although the improvement is less pronounced. The only exception is the magnesium cation, for which the force RMSE increases upon application of the Δ ML correction. This behaviour can be attributed to aforementioned different number of oxygen atoms in the coordination shell of magnesium in vacuum and in water. As a result, the Δ ML correction, which is trained on vacuum structures with incomplete magnesium coordination, overcorrects magnesium forces in aqueous solution and thereby increases the magnesium force RMSE. For phosphorus atoms, the Δ ML correction has only a minimal impact on the atomic force correction.

Table 9.1.: RMSE values of atomic forces by chemical element in vacuum and water for system 5. Shown are results for DFTB and DFTB- Δ ML relative to DFT reference forces. RMSE values are given in $\text{kcal mol}^{-1} \text{\AA}^{-1}$.

Element	DFTB vs DFT		DFTB- Δ ML vs DFT		Improvement	
	RMSE _{vac}	RMSE _{aq}	RMSE _{vac}	RMSE _{aq}	RMSE _{vac}	RMSE _{aq}
H	3.694	3.046	1.768	2.402	52.1%	21.1%
C	14.692	12.711	4.535	5.974	69.1%	53.0%
N	20.516	17.661	5.884	8.029	71.3%	54.5%
O	15.486	26.214	2.542	14.237	83.6%	45.7%
Mg	18.913	8.424	3.300	11.557	82.6%	−37.2%
P	12.111	18.393	5.353	17.248	55.8%	6.2%

Metadynamics Simulations in Water

Running the DFTB- Δ ML/MM simulations in water shows a stable performance, with all 16 walkers remaining active throughout the simulations. The resulting FES spans the same region of CV space as the one obtained from DFTB/MM simulations (refer to Figure 9.5), indicating that the training data generated in chapter 8 cover all relevant configurations. Focusing on the differences in ΔG and the barrier height an increase of $10\text{--}12 \text{ kcal mol}^{-1}$ is observed for both values compared to DFTB/MM simulations. This indicates that the reactant state is stabilised and lowered by approximately $10\text{--}12 \text{ kcal mol}^{-1}$. While the position of the transition state remains comparable, the overall shape of the FES is altered in the region around $(4.0 \mid 2.7)$.

In order to identify the origin of these changes in the FES, the overcorrection of forces acting on the magnesium cation and the ATP oxygen atoms are examined. While the three ATP oxygen atoms that coordinate the magnesium cation at the beginning of the DFTB/MM simulation stay

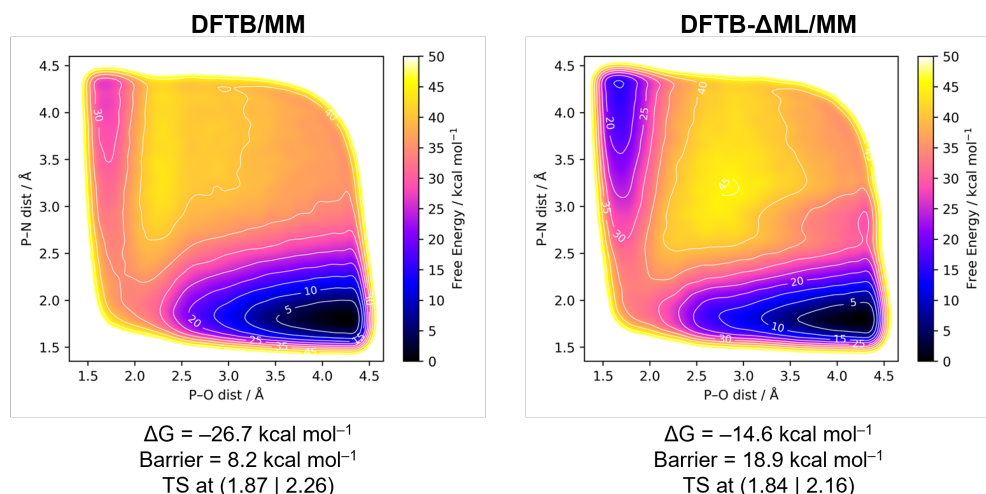


Figure 9.5.: Comparison of FESs generated with DFTB/MM (left) and DFTB- Δ ML/MM (right) metadynamics simulations in water for system 5. In the DFTB- Δ ML/MM FES, an artefact can be observed around (4.0 | 2.7).

coordinated throughout, the coordination sphere changes in the DFTB- Δ ML/MM. Specifically, two β -phosphoryl oxygen atoms interchange their coordination behaviour for 5% of the total simulation time. It is crucial to note that this newly coordinating β -phosphoryl oxygen is the one used to define the P-O distance CV. Consequently, the phosphoryl back-transfer is suppressed, reducing the population of the reactant state and thereby increasing its free energy.

More severely, the observed change in the FES around (4.0 | 2.7) cannot be attributed to the altered magnesium coordination, but instead arises from a misguided phosphorylation event. Rather than phosphorylating the N_ϵ atom, the C_ϵ atom forms a stable interaction with the γ -phosphoryl group in 2 out of 16 walkers, with a P-C distance of approximately 2.1 Å that persists for about 0.5 ns. This interaction effectively leads to the formation of a transient, chemically implausible penta-coordinated carbon centre.

9.3.2. Model System for Creatine Kinase

In system 8 (refer to Figure 9.1), the histidine side chain of system 5 is replaced by creatine. Together with triphosphate- Mg^{2+} -complex it serves as a model system for creatine kinase. The carboxylate group in creatine increases the net charge of the model system compared to system 5, resulting in a net charge of -3 instead of -2 .

Single-Point Energies

Regarding the energies of system 8 (refer to Figure B.2), the DFTB- Δ ML calculations in vacuum yield an R^2 value of 0.989 and an RMSE of $2.361 \text{ kcal mol}^{-1}$, not quite achieving the performance of system 5. This is mainly caused by disagreement in a few high-energy structures. In aqueous solution, the improvement is also less pronounced compared to system 5, resulting in an R^2 value below 0.9 and an RMSE of $14.029 \text{ kcal mol}^{-1}$.

Single-Point Forces

For the forces of system 8, the improvement in vacuum is again slightly smaller than for system 5 with an improvement of 68%, while the improvement for the water calculations is comparable with 40% (refer to Figure B.3). For element-wise RMSE values in the forces refer to Table 9.2, for

element-wise box-plots and standard deviations refer to Figure B.4 and Table B.2, respectively. Again, the largest element-wise RMSE improvements are detected for the magnesium cation and oxygen atoms. While for phosphorus atoms in system 5, the RMSE and standard deviation were only slightly reduced (6.2% and 1.7%), they are even slightly increased in system 8 (-0.9% and -5.3%). While the improvement for hydrogen atoms in aqueous solution is larger than in system 5, the RMSE for phosphorus slightly increases compared to uncorrected DFTB. For carbon and nitrogen, the RMSE improvement is 16–19 percentage points smaller than in system 5, both in vacuum and in water. Nevertheless, the resulting RMSE values for DFTB- Δ ML are comparable between the two systems, even though the initial DFTB RMSE values in system 5 were larger.

Table 9.2.: RMSE values of atomic forces by chemical element in vacuum and water for system 8. Shown are results for DFTB and DFTB- Δ ML relative to DFT reference forces. RMSE values are given in $\text{kcal mol}^{-1} \text{\AA}^{-1}$.

Element	DFTB vs DFT		DFTB- Δ ML vs DFT		Improvement	
	RMSE _{vac}	RMSE _{aq}	RMSE _{vac}	RMSE _{aq}	RMSE _{vac}	RMSE _{aq}
H	4.431	4.030	2.252	2.704	49.2%	32.9%
C	10.832	16.307	5.338	10.298	50.7%	36.8%
N	13.011	16.166	5.815	10.338	55.3%	36.1%
O	16.344	27.030	3.423	14.130	79.1%	47.7%
Mg	18.379	8.130	3.575	11.243	80.5%	-38.3%
P	10.284	17.739	4.935	17.895	52.0%	-0.9%

Metadynamics Simulations in Water

As in the DFTB- Δ ML/MM metadynamics simulation of system 5, all walkers remain stable. However, the resulting FES again exhibits an artefact, appearing as a minimum in the upper right corner (refer to Figure 9.6). This minimum arises from phosphorylation of one of the carboxylate oxygen atoms instead of the guanidino group in three of the walkers. In addition, the β -phosphoryl oxygen used in the CV definition enters the magnesium coordination sphere for 76% of the total simulation time, compared to only 5% in system 5. In one walker, this leads to a phosphoryl back-transfer to

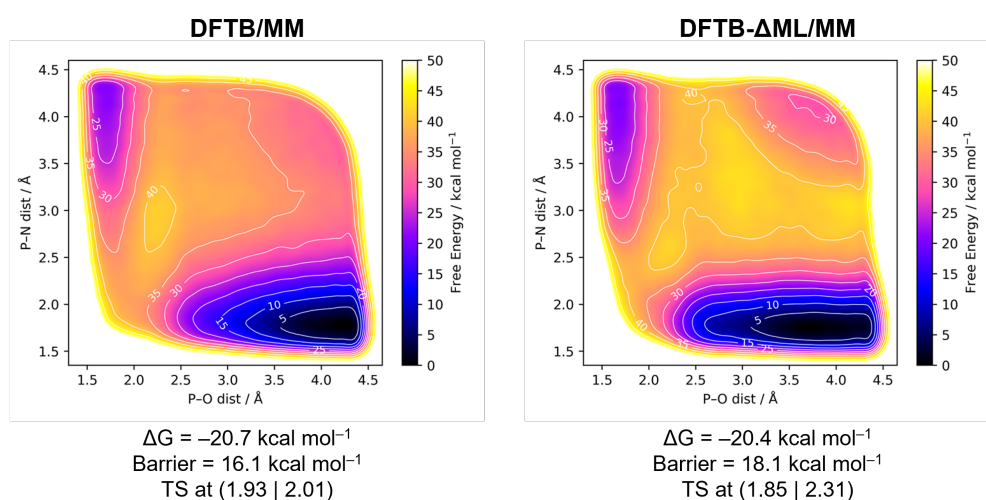


Figure 9.6.: Comparison of FESs generated with DFTB/MM (left) and DFTB- Δ ML/MM (right) metadynamics simulations in water for system 8. In the DFTB- Δ ML/MM FES, an artefact can be observed as an additional minimum in the upper right corner.

the wrong β -phosphoryl oxygen. As a result, although the γ -phosphoryl group remains bound to the diphosphate and is therefore formally in the reactant state, it appears at an incorrect position on the FES. This is a consequence of the chosen CV, which is defined as the distance between the γ -phosphorus and a specific β -phosphoryl oxygen. In another walker, the β -phosphoryl oxygen leaves the coordination sphere. Since the γ -phosphoryl remains coordinated to the magnesium cation, this geometry becomes unsuitable for a back-transfer. In contrast to system 5, no penta-coordinated carbon atom is observed. Consequently, while the artefacts observed in system 8 are undesirable, they correspond to chemically plausible side reactions. Overall, the increased number of artefacts in system 8 can likely be attributed to the increased RMSE in phosphorus atomic forces and to the presence of additional carboxylate oxygen atoms, which introduce further possibilities for side reactions. Furthermore, with a CV that treats the oxygen atoms as indistinguishable, no back-transfer to an incorrect oxygen atom could occur, as all oxygen atoms would be treated equivalently.

9.3.3. Computational Cost

Regarding the computational cost of the single-point calculations (refer to Table 9.3), it is evident that the water calculations converge faster compared to those in vacuum, presumably due to the stabilisation of the charges in aqueous solution. The relative increase in computational cost caused by the Δ ML correction is comparable in vacuum and water with factor of 2.39–3.83. For the metadynamics simulations in water, the Δ ML correction increases the runtime by a factor of 3.46–6.21 (refer to Table 9.4).

Table 9.3.: Mean CPU times per single-point calculation in vacuum and water. The relative computational cost is defined as the ratio of the CPU times of DFTB- Δ ML and DFTB. All times are given in s.

System	single-point vacuum		single-point water		Rel. comp. cost	
	DFTB	DFTB- Δ ML	DFTB	DFTB- Δ ML	vacuum	water
5 Histidine	0.258	0.616	0.143	0.499	2.39	3.49
8 Creatine	0.214	0.735	0.184	0.704	3.43	3.83

Table 9.4.: Mean computational cost of metadynamics simulations, averaged over all walkers, in days/ns. The relative computational cost is defined as the ratio of DFTB- Δ ML and DFTB.

System	DFTB/MM	DFTB- Δ ML/MM	Rel. comp. cost
5 Histidine	0.920	3.185	3.46
8 Creatine	0.665	4.132	6.21

To put these results into context, these values are compared to the work of Giese et al.²²⁷, who presented a Δ ML approach based on the graph neural network MACE, trained to correct O-phosphorylation reaction model systems that mimic the RNA transphosphorylation mechanism. Giese et al. demonstrate that their Δ ML model only causes a small increase in computational cost of a factor of 1.28 when implemented on GPUs. However, executing it on CPUs results in a significant slowdown by a factor of 30, restricting its applicability for CPU-based production simulations. In contrast, the here presented Δ ML approach remains computationally feasible on CPU architectures.

9.4. Conclusions

To validate the transferability of the vacuum-trained Δ ML HDNNP presented in chapter 8 to water, single-point calculations were performed in vacuum and in water, and DFTB- Δ ML/MM metadynamics simulations were carried out in aqueous solution.

The shape of the single-point energy distributions obtained in vacuum and aqueous solution differ significantly from the DFT reference. For optimal transferability of a vacuum-trained Δ ML model, a closer correspondence between the uncorrected DFTB energies in both environments would be desirable. The observed discrepancy could be caused by the highly charged character of the system, which can lead to overpolarisation in DFTB, especially at the QM/MM interface. As a consequence, scaling of the MM point charges could present a promising approach to reduce these artefacts.

The DFTB- Δ ML single-point forces revealed an overcorrection for magnesium and oxygen atoms in aqueous solution. This behaviour is likely attributable to differences in the coordination environment. In water, the magnesium cation is hexacoordinated by the triphosphate and three water molecules. However, the training data were generated in vacuum and therefore did not include water molecules, resulting in an incomplete coordination sphere where magnesium is coordinated only by the triphosphate. Consequently, inclusion of all Mg^{2+} -coordinating water molecules in the training data generation is expected to improve the description of magnesium–oxygen interactions.

DFTB- Δ ML/MM metadynamics simulations in aqueous solution exhibited artefacts and side reactions, likely arising from the issues discussed above. In the following chapter, both the scaling of MM point charges and the inclusion of all Mg^{2+} -coordinating molecules in the training data are evaluated in the protein environment.

On a positive note, notable differences in computational performance are observed when comparing the Δ ML approach reported by Giese et al.²²⁷. While their model increases the computational cost by a factor of approximately 30 on CPUs, it is only 1.28 times slower when executed on GPUs. In contrast, the Δ ML HDNNP presented in this work shows a substantially smaller CPU-time increase, corresponding to a runtime ratio of 3.46–6.21 relative to uncorrected DFTB, demonstrating significantly faster performance on CPU-based architectures.

To summarise, although several challenges remain, this chapter demonstrated that the transfer of the vacuum-trained Δ -ML model to aqueous solution does not produce force outliers and therefore enables stable simulations. Furthermore, the transfer to water provided valuable insights into the limitations of the current approach and identified promising strategies for its application in protein environments. Together with the method's comparatively low computational cost, these results demonstrate the general feasibility of the Δ ML approach. The challenges identified in this chapter, namely the overpolarisation at the QM/MM interface and the incomplete description of the magnesium coordination sphere, will be addressed in the next chapter for the transfer to a protein environment.

10. Autophosphorylation in a *trans*-Histidine Kinase

Author Contributions The work in this chapter was carried out in collaboration with Christian Schmidt.¹⁷⁴ Christian Schmidt set up the neural network, coded the descriptors, training and inference workflows and end-to-end, versioned data pipeline. Lukas Petersen contributed to design choices, implementation details, and practical troubleshooting. The coordinates and partial charges of the atoms of the MM region are retrieved from QM/MM simulations based on a protocol from Katharina Spies.²²⁶ Mayukh Kansari prepared the histidine kinase structure and performed the MM equilibration. Lena Eichinger performed the QM/MM setup and equilibration, as well as the initial metadynamics simulations from which the structures for the single-point calculations were extracted during her Master's thesis.¹⁷ The single-point calculations in vacuum and water, as well as the metadynamics simulations and their analyses were performed by her in the course of this work.

10.1. Introduction

Whereas higher organisms can withstand environmental fluctuations due to their multicellular nature, unicellular organisms such as bacteria and archaea must respond to environmental changes immediately to ensure their survival.²³⁰ This rapid adaptation is mediated by two-component systems (TCSs), which are among the most important signal transduction systems in bacteria and consist of a sensory histidine kinase (HK) and a response regulator (RR).²³¹ The membrane-bound HK performs an autophosphorylation upon detection of changes in environmental stimuli, such as temperature, pH or nutrient concentration, and the RR initiates the corresponding cellular response by controlling gene expression.²³² In order to respond to different environmental signals, bacteria on average encode 52 TCSs,²³⁰ while mammals, including humans, do not encode TCSs in their genome, making TCSs promising targets for antimicrobial agents. As TCS inhibitors operate via a different mechanism to conventional antibiotics, they are particularly useful in the context of antibiotic resistance.²³² In addition, given that TCSs regulate the expression of many antibiotic resistance determinants, combining conventional antibiotics with TCS-targeted drugs may provide an effective approach to antimicrobial treatment.²³³

Upon detection of an environmental signal by a HK, the signal is transmitted to the interior of the cell via conformational changes, ultimately inducing autophosphorylation within the catalytic core. Structurally, HKs can be divided into five distinct domains (refer to Figure 10.1) with the signal detection domains typically located at the N-terminus of the HK.²³¹ The periplasmic sensor domain is responsible for detecting changes in the cellular environment. One example is the HK CitA from *Klebsiella pneumoniae*, which binds citrate within its ligand-binding pocket.²³⁴ However, signal detection is not restricted to the extracellular or periplasmic space.²³¹ The transmembrane

domain of the thermosensor HK DesK from *Bacillus subtilis* detects changes in temperature through membrane fluidity,^{231,235} while the cytoplasmic sensor domain of the kinase NreB found in different species can detect changes in oxygen concentration with an iron–sulphur cluster.^{231,236} As soon as the signal reaches the catalytic core, a conformational transition is induced, bringing the HK from an inactive to an active state. This moves one of the catalytic and ATP-binding (CA) domains closer to a dimerisation and histidine phosphotransfer (DHp) domain, which brings the reacting molecules ATP and histidine in close proximity and an autophosphorylation is initiated.²³¹ The RR, the second component of the TCS, can now be phosphorylated at an aspartate residue, which initiates a specific cellular response to the signal detected by the HK, by altering DNA transcription through interaction of the RR with promoter regions.^{230,232}

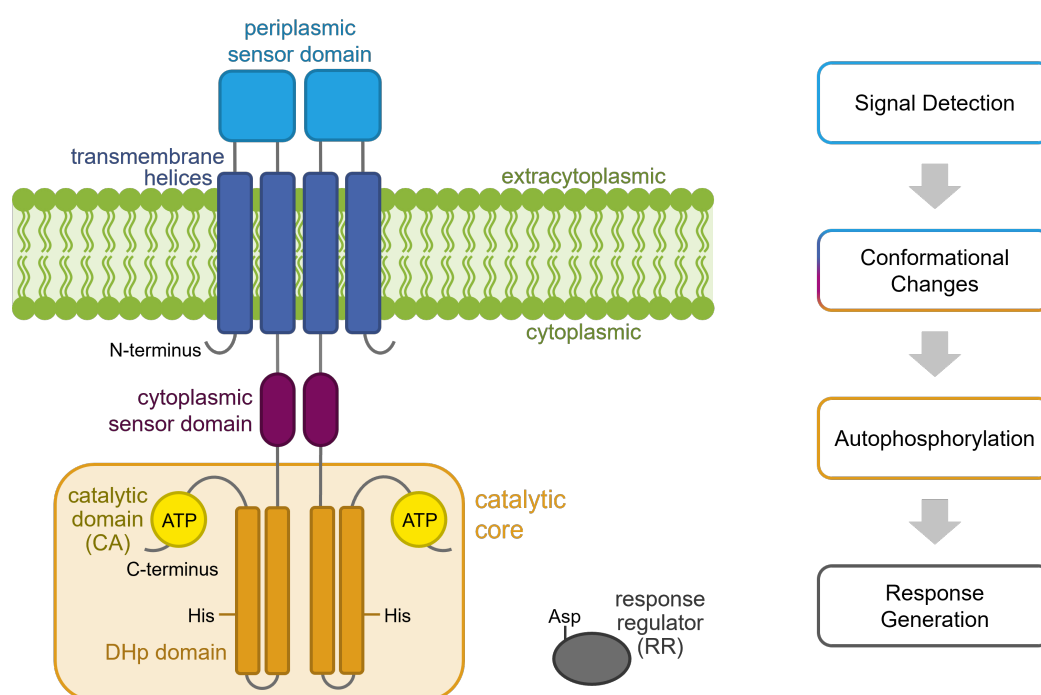


Figure 10.1: Schematic representation of a TCS consisting of a membrane-bound HK and a cytoplasmic, mobile RR, adapted from References [231, 237]. The corresponding signal transduction cascade can be found on the right. After an environmental signal is detected by the HK, a sequence of conformational changes transmits it to the interior of the cell and to the catalytic core of the histidine kinase (HK), where an autophosphorylation of a conserved histidine residue is performed. Subsequent transphosphorylation of a conserved aspartate residue on the response regulator (RR) enables the RR to regulate gene expression and thereby initiate the appropriate cellular response.

During autophosphorylation, the γ -phosphoryl group of ATP is transferred to the N_ϵ of a conserved histidine residue (refer to Figure 10.2). The resulting phosphorylated histidine, also known as phosphohistidine (pHis), is a common intermediate in various enzymatic reactions. Subsequent to autophosphorylation, the pHis is deprotonated by a glutamate in the vicinity, resulting in a deprotonated pHis product.^{15,238,239} HKs are homodimers with each monomer containing a DHp and a CA domain (refer to Figure 10.2 top). Depending on the TCS, one of two different autophosphorylation mechanisms can occur: In *cis*-HKs, the ATP bound to one CA domain phosphorylates the histidine in the DHp domain of the same monomer. In contrast, in *trans*-HKs, which are the focus of this work, the histidine in the DHp domain of the opposing monomer is phosphorylated.⁷

Experimental studies have shown that *cis*- or *trans*-directionality depends on the loops at the base of the four-helix bundle of the DHp domain, which position the CA domain closer to either the

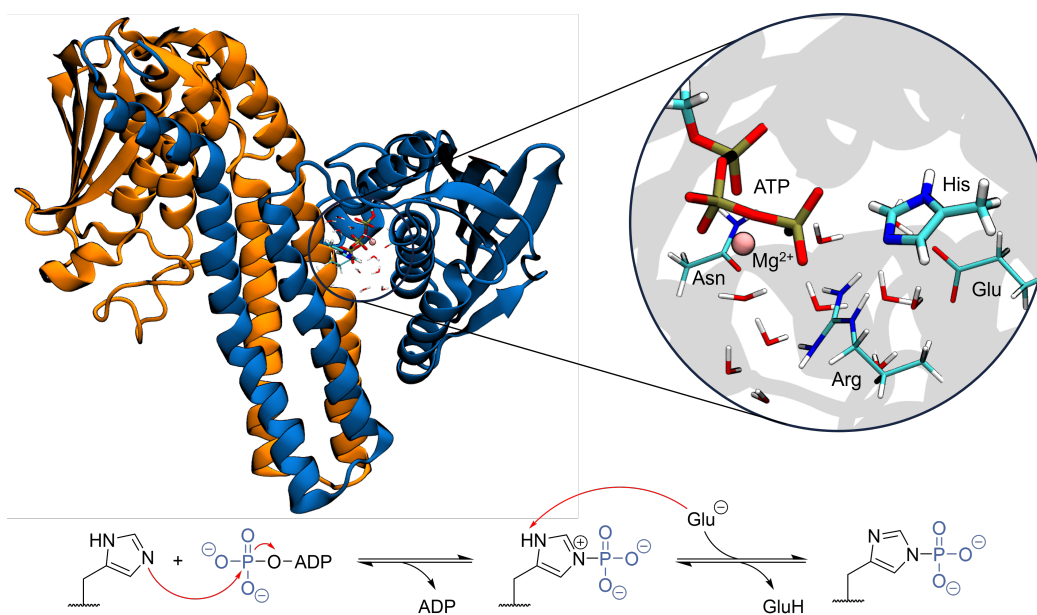


Figure 10.2.: Structure of the investigated *trans*-HK (PDB ID: 5LFK)²²⁵ with enlarged QM region and reaction scheme. Top: The monomers of the homodimeric HK are coloured in orange and blue. The enlarged region shows all active site molecules that are part of the QM region. While the histidine ATP-Mg²⁺-complex and glutamate take part in the autophosphorylation and subsequent deprotonation, the water molecules and asparagine are included as part of the coordination sphere. The arginine was included to stabilise the protonated pHis intermediate. Bottom: Reaction equation showing the autophosphorylation of the histidine followed by the deprotonation by a nearby glutamate.

histidine on the same chain (*cis*) or on the opposite chain (*trans*). This is supported by a study showing that replacing the loop region of a *cis*-HK with that of a *trans*-HK switches the catalytic mechanism, leading to *trans*-autophosphorylation and the formation of pHis.⁷ The mechanism of the autophosphorylation on the other hand is independent of the *cis*- or *trans*-directionality.⁸ Although pHis has been known for over six decades, its chemical instability makes it difficult to study *in vivo* and *in vitro*.²⁴⁰ These limitations make computational approaches particularly valuable. QM/MM-based simulations have enabled detailed investigation of the mechanism and energetics of autophosphorylation in a *cis*-HK, allowing the determination of free-energy barriers of both the autophosphorylation reaction and the preceding conformational transition. The autophosphorylation step proceeds via a penta-coordinated transition state with a barrier of 8 kcal mol⁻¹, and the combined barrier, including the conformational transition, is in agreement with the experimentally determined rate.¹⁵ In the *trans*-HK CpxA from *Escherichia coli*, which is also investigated in this work, the barrier was found to be 15 kcal mol⁻¹. However, the authors note that the difference in barrier height between these two HKs might be a result of the different collective variables (CVs) that have been employed.¹⁶

10.1.1. Motivation and Aim of the Project

Kinase-catalysed phosphorylation reactions in proteins are challenging to describe computationally. QM/MM methods combined with enhanced sampling allow simulations on the relevant timescales when computationally efficient semi-empirical approaches such as DFTB are employed, but their accuracy can be limited. Δ -machine learning corrections offer a strategy to improve the accuracy of DFTB calculations while largely retaining their computational efficiency.

In this context, the HDNNP introduced in chapter 8 is applied to QM/MM simulations of the autophosphorylation reaction catalysed by the *trans*-HK CpxA from *Escherichia coli*. As demonstrated by the DFTB- Δ /MM simulations in aqueous solution presented in chapter 9, it may be advantageous to include the complete coordination sphere of the magnesium cation during training, as well as to scale the MM point charges in order to reduce overbinding at the QM/MM boundary. Both strategies are therefore investigated in the protein environment.

Including all QM water molecules together with the Mg-coordinating asparagine residue increases the number of atoms in the training set by factor of 2.3. To keep the computational cost manageable, the number of symmetry functions, as well as the radial and angular cut-offs, are reduced accordingly. To evaluate the impact of these adjustments, single-point DFTB and DFTB- Δ /MM energy and force calculations are conducted and compared with the DFT reference, both in vacuum and within an electrostatic embedding of the protein environment. In the latter case, the protein is represented by fixed point charges to mimic a QM/MM-like setup.

Most importantly, QM/MM metadynamics simulations are performed for the protein system using DFTB and DFTB- Δ ML as the QM methods, in order to evaluate the impact of the Δ ML correction on the dynamics and free-energy landscape of QM/MM simulations in the protein environment.

10.2. Methods

Structure Preparation and MM Equilibration

The crystal structure of a hemi-phosphorylated CpxA dimer (PDB ID: 5LFK)²²⁵, an HK from the *Escherichia coli* signalling system, was used as starting structure. Missing loop segments were rebuilt using UCSF Chimera²⁴¹. The MM equilibration was carried out using GROMACS 5.1^{200–204}, employing the AMBER99SB-ILDN force field^{198,199} to describe the protein and TIP3P water¹⁹⁷ for solvent representation. The integration of the equations of motion was performed with the Leap Frog algorithm⁵³ using a time step of 1 fs. All bonds involving hydrogen atoms were constrained with the LINCS algorithm²⁴². Periodic boundary conditions were applied and the electrostatic interactions were treated with the particle mesh Ewald method²⁰⁹. The cut-off for short-ranged electrostatic and Van der Waals interactions was placed at 1 nm. The energy of the system was minimised using the steepest descent algorithm, which was followed by an NVT equilibration for 10 ns at 300 K with the velocity rescaling thermostat⁵⁵.

Preparation of the QM/MM Simulation

The QM region was modelled to contain the side chains of Asn145, His248, Glu249 and Arg251, the Mg²⁺ cation, the phosphoryl groups of the ATP and ten water molecules (refer to Figure 10.2). Hydrogen link atoms were placed at the QM/MM interface and the QM–MM interactions were treated with the electronic embedding scheme.²⁴³ The QM region was described with the semi-empirical DFTB3 method¹¹⁷ and the 30B parameter set^{192,195,196} and specific reaction parameters (SRPs) OPhyd¹⁹² and PNmod¹⁶. QM/MM equilibrations and simulations were performed using our local implementation of QM/MM in GROMACS 2020.2^{200–204}, which is available from a public repository.²⁰⁷ Restraints were placed using the open-source, community-developed PLUMED library²¹⁰, version 2.5.1²¹¹. In a first step, the whole system was slowly heated from 10 to 300 K for

25 ps with a time step of 0.5 fs, using simulated annealing. Afterwards, the QM equilibration was performed at 300 K for 10 ns with a time step of 1 fs.

Metadynamics Simulation

The two-dimensional metadynamics simulation was calculated along two collective variables (CVs) to describe the phosphoryl transfer and subsequent proton transfer, and are defined as

$$\text{phosTrans} = d(\text{P-O}) - d(\text{P-N}) \quad (10.1)$$

$$\text{protTrans} = d(\text{N-H}) - d(\text{O-H}), \quad (10.2)$$

with the respective atomic distances defined according to Figure 10.3. Gaussians with a width of 0.02 nm and a constant height of 0.8 kJ were deposited every 0.5 ps and exchanged between the individual walkers every 1 ps. The simulation was performed in a multiple-walker approach with 24 walkers, each running for 1 ns.

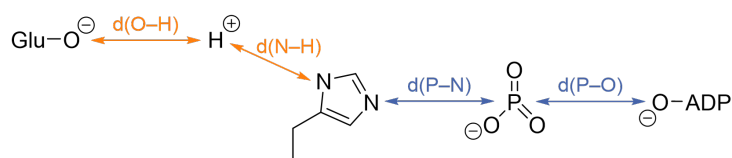


Figure 10.3.: Schematic representation of the collective variables (CVs) used for the metadynamics simulations of HK.

Training and Single-Point Calculations in Vacuum and Protein Environment

From these metadynamics simulations, 500 structures for the single-point calculations were extracted randomly and uniformly from the relevant regions across the free energy surface (FES) using MDAnalysis (version 2.6.1)^{228,229}. Since only the phosphoryl transfer reaction is corrected by the ΔML model, the structure extraction was restricted to configurations with a maximum N–H distance of 1.1 Å. Single-point calculations were performed using DFT ($\omega\text{B97M-V/ma-def2-TZVP}$)^{101,107,109}, DFTB (DFTB3/3OB^{192,195,196} with OPhyd¹⁹² and PNmod¹⁶ SRPs), and the corresponding DFTB- ΔML approach. As ΔML models, two different subsets of the HK active site were employed. The first one is the HK system described in chapter 8, which includes only the histidine side chain and the triphosphate- Mg^{2+} complex (‘small QM subset’). In addition, a separate

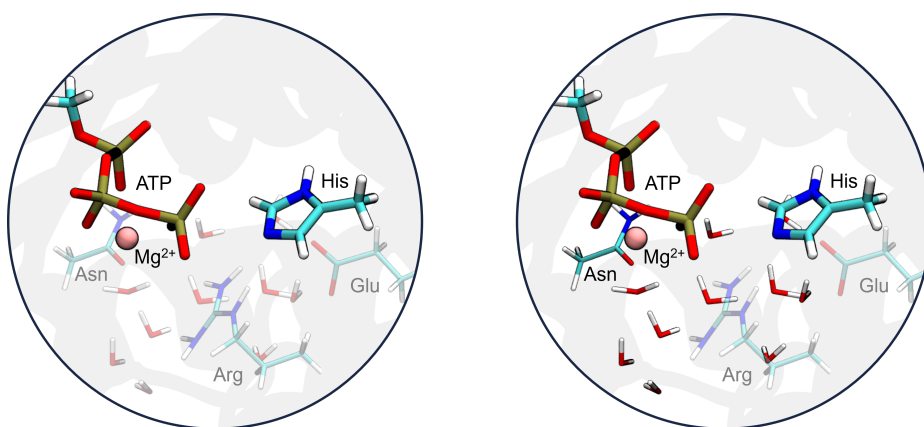


Figure 10.4.: Subsets of the HK QM region corrected during DFTB- ΔML calculations. Solid molecules are corrected by the ΔML model, whereas transparent molecules are included in the QM region but are not corrected. The protein is shown in grey. Left: The small QM subset corresponds to the HK system trained in chapter 8 and aims to correct only the reacting species. Right: The big QM subset includes the QM water and asparagine to represent the complete coordination sphere of the magnesium cation.

model was trained in this chapter, that also includes the QM water molecules and the asparagine side chain, thereby representing a larger subset of the QM region and the complete coordination sphere of the magnesium cation ('big QM subset', refer to Figure 10.4).

While the network specifications described in Section 8.2 were used for the small QM subset, a variant with a reduced number of symmetry functions was additionally applied for the big QM subset to address the increased computational cost associated with the larger QM region. The radial symmetry functions used 5 R_s values ($R_s \in \{0.0, 1.0, 2.0, 3.0, 4.0\} \text{ \AA}$) and 4 η values ($\eta \in \{0.08, 0.16, 0.3, 0.5\} \text{ \AA}^{-2}$), yielding 20 radial features per atom. The angular symmetry functions used $\lambda \in \{-1, 1\}$, $\zeta \in \{1, 4, 16\}$, and the same 4 η values, resulting in 24 angular features per atom.

As in chapters 8 and 9, DFT calculations were performed using ORCA²¹², version 5.0.5²¹³, while DFTB calculations were carried out using DFTB+^{205,206} in our local implementation²⁰⁸. For both DFT and DFTB energies, the atomisation energies (refer to Equation (8.1)) were calculated and used for comparison between the methods. Atomic force differences ΔF were computed according to Equation (9.1).

In addition to vacuum calculations, a QM/MM electrostatic embedding was applied by representing the surrounding protein, solvent, and ions as fixed point charges, thereby mimicking the electrostatic environment of a QM/MM setup. For all calculations, energies and atomic forces were computed.

Metadynamics Simulations with DFTB/MM and DFTB- Δ ML/MM

To evaluate the effect of the Δ ML model on the dynamics, two-dimensional metadynamics simulations were performed. As CVs, the distance between the atoms of the P–O and P–N bonds formed and broken during the phosphoryl transfer were used ($d(\text{P–O})$ and $d(\text{P–N})$ in Figure 10.3). Gaussians with a width of 0.01 nm and were deposited every 0.5 ps and exchanged between the individual walkers every 1 ps. An initial height of 1.2 kJ mol^{-1} and bias factor of 20 were chosen to ensure a controlled exploration of the FES. The simulation was performed in a multiple-walker approach with each of the 24 walkers running for 1 ns, resulting in a total simulation time of 24 ns.

The simulations were started from different starting structures: the equilibrated starting structure, which was also used for the phosTrans/protTrans metadynamics simulation described above, and a second structure extracted after 1 ns from that simulation. The two starting structures mainly differed in the coordination of the magnesium cation, allowing its influence on the phosphoryl transfer to be evaluated. As a second parameter, the effect of placing a flat-bottom restraint at 170 deg of the N–P–O angle was assessed. This gives a lower bound to the angle and effectively restricts it to a near-linear geometry. Combining the two starting structures with the presence or absence of the angular restraint resulted in four different variants of the metadynamics simulations. Each variant was performed as DFTB- Δ ML/MM simulation as well as DFTB/MM simulation as a reference, resulting in eight metadynamics simulations total. For the DFTB- Δ ML/MM, the big QM subset was corrected using the reduced set of symmetry functions, as described above.

Analysis

The visualisation software VMD (version 1.9.4)²⁴⁴ was used to perform trajectory analyses and to generate figures for this work. Distance calculations were carried out with PLUMED²¹⁰, version 2.5.1²¹¹. All data analysis and visualisation were performed in Python 3.11²⁴⁵ using NumPy (version 1.26.0)²⁴⁶, pandas (version 2.1.3)^{247,248}, and Matplotlib (version 3.8.2)²⁴⁹.

10.3. Results and Discussion

In this section, the performance and applicability of the Δ ML model are discussed. First, the model is validated using single-point energy and force calculations. In a second step, DFTB/MM and DFTB- Δ ML/MM metadynamics simulations are performed in order to evaluate the influence of the Δ ML model during simulation and on the resulting FESs.

10.3.1. Optimising Δ ML Corrections for Protein QM/MM Simulations

In chapter 9, several artefacts were detected when transferring the vacuum-trained Δ ML model to aqueous solution for system 5. System 5 represents a model of the HK active site and comprises a histidine side chain and triphosphate- Mg^{2+} complex. In particular, the inclusion of only half of the magnesium coordination sphere led to an overcorrection of the magnesium atomic forces. In addition, the energy distributions obtained from DFTB calculations in vacuum and in water exhibited different shapes, which reduced the transferability of the Δ ML model. These observations led to the hypothesis that the reduced transferability originates from an overpolarisation at the QM/MM boundary.

In the HK simulation, the QM region is larger and includes, in addition to the histidine and triphosphate- Mg^{2+} complex, the amino acids glutamate, arginine, asparagine and ten water molecules (refer to Figure 10.2). As this results in a QM region with 98 atoms, only a subset is corrected by the network to ensure the computational cost remains manageable.

To this end, two different subsets of the QM region are tested (refer to Figure 10.4). In the first subset, referred to as the ‘small QM subset’, only the reacting molecules are included in the training and corrected in DFTB- Δ ML calculations. The second model, referred to as the ‘big QM subset’, also includes the QM water molecules and the asparagine side chain in order to capture the complete coordination sphere of the magnesium cation.

Training and Correcting the Small QM Subset

The small QM subset corresponds to the QM region of system 5. Consequently, similar artefacts are observed when transferring the Δ ML model trained on system 5 to water (refer to Section 9.3.1), and when transferring the Δ ML model trained on the small QM subset to the protein environment. For vacuum single-point calculations, all 98 atoms in the QM region are considered. However, as the network was only trained on a small subset of the QM region, only this part is corrected by the Δ ML model. Consequently, the model’s predictive quality is already lower in the vacuum calculations.

Consequently, the error in the vacuum forces is reduced by 50.6%, and by 38.0% when additionally including protein point charges, with force errors evaluated exclusively for atoms subject to the Δ ML correction (refer to Table C.1). Of all the investigated atoms, the quality of the magnesium atomic forces were the only ones to decrease, by -11.5% in vacuum and -14.0% in the protein environment.

When the Δ ML model trained on the small QM subset is applied to the big QM subset, the overall error increases by 393.1% in the protein (refer to Table C.2). This is mainly caused by the magnesium atomic forces, where the RMSE increases by 2010.8%, presumably due to the increased number of QM oxygen atoms coordinated to the magnesium.

Training and Correcting the Big QM Subset

The big QM subset, which includes water molecules and the asparagine side chain, comprises all oxygen atoms coordinating the magnesium cation. When the network trained on these structures is applied in DFTB- Δ ML single point calculations, the RMSE in atomic forces is reduced by 57.1% in vacuum and by 50.0% in the protein environment (refer to Table C.3). Notably, whereas magnesium atomic force RMSE was increased for the small QM subset, they improve when trained and correcting the big QM subset with RMSE reductions of 38.0% in vacuum and 34.8% in the protein. This clearly demonstrates the importance of including the full coordination sphere of the magnesium cation.

Reducing the Computational Time

Since the big QM subset contains 69 atoms compared to 30 atoms in the small subset, the computational cost is correspondingly expected to increase by approximately a factor of two. To address this, several strategies were explored in order to accelerate the Δ ML model, including a reduction of the radial and angular cut-off radii and a reduction in the number of symmetry functions.

For the reduction of the cut-off radius, a factor of 0.5 was chosen, reducing the radial cut-off from 10 to 5 Å and the angular cut-off from 6 to 3 Å. While this limits the increase in computational cost to a factor of 1.84 relative to uncorrected DFTB (compared to 2.56 for the full cut-off), the improvement in atomic forces is also reduced: only a 28.4% reduction in RMSE is achieved, which is even lower than for the small QM subset with full cut-off (refer to Table 10.1).

In a second attempt, the cut-off radius was unchanged (10 Å for radial, 6 Å for angular), but the number of symmetry functions was reduced from 35 to 20 for the radial part and 50 to 24 for the angular ones (refer to Section 10.2 for details). Surprisingly, this not only reduces the increase in computational cost to a factor of 1.76 (corresponding to a speed-up of approximately 1.5 relative to using all symmetry functions), but also improves the RMSE of the atomic forces by 51.0%, which is even slightly better than the results from the Δ ML model with the full symmetry function set. This demonstrates the importance of optimising the symmetry functions.^{165,167}

Table 10.1.: Mean CPU times per single-point calculation in the protein. The relative computational cost is defined as the ratio of the CPU times of DFTB- Δ ML and DFTB. All times are given in s. The improvement in the RMSE of atomic forces is evaluated for the QM atoms corrected by the Δ ML model, relative to uncorrected DFTB.

QM subset	CPU times single-point calc. in protein			RMSE
	DFTB	DFTB- Δ ML	Rel. comp. cost	Improvement
Small	1.261	1.628	1.29	38.0%
Big	1.269	3.243	2.56	50.0%
Big (smaller cut-off)	1.269	2.340	1.84	28.4%
Big (less sym. func.)	1.269	2.242	1.76	51.0%

Reducing Overpolarisation at the QM/MM Boundary

As observed for aqueous solution in Sections 9.3.1 and 9.3.2, the force corrections obtained from vacuum calculations remain more effective compared to those in the protein environment. The initial assumption was that the repulsive part of DFTB mainly contributes to the error. As it is unaffected by the environment, training in vacuum and transferring to a different environment was considered

appropriate. However, differences in the shape of the single-point energy distributions in vacuum and water/protein indicate that the electronic contributions also play a role (refer to Figure C.1).

This behaviour was attributed to overpolarisation effects, as DFT can accommodate external electrostatic fields through a continuous redistribution of the electron density, whereas polarisation in DFTB is mediated by atom-centred SCC charges, resulting in an increased sensitivity to electrostatic interactions at the QM/MM boundary. To address this issue, the electrostatic interactions between the QM and MM region were scaled down by a factor of 0.75 in the DFTB and DFTB- Δ ML calculations by uniformly reducing the MM point charges, thereby compensating for the missing polarisation of the MM region. This scaling factor (approximately $1/\sqrt{\epsilon_{el}}$) incorporates the electronic dielectric constant of the environment in order to reduce overpolarisation.^{250,251} Applying the charge scaling factor to single-point calculations with the reduced symmetry function set further increases the improvement of the atomic force RMSE to 54.8%. This is almost four percentage points more than without the scaling, bringing it close to the vacuum improvement of 57.8%. This observation supports the assumption that DFTB is more sensitive to strong MM point charges near the QM/MM boundary than DFT.

Based on the single-point benchmark against DFT reference calculations, DFTB- Δ ML/MM metadynamics simulations were performed using the network trained on the big QM subset with the reduced set of symmetry functions, while scaling the MM charges by a factor of 0.75.

10.3.2.DFTB- Δ ML/MM Simulations for Mechanistic Modelling

Before discussing the results, a methodological detail of the metadynamics simulations must be addressed. For the metadynamics simulations of the histidine model system in water (refer to Section 9.3.1), a bias factor of 50 was employed. After 16 ns, the FES had been filled up to 57 kcal mol⁻¹. In contrast, for the protein HK simulations in this chapter, a lower bias factor of 20 was deliberately chosen to avoid potential overfilling of the FES, which could result in unwanted side reactions and can occur when the bias factor is set too high. Based on the inverse scaling of the effective barrier with the bias factor, the expected total filling height would be approximately 23 kcal mol⁻¹. However, after the same simulation time of 16 ns, the FES in the HK DFTB- Δ ML/MM simulations was filled up to 44 kcal mol⁻¹, even though identical CVs were employed and the barrier height is similar. Therefore, this difference cannot be attributed solely to the choice of bias factor. Instead, it reflects the different configurational freedom of the two systems. In aqueous solution, a larger region of CV space is accessible, whereas in the protein environment, the surrounding residues restrict the reaction pathway, thereby limiting the accessible free-energy space. Since the overfilled FES in the HK simulation introduced artefacts in the phosphorylation reaction, possibly due to the crossing of an additional barrier, the analysis presented here focuses only on the initial 10 ns (corresponding to approximately 0.42 ns per walker) rather than the full 24 ns planned.

Eight metadynamics simulations were performed. The simulations differ with respect to the QM method used in the QM/MM calculations (DFTB or DFTB- Δ ML), the presence or absence of a restraint applied to the N-P-O angle, and the starting structure. The restraint was placed at 170°, effectively constraining the system towards the linear transition state expected for an S_N2 mechanism.

The two starting structures mainly differ in the coordination environment of the magnesium cation. In the first structure, the α - and γ -phosphoryl groups of ATP coordinate the magnesium cation, together with the asparagine residue and two QM water molecules. In the second structure, the β -phosphoryl group is additionally included in the coordination sphere, replacing one water molecule. Accordingly, the structures are denoted 2PC and 3PC, referring to coordination by two or three phosphoryl groups, respectively.

A two dimensional FES is exemplarily shown for the DFTB- Δ ML/MM case (3PC, with angle restraint) in Figure 10.5. Table 10.2 shows the coordinates and relative energies of the reactant, product, and transition states obtained from all eight metadynamics simulations. The respective 1D FES projections onto the two CVs are shown in Figure C.2. The DFTB- Δ ML/MM simulation without an angle restraint and starting from the 2PC structure is excluded from further discussion, as it had not yet converged after 10 ns of simulation time. The other seven simulations showed near-convergence after 10 ns (refer to Figure C.3).

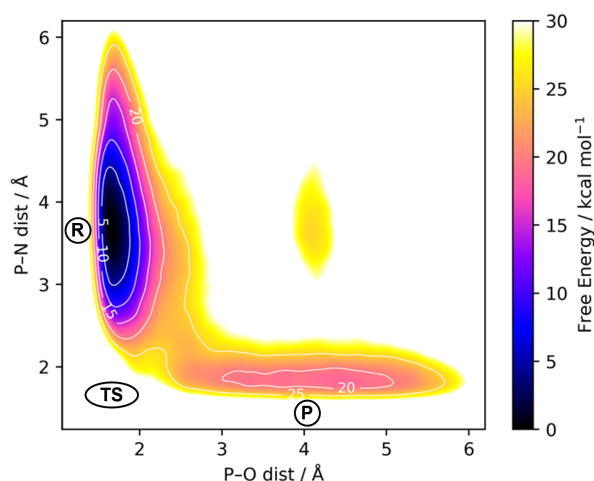


Figure 10.5.: FES of a HK phosphoryl transfer simulation, exemplarily shown for the DFTB- Δ ML/MM simulation with angle restrained, which was started from the 3PC structure. The position of the reactant, transition state and product are indicated.

Influence of DFTB vs DFTB- Δ ML as QM Method

For DFTB/MM simulations, activation barriers range from 11.1 to 22.9 kcal mol⁻¹, whereas DFTB- Δ ML increases them substantially to 23.5–35.1 kcal mol⁻¹ (refer to Table 10.2). This corresponds to an increase of 12.4–18.4 kcal mol⁻¹, approximately doubling the barrier height. The same trend can be observed for the reaction free energies ΔG . DFTB predicts nearly the same energy for reactant and product state ($\Delta G = -1.1$ – -0.7 kcal mol⁻¹), while DFTB- Δ ML consistently yields strongly endergonic profiles ($\Delta G = 18.7$ – 24.1 kcal mol⁻¹). In the histidine test system in water, the barrier was increased by 10.7 kcal mol⁻¹ and the ΔG was increased by 12.1 kcal mol⁻¹, therefore showing similar behaviour as observed in the protein.

Thus, the Δ ML correction systematically shifts the reaction energetic upwards, destabilising both the product and the transition state and thereby increasing both the kinetic barrier and the reaction free energy. From another perspective, this can be interpreted as a relative stabilisation of the reactant state. Possible reasons could be a weakening of the formed P–N bond or a change in magnesium coordination (as discussed on the next pages).

Table 10.2.: Coordinates and energies of reactant, product, and transition state for different methods, starting structures and angle restraints. In the ‘Angle’ column, ‘yes’ indicates that a restraint at 170° was applied to the N–P–O angle. In the ‘Structure’ column, 2PC and 3PC specify whether the magnesium cation in the starting structure is coordinated by two or three phosphoryl groups, respectively. The simulation shown in grey was not converged after 10 ns of simulation time. The first coordinate is the P–O distance, the second the P–N distance. The transition state energy is given relative to the reactant state energy. Coordinates are given in Å, energies in kcal mol^{−1}.

QM method	Structure	Angle	Reactant state		Product state		Transition state	
			Coordinates	Energy	Coordinates	Energy	Coordinates	Energy
DFTB	2PC	yes	1.68 3.88	0.0	4.39 1.84	0.3	2.28 2.44	18.6
DFTB	2PC	no	1.63 4.06	0.8	4.66 1.81	0.0	2.37 2.51	22.9
DFTB	3PC	yes	1.70 3.46	1.1	4.13 1.84	0.0	2.23 2.40	11.1
DFTB	3PC	no	1.72 3.83	0.0	4.39 1.86	0.7	2.24 2.35	12.7
DFTB-ΔML	2PC	yes	1.63 4.20	0.0	4.41 1.77	21.1	2.47 2.23	35.1
DFTB-ΔML	2PC	no	1.62 4.07	0.0	4.23 1.78	20.7	2.35 2.21	37.1
DFTB-ΔML	3PC	yes	1.67 3.61	0.0	4.38 1.84	18.7	2.48 2.19	23.5
DFTB-ΔML	3PC	no	1.63 4.00	0.0	4.30 1.81	24.1	2.55 2.27	31.1

Structurally, this energetic shift is accompanied by a systematic change in the transition state geometry. While in plain DFTB the transition state is located at (2.23–2.37 | 2.35–2.51), the DFTB-ΔML transition states are shifted to (2.47–2.55 | 2.19–2.27). This indicates that the transition state becomes more product-like, with significantly increased P–O dissociation and more pronounced P–N bond formation compared to plain DFTB.

Since DFTB-ΔML is designed to approach DFT-level accuracy, this observation indicates that plain DFTB underestimates the activation barrier and overstabilises the product state relative to DFT. Equivalently, this may be interpreted as an underestimation of reactant state stabilisation in the DFTB calculation.

Structural Effect: Magnesium Cation Coordinated by two or three Phosphoryl Groups

Not only the method, but also the change in magnesium coordination exerts a pronounced influence on the reaction energetics (refer to Table 10.2). For DFTB/MM simulations, the 3PC setup exhibits substantially reduced barriers, approximately 40–45% lower than in the 2PC simulations, while the free energy difference between reactant and product remains largely unaffected. In DFTB-ΔML/MM simulations (excluding non-converged cases), the 3PC setup likewise lowers the barrier by approximately 33% compared to 2PC, again with only minor changes in the product free energy.

Across all eight simulations, the average number of oxygen atoms coordinating the magnesium cation ranges between 4.1 and 4.7 (refer to Table 10.3). Nevertheless, the specific identity of the coordinating phosphoryl group appears to influence the reaction barrier. Interestingly, in the 2PC simulations, the β-phosphoryl group transiently entered the coordination sphere in approximately one quarter of the walkers, whereas in the 3PC simulations the β-phosphoryl group remained stably coordinated throughout. This further supports the importance of full phosphoryl coordination.

These observations suggest that coordination by all ATP phosphoryl groups (3PC) preferentially stabilises the transition state relative to the reactant state, likely by improving the relative orientation

Table 10.3.: Average number of oxygen atoms (of ATP, asparagine, water) coordinating the magnesium cation for different methods, starting structures, and with or without an angle restraint. An oxygen atom is considered as coordinated if it is within 2.3 Å of the magnesium cation. In the ‘Angle’ column, ‘yes’ indicates that a restraint was applied to the N–P–O angle, restricting it to a minimum value of 170°. In the ‘Structure’ column, 2PC and 3PC specify whether the magnesium cation in the starting structure is coordinated by two or three phosphoryl groups, respectively. The simulation shown in grey was not converged after 10 ns of simulation time.

Structure	Angle restraint	DFTB	DFTB-ΔML
2PC	yes	4.4 ± 0.4	4.7 ± 0.4
2PC	no	4.3 ± 0.4	4.7 ± 0.4
3PC	yes	4.1 ± 0.4	4.9 ± 0.4
3PC	no	4.1 ± 0.3	4.7 ± 0.4

of the reactant species before the reaction takes place. Consequently, full phosphoryl coordination of magnesium appears to enhance the catalytic activity of the histidine kinase.

Effect of the Angular Restraint

Initially, a restraint was placed at 170° on the N–P–O angle, following the approach of Kansari et al.¹⁶, who investigated the same *trans*-HK structure and applied a restraint at 160°. The intention was to enforce a near-linear S_N2-like transition state.

As shown in Table 10.2, across both methods and starting structures, removal of the N–P–O angle restraint leads to higher activation barriers compared to simulations with the restraint, increasing the barrier by a factor of 1.14–1.32. Inspection of Figure 10.6 reveals that, even without the angle restraint, the transition state remains linear. However, the reactant and product states are located at substantially smaller N–P–O angles. In addition, Table 10.2 shows that most reactant and product state exhibits approximately 1.1-times larger P–O and P–N distances, respectively, when no angle restraint is applied. This indicates that, without the angular restriction, the reactant and product structures can relax more fully within the protein environment. Consistently, the reactant state is shifted to higher values along the phosphoryl-transfer coordinate.

Moreover, in the region where the reactant and product states are found in the restrained simulations, little to no sampling is observed in the simulations without the angle restraint, particularly for the reactant state (refer to Figure 10.6). This suggests that the angular restraint artificially constrains the reactant and product geometries, preventing full relaxation and thereby increasing their relative energies. As a consequence, the activation barrier appears lower in the restrained simulations because the reactant state is destabilised and higher in energy.

In contrast, without the angular restriction, the protein environment can adapt more freely, allowing the reactant state to relax and become energetically stabilised. Together with the observation that the linear S_N2 transition state is still accessible, this indicates that simulations without an angle restraint provide a more realistic description of the reaction pathway.

Setting the Results into Context

Kansari et al.¹⁶ performed DFTB/MM metadynamics simulations starting from the same PDB structure and using the same CVs as in this work, and reported a barrier of 15 kcal mol^{−1}, consistent with

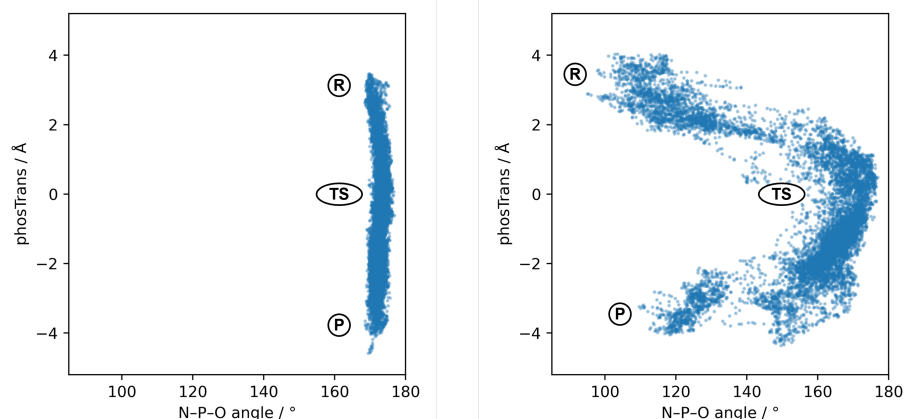


Figure 10.6.: Relation between phosphoryl transfer and N-P-O angle, shown for the DFTB/MM metadynamics started from 3PC. Left: with angle restraint at 170° , right: no angle restraint applied. The phosTrans coordinate is defined as the difference between the distances $d(\text{P-O}) - d(\text{P-N})$.

the value obtained by Marsico et al.²⁵². This barrier lies within the range of the DFTB/MM barriers obtained in this chapter (refer to Table 10.2).

For the reaction free energy, it is noteworthy that Kansari et al. reported two product minima. One minimum was found at a P-O distance of approximately $3.4 \text{ \AA}$ and approximately 4 kcal mol^{-1} higher in energy than the reactant state, whereas a second minimum at approximately $5 \text{ \AA}$ was reported to be approximately -5 kcal mol^{-1} relative to the reactant, which was attributed to a structural rearrangement. In contrast, only a single product minimum is observed in all eight simulations reported here, and the ΔG obtained from the DFTB/MM simulations falls between the two product minima reported by Kansari et al.

Because DFTB- Δ ML is fitted towards DFT reference data and has shown promising results in the single-point calculations (refer to Section 10.3.1) and similar behaviour in water (refer to chapter 9), the DFTB- Δ ML/MM results are expected to be more reliable in terms of underlying energetics. However, the DFTB- Δ ML/MM simulations predict a comparatively high-energy product state, which suggests limited product stability on the sampled time scale. In order to compare the obtained barriers to experimental rates, the transition state theory (TST) is used, which is given as

$$k = \frac{k_B T}{h} \exp\left(-\frac{\Delta G^\ddagger}{RT}\right), \quad (10.3)$$

with the rate constant k , the Boltzmann constant k_B , the temperature T , the Planck constant h , the barrier height $\Delta G^\ddagger$, and the universal gas constant R . This indicates that the DFTB- Δ ML/MM barriers of $23.5\text{--}35.1 \text{ kcal mol}^{-1}$ correspond to time scales ranging from hours to thousands of years. This is in vast contrast to experimentally determined time scales on the order of seconds,²⁵³ which correspond to an effective barrier of approximately 17 kcal mol^{-1} .

Several factors in protein QM/MM simulations can influence the energetics of phosphoryl transfer reactions. Conformational rearrangements preceding the chemical step and subsequent proton transfer events may contribute to the overall kinetics, although it appears unlikely that these effects alone account for a discrepancy of this magnitude. Importantly, the present simulations also show that subtle structural differences can substantially affect the energetics, for example through the identity

of the oxygen atoms coordinating the magnesium cation, even when the average coordination number is similar. In the DFTB- Δ ML/MM simulations, the average magnesium coordination increases by approximately 0.3–0.6 oxygen atoms relative to the DFTB/MM simulations, raising the question of whether the magnesium coordination environment in the *trans*-HK active site is accurately represented and whether additional sampling or alternative starting structures are required.

Lassila et al.⁶ point out that enzymatic phosphoryl transfer does not substantially alter the nature of the transition state when going from water to the protein environment, but instead mainly stabilises the transition state, thereby lowering the barrier and facilitating catalysis. Accordingly, the hexacoordinated state of magnesium in solution suggests that a similar coordination geometry may be expected in the protein active site, as supported by structural and computational studies of histidine kinases.^{15,254,255} While magnesium can adopt a tetracoordinated geometry, it shows a strong preference for octahedral hexacoordination.^{256,257} Against this background, the DFTB- Δ ML/MM simulations indeed show an improvement towards a more complete coordination of the magnesium cation. However, only five rather than six oxygen atoms coordinate the magnesium cation. This may indicate limitations of the current structural model, which could contribute to the instability of the transition state and product by not fully satisfying the geometric requirements for hexacoordination. Furthermore, the Δ ML model was trained primarily on near-tetracoordinated magnesium environments, which may limit its ability to accurately describe fully hexacoordinated geometries. Future work should therefore aim to extend the training data to better capture fully hexacoordinated magnesium environments.

In addition, removing the N–P–O angle restraint was found to have a strong impact on the location of the reactant and product minima, suggesting that further slow structural rearrangements may be required to access a more stable product configuration. This interpretation is consistent with the two product minima reported by Kansari et al., where the second minimum was almost 10 kcal mol^{−1} lower than the first. It is therefore possible that the structural rearrangement leading to a lower-energy product was not accessible within the 10 ns time scale considered here. For extended simulations, it may be necessary to adjust sampling parameters (e.g., using a lower bias factor) to enable longer simulations and to better capture slow relaxation processes.

Overall, the strong sensitivity to the protein environment implies that small differences in local structure can lead to substantial changes in the free energy surface, which can also hinder convergence. Consequently, the focus of the present comparison is placed primarily on qualitative trends across methods and setups, rather than on absolute quantitative agreement with experiment.

10.3.3. Computational Cost

In Table 10.1, it was shown that DFTB- Δ ML/MM single-point calculations reduce the atomic force RMSE by more than 50%, while increasing the computational time by only a factor of 1.76. For the metadynamics simulations, Table 10.4 demonstrates that the computational time increases only slightly more than twofold (by a factor of 2.22). Notably, this corresponds to a smaller relative overhead than observed for simulations in water (refer to Section 9.3.3), even though more than twice as many QM atoms are corrected in the enzyme system (30 atoms in the histidine test system in water compared to 69 atoms in the protein). This can be explained by the fact that, in the enzyme simulations, both the QM and MM regions are substantially larger. As a consequence, the

Δ ML correction constitutes a smaller fraction of the total computational workload, and the relative increase in wall-clock time is therefore reduced.

Table 10.4.: Mean computational cost of metadynamics simulations, averaged over all walkers, in days/ns. The relative computational cost is defined as the ratio of DFTB- Δ ML/MM and DFTB/MM.

DFTB/MM	DFTB- Δ ML/MM	Rel. comp. cost
5.89	13.06	2.22

In comparison to the work of Giese et al.²²⁷, who applied a Δ ML correction based on the graph neural network MACE to O-phosphorylation model systems mimicking RNA transphosphorylation, the computational overhead observed in this work remains modest. Giese et al. reported that their approach is approximately 1.28 times slower than DFTB on GPUs; however, when executed on CPUs, the slowdown is increased to a factor of 30.

In the present work, the Δ ML correction increases the computational cost by only a factor of 2.22 relative to DFTB on CPUs, showing that accurate Δ ML/MM simulations remain feasible on standard CPU hardware, without requiring specialised GPU acceleration.

Furthermore, while Giese et al. demonstrated extrapolation abilities by learning predominantly in the minima and accurately predicting intermediate regions of the free energy surface, their results were obtained for non-enzymatic model reactions in aqueous solution. In contrast, the present work addresses a substantially more complex enzymatic environment, including explicit QM/MM coupling and magnesium coordination, which proved to be the most challenging component of the system.

10.4. Conclusions

In chapter 8, a vacuum-trained Δ ML HDNNP was developed, and in chapter 9, its transferability to aqueous solution was assessed. In this chapter, the Δ ML approach was extended to a protein environment, specifically a *trans*-histidine kinase (*trans*-HK), addressing whether a vacuum-trained model can be reliably transferred to a complex enzymatic QM/MM setting.

Single-Point Calculations

Chapter 9 showed that an accurate description of the Mg^{2+} coordination sphere is crucial for transfer from vacuum to solvated systems, requiring inclusion of all oxygen atoms in the first coordination shell. The protein results in this chapter confirm this finding. Including the complete first coordination shell substantially reduced the atomic force RMSE, particularly for magnesium, highlighting the decisive role of its local electrostatic environment for transferability. Although this increases the size of the corrected QM subset and thus the computational cost, efficiency can be recovered by adjusting the symmetry-function setup. Whereas reducing the cut-off radius increased the RMSE, reducing the number of symmetry functions maintained accuracy while accelerating the calculations by a factor of approximately 1.5. This offers a practical compromise between computational cost and predictive performance for biomolecular QM/MM applications.

A further key factor limiting transferability was QM/MM overpolarisation. DFTB is particularly sensitive to the electrostatic embedding, and unscaled MM charges led to systematic overpolarisation that impaired transfer from vacuum to water. Uniform scaling of all MM point charges

by a factor of 0.75 further reduced the atomic force RMSE, yielding an overall improvement of 54.8% compared to DFTB/MM. This highlights the decisive role of the electrostatic treatment at the QM/MM boundary. However, uniform scaling is a coarse approximation, as primarily charges close to the QM/MM interface contribute to overpolarisation. A more physically motivated alternative would be a distance-dependent scheme, such as a Klopman–Ohno formulation²⁵⁸, which may improve transferability without weakening long-range electrostatics unnecessarily.

QM/MM Metadynamics in the Protein Environment

Building on these insights, the DFTB- Δ ML/MM approach was applied to enzymatic phosphoryl transfer simulations. The corrected potential remained stable in well-tempered metadynamics simulations exceeding 10 ns, demonstrating successful transfer to a complex protein environment. The computational cost increased only by a factor of 2.2 compared to standard DFTB/MM, allowing extended sampling at moderate cost.

In addition to the influence of the Δ -learning correction, structural factors were investigated systematically by comparing two starting structures that differed in the magnesium coordination motif. Although both featured approximately the same total number of coordinating oxygen atoms, the identity of the coordinating phosphoryl groups proved important. A configuration in which all three ATP phosphoryl groups coordinated the magnesium cation resulted in free-energy barriers approximately 40–45% lower than a setup with only two of the triphosphate coordinating. This indicates that complete coordination of the triphosphate moiety to magnesium plays an important role in facilitating catalysis in histidine kinases.

Interestingly, the DFTB- Δ ML/MM simulations increased the average magnesium coordination number to approximately 4.7, indicating a strengthening of Mg^{2+} –O interactions relative to DFTB/MM. Although this shifts the system closer to the expected hexacoordinated state of magnesium, a predominantly pentacoordinated environment is unlikely to correspond to the most stable coordination geometry. The absence of full hexacoordination may arise from structural constraints within the protein environment and requires further investigation.

The application of a flat-bottom restraint on the N–P–O angle was found not to be essential for enforcing an $\text{S}_{\text{N}}2$ -like transition state. Even without angular restraints, the transition state remained nearly linear. Instead, the restraint primarily destabilised the reactant state through geometric constraint, thereby artificially decreasing the relative barrier height. This highlights the importance of ensuring that any introduced restraints are physically reasonable and that their influence on the free-energy landscape is carefully assessed.

In comparison to the work from the group of Darrin York, who demonstrated improved accuracy for phosphoryl-transfer model reactions in aqueous solution using Δ -learning with MACE,²²⁷ the Δ -learning strategy presented here is applied to an explicit enzymatic environment, including full QM/MM coupling and magnesium coordination in a histidine kinase, and employs a comparatively high-level DFT reference (ω B97M-V/ma-def2-TZVP). Nevertheless, the computational overhead introduced by the correction remains moderate, enabling nanosecond-scale metadynamics simulations in the protein. At the same time, the remaining deviations in absolute barrier heights emphasise that achieving quantitative agreement in enzymatic systems remains challenging and depends not only on the electronic-structure description but also on adequate sampling and a balanced treatment of electrostatics.

Quantitative Limitations and Outlook

Despite the substantial reduction of atomic force errors by 50% compared to DFTB/MM, the calculated free-energy barriers remain higher than experimental estimates. This indicates that improvements in local forces alone do not guarantee quantitatively accurate reaction energetics in complex enzymatic systems. Accordingly, the present results are interpreted primarily in terms of qualitative mechanistic trends.

Several factors may contribute to the overestimation of the barrier and the product free energy. The accessible simulation time may still be insufficient to capture slow structural rearrangements, as conformational relaxation has been shown to stabilise product states by up to 10 kcal mol⁻¹.¹⁶ In addition, magnesium coordination exerts a strong influence on the barrier height. Although the Δ -learning correction increases the average coordination number, incomplete hexacoordination may lead to an artificial destabilisation of certain states. Future work should therefore focus on refining the electrostatic embedding, for instance through distance-dependent charge scaling schemes rather than uniform scaling, and on further structural investigations of the magnesium coordination environment.

11. Understanding the Catalytic Mechanism of Creatine Kinase

11.1. Introduction

Creatine kinases (CKs) are essential enzymes in vertebrate energy metabolism as they catalyse the reversible transfer of the γ -phosphoryl group from ATP to creatine to form phosphocreatine (pCr, refer to Figure 11.1).²⁵⁹ This reaction plays a central role in buffering and rapidly regenerating ATP levels, particularly in tissues with high and fluctuating energy demands, such as skeletal muscle, cardiac muscle, and brain.^{260,261} Maintaining a constantly high intracellular ATP concentration would interfere with ATP-dependent metabolic processes. Therefore, in such tissues, cells do not rely solely on ATP, but also store energy in the form of the phosphagen phosphocreatine. In invertebrates, arginine phosphate, which is formed by arginine kinase (AK), plays a similar role. As with other kinases in the guanidino kinase family, AK and CK show strong homology in sequence and structure, suggesting a shared evolutionary origin.^{259,262}

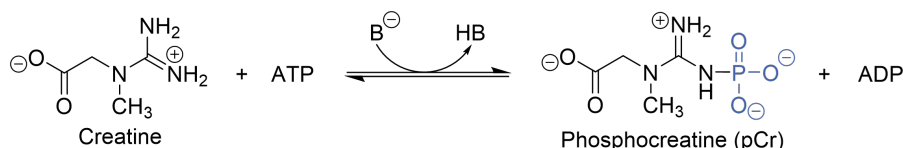


Figure 11.1.: Reaction equation of the phosphoryl transfer catalysed by CKs.

Because its phosphoryl-group transfer potential is higher than that of ATP, pCr enables the rapid phosphorylation of ADP to ATP.²⁶³ This allows cells to address rapid changes in energy demand with greater efficiency compared to slower ATP-generating pathways, such as oxidative phosphorylation. The CK/pCr system functions as a short-term energy buffer, thereby maintaining cellular energy homeostasis and supporting sustained activity in ATP-dependent processes.²⁶⁰

CKs exist in tissue-specific isoforms grouped into cytosolic and mitochondrial types. Cytosolic CKs include muscle-type (M-CK) and brain-type (B-CK) isoforms, which form active homo- or heterodimers (MM-CK, BB-CK, MB-CK). Mitochondrial CKs (MtCK or Mi-CK) comprise of two subtypes: uMtCK (ubiquitous/acidic, also known as Mi_a-CK) and sMtCK (sarcomeric/basic, also known as Mi_b-CK), which usually assemble into functional octamers formed from dimeric subunits, and are located in the intermembrane space of mitochondria.^{264,265}

MtCKs catalyse creatine phosphorylation at ATP production sites in the mitochondria, producing pCr that diffuses to areas of high energy demand. There, cytosolic CK regenerates ATP from pCr. This spatial separation together with the faster diffusion of pCr compared to ATP, forms the basis of the phosphocreatine shuttle (refer to Figure 11.2), which maintains local ATP levels and supports cellular energy homeostasis. This micro-compartmentation of creatine kinase isoforms offers

several advantages. In the vicinity of ATP-consuming ATPases, it enables a high ATP-to-ADP concentration ratio to be maintained. In contrast, within mitochondria, a low ATP-to-ADP ratio ensures the availability of ADP required for efficient oxidative phosphorylation.²⁶⁵

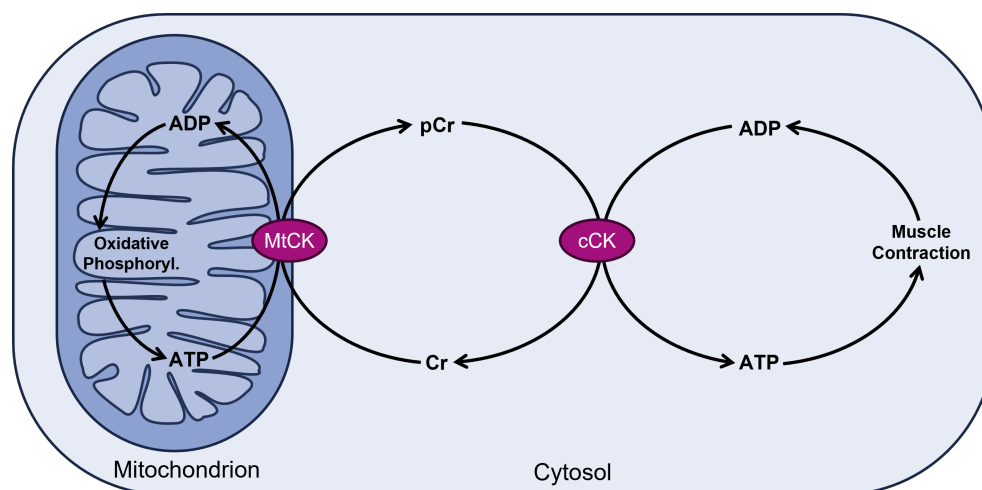


Figure 11.2.: Schematic representation of the phosphocreatine shuttle, adapted from References [265, 266]. In the mitochondrial intermembrane space, MtCK catalyses the formation of pCr using ATP generated during oxidative phosphorylation as a substrate. From there, the created pCr then diffuses to areas of the cell with high energy demand, where cytosolic CK isoforms regenerate ATP by transferring the phosphate group from pCr to ADP. The phosphocreatine shuttle thus facilitates the efficient distribution of energy across cellular compartments, thereby ensuring that local ATP concentrations remain sufficient at key functional sites.

Given their key role in energy metabolism and high abundance in muscle, brain, and heart, CKs are also widely used as a clinical biomarker.²⁶⁵ Tissue damage in these organs leads to elevated CK levels in the blood, aiding the diagnosis of conditions such as myocardial infarction (heart attack),²⁶⁷ muscle damage,²⁶⁸ and neuromuscular disorders.²⁶⁹ uMtCK is known to be overexpressed in various types of cancer cells.²⁶⁵ In Hodgkin lymphoma cells, inhibition of uMtCK by the creatine analogue cyclocreatine has an anti-proliferative effect without inducing apoptosis. This highlights the potential therapeutic relevance of targeting CK-dependent energy metabolism in apoptosis-resistant tumour cells.²⁷⁰ A similar metabolic effect is observed in breast cancer, where uMtCK is also overexpressed,^{271,272} and the tumour cells' reliance on uMtCK activity again makes cell growth vulnerable to cyclocreatine inhibition.^{14,271} These observations motivate a detailed mechanistic understanding of CK-dependent energy metabolism, particularly in the context of therapeutic strategies.

In CKs, conformational changes of two flexible loops near the active site play a critical role in substrate recognition and catalysis. Crystallographic and small-angle X-ray scattering studies have shown that these loops close over the active site, with the closed conformation being stabilised by a histidine and aspartate forming a 'latch'. This conformational transition is initiated upon substrate binding, reducing solvent access while aligning key residues as well as ATP and creatine for phosphoryl transfer.⁹⁻¹² This transition is accompanied by a measurable decrease in the protein's radius of gyration in the presence of Mg^{2+} -ATP or a transition state analogue, supporting the idea that loop closure is essential for catalysis.¹⁰

Figure 11.3 (left) shows a uMtCK dimer with an enlarged view of the active site, containing the ligand Mg^{2+} -ATP and creatine. The depicted Glu227 residue acts as the catalytic base in the proton transfer reaction.^{13,14,273} During the chemical reaction, a nearby base deprotonates creatine in a first

step.¹³ As a result, the guanidino nitrogen becomes nucleophilic, thereby enabling direct phosphorylation via formation of a P–N bond, accompanied by cleavage of the P–O bond to ADP (refer to Figure 11.3, right).

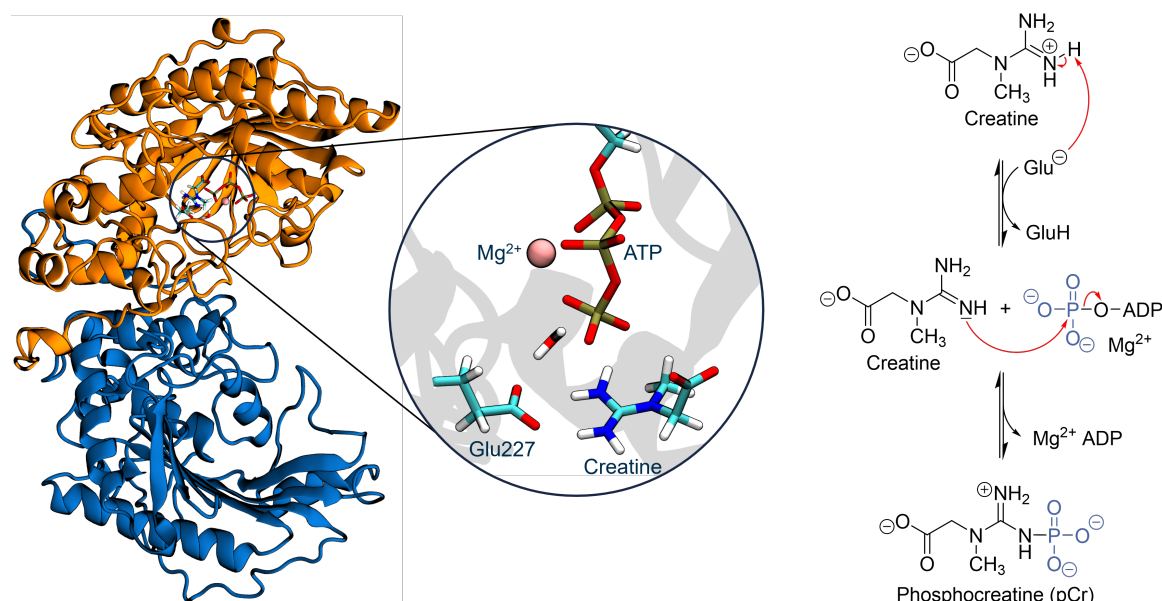


Figure 11.3.: Creatine kinase dimer from the octameric uMtCK, with enlarged active site consisting of the triphosphate of ATP, creatine, and a magnesium cation (left) and the chemical equation of the catalysed reaction (right).

In addition to the residues directly involved in proton and phosphoryl transfer, a conserved cysteine residue in the active site (Cys278 in uMtCK) has been proposed to influence catalysis.²⁷⁴ While early modification and mutagenesis studies classified this residue as essential,^{275–277} later site-directed mutagenesis studies demonstrated that it is neither directly involved in the chemical step nor essential for catalysis. Nevertheless, Cys278 contributes to substrate binding and its mutation leads to reduced catalytic activity.^{274,278} Structural evidence from X-ray crystallography shows that the corresponding cysteine (Cys283 in other CK isoforms) interacts with the non-nucleophilic η -nitrogen of creatine and has been suggested to contribute to stabilising and orienting the substrate for optimal catalysis.^{12,259}

11.1.1. Motivation and Aim of the Project

Although structural studies provide a general understanding of the reaction mechanism, the energetic landscape and the influence of surrounding amino acids remain unclear. Previous computational studies have aimed to shed light on the catalytic mechanism of creatine kinase.¹³ However, by focusing on the potential energies of only a limited number of structures, these approaches neglect the impact of protein dynamics. To address these limitations, this study employs quantum mechanics/molecular mechanics (QM/MM) simulations with DFTB3 as the QM method, combined with metadynamics as an enhanced sampling technique to gain deeper insight into the catalytic mechanism of uMtCK.

This work aims to simulate the opening and closing of the loops surrounding the active site, investigate the phosphorylation and deprotonation of creatine and discuss potential pitfalls in selecting collective variables (CVs) for metadynamics simulations. Furthermore, we investigate the influence

of the protonation state of the active-site cysteine on the phosphoryl transfer to better understand the reaction pathway of the phosphoryl transfer. Previous studies have reported conflicting mechanistic descriptions: while experimental studies suggest that the phosphoryl transfer proceeds via an associative S_N2 -like transition state,²⁷³ computational studies propose a dissociative S_N1 -like mechanism.¹³ Understanding how the protonation state of the active-site cysteine affects the reaction pathway may help to address these discrepancies.

Finally, we aim to provide an estimate of the reaction free energy ΔG for the entire catalytic cycle: from an enzyme with unbound ligands, through catalysis, to the post-catalysis state when the products are released and the enzyme returns unaltered to its initial state.

11.2. Methods

Structure Preparation

The crystal structure of human ubiquitous mitochondrial CK (PDB ID: 1QK1)²⁷⁹ was used as a starting structure. It provides the complete protein including all amino acids, but without bound ligands. To obtain a structure with bound ligands, the structures of a CK from the Pacific electric ray *Torpedo californica* (PDB ID: 1VRP)¹² and from European rabbit muscle in the R134K mutant (PDB ID: 1U6R)²⁶² were selected, as these contain bound transition-state analogue complexes (TSACs) consisting of ADP, NO_3^{2-} , Mg^{2+} , and creatine, similar to the approach used by Li et al.¹³

The protein structure from chain B of 1VRP were aligned with chain A of 1QK1 using Swiss-PdbViewer²⁸⁰, version 4.1.0. From this alignment, the positions of the ligands were taken from 1VRP, while the protein structure remained unchanged from 1QK1. Chain B from 1U6R was included as a secondary verification for the ligand placement. Hydrogen atoms were then added to the ligands using Avogadro 1.2.0^{193,194} at high pH, ensuring the correct protonation of creatine.

The protein was parametrised using the AMBER99SB-ILDN force field^{198,199}, and TIP3P water¹⁹⁷ was used for solvation. Creatine was parametrised with GAFF²⁸¹, and ATP parameters were taken from Meagher et al.²⁸². Electrostatic interactions were treated with the Particle Mesh Ewald (PME) method²⁰⁹, with a short-range cut-off of 1 nm. The Van der Waals interactions were also truncated at 1 nm, and the LINCS algorithm was used to constrain bonds involving hydrogen atoms. All simulations were carried out using our local implementation of QM/MM in GROMACS 2020.2^{200–204} and DFTB+^{205,206}, which is available from a public repository^{207,208}.

Energy minimisation was performed using the steepest descent algorithm, with convergence set to a maximum force of $1000.0 \text{ kJ mol}^{-1} \text{ nm}^{-1}$. Following the minimisation, NVT equilibration was performed using the leap-frog integrator⁵³ with a time step of 1 fs, at a temperature of 300 K, for a duration of 10 ns using the velocity rescaling thermostat⁵⁵. NpT equilibration was performed for an additional 10 ns using the Parrinello–Rahman barostat⁶⁰, with a 1 fs time step. Since the starting structure had no ligands bound, moving restraints were applied to close the active site around the ligands, mimicking the closed state for the next simulation steps. Restraints were placed using the open-source, community-developed PLUMED library²¹⁰, version 2.5.1²¹¹.

MM Metadynamics Simulations to Simulate the Loop Opening/Closing Transition

The starting point for the loop opening/closing transition was the equilibrated ligand-bound uMtCK structure, which reflects the closed conformation of the active site. 1D metadynamics simulations

were performed to capture the dynamics of the loop opening and closing events. The distance between the two inter-loop hydrogen bond forming atoms (H_e of His61 and the carboxylate oxygen of Asp321) was employed as collective variable (CV) to facilitate the transition. The simulation time step was set to 2 fs, and the simulations were run for 200 ns.

QM/MM Preparation for the Chemical Steps

For the QM/MM simulations of the chemical steps, the system was divided into a quantum mechanical (QM) region and a molecular mechanics (MM) region. The QM region, which contains the active site with creatine, triphosphate of ATP, magnesium cation, the side chain of Glu227 as proton acceptor, and eight water molecules, was simulated using DFTB3 with the 3OB parameter set^{192,195,196} and the specific reaction parameterisations (SRPs) OPhyd¹⁹² and PNmod¹⁶. The MM region was described using TIP3P water and the AMBER99SB-ILDN force field. Electrostatic interactions between the QM and MM regions were treated using the electronic embedding scheme. The system was equilibrated with an initial energy minimisation followed by a short equilibration phase using the velocity rescaling thermostat to maintain stability.

QM/MM Metadynamics Simulations of the Chemical Steps

To investigate the chemical reaction in the QM/MM framework, multiple-walker well-tempered metadynamics simulations were carried out.

Phosphoryl transfer. For the phosphoryl transfer, the P–N and P–O distances were selected as CVs, with the respective atomic distances defined according to Figure 11.4. Gaussians with a width of 0.02 nm were deposited and exchanged between the individual walkers every 0.5 ps. An initial height of 1.2 kJ mol^{-1} and bias factor of 50 were chosen to ensure a controlled exploration of the FES. The simulation was performed with a time step of 1 fs and with each of the 24 walkers running for 1 ns.

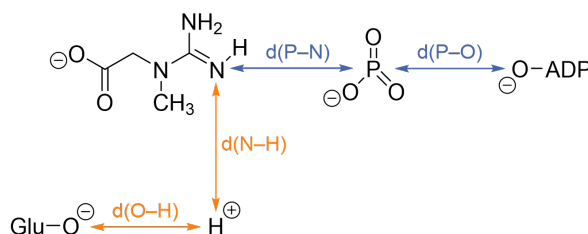


Figure 11.4.: Schematic representation of the collective variables (CVs) used for the metadynamics simulations of CK.

Direct proton transfer. For the direct proton transfer, the N–H and O–H distances were used as CVs (refer to Figure 8.4).

Water-mediated proton transfer. For the initial water-mediated proton transfer, one water molecule was included in the CV definition. The two proton transfer CVs were defined as

$$\text{profTrans}_1 = d(O_{H_2O}-H_{H_2O}) - d(O_{Glu227}-H_{H_2O}) \quad (11.1)$$

$$\text{profTrans}_2 = d(N_{Cre}-H_{Cre}) - d(O_{H_2O}-H_{Cre}), \quad (11.2)$$

with the respective atomic distances defined according to Figure 11.5.

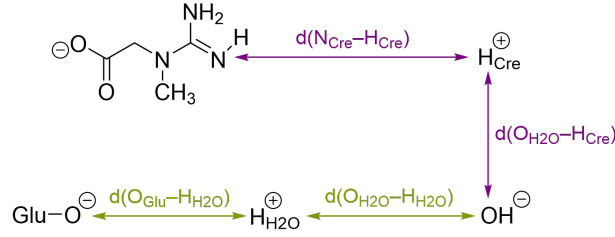


Figure 11.5.: Schematic representation of the CVs used for the water-mediated proton transfer.

Water-mediated proton transfer with mCEC as CV. In the final proton transfer simulation, an alternative collective variable, the modified centre of excess charge (mCEC) coordinate introduced by König et al.²⁸³, was employed. The mCEC coordinate ξ calculates the spatial location of the excess proton being transferred, and is defined as

$$\xi = \sum_{i=1}^{N_H} \mathbf{r}^{H_i} - \sum_{j=1}^{N_X} w^{X_j} \mathbf{r}^{X_j} - \sum_{i=1}^{N_H} \sum_{j=1}^{N_X} f_{sw}(d^{H_i, X_j}) (\mathbf{r}^{H_i} - \mathbf{r}^{X_j}), \quad (11.3)$$

with the number of hydrogen atoms N_H and heavy atoms N_X , and their respective coordinates $\mathbf{r}^{H_i}$ and $\mathbf{r}^{X_j}$. The switching function f_{sw} is given as

$$f_{sw}(d^{H_i, X_j}) = \left(1 + \exp \left(\frac{d^{H_i, X_j} - r_{sw}}{d_{sw}} \right) \right)^{-1}, \quad (11.4)$$

with the switching radius r_{sw} and switching width d_{sw} and the distance d^{H_i, X_j} between a hydrogen atom H and a heavy atom X. Here, the parameters were chosen as $r_{sw} = 0.125$ and $d_{sw} = 0.006$. For a visualisation, refer to Figure D.1. The mCEC vector ξ can also be expressed as a scalar, which is obtained as

$$\zeta = \frac{d_{\xi, D}}{d_{\xi, D} + d_{\xi, A}}. \quad (11.5)$$

ζ was employed as mCEC CV in this work. As donor, the N_η of the creatine guanidinium group was selected. For the acceptor, the carbon atom of the glutamate was chosen in order to capture a protonation of either carboxylate oxygen atoms.

Analysis

The visualisation software VMD (version 1.9.4)²⁴⁴ was used to perform trajectory analyses and to generate figures for this work. Distance calculations were carried out with PLUMED²¹⁰, version 2.5.1²¹¹. All data analysis and visualisation were performed in Python 3.11²⁴⁵ using NumPy (version 1.26.0)²⁴⁶, pandas (version 2.1.3)^{247,248}, and Matplotlib (version 3.8.2)²⁴⁹.

11.3. Results and Discussion

In this section, the conformational transition formed by the latching of two flexible loops, which close the active site to retain the reacting molecules (refer to Section 11.3.1), as well as the subsequent catalytic steps comprising phosphoryl and proton transfer (refer to Section 11.3.2), are discussed.

11.3.1. Conformational Transition

The conformational transition in CKs is formed by the motion of two flexible loops near the active site, comprising of residues 60–65 and 316–326. The closed state is stabilised by a hydrogen bond between the two loops, formed between the amino acids His61 and Asp321.^{9–12}

The starting structure did not contain any bound ligands. Upon addition of ATP and creatine and subsequent equilibration, the substrates remained near the active site only transiently. In unbiased MD simulations, both ligands dissociated within a few nanoseconds, consistent with the idea that a fully open loop conformation is not sufficient to retain the substrates. However, when loop closure was enforced using a moving restraint and stabilised by a hydrogen bond between His61 and Asp321, the ligands remained bound throughout the simulation. This supports the proposed role of loop closure in creating a catalytically competent environment and reinforces previous structural and biophysical studies highlighting the functional relevance of these conformational changes.^{13,284}

To further explore the energetics of this conformational transition, metadynamics simulations were carried out. For this, the distance between the two interacting amino acids, His61 and Asp321 was employed as CV. The simulated inter-loop connection positioned the reacting molecules in closer proximity (refer to Figure 11.6, right). The resulting FES (refer to Figure 11.6, left) reveals a barrier of approximately 3 kcal mol^{-1} separating the open and closed states, which is within typical energetic range associated with breaking or forming a hydrogen bond.²⁸⁵ This moderate barrier suggests that loop motion is not overly restricted, instead acting as a flexible gate enabling quick transitions between the open and closed states during catalysis.

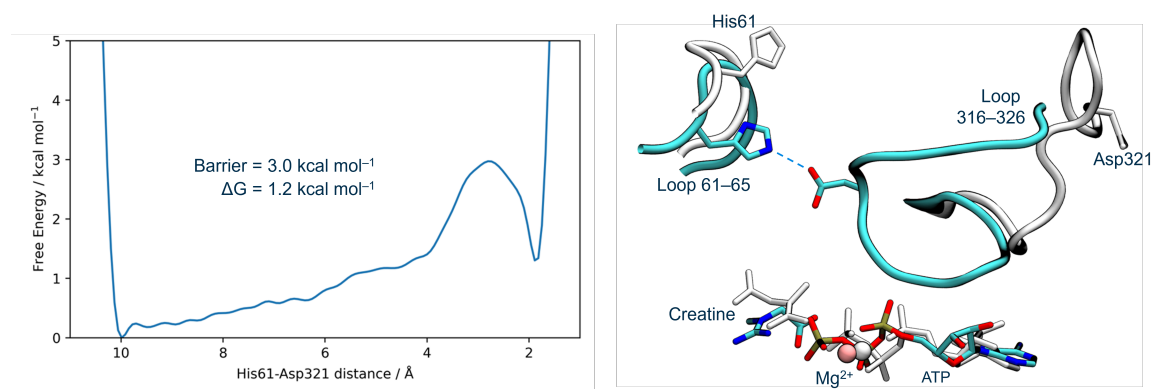


Figure 11.6.: The 1D FES of the loop-closing transition reveals a low barrier of 3 kcal mol^{-1} (left). Loops 61–65 and 316–326 can be connected by a hydrogen bond between His61 and Asp321, effectively forming a ‘latch’ over the active site (right). The open state is shown in white and the closed state in colour. In the closed state, creatine and ATP move into closer proximity as needed for catalysis.

Once the loops are closed, the active site adopts a conformation that enables the chemical reaction to take place. Initially, as expected under physiological pH, creatine is present in its protonated form with a guanidinium group. In its phosphorylated form, however, creatine is deprotonated. To model the proton transfer, Glu227 was selected as proton acceptor, in line with previous literature suggestions.^{13,14,273}

11.3.2. Chemical Steps

The catalysed chemical reaction comprises two steps: a phosphoryl transfer and a proton transfer reaction. The first mechanistic question to address was therefore which of these two events occurs first.

To investigate this, the phosphoryl transfer was simulated with the P–N and P–O distances as CVs (refer to Figure 11.4). Two scenarios were considered for the simulation of the phosphoryl transfer. In the first scenario, the simulation was initiated from the pre-proton transfer state, in which the proton remains bound to creatine, resulting in a positively charged guanidinium group. In the second scenario, the starting structure represented the state following the initial proton transfer to Glu227, leaving creatine in its neutral guanidino form. In the simulation starting from the protonated guanidinium state, no product formation was observed (refer to Figure 11.7). This clearly indicates that the proton transfer must precede the phosphoryl transfer. Chemically, this is reasonable, as the guanidinium group already possesses a fully occupied valence shell and is therefore not suitable to act as a nucleophile. Furthermore, this order of chemical steps is in accordance with previous computational studies.¹³ A similar qualitative trend was also observed in simulations in aqueous solution in chapter 8, where the creatine had to be deprotonated for a phosphoryl transfer to take place.

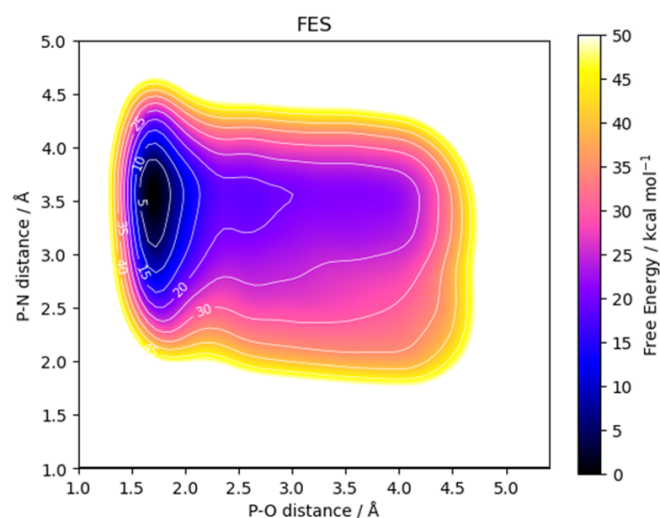


Figure 11.7.: FES of phosphoryl transfer to a protonated guanidinium. The absence of a minimum in the expected product region around (4.0 | 1.8) indicates that no product is formed. This can be attributed to the fully occupied valence shell of the guanidinium group.

11.3.2.1. Proton Transfer

For the simulation of the proton transfer, Glu227 was selected as proton acceptor, as previous experimental and computational investigations have identified this residue as the most plausible catalytic base.^{13,14,273}

Direct Proton Transfer

Therefore, the N–H and O–H distances corresponding to the bonds being broken and formed during the reaction were initially selected as CVs (refer to Figure 11.4). Unfortunately, the resulting free energy surface (FES) revealed an unusually high barrier for the proton transfer from creatine to Glu227 of approximately 19 kcal mol^{-1} , and a free energy difference (ΔG) of 20 kcal mol^{-1} . This results in a negligible back-transfer barrier of below 1 kcal mol^{-1} and a high-energy, unstable product state. Investigation of the corresponding structures show a relatively long distance of 3–4 Å between the creatine proton and the glutamate oxygen, a geometry that appears unfavourable for direct proton transfer. This raises the question of whether the proton transfer may instead proceed via a water-mediated mechanism.

Water-Mediated Proton Transfer

Consequently, in a subsequent metadynamics simulation, a water-mediated proton transfer was simulated (refer to Figure 11.8). For this, the reaction was biased along two different proton transfer CVs (refer to Equations (11.1) and (11.2)). This simulation revealed that, in a first step, Glu227 deprotonates a nearby QM water molecule, generating a hydroxide ion, which subsequently deprotonates creatine in a second step. This hydroxide mediated mechanism is also known as ‘proton hole’ transfer.²⁸⁶

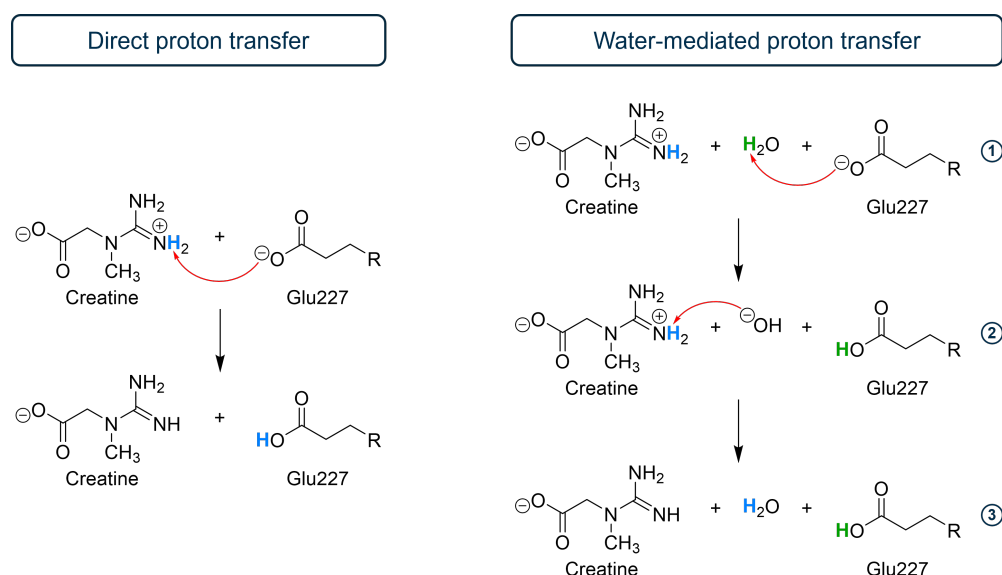


Figure 11.8.: Comparison of a direct proton transfer pathway and a water-mediated mechanism, in which Glu227 first deprotonates a water molecule, generating a hydroxide ion that subsequently deprotonates creatine.

A limitation of employing simple CVs defined solely by distances between two specific atoms is that they implicitly assume that exactly this bond is formed or broken. In the present simulation, however, it was observed that occasionally a different hydrogen atom of the water molecule was transferred. Although the product is chemically identical, alternative proton transfer events cannot be captured by the chosen CV, which results in misleading artefacts in the FES.

Water-Mediated Proton Transfer with mCEC as CV

As an alternative CV, the modified centre of excess charge (mCEC) coordinate ζ introduced by König et al.²⁸³ was employed. This coordinate has previously been applied successfully to describe proton transfer processes in different proteins.^{226,287} The incorporated switching function (refer to Equation (11.4)) enables a continuous transition between ‘bound’ and ‘unbound’ states of a hydrogen–heavy atom pair based on their interatomic distance, and serves as a weight for bond-like configurations.

The obtained FES is shown in Figure 11.9. With the mCEC coordinate ζ as CV, values near zero correspond to the reactant state before any proton transfer has taken place (state 1, as given in Figure 11.8). The product state is found near $\zeta = 0.75$ and corresponds to both proton transfer steps being completed (state 3, as given in Figure 11.8). The maximum value remains below 1 since the carboxylate carbon was selected as reference, allowing both oxygen atoms of the carboxylate group to act equivalently as proton acceptors.

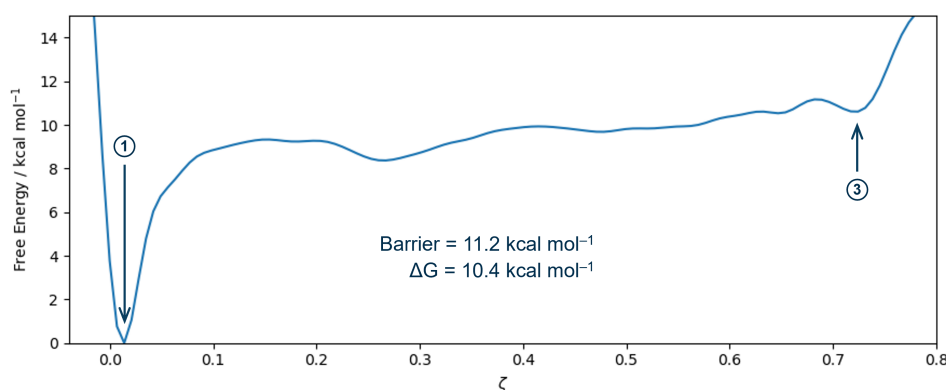


Figure 11.9.: FES of the CK proton transfer along the mCEC coordinate ζ . States 1 and 3 correspond to the states before proton transfer and after completion of both proton-transfer steps, respectively (refer to Figure 11.8).

The FES reveals a free energy difference ΔG of approximately 10 kcal mol^{-1} and a barrier of approximately 11 kcal mol^{-1} . Again, the almost negligible back-transfer barrier suggests that the product state is not stabilised. Further analysis is required to determine which destabilising factors in the immediate environment contribute to this behaviour. Nevertheless, both the barrier and ΔG are reduced by nearly 10 kcal mol^{-1} relative to the direct proton transfer. This supports the hypothesis that the reaction proceeds via a water-mediated mechanism.

11.3.2.2. Phosphoryl Transfer

After the proton transfer, the phosphoryl transfer can take place. For the metadynamics simulations, the P–N and P–O distances were employed as CV (refer to Figure 11.4). Whereas the transfer of the phosphoryl group to the positively charged guanidinium group of creatine resulted in no product formation (refer to Figure 11.7), the phosphoryl transfer to the neutral (deprotonated) guanidino group proceeds successfully. In the latter case, the phosphoryl transfer becomes not only feasible but also thermodynamically favourable, with a ΔG of approximately -9 kcal mol^{-1} (refer to Figure 11.10). The minimum energy path proceeds through a pentavalent S_N2 -like transition state with a barrier height of approximately 18 kcal mol^{-1} .

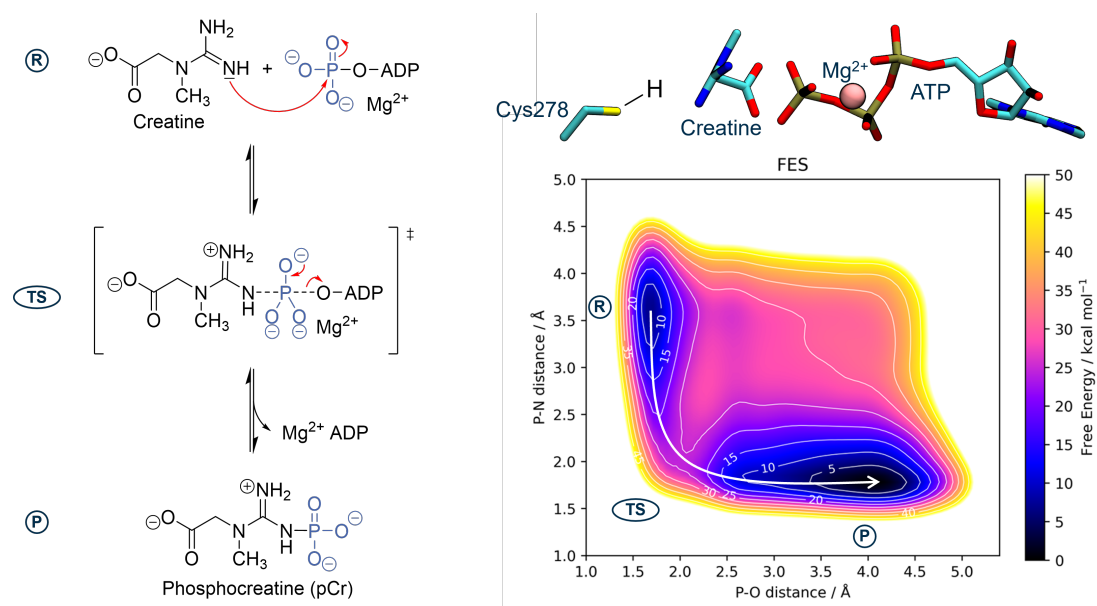


Figure 11.10.: Influence of the protonation state of Cys278 on the phosphoryl transfer reaction. With Cys278 in its thiol form, the phosphoryl transfer proceeds via an S_N2 -like mechanism. Left: Reaction equation with pentavalent transition state. Top right: Position of Cys278 near the substrates in the active site of uMtCK. Bottom right: FES obtained when Cys278 is modelled as a thiol, revealing an S_N2 -like mechanism with a barrier of 18 kcal mol^{-1} and a ΔG of -7 kcal mol^{-1} .

To investigate the potential influence of local redox conditions on catalysis, a cysteine residue (Cys278), located in the creatine-binding pocket, was modelled in both its neutral thiol and deprotonated thiolate forms. When present in its deprotonated thiolate form, this cysteine introduces an additional negative charge into the active site environment. The FES with the thiolate present (refer to Figure 11.11) indicates that while the reaction barrier for phosphoryl transfer remains largely unchanged compared to the FES of neutral Cys278, the overall thermodynamic stability of the pCr product is reduced by approximately 4 kcal mol^{-1} , yielding a ΔG of only -3 kcal mol^{-1} . This effect is likely due to electrostatic repulsion between the negatively charged thiolate and the phosphate group or subtle disruption of the local Mg^{2+} coordination environment. No major conformational shifts in the loops or substrate orientation were observed, though subtle changes in the solvation environment may also contribute.

Notably, the equivalent cysteine residue in human muscle CK (Cys283) has an unusually low pK_a of approximately 5.7,²⁸⁸ which is about three pH units lower the pK_a of a regular cysteine residue with 9.1.^{289,290} This is attributed to hydrogen bonds within the conserved cysteine-proline-serine (CPS) motif.²⁸⁸ MD simulations indicate that the ionised thiolate form is necessary to maintain substrate alignment during catalysis.²⁸⁸ The low pK_a of Cys283 in human muscle CK and its role in optimal creatine binding have also been demonstrated by site-directed mutagenesis, with mutations resulting in a reduced activity, although catalysis is not abolished.²⁷⁴

Previous studies indicate that the active-site cysteine in muscle CK predominantly exists as a deprotonated thiolate during catalysis, facilitating substrate binding and proper alignment.^{274,288} Our findings suggest that in uMtCK the thiolate form of Cys278 may introduce local electrostatic strain that destabilises the pCr product. However, this product destabilisation may facilitate product release and turnover by lowering product affinity, thus minimising product inhibition and improving

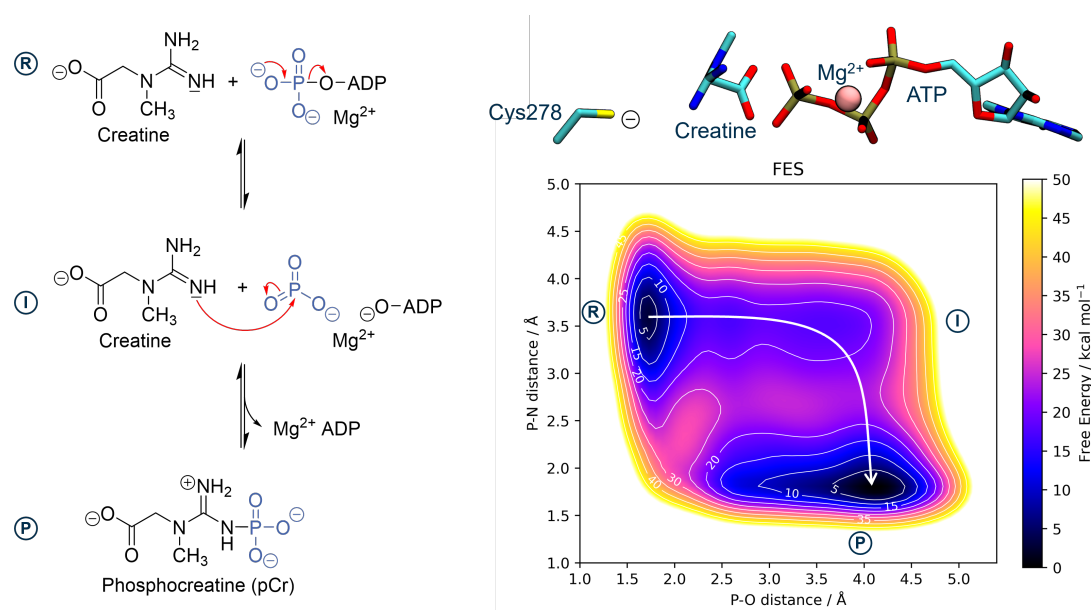


Figure 11.11.: Influence of the protonation state of Cys278 on the phosphoryl transfer reaction. With Cys278 in its thiolate form, the phosphoryl transfer proceeds via an S_N1 -like mechanism. Left: Reaction equation with metaphosphate intermediate. Top right: Position of Cys278 near the substrates in the active site of uMtCK. Bottom right: FES obtained when Cys278 is modelled as a thiolate, revealing an S_N1 -like mechanism and a slight destabilisation of the pCr product compared to the FES with the cysteine in its thiol form. The barrier has a height of 19 kcal mol^{-1} , the ΔG is obtained as -3 kcal mol^{-1} .

catalytic efficiency. Therefore, these catalytic advantages support the hypothesis that Cys278 predominantly exists in its deprotonated thiolate form during catalysis. Nevertheless, further investigation is required to identify the precise role of this cysteine residue during uMtCK catalysis.

Although the thiol and thiolate forms of Cys278 lead to slightly different electrostatic environments, the calculated phosphoryl-transfer barriers remain comparatively high, with values of 18 kcal mol^{-1} and 19 kcal mol^{-1} , respectively. While these barriers are slightly lower than the $20.5 \text{ kcal mol}^{-1}$ obtained for the corresponding non-enzymatic reaction in water (refer to Table 8.1), they still appear high for an enzymatic reaction. This may indicate that additional stabilising factors are not fully captured in the present simulations, or that the initial structure and ligand placement do not fully represent the catalytically optimal configuration of the active site.

11.3.3. Setting the Results into Context

The barriers obtained for the conformational transition and chemical steps are summarised in Figure 11.12. In order to release the products after successful catalysis, the loops must open again, which is here considered as the reverse of the loop-closing process. The overall barrier that needs to be overcome is mainly caused by the chemical steps. Due to the small back-transfer barrier of the proton transfer, the two steps must be considered together, amounting to 29 kcal mol^{-1} . This is not only higher than the experimentally determined barrier of approximately $16.7 \text{ kcal mol}^{-1}$,¹³ but when the transition state theory (TST, refer to Equation (6.2)) is used to estimate the corresponding rate constant, it corresponds to a time scale of seven years. Since the total barrier is composed of several steps, potential sources of error can arise in each of them, contributing to the overall deviation from experimental data. In general, the surrounding environment can have a substantial impact, as observed for the histidine kinase (HK) in chapter 10 and for CK investigated here. Because the

active sites in both kinases are highly charged, even small changes in surrounding residues may alter the electrostatic environment, as observed in this chapter for Cys278. This could be caused, for example, by an unfavourable placement of the ligands in the initial structure, necessary side-chain movement that was not captured on the simulated timescale, or incorrect protonation states of surrounding amino acids. While the effect of the protonation state of Cys278 was evaluated for the phosphoryl transfer step, this has not yet been investigated for the proton transfer and should therefore be considered in future simulations.

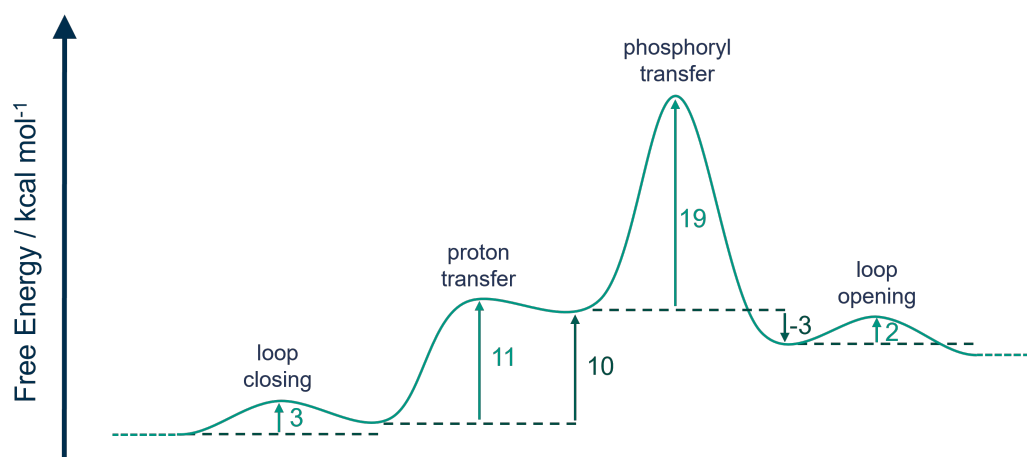


Figure 11.12.: Free energy profile combining conformational transitions and chemical steps. For the phosphoryl transfer, the energies obtained with Cys278 in its thiolate form are used, as the negative ΔG may facilitate product release and turnover.

However, not only the limitations of the chemical step but also those of the conformational transition must be discussed. While imperfect substrate alignment can certainly affect the chemical steps, it can also influence the movement of the loops into the active site. In this work, the interaction between His61 and Asp321 (refer to Figure 11.6) was modelled as a hydrogen bond, which is also supported by literature.¹³ However, other studies describe a salt bridge between the amino acids of the two loops.¹⁴ Furthermore, since the loops play an important role in retaining the substrates in the active site and aligning them for catalysis, a back-transfer barrier of only 2 kcal mol⁻¹ suggests that the full conformational process may not have been captured. For example, a final rearrangement step may not have been sampled. Additionally, while in this work the loops were closed simultaneously, Demir et al.¹⁴ report that one loop first closes over ATP, followed by the second loop closing over creatine. Moreover, to obtain a more reliable estimate of the back-transfer barrier, simulations should be performed with the products bound in the active site.

11.3.4. ΔG Estimation of the Complete Process

In order to estimate the ΔG of the entire process, all steps must be considered from empty enzyme and unbound ligands to after the catalysis is finished and the enzyme emerges unchanged from the reaction. Three steps are missing at this point: the binding of the reactants, the release of the products and the proton transfer from the glutamate to the environment, bringing the protein back into its initial state. The ΔG of the last step can be estimated from the acidity constant pK_a via

$$\Delta G = -RT \ln(K) = -RT \Delta pK_a. \quad (11.6)$$

For the deprotonation of glutamate, $\Delta G \approx 6 \text{ kcal mol}^{-1}$ is obtained, under the assumption that water functions as proton acceptor.

The energies of the product binding and reactant release on the other hand cannot be computed via Equation (11.6). Instead, a thermodynamic cycle can be used (refer to Figure 11.13). Here, advantage is taken of the fact that free energy is a state function and therefore the path from reactant to product is irrelevant. The thermodynamic cycle enables to calculate $\Delta\Delta G$ as

$$\begin{aligned}\Delta G_{\text{protein}} + \Delta G_{\text{product-release}} - \Delta G_{\text{water}} - \Delta G_{\text{reactant-release}} &= 0 \text{ kcal mol}^{-1} \\ \Delta\Delta G = \Delta G_{\text{protein}} - \Delta G_{\text{water}} &= \Delta G_{\text{reactant-release}} - \Delta G_{\text{product-release}}.\end{aligned}\quad (11.7)$$

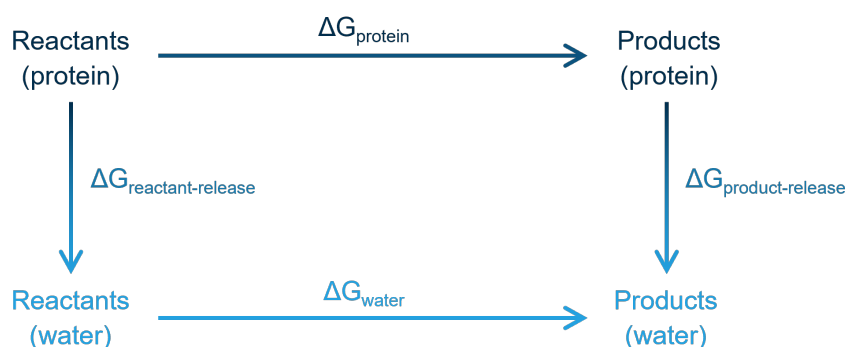


Figure 11.13.: Thermodynamic cycle relating the reaction free energies in the protein and in aqueous solution.

The phosphorylation of creatine was simulated in water and yielded a ΔG of $-19 \text{ kcal mol}^{-1}$ (refer to Table 8.1). The free energy of the proton transfer from creatine to water can again be estimated using Equation (11.6), yielding $\Delta G \approx 17 \text{ kcal mol}^{-1}$. Consequently, for water, this results in $\Delta G_{\text{water}} \approx -2 \text{ kcal mol}^{-1}$.

In this chapter, ΔG values of the proton and phosphoryl transfer were obtained as 10 kcal mol^{-1} and -3 kcal mol^{-1} , respectively. This is followed by a proton transfer from glutamate to water of approximately 6 kcal mol^{-1} . In total, for the enzyme-catalysed reaction, this sums up as $\Delta G_{\text{protein}} \approx 13 \text{ kcal mol}^{-1}$.

Plugging these two values into Equation (11.7) results in

$$\begin{aligned}\Delta\Delta G &= \Delta G_{\text{reactant-release}} - \Delta G_{\text{product-release}} \\ &= 13 \text{ kcal mol}^{-1} - (-2 \text{ kcal mol}^{-1}) = 15 \text{ kcal mol}^{-1},\end{aligned}\quad (11.8)$$

indicating that the reactants bind more strongly to the protein than the products. This could provide the necessary driving force for catalysis and also prevent product inhibition. Experimental studies likewise report that the reactants bind more strongly to the protein than the products, but their obtained binding energies correspond to lower $\Delta\Delta G$ values of approximately $3\text{--}6 \text{ kcal mol}^{-1}$. These values were obtained using $\Delta G_{\text{binding}} = -RT \ln K_d$, resulting in $\Delta G_{\text{reactant-binding}}$ and $\Delta G_{\text{product-binding}}$ values of approximately $14\text{--}15 \text{ kcal mol}^{-1}$ and $9\text{--}12 \text{ kcal mol}^{-1}$, respectively.¹⁴ They further report that creatine binds more strongly when the transition-state analogue complex (TSAC), formed by

ADP and NO_3^- , is already bound. Differences between the values obtained from the thermodynamic cycle and the experimental data likely originate from uncertainties associated with the individual simulation steps described above.

11.4. Conclusions

In this work, QM/MM metadynamics simulations were employed to investigate key steps of the catalytic cycle of uMtCK, including loop dynamics, phosphoryl transfer, and proton transfer reactions. By combining the semi-empirical DFTB3 method with enhanced sampling techniques, the aim was to obtain a dynamic and energetic insights of the catalytic mechanism beyond static structural models.

Conformational Transition. The calculated free-energy barrier for loop closing around the active site is consistent with the energetic scale expected for hydrogen bond formation and breaking. However, the reverse barrier for loop opening is so low that it is unlikely that the loops remain closed throughout the entire catalytic cycle, even though the closed conformation would be necessary to properly align the substrates for efficient catalysis. This suggests that the chosen CV may not fully capture the structural rearrangements associated with loop closure. Furthermore, it has been discussed in the literature whether the two amino acids actually form a hydrogen bond or rather a salt bridge.¹⁴ Future work should therefore investigate alternative scenarios, such as simulations with a protonated histidine to test salt-bridge formation, as well as a sequential loop-closing mechanism as recently proposed.¹⁴

Proton Transfer The proton-transfer simulations highlight the methodological challenges associated with selecting suitable collective variables (CVs). When assuming a direct proton transfer, the calculated barrier was unrealistically high. Switching to a water-mediated pathway reduced the barrier substantially, indicating that proton transfer may proceed via a bridging water molecule. However, the initial CVs were not able to capture a chemically reasonable proton exchange along this pathway. While the final CV can account for this, the almost negligible back-transfer barrier suggests that the deprotonated creatine is not stabilised in the present simulations. Further analysis is therefore required to identify destabilising factors in the immediate environment that may influence the proton-transfer step. In particular, the protonation state of Cys278 should be investigated, as it was only considered in its thiolate form in the proton-transfer simulations performed here.

Phosphoryl Transfer The phosphoryl transfer simulations further highlight the importance of the protonation state of the active-site cysteine. Previous studies suggest that the catalytic cysteine in muscle CK predominantly exists in its deprotonated thiolate form during catalysis, facilitating substrate binding and proper alignment.^{274,288} In agreement with this hypothesis, our results indicate that the thiolate form of Cys278 in uMtCK may introduce local electrostatic strain that destabilises the phosphocreatine product, which could facilitate product release and turnover by lowering product affinity and thereby reducing product inhibition. Nevertheless, the calculated phosphoryl-transfer barrier of 19 kcal mol^{-1} for the thiolate form remains comparatively high. Although slightly lower than the $20.5 \text{ kcal mol}^{-1}$ obtained for the corresponding non-enzymatic reaction in water (refer to Table 8.1), it may indicate that additional stabilising interactions or an optimal ligand arrangement are not fully captured in the present simulations.

Comparison to Experiment and Limitations. While the calculated phosphoryl-transfer barrier alone is in good agreement with experimentally observed turnover rates corresponding to barriers of approximately 17 kcal mol^{-1} ,¹³ the almost negligible back-transfer barrier of the proton-transfer step provides little stabilisation of the deprotonated creatine. Consequently, both steps must be considered together. This results in a combined barrier of approximately 39 kcal mol^{-1} for the complete chemical reaction, corresponding to a timescale of years. This timescale is incompatible with the biological role of creatine kinase in rapid cellular energy distribution and with experimentally measured turnover rates occurring on the millisecond timescale.¹⁴ This discrepancy highlights the need for further investigation of both chemical steps.

ΔG Estimation of the Complete Process. Finally, the overall reaction free energy ΔG of the complete catalytic cycle was estimated, considering all steps from the unbound enzyme and reactants, through catalysis, to the release of the products and restoration of the enzyme to its initial state. The analysis suggests that the reactants bind more strongly to the enzyme than the products, which may facilitate catalysis by promoting product release and preventing product inhibition. Using a thermodynamic cycle, a difference of approximately 15 kcal mol^{-1} between reactant and product binding was obtained. While this qualitative trend agrees with experimental observations, the calculated value exceeds experimentally derived estimates of $3\text{--}6 \text{ kcal mol}^{-1}$.¹⁴ These discrepancies likely arise from uncertainties associated with the individual free-energy contributions entering the thermodynamic cycle, including the energies obtained through simulation, as well as the estimations from pK_a values. Nevertheless, the results support the general mechanistic picture that stronger binding of the reactants relative to the products provides a thermodynamic driving force for efficient enzyme turnover.

Overall, the present QM/MM metadynamics simulations provide insights into several key steps of the catalytic cycle of uMtCK and highlight both mechanistic aspects and methodological challenges associated with modelling enzymatic phosphoryl transfer. While the obtained results provide insight into the reaction pathway, and allow for a qualitative interpretation of the catalytic energetics, they also reveal limitations of the present modelling approach and point to directions for future investigations. In particular, the interplay between proton transfer, phosphoryl transfer, and active-site electrostatics appears to play a critical role in determining the reaction energetics. As also discussed in chapter 10, simulations of enzymatic reactions are highly sensitive to the protein environment, where even small changes, such as the presence of a negatively charged cysteine, can significantly influence the reaction barrier. Further investigations, including improved sampling, alternative CVs, and refined descriptions of the active-site environment, will therefore be necessary to obtain a more complete and quantitatively accurate description of the catalytic mechanism of creatine kinase. Moreover, the simulations performed in this work could form a basis for future Δ -learning corrections of semi-empirical QM/MM simulations.

Part IV.

Summary and Outlook

12. Summary and Outlook

The work presented in this thesis aims to advance the mechanistic understanding of kinase-catalysed N-phosphorylation reactions using multiscale molecular simulations. Phosphoryl transfer reactions play a central role in biological energy metabolism and signalling, yet their computational investigation remains challenging as the relevant processes span a wide range of time scales. While enzymatic turnover may occur on microsecond to millisecond time scales, atomistic simulations typically access only microseconds, and the use of QM/MM methods further restricts accessible sampling times. To enable extensive sampling of phosphoryl transfer reactions in complex environments, the semi-empirical density-functional tight-binding method (DFTB) was employed as an efficient quantum mechanical description. However, although DFTB provides the required computational speed, its accuracy for phosphoryl transfer reactions is limited compared to density functional theory (DFT). To overcome this limitation, a Δ -machine learning (Δ ML) approach based on high-dimensional neural network potentials (HDNNP) was employed to correct DFTB energies towards DFT accuracy while retaining the computational efficiency of DFTB.

In chapter 8, the Δ ML approach was first evaluated for N-phosphorylation reactions in vacuum. Training data were generated for a diverse set of model systems undergoing phosphoryl transfer, including small test systems with different leaving groups, as well as biologically relevant molecules. When trained and tested on small systems, the resulting models achieved high accuracy, with minimum R^2 values of 0.969 and RMSE values of 1 kcal mol^{-1} or lower, reaching chemical accuracy. Application of the Δ ML correction to vacuum single-point calculations substantially improves the accuracy and precision of DFTB energies, reducing their RMSE by up to 93%. Although the models were trained only on energies, the atomic forces obtained as energy gradients were also significantly improved, with reductions in the force RMSE of more than 70%. Cross-validations further demonstrated that the approach can improve both energies and forces for systems not included in the training data, although the performance depends on chemical similarity and charge consistency between training and target systems. These results establish the Δ ML correction as a promising strategy for improving the description of phosphoryl transfer reactions while retaining the computational efficiency of DFTB.

In chapter 9, the application of the Δ ML model is extended to QM/MM metadynamics simulations of phosphoryl transfer in aqueous solution. While the DFTB- Δ ML/MM simulations reveal artefacts and side reactions, analysing these issues provides important insights for subsequent applications in protein environments. In particular, differences between vacuum and aqueous single-point energy distributions suggest an overpolarisation in DFTB, especially at the QM/MM interface. Furthermore, analysis of DFTB- Δ ML single-point forces indicates an overcorrection for magnesium and oxygen atoms in solution. This behaviour is attributed to the incomplete representation of the magnesium coordination sphere in the vacuum training data, emphasising the need to include all coordinating molecules when applying the model to systems with metal ions. These observations provide valuable guidance for the subsequent application of the Δ ML approach in enzymatic environments.

In chapter 10, building on the observations from the aqueous simulations, transferability was improved by including the complete Mg^{2+} coordination sphere in the corrected QM region and by scaling the MM charges by a factor of 0.75 to mitigate QM/MM overpolarisation. Together, these adjustments reduced the atomic force RMSE by more than 50% compared to DFTB/MM, underlining the importance of accurately describing the local electrostatic environment of the metal centre. Using this improved setup, stable DFTB- ΔML /MM well-tempered metadynamics simulations exceeding 10 ns were achieved with only moderate increase in computational cost by a factor of approximately 2.2 compared to standard DFTB/MM. The simulations further provided mechanistic insight, showing that coordination of all three ATP phosphoryl groups to Mg^{2+} lowered the free-energy barrier by approximately 40–45%, indicating that complete triphosphate coordination facilitates catalysis. In addition, the transition state remained $\text{S}_{\text{N}}2$ -like even without angular restraints on the N–P–O angle, suggesting that such restraints are not required and even may distort the free-energy landscape. Despite the substantial improvement in atomic forces, the calculated free-energy barriers remained higher than experimental estimates, highlighting remaining challenges related to sampling, electrostatics, and structural aspects such as the magnesium coordination environment.

In chapter 11, QM/MM metadynamics simulations were employed to investigate key steps of the catalytic cycle of a mitochondrial creatine kinase, including loop opening and closing, proton transfer, and phosphoryl transfer. The calculated free-energy barrier for loop closing was consistent with the energetic scale expected for hydrogen bond formation, although the low barrier for loop opening suggests that the chosen collective variable (CV) may not fully capture the structural rearrangements associated with loop closure. Proton transfer simulations indicated that a direct proton transfer from creatine to the active-site glutamate is unlikely, whereas a water-mediated pathway substantially lowered the barrier, emphasising the importance of choosing flexible CVs for describing the proton transfer step. The phosphoryl transfer simulations further emphasised the influence of active-site electrostatics, particularly the protonation state of the active-site cysteine, whose thiolate form may introduce local electrostatic strain that destabilises the phosphocreatine product and thereby facilitate product release. Considering the proton- and phosphoryl transfer steps together yields an overall barrier significantly higher than expected from experimental turnover rates, indicating limitations of the present modelling approach, possibly due to insufficient sampling of slow conformational rearrangements or an incomplete description of the electrostatic environment. In addition, estimation of the overall reaction free energy using a thermodynamic cycle suggested stronger binding of the reactants than the products, facilitating product release. Although the calculated difference exceeds experimental estimates, the qualitative trend agrees with experiment. Overall, the simulations provide qualitative mechanistic insight into the catalytic cycle of creatine kinase. Future work should therefore focus on improved sampling, refined CVs, and a more accurate description of the active-site environment to obtain a more quantitatively reliable description of the reaction mechanism. Furthermore, the present work establishes a foundation for future ΔML applications to large, biologically relevant systems.

Overall, the results presented in this thesis demonstrate how multiscale simulations can provide detailed mechanistic insight into kinase-catalysed N-phosphorylation reactions. By combining enhanced sampling, QM/MM simulations, and Δ -machine learning corrections, this work contributes to a deeper understanding of the fundamental chemical processes governing these enzymatic reactions and establishes a foundation for future studies of complex biological phosphorylation processes.

Part V.

Appendix

A. Δ ML Model for N-Phosphoryl Transfer Reactions in Gas Phase

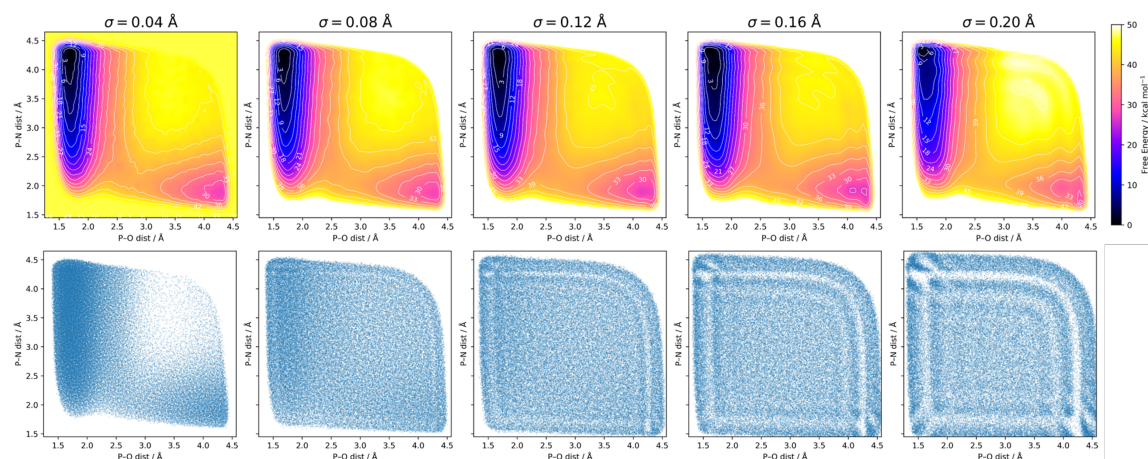


Figure A.1.: Influence of the Gaussian width σ on the FES (top) and the Gaussian deposition distribution (bottom) for all five tested widths. Reducing the Gaussian width (right to left) suppresses the band-like pattern observed at bigger Gaussian widths. Energy values are displayed up to 50 kcal mol^{-1} , with areas exceeding this threshold and areas not explored during the simulation indicated in white. For $\sigma = 0.04 \text{ Å}$, the surrounding area of the FES appears in yellow, as the FES could not be filled up to 50 kcal mol^{-1} within the same simulation time due to the smaller Gaussian width.

Table A.1.: Atomic energies used for the calculation of atomisation energies. DFT energies were calculated at the ω B97M-V/ma-def2-TZVP level, and DFTB energies were obtained using DFTB3 with the 3OB parameter set. All energies are given in E_h .

Element	DFT	DFTB
C	−37.846619775	−1.44397274
N	−54.593097185	−2.21694954
H	−0.494116048	−0.27966183
O	−75.078060060	−3.14196770
P	−341.242118381	−1.69826234
Mg	−200.059614340	−0.34535480

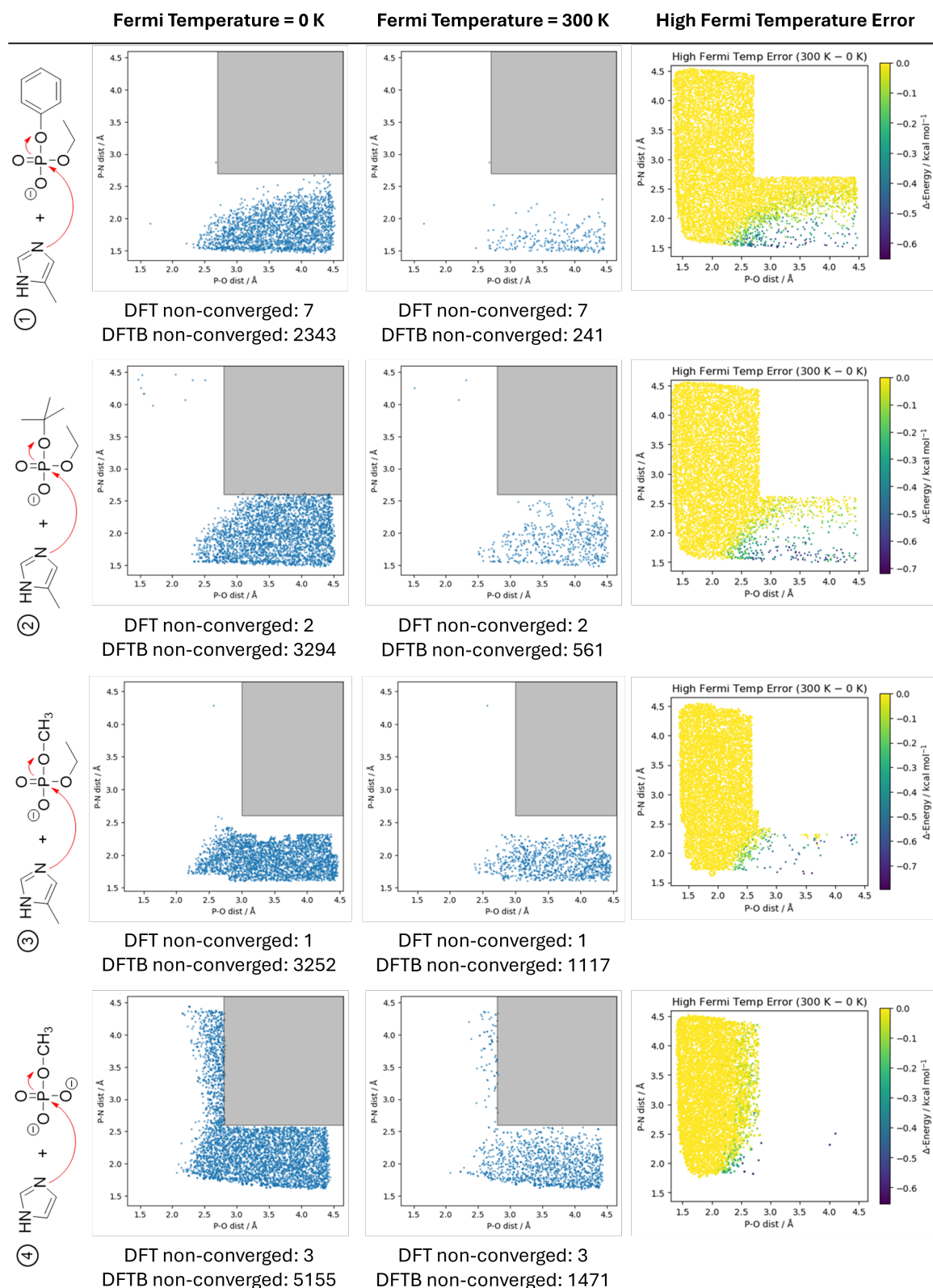


Figure A.2.: Non-converged geometries of systems 1–4 at a Fermi temperature of 0 K (left) and 300 K (middle) and the error introduced by increasing the Fermi temperature (right). Grey areas indicate regions of the FES that were excluded from the single-point calculations as they are not part of the minimum energy path. Regardless of the chosen Fermi temperature, most convergence issues occur in the product state, where the leaving group is poorly stabilised. Increasing the Fermi temperature improves convergence due to fractional orbital occupations but also introduces a small energy error, which remains within the expected DFTB accuracy range. A total of 12,000 single-point calculations were performed for each method and system.

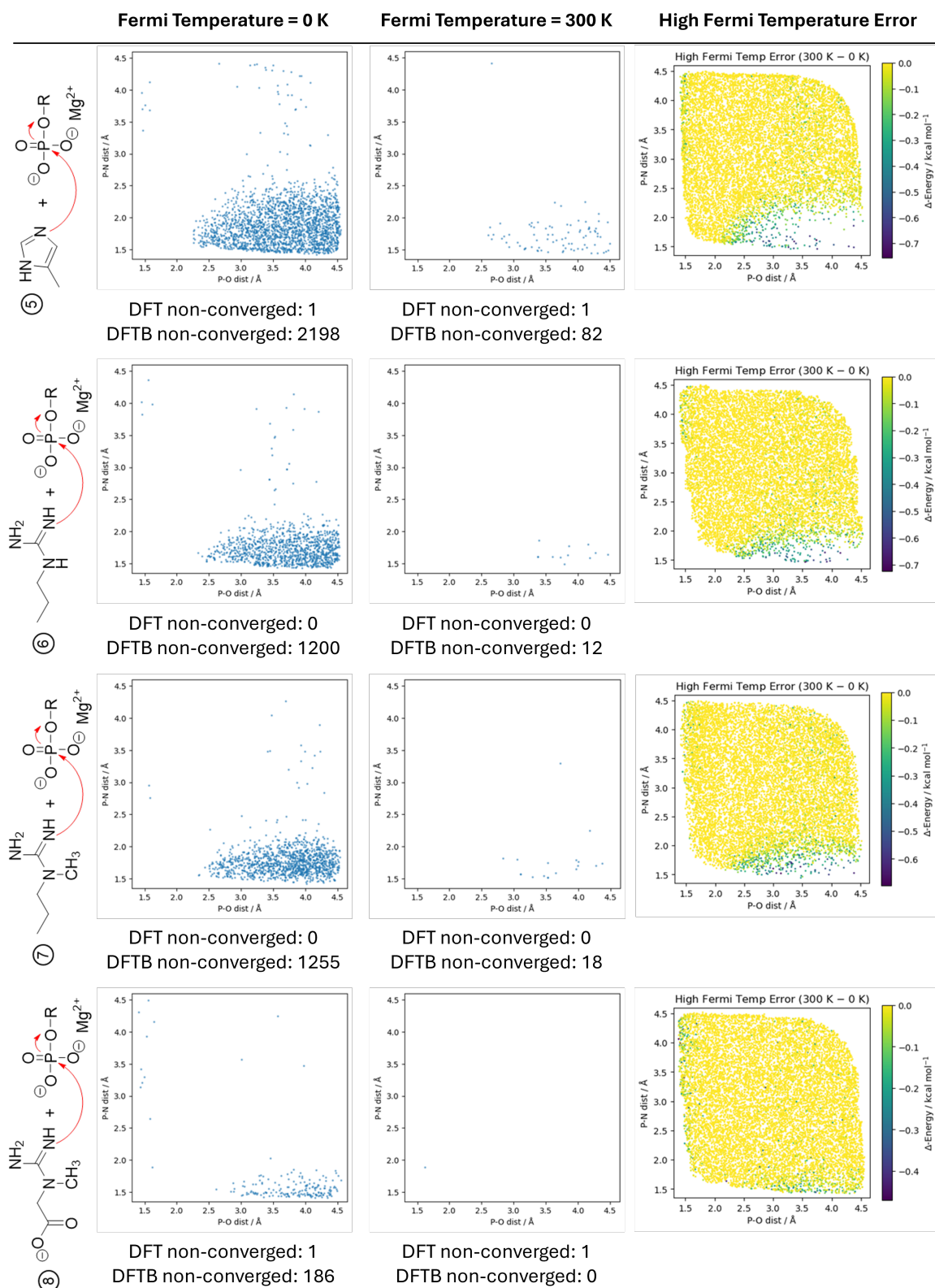


Figure A.3: Non-converged geometries of systems 5–8 at a Fermi temperature of 0 K (left) and 300 K (middle) and the error introduced by increasing the Fermi temperature (right). Regardless of the chosen Fermi temperature, most convergence issues occur in the product state, where the leaving group is poorly stabilised. Increasing the Fermi temperature improves convergence due to fractional orbital occupations but also introduces a small energy error, which remains within the expected DFTB accuracy range. A total of 12,000 single-point calculations were performed for each method and system.

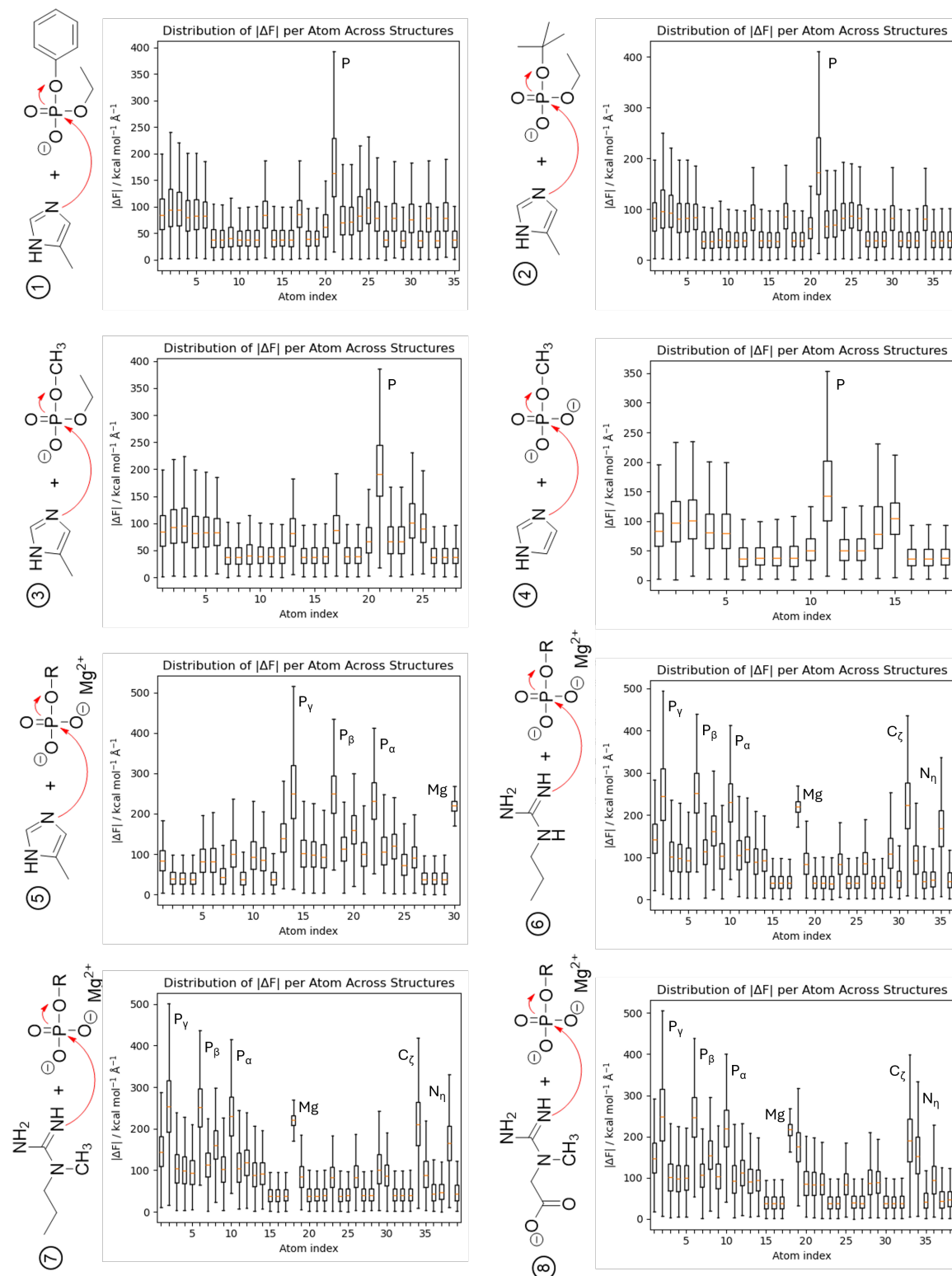


Figure A.4. Per-atom distribution of absolute force differences $|\Delta F|$ for all model systems. The main deviations are observed for the reacting phosphorus atom.

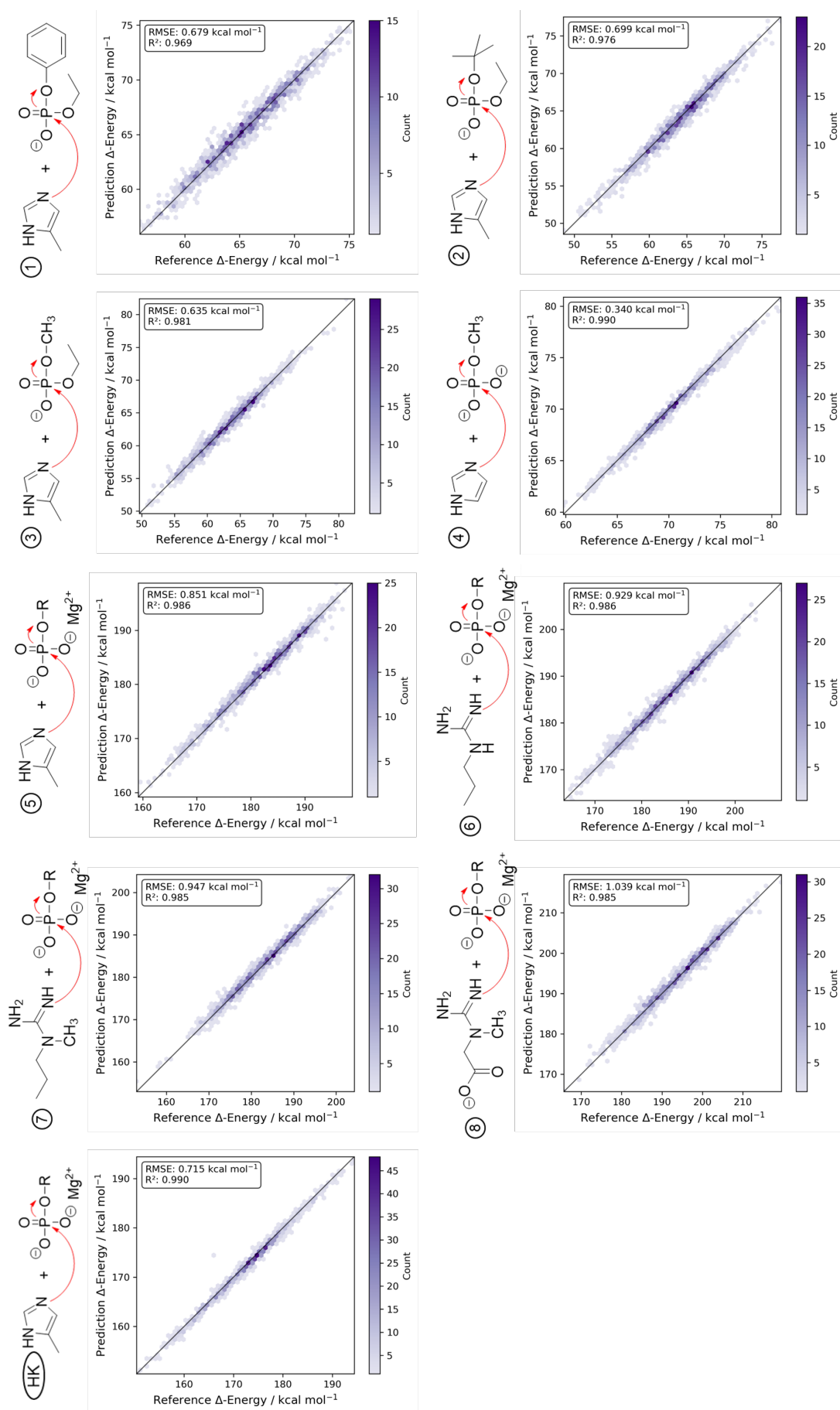


Figure A.5.: Parity plots comparing predicted and reference Δ -energies for single-system cases (trained and tested on the same system). R² values of 0.969 and above indicate good predictive performance.

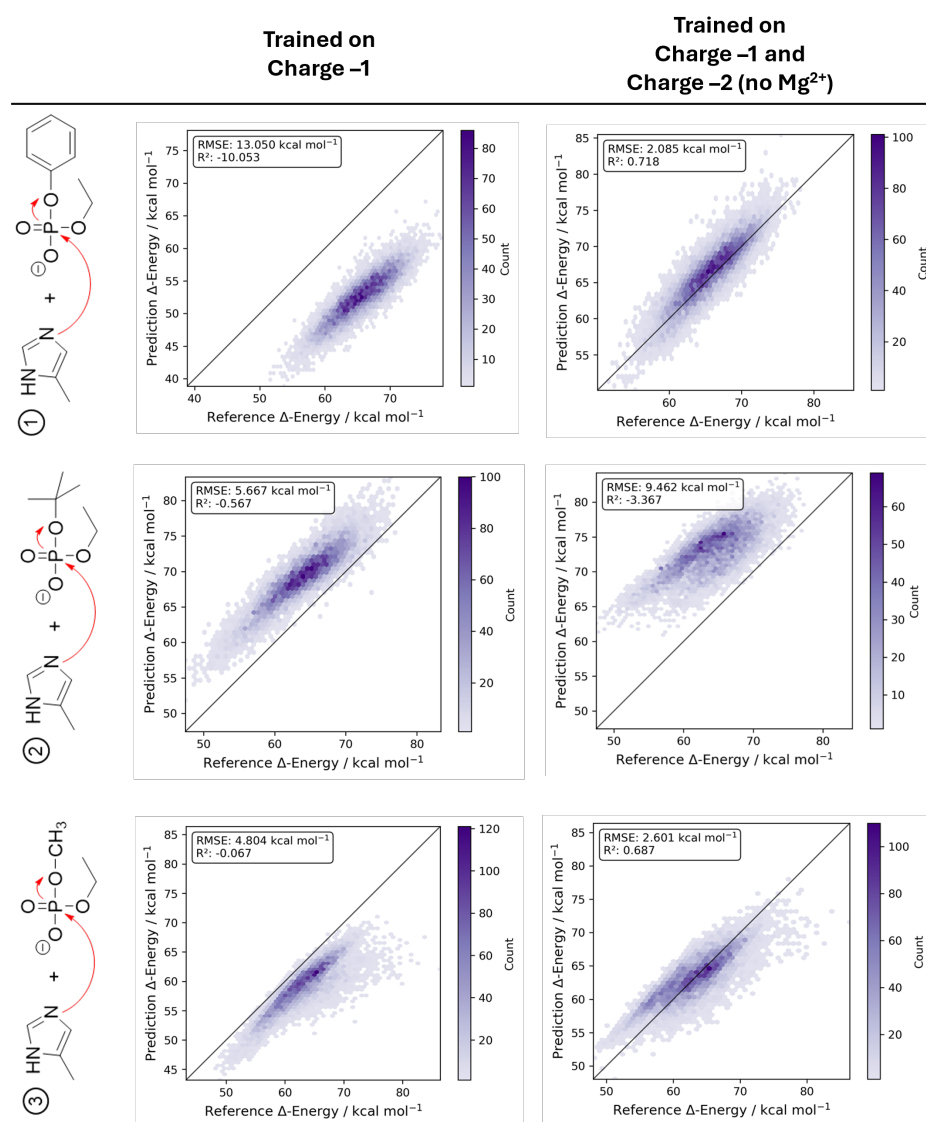


Figure A.6.: Cross-validations for systems with charge -1 (systems 1–3). In the first column, the training was performed on two systems with charge -1 and cross-validated on the remaining system shown on the left. In the second column, system 4 was additionally included in the training.

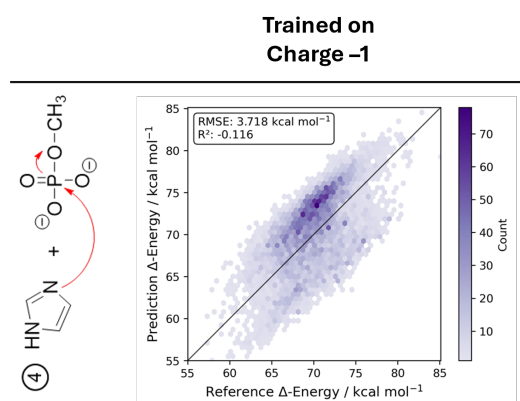


Figure A.7.: Cross-validations for systems with charge -2 and no Mg^{2+} (system 4). The training was performed on system with charge -1 (systems 1–3) and cross-validated on system 4.

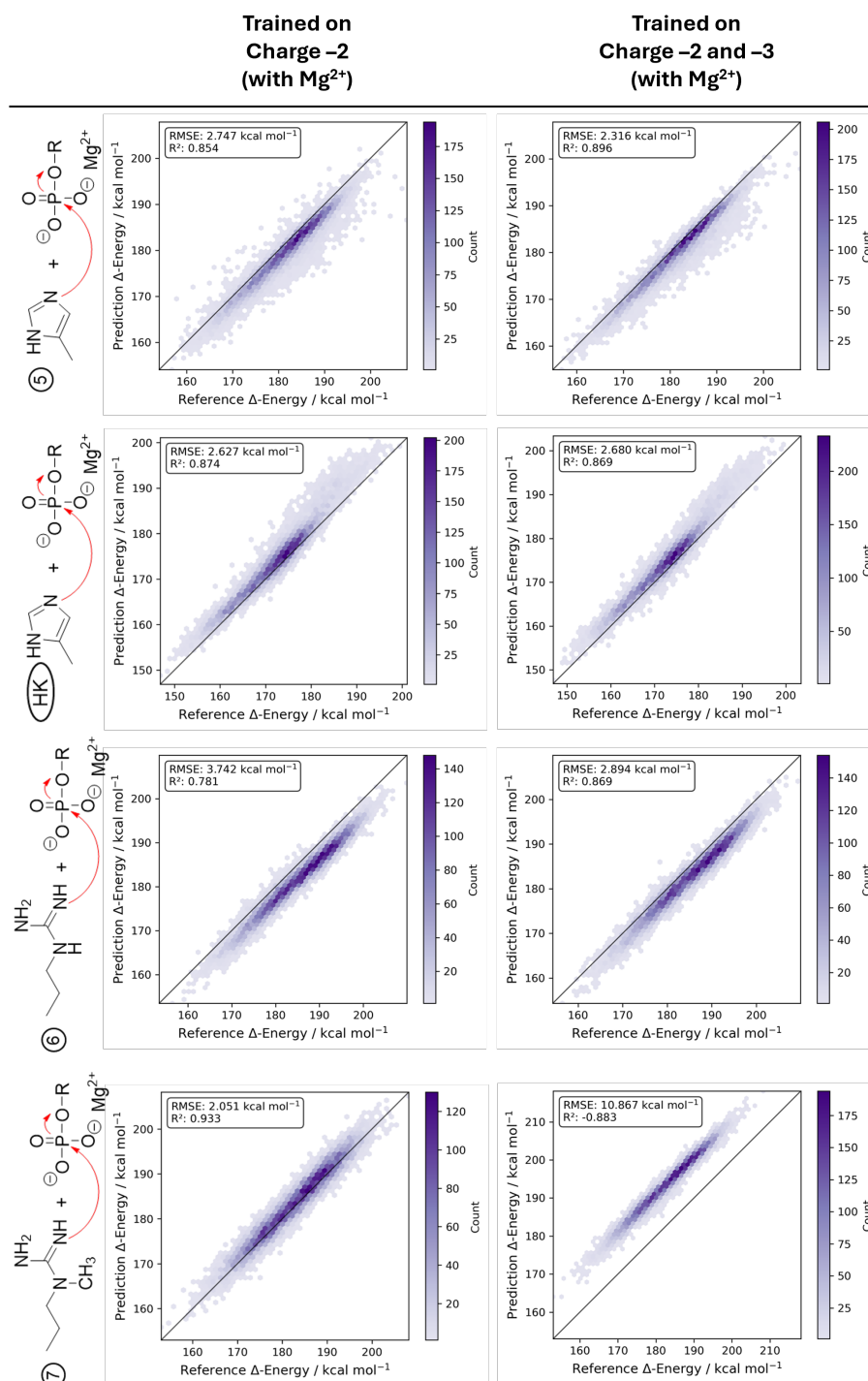


Figure A.8.: Cross-validations for systems with charge -2 and with Mg^{2+} (systems 5–7 and HK). In the first column, the training was performed on three systems with charge -2 and cross-validated on the remaining system shown on the left. In the second column, system 8 was additionally included in the training.

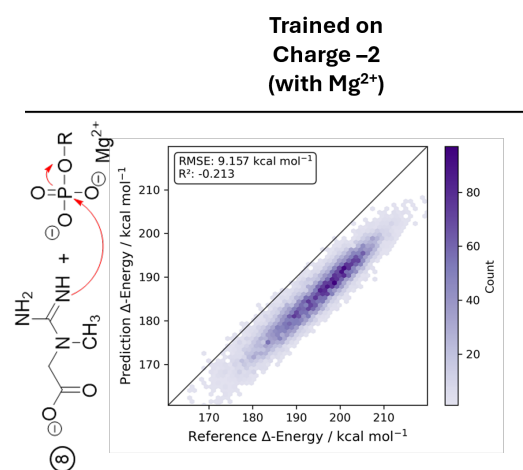


Figure A.9.: Cross-validations for systems with charge -3 (system 8). The training was performed on system with charge -2 with Mg^{2+} (systems 5–7 and HK) and cross-validated on system 8.

B. Transferability of Δ ML Models from Vacuum to Water

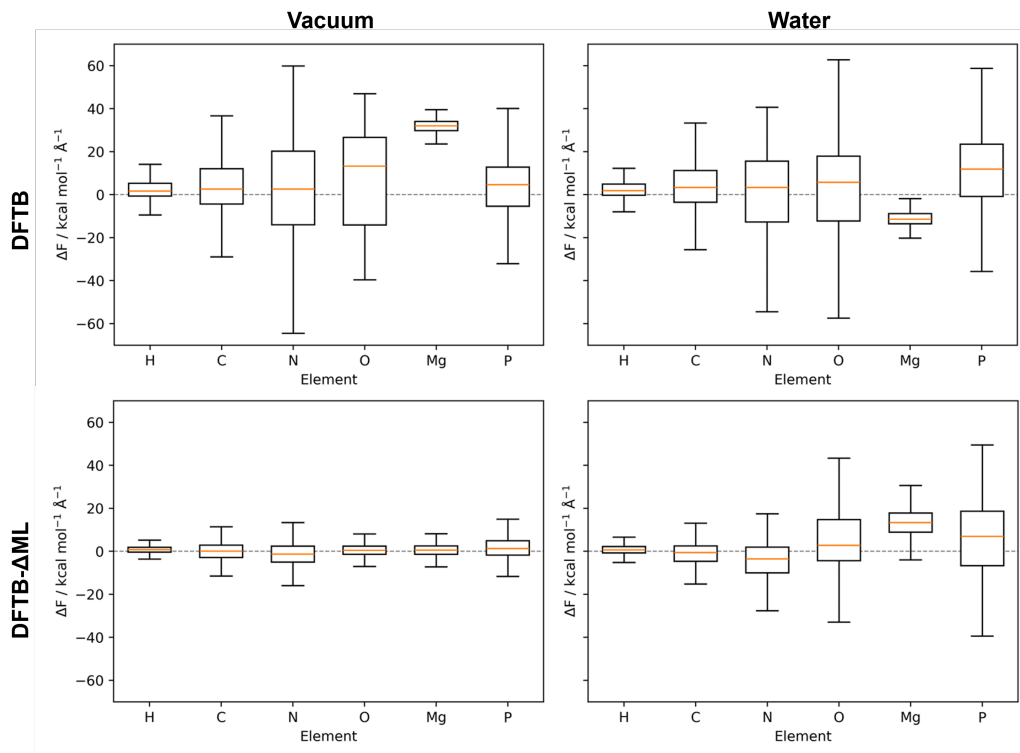


Figure B.1.: Element-wise distribution of force differences ΔF in vacuum and water for system 5. The biggest difference between vacuum and water is obtained for the magnesium cation. In vacuum, the magnesium cation is only coordinated by the triphosphate, while in water, the surrounding oxygen point charges are also part of the coordination sphere, leading to a full hexa-coordination.

Table B.1.: Standard deviation (SD) of atomic forces by chemical element in vacuum and water for system 5. Shown are results for DFTB and DFTB- Δ ML relative to DFT reference forces. Values are given in $\text{kcal mol}^{-1} \text{\AA}^{-1}$.

Element	DFTB vs DFT		DFTB- Δ ML vs DFT		Improvement	
	SD_{vac}	SD_{aq}	SD_{vac}	SD_{aq}	SD_{vac}	SD_{aq}
H	4.367	3.463	2.219	2.961	49.2%	14.5%
C	15.219	13.705	5.255	6.725	65.5%	50.9%
N	23.274	19.458	6.786	9.350	70.8%	51.9%
O	21.076	25.216	3.184	15.252	84.9%	39.5%
Mg	3.286	3.605	3.212	7.285	2.3%	−102.1%
P	14.660	17.045	6.536	16.752	55.4%	1.7%

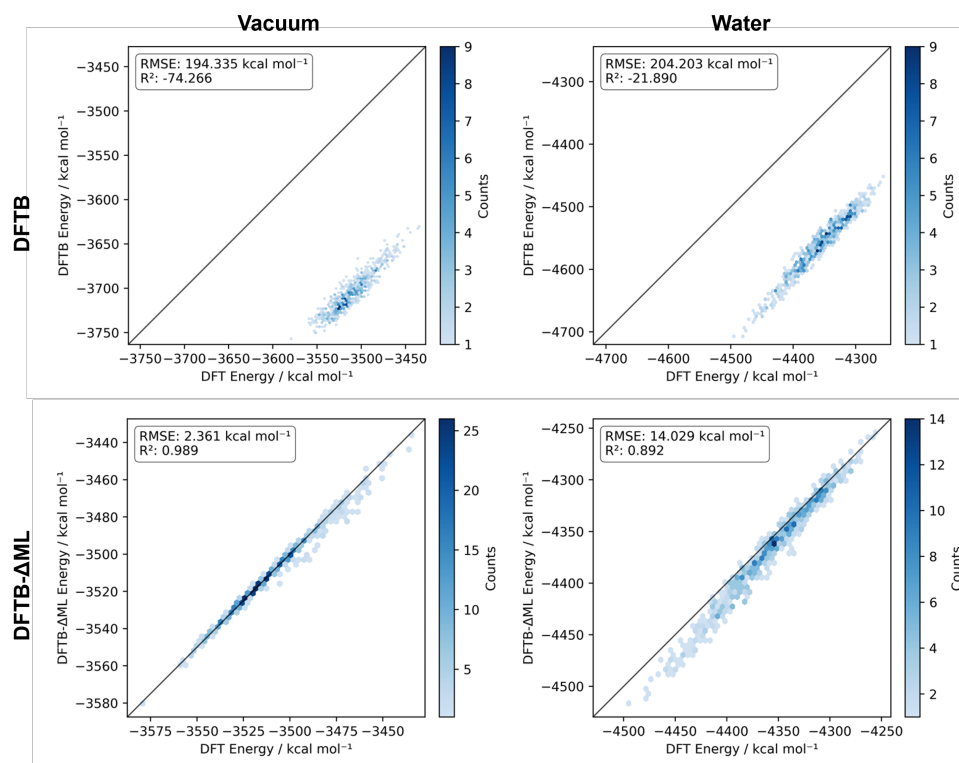


Figure B.2.: Parity plots illustrating the transferability of the vacuum-trained Δ -ML correction to water for the total energies of system 8. Shown are total energies obtained with DFTB (top row) and DFTB- Δ ML (bottom row) compared against DFT reference energies in vacuum (left) and aqueous solution (right).

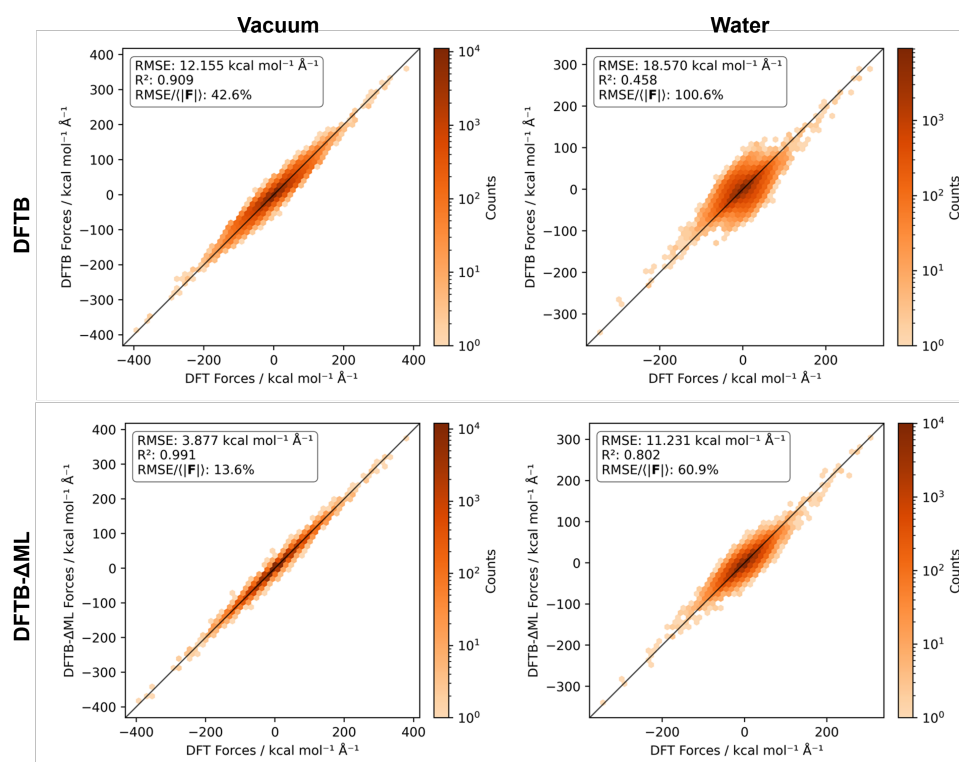


Figure B.3.: Parity plots illustrating the transferability of the vacuum-trained Δ -ML correction to water for the atomic forces of system 8. Shown are atomic forces obtained with DFTB (top row) and DFTB- Δ ML (bottom row) compared against DFT reference forces in vacuum (left) and aqueous solution (right). In vacuum, the Δ -ML correction reduced the error by 68%, in water by 40%. The forces were not included in the training, but are obtained as the gradient of the respective total energies.

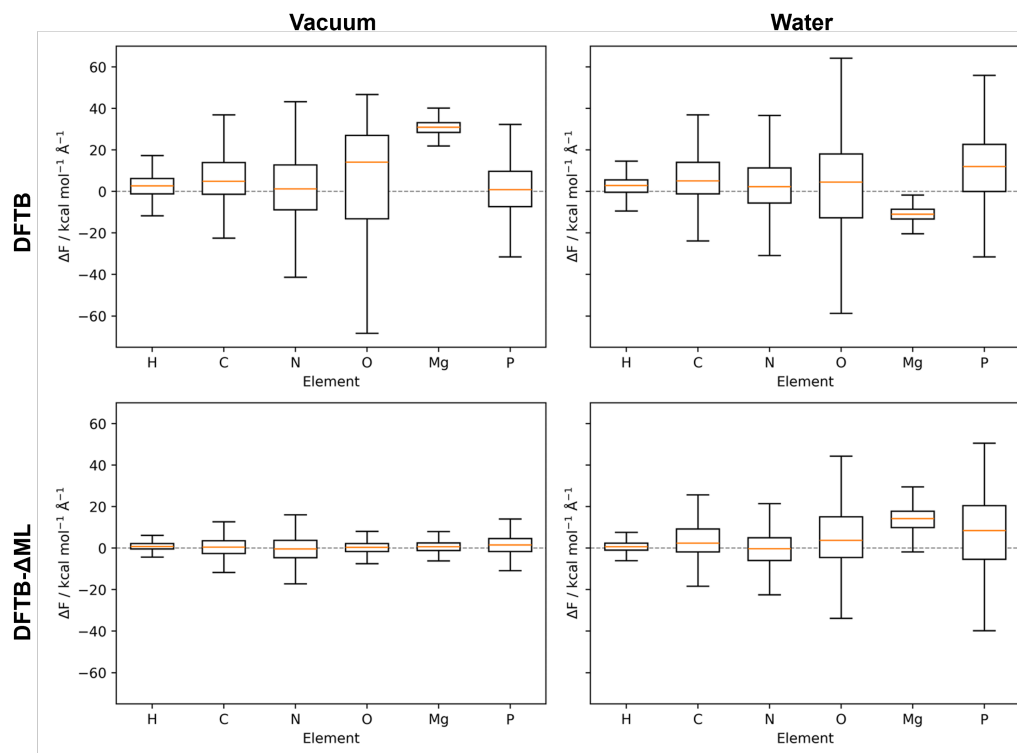


Figure B.4.: Element-wise distribution of force differences ΔF in vacuum and water for system 8. The difference Δ -force distributions between oxygen in vacuum and water in uncorrected DFTB could be attributed to the negatively charged carboxylate group. While the negative charge of the triphosphate oxygens is stabilised by the magnesium cation, the carboxylate has no counter charge in its proximity. The biggest difference between vacuum and water is obtained for the magnesium cation. In vacuum, the magnesium cation is only coordinated by the triphosphate, while in water, the surrounding oxygen point charges are also part of the coordination sphere, leading to a full hexa-coordination.

Table B.2.: Standard deviation (SD) of atomic forces by chemical element in vacuum and water for system 8. Shown are results for DFTB and DFTB- Δ ML relative to DFT reference forces. Values are given in $\text{kcal mol}^{-1} \text{\AA}^{-1}$.

Element	DFTB vs DFT		DFTB- Δ ML vs DFT		Improvement	
	SD _{vac}	SD _{aq}	SD _{vac}	SD _{aq}	SD _{vac}	SD _{aq}
H	4.938	4.620	2.695	3.185	45.4%	31.1%
C	11.519	15.874	5.750	10.805	50.1%	31.9%
N	16.512	16.379	6.959	10.210	57.9%	37.7%
O	21.142	26.417	4.271	15.216	79.8%	42.4%
Mg	3.528	3.661	3.233	6.975	8.4%	−90.5%
P	11.039	15.890	4.884	16.738	55.8%	−5.3%

C. Autophosphorylation in a *trans*-Histidine Kinase

Table C.1.: RMSE values of atomic forces by chemical element of the QM region of the *trans*-HK in vacuum and protein, trained and applied to the small QM subsystem. Shown are results for DFTB and DFTB- Δ ML relative to DFT reference forces. RMSE values are given in kcal mol⁻¹ Å⁻¹.

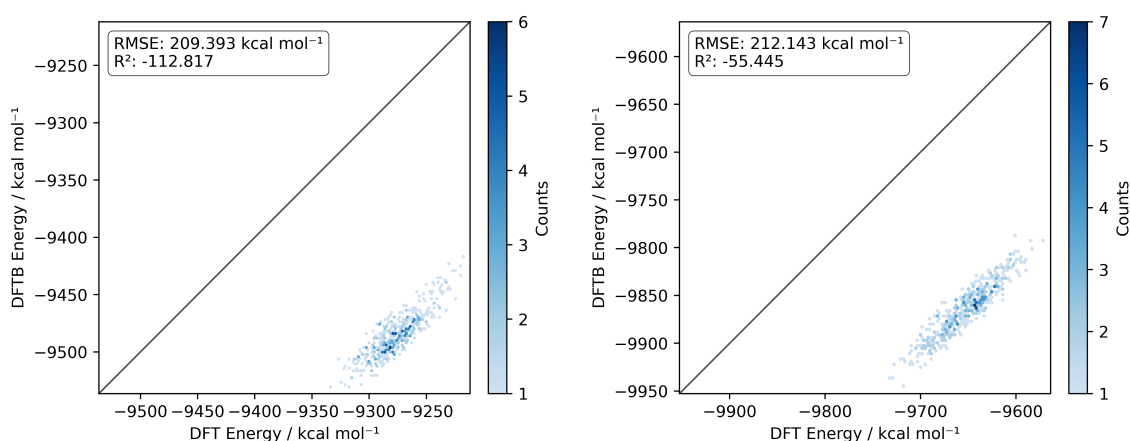
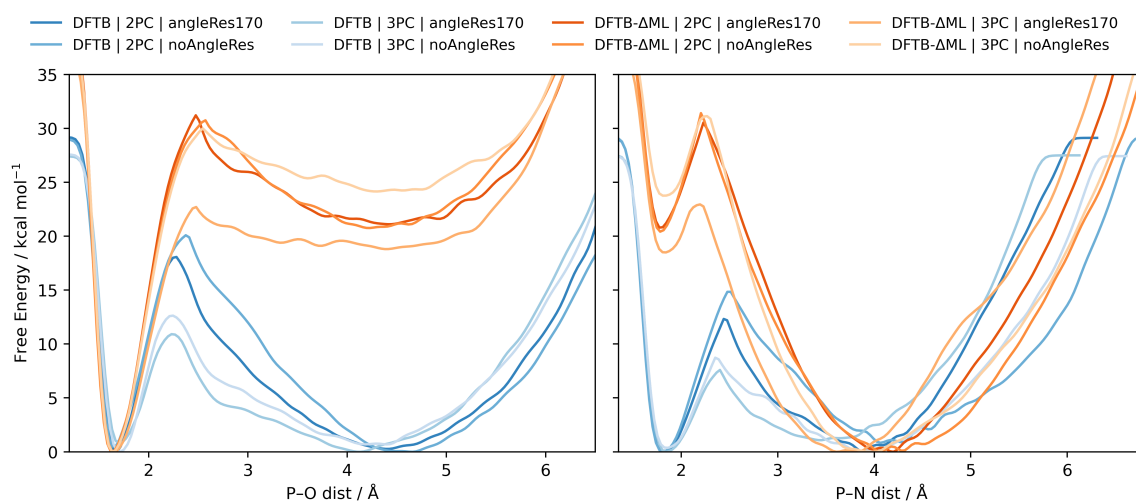
Element	DFTB vs DFT		DFTB- Δ ML vs DFT		Improvement	
	RMSE _{vac}	RMSE _{prot}	RMSE _{vac}	RMSE _{prot}	RMSE _{vac}	RMSE _{prot}
H	4.858	4.806	2.169	2.974	55.4%	38.1%
C	12.297	11.942	4.401	5.011	64.2%	58.0%
N	16.498	16.601	7.014	6.908	57.5%	58.4%
O	13.303	15.066	4.786	7.837	64.0%	48.0%
Mg	8.591	8.426	9.583	9.604	-11.5%	-14.0%
P	11.181	10.874	7.463	9.954	33.3%	8.5%
overall	9.958	10.634	4.918	6.594	50.6%	38.0%

Table C.2.: RMSE values of atomic forces by chemical element of the QM region of the *trans*-HK in vacuum and protein, trained on the small and applied to the big QM subsystem. Shown are results for DFTB and DFTB- Δ ML relative to DFT reference forces. RMSE values are given in kcal mol⁻¹ Å⁻¹.

Element	DFTB vs DFT		DFTB- Δ ML vs DFT		Improvement	
	RMSE _{vac}	RMSE _{prot}	RMSE _{vac}	RMSE _{prot}	RMSE _{vac}	RMSE _{prot}
H	4.858	4.806	16.689	16.706	-243.5%	-247.6%
C	12.297	11.942	89.123	89.157	-624.8%	-646.6%
N	16.498	16.601	54.301	54.236	-229.1%	-226.7%
O	13.303	15.066	62.835	63.111	-372.3%	-318.9%
Mg	8.591	8.426	177.849	177.848	-1970.2%	-2010.8%
P	11.181	10.874	13.149	12.854	-17.6%	-18.2%
overall	9.958	10.634	52.334	52.437	-425.5%	-393.1%

Table C.3.: RMSE values of atomic forces by chemical element of the QM region of the *trans*-HK in vacuum and protein, trained and applied to the big QM subsystem. Shown are results for DFTB and DFTB- Δ ML relative to DFT reference forces. RMSE values are given in kcal mol⁻¹ Å⁻¹.

Element	DFTB vs DFT		DFTB- Δ ML vs DFT		Improvement	
	RMSE _{vac}	RMSE _{prot}	RMSE _{vac}	RMSE _{prot}	RMSE _{vac}	RMSE _{prot}
H	4.858	4.806	2.866	3.173	41.0%	34.0%
C	12.297	11.942	5.501	6.269	55.3%	47.5%
N	16.498	16.601	7.030	7.377	57.4%	55.6%
O	13.303	15.066	4.147	6.159	68.8%	59.1%
Mg	8.591	8.426	5.329	5.494	38.0%	34.8%
P	11.181	10.874	8.740	10.680	21.8%	1.8%
overall	9.958	10.634	4.270	5.316	57.1%	50.0%

**Figure C.1.:** Parity plots illustrating the transferability between DFTB calculations in vacuum and DFTB/MM calculations in the protein. The shape difference of the distributions indicates limited transferability.**Figure C.2.:** 1D projections of the eight HK metadynamics simulations with different methods, starting structures and angle walls. 2PC and 3PC specify whether the magnesium cation in the starting structure is coordinated by two or three phosphoryl groups, respectively. DFTB- Δ ML | 2PC | noAngleRes was excluded from the discussion, as it was not converged after 10 ns of simulation time.

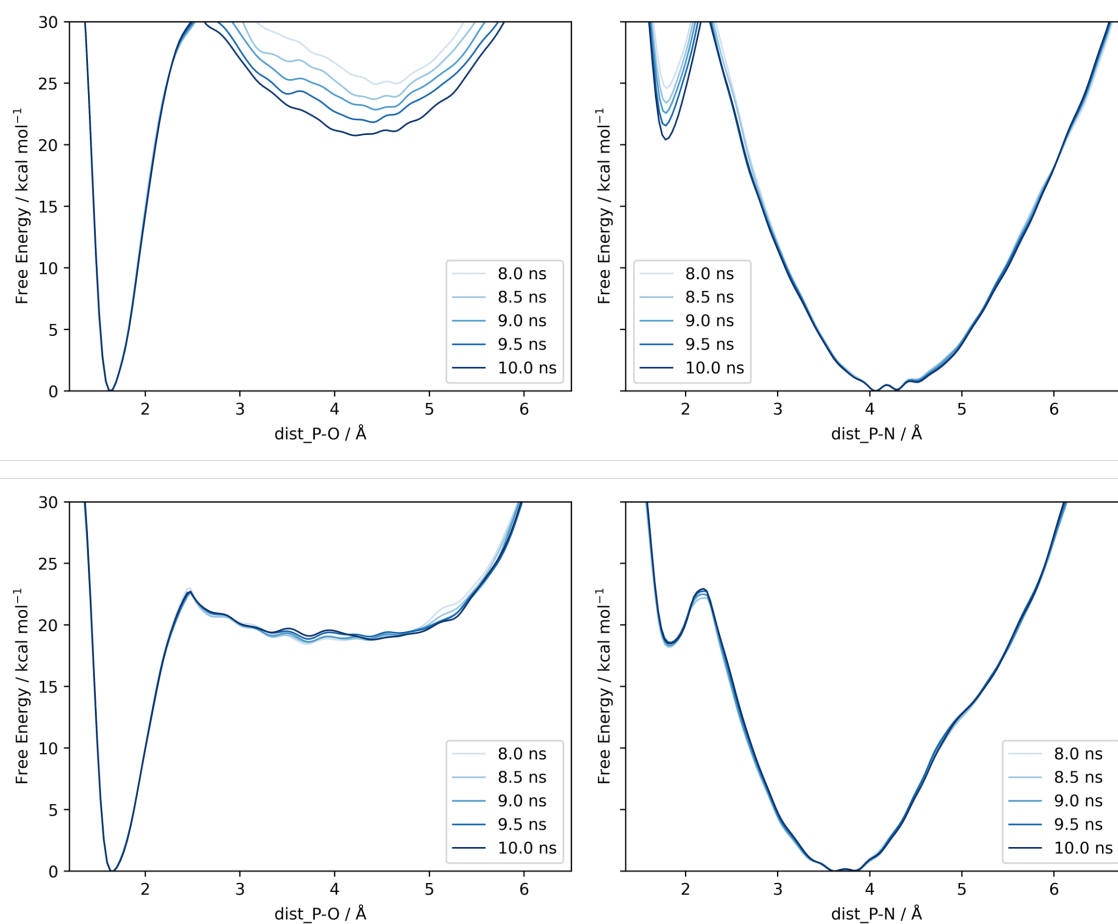


Figure C.3.: 1D projections of the FES onto the two CVs for different time steps. Top: DFTB-ΔML | 2PC | noAngleRes was excluded from the discussion, as it was not converged after 10 ns of simulation time. Bottom: Near-convergence shown exemplarily for DFTB-ΔML | 3PC | angleRes.

D. Understanding the Catalytic Mechanism of Creatine Kinase

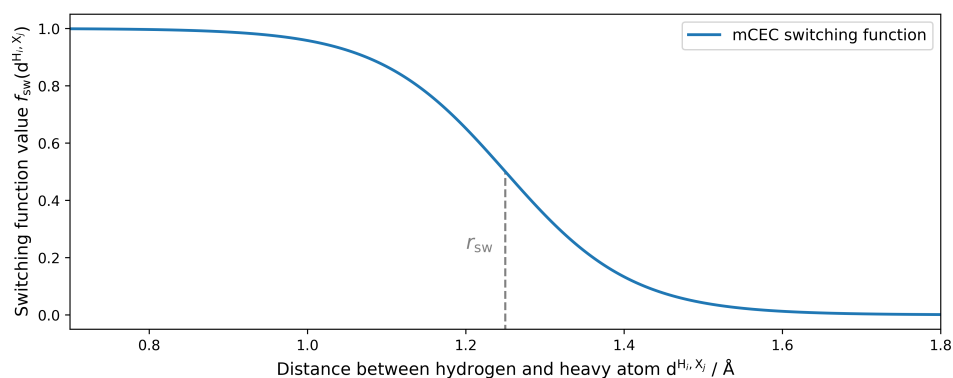


Figure D.1.: Switching function used in the mCEC definition. The switching function $f_{sw}(d_{H,X})$ enables a continuous transition between ‘bound’ and ‘unbound’ states of a hydrogen atom H and a heavy atom X based on their distance $d_{H,X}$. It therefore acts as a weight for bound-like configurations. The dashed line indicates the switching distance r_{sw} .

Use of Generative AI Tools

In accordance with the guidelines of the Karlsruhe Institute of Technology (KIT), this section documents the use of generative artificial intelligence (AI) tools during the preparation of this thesis.

Several AI-based systems were employed to support the research and writing process. These included OpenAI GPT models (e.g. GPT-5 Plus), DeepL Write/Translator, Perplexity, and GitHub Copilot. The tools were employed for tasks such as software development support (e.g. debugging assistance and code suggestions), language refinement and stylistic improvement of text passages, conceptual exploration of ideas, and assistance in identifying relevant literature.

The use of these tools was strictly assistive. All generated outputs, including code and text, were critically reviewed, verified, and revised by the author before being incorporated into the thesis. Responsibility for the content, accuracy, interpretation of results, and compliance with scientific standards remains solely with the author.

No AI system was used to generate research results, fabricate sources, or replace independent scientific reasoning.

List of Figures

1.1.	Structures of histidine kinase and creatine kinase with enlarged active site and reaction equation.	3
2.1.	Overview of different computational methods.	9
3.1.	Comparison of Morse and harmonic potentials.	12
3.2.	Schematic representation of force field contributions.	13
4.1.	Schematic representation of the QM/MM concept.	19
4.2.	Link atoms in the QM/MM concept.	21
5.1.	Schematic representation of Jacob's ladder of DFT.	29
5.2.	Number of DFT publications per year (1990–2025).	31
6.1.	Schematic representation of a free energy surface with two minima A and B.	35
6.2.	Schematic representation of different enhanced sampling methods.	37
6.3.	Schematic representation of the metadynamics method.	39
6.4.	Schematic representation of standard metadynamics versus well-tempered metadynamics.	40
7.1.	Structure of a second-generation HDNNP.	42
7.2.	Visualisation of radial symmetry functions.	43
7.3.	Visualisation of angular symmetry functions.	44
8.1.	Schematic representation of roles of phosphorus in biochemistry.	47
8.2.	Chemical structures of inorganic phosphate and adenosine triphosphate (ATP).	48
8.3.	Chemical structures of phosphomonoesters and phosphoramidates with examples.	49
8.4.	Schematic representation of the collective variables (CVs) used for the well-tempered metadynamics simulations.	52
8.5.	Influence of the Gaussian width σ on the FES and the Gaussian deposition distribution.	53
8.6.	Projection of the two-dimensional FESs onto the two CVs, simulated at different Gaussian widths σ	54
8.7.	Development of the 1D projection of the FES with $\sigma = 0.20 \text{ \AA}$ over time.	55
8.8.	Reaction equations of the simulated phosphorylation reactions.	56
8.9.	Moieties for reaction equations abbreviated in Figure 8.8.	56
8.10.	Free energy surface (FES), ΔG , and barrier height for system 5.	57
8.11.	Free energy surfaces (FES), ΔG , and barrier heights for systems 1–4.	58
8.12.	Free energy surfaces (FES), ΔG , and barrier heights for systems 5–8.	59
8.13.	Schematic representation of S_N2 - and S_N1 -like reaction mechanisms, illustrated for the phosphorylation of a guanidino group.	60
8.14.	Effect of Fermi temperature on convergence and errors.	62
8.15.	Box plot representation of the Δ -energy distributions for the different model systems and their standard deviations.	63
8.16.	Δ -Energies across the FESs of the model systems.	64
8.17.	Per-atom distribution of absolute force differences $ \Delta F $ for system 8.	65

8.18.	Per-atom distribution of absolute force differences $ \Delta F $ for system 5 and protein.	66
8.19.	R^2 and RMSE values for single-system cases.	67
8.20.	Histogram of Δ -energies comparing DFTB and DFTB- Δ ML for the single-system case applied to vacuum single-point calculations of system 5.	67
8.21.	Cross-validations for systems 3.	68
8.22.	Cross-validations for systems 7.	69
8.23.	Cross-validation for systems 8.	70
8.24.	Influence of single-system training and cross-validation on DFTB- Δ ML energies and forces of system 5.	70
8.25.	Influence of single-system training and cross-validation on DFTB- Δ ML energies and forces of system 8.	71
9.1.	Test systems considered for water simulations.	77
9.2.	Parity plots illustrating the transferability of the vacuum-trained Δ ML correction to water for the total energies of system 5.	78
9.3.	Parity plots illustrating the transferability of the vacuum-trained Δ ML correction to water for the atomic forces of system 5.	79
9.4.	Magnesium coordination in vacuum and water.	80
9.5.	Comparison of FESs generated with DFTB/MM and DFTB- Δ ML/MM metadynamics simulations in water for system 5.	81
9.6.	Comparison of FESs generated with DFTB/MM and DFTB- Δ ML/MM metadynamics simulations in water for system 8.	82
10.1.	Schematic representation of a TCS.	86
10.2.	Structure of the investigated <i>trans</i> -HK with enlarged QM region and reaction scheme.	87
10.3.	Schematic representation of the CVs used for the well-tempered metadynamics simulations of HK.	89
10.4.	Subsets of the HK QM region corrected during DFTB- Δ ML calculations.	89
10.5.	FES of a HK phosphoryl transfer simulation.	94
10.6.	Relation between phosphoryl transfer and N-P-O angle with and without angle restraint.	97
11.1.	Reaction equation of the phosphoryl transfer catalysed by CKs.	103
11.2.	Schematic representation of the phosphocreatine shuttle.	104
11.3.	CK structure with enlarged active site and chemical equation.	105
11.4.	Schematic representation of the collective variables (CVs) used for the well-tempered metadynamics simulations of CK.	107
11.5.	Schematic representation of the CVs used for the water-mediated proton transfer.	108
11.6.	FES of the loop-closing transition.	109
11.7.	FES of phosphoryl transfer to a protonated guanidinium.	110
11.8.	Direct versus water-mediated proton transfer.	111
11.9.	FES of the CK proton transfer along the mCEC coordinate ζ	112
11.10.	Influence of the thiol form of Cys278 on the CK phosphoryl transfer reaction.	113
11.11.	Influence of the thiolate form of Cys278 on the CK phosphoryl transfer reaction.	114
11.12.	Free energy profile combining conformational transitions and chemical steps.	115
11.13.	Thermodynamic cycle relating the reaction free energies in the protein and in aqueous solution.	116

A.1.	Influence of the Gaussian width σ on the FES and the Gaussian deposition distribution for all five tested widths.	125
A.2.	Effect of Fermi temperature on convergence and errors for systems 1–4.	126
A.3.	Effect of Fermi temperature on convergence and errors for systems 5–8.	127
A.4.	Per-atom distribution of absolute force differences $ \Delta F $ for all model systems.	128
A.5.	Parity plots comparing predicted and reference Δ -energies for single-system cases.	129
A.6.	Cross-validations for systems with charge -1	130
A.7.	Cross-validations for systems with charge -2 and no Mg^{2+}	130
A.8.	Cross-validations for systems with charge -2 and with Mg^{2+}	131
A.9.	Cross-validations for systems with charge -3	132
B.1.	Element-wise distribution of force differences ΔF in vacuum and water for system 5.	133
B.2.	Parity plots illustrating the transferability of the vacuum-trained Δ -ML correction to water for the total energies of system 8.	134
B.3.	Parity plots illustrating the transferability of the vacuum-trained Δ -ML correction to water for the atomic forces of system 8.	134
B.4.	Element-wise distribution of force differences ΔF in vacuum and water for system 8.	135
C.1.	Parity plots illustrating the transferability between DFTB calculations in vacuum and DFTB/MM calculations in the protein.	138
C.2.	1D projections of the eight HK metadynamics simulations.	138
C.3.	1D projections of the FES onto the two CVs for different time steps.	139
D.1.	Switching function used in the mCEC definition.	141

List of Tables

8.1.	Coordinates and energies of reactant, product, and transition state.	61
8.2.	Mean CPU times per single-point calculation and relative computational cost for the different model systems.	72
9.1.	RMSE values of atomic forces by chemical element in vacuum and water for system 5.	80
9.2.	RMSE values of atomic forces by chemical element in vacuum and water for system 8.	82
9.3.	Mean CPU times per single-point calculation in vacuum and water and respective relative computational cost.	83
9.4.	Mean computational cost of metadynamics simulations.	83
10.1.	Mean CPU times per single-point calculation in the protein, respective relative computational cost and RMSE improvement.	92
10.2.	Coordinates and energies of reactant, product, and transition state for different methods, starting structures and angle restraints.	95
10.3.	Average number of oxygen atoms coordinating the magnesium cation.	96
10.4.	Mean computational cost of metadynamics simulations.	99
A.1.	Atomic energies used for the calculation of atomisation energies.	125
B.1.	Standard deviation of atomic forces by chemical element in vacuum and water for system 5.	133
B.2.	Standard deviation of atomic forces by chemical element in vacuum and water for system 8.	135
C.1.	RMSE values of atomic forces by chemical element of the QM region of the <i>trans</i> -HK in vacuum and protein, trained and applied to the small QM subsystem.	137
C.2.	RMSE values of atomic forces by chemical element of the QM region of the <i>trans</i> -HK in vacuum and protein, trained on the small and applied to the big QM subsystem.	137
C.3.	RMSE values of atomic forces by chemical element of the QM region of the <i>trans</i> -HK in vacuum and protein, trained and applied to the big QM subsystem. . .	138

Bibliography

- [1] Jack W. F. Nicholls et al. 'On the potential roles of phosphorus in the early evolution of energy metabolism'. English. In: *Frontiers in Microbiology* 14 (Aug. 2023). ISSN: 1664-302X. DOI: 10.3389/fmicb.2023.1239189.
- [2] Fatima Ardito et al. 'The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review)'. In: *International Journal of Molecular Medicine* 40.2 (Aug. 2017), pp. 271–280. ISSN: 1107-3756. DOI: 10.3892/ijmm.2017.3036.
- [3] Joanna Cieřła, Tomasz Frączyk and Wojciech Rode. 'Phosphorylation of basic amino acid residues in proteins: important but easily missed'. eng. In: *Acta Biochimica Polonica* 58.2 (2011), pp. 137–148. ISSN: 1734-154X.
- [4] Paul V. Attwood. 'Histidine kinases from bacteria to humans'. In: *Biochemical Society Transactions* 41.4 (July 2013), pp. 1023–1028. ISSN: 0300-5127. DOI: 10.1042/BST20130019.
- [5] W. Ross Ellington. 'Evolution and Physiological Roles of Phosphagen Systems'. en. In: *Annual Review of Physiology* 63. Volume 63, 2001 (Mar. 2001), pp. 289–325. ISSN: 0066-4278, 1545-1585. DOI: 10.1146/annurev.physiol.63.1.289.
- [6] Jonathan K. Lassila, Jesse G. Zalatan and Daniel Herschlag. 'Biological Phosphoryl-Transfer Reactions: Understanding Mechanism and Catalysis'. en. In: *Annual Review of Biochemistry* 80. Volume 80, 2011 (July 2011), pp. 669–702. ISSN: 0066-4154, 1545-4509. DOI: 10.1146/annurev-biochem-060409-092741.
- [7] Orr Ashenberg, Amy E. Keating and Michael T. Laub. 'Helix Bundle Loops Determine Whether Histidine Kinases Autophosphorylate in *cis* or in *trans*'. In: *Journal of Molecular Biology* 425.7 (Apr. 2013), pp. 1198–1209. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2013.01.011.
- [8] Patricia Casino, Laura Miguel-Romero and Alberto Marina. 'Visualizing autophosphorylation in histidine kinases'. en. In: *Nature Communications* 5.1 (Feb. 2014), p. 3258. ISSN: 2041-1723. DOI: 10.1038/ncomms4258.
- [9] Karin Fritz-Wolf et al. 'Structure of mitochondrial creatine kinase'. en. In: *Nature* 381.6580 (May 1996), pp. 341–345. ISSN: 1476-4687. DOI: 10.1038/381341a0.
- [10] Michael Forstner et al. 'Changes of creatine kinase structure upon ligand binding as seen by small-angle scattering'. In: *Journal of Molecular Structure. Horizons in Small Angle Scattering From Mesoscopic Systems* 383.1 (Sept. 1996), pp. 217–222. ISSN: 0022-2860. DOI: 10.1016/S0022-2860(96)09289-7.
- [11] Wolfgang Kabsch and Karin Fritz-Wolf. 'Mitochondrial creatine kinase - a square protein'. In: *Current Opinion in Structural Biology* 7.6 (Dec. 1997), pp. 811–818. ISSN: 0959-440X. DOI: 10.1016/S0959-440X(97)80151-0.
- [12] Sushmita D. Lahiri et al. 'The 2.1 Å Structure of Torpedo californica Creatine Kinase Complexed with the ADP-Mg²⁺-NO₃-Creatine Transition-State Analogue Complex,' in: *Biochemistry* 41.47 (Nov. 2002), pp. 13861–13867. ISSN: 0006-2960. DOI: 10.1021/bi026655p.
- [13] Qianjie Li et al. 'Insights into the Phosphoryl Transfer Mechanism of Human Ubiquitous Mitochondrial Creatine Kinase'. en. In: *Scientific Reports* 6.1 (Dec. 2016), p. 38088. ISSN: 2045-2322. DOI: 10.1038/srep38088.
- [14] Merve Demir et al. 'Structural basis for substrate binding, catalysis, and inhibition of cancer target mitochondrial creatine kinase by a covalent inhibitor'. en. In: *Structure* 33.4 (Apr. 2025), 786–797.e3. ISSN: 09692126. DOI: 10.1016/j.str.2025.01.008.
- [15] Mayukh Kansari et al. 'Mechanism of activation and autophosphorylation of a histidine kinase'. en. In: *Communications Chemistry* 7.1 (Sept. 2024), p. 196. ISSN: 2399-3669. DOI: 10.1038/s42004-024-01272-6.
- [16] Mayukh Kansari, Lena Eichinger and Tomáš Kubař. 'Extended-sampling QM/MM simulation of biochemical reactions involving P–N bonds'. In: *Physical Chemistry Chemical Physics* 25.14 (2023), pp. 9824–9836. DOI: 10.1039/D2CP05890A.

- [17] Lena Eichinger. 'Exploring the Mechanism of Autophosphorylation in a trans-Histidine Kinase using QM/MM Simulations'. Master's Thesis. Karlsruher Institut für Technologie (KIT), Aug. 2022.
- [18] Rami Gherib, Hisham M. Dokainish and James W. Gault. 'Multi-Scale Computational Enzymology: Enhancing Our Understanding of Enzymatic Catalysis'. en. In: *International Journal of Molecular Sciences* 15.1 (Jan. 2014), pp. 401–422. ISSN: 1422-0067. DOI: 10.3390/ijms15010401.
- [19] Maria J. Ramos and Pedro A. Fernandes. 'Computational Enzymatic Catalysis'. In: *Accounts of Chemical Research* 41.6 (June 2008), pp. 689–698. ISSN: 0001-4842. DOI: 10.1021/ar7001045.
- [20] John A. Keith et al. 'Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems'. In: *Chemical Reviews* 121.16 (Aug. 2021), pp. 9816–9872. ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.1c00107.
- [21] Arup Chakraborty et al. 'Review of Computational Studies of NCM Cathode Materials for Li-ion Batteries'. en. In: *Israel Journal of Chemistry* 60.8-9 (2020). Pp. 850–862. ISSN: 1869-5868. DOI: 10.1002/ijch.201900116.
- [22] Yawen Sun et al. 'Boosting the Optimization of Lithium Metal Batteries by Molecular Dynamics Simulations: A Perspective'. en. In: *Advanced Energy Materials* 10.41 (2020). P. 2002373. ISSN: 1614-6840. DOI: 10.1002/aenm.202002373.
- [23] Ying Ma. 'Computer Simulation of Cathode Materials for Lithium Ion and Lithium Batteries: A Review'. en. In: *ENERGY & ENVIRONMENTAL MATERIALS* 1.3 (2018). Pp. 148–173. ISSN: 2575-0356. DOI: 10.1002/eem2.12017.
- [24] Scott A. Hollingsworth and Ron O. Dror. 'Molecular Dynamics Simulation for All'. In: *Neuron* 99.6 (Sept. 2018), pp. 1129–1143. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2018.08.011.
- [25] Hans Martin Senn and Walter Thiel. 'QM/MM Methods for Biomolecular Systems'. en. In: *Angewandte Chemie International Edition* 48.7 (2009). Pp. 1198–1229. ISSN: 1521-3773. DOI: 10.1002/anie.200802019.
- [26] Jacob D. Durrant and J. Andrew McCammon. 'Molecular dynamics simulations and drug discovery'. In: *BMC Biology* 9.1 (Oct. 2011), p. 71. ISSN: 1741-7007. DOI: 10.1186/1741-7007-9-71.
- [27] Jana Khandogin and Charles L. Brooks. 'Linking folding with aggregation in Alzheimer's β -amyloid peptides'. In: *Proceedings of the National Academy of Sciences* 104.43 (Oct. 2007), pp. 16880–16885. DOI: 10.1073/pnas.0703832104.
- [28] Chun Wu and Joan-Emma Shea. 'Structural Similarities and Differences between Amyloidogenic and Non-Amyloidogenic Islet Amyloid Polypeptide (IAPP) Sequences and Implications for the Dual Physiological and Pathological Activities of These Peptides'. en. In: *PLOS Computational Biology* 9.8 (Aug. 2013), e1003211. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003211.
- [29] Aashish Manglik et al. 'Structure-based discovery of opioid analgesics with reduced side effects'. en. In: *Nature* 537.7619 (Sept. 2016), pp. 185–190. ISSN: 1476-4687. DOI: 10.1038/nature19112.
- [30] John D. McCorvy et al. 'Structure-inspired design of β -arrestin-biased ligands for aminergic GPCRs'. en. In: *Nature Chemical Biology* 14.2 (Feb. 2018), pp. 126–134. ISSN: 1552-4469. DOI: 10.1038/nchembio.2527.
- [31] V. Spahn et al. 'A nontoxic pain killer designed by modeling of pathological receptor conformations'. eng. In: *Science (New York, N.Y.)* 355.6328 (Mar. 2017), pp. 966–969. ISSN: 1095-9203. DOI: 10.1126/science.aai8636.
- [32] Adrian J. Mulholland. 'Modelling enzyme reaction mechanisms, specificity and catalysis'. In: *Drug Discovery Today* 10.20 (Oct. 2005), pp. 1393–1402. ISSN: 1359-6446. DOI: 10.1016/S1359-6446(05)03611-1.
- [33] Kwangho Nam et al. 'Perspectives on Computational Enzyme Modeling: From Mechanisms to Design and Drug Development'. In: *ACS Omega* 9.7 (Feb. 2024), pp. 7393–7412. DOI: 10.1021/acsomega.3c09084.
- [34] Juan R Perilla et al. 'Molecular dynamics simulations of large macromolecular complexes'. In: *Current Opinion in Structural Biology*. Theory and simulation/Macromolecular machines and assemblies 31 (Apr. 2015), pp. 64–74. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2015.03.007.
- [35] David E. Shaw et al. 'Anton, a special-purpose machine for molecular dynamics simulation'. In: *Commun. ACM* 51.7 (July 2008), pp. 91–97. ISSN: 0001-0782. DOI: 10.1145/1364782.1364802.
- [36] Frank Jensen. *Introduction to computational chemistry*. Third edition. Chichester, UK ; Hoboken, NJ: John Wiley & Sons, 2017. ISBN: 978-1-118-82599-0.

- [37] Andrew R. Leach. *Molecular modelling: principles and applications*. 2nd ed. Harlow, England ; New York: Prentice Hall, 2001. ISBN: 978-0-582-38210-7.
- [38] Christopher J. Cramer. *Essentials of computational chemistry: theories and models*. 2nd ed. Chichester, West Sussex, England ; Hoboken, NJ: Wiley, 2004. ISBN: 978-0-470-09182-1 978-0-470-09181-4.
- [39] Amir Mirzanejad and Sergey A. Varganov. 'Derivation of Morse potential'. In: *Molecular Physics* 123.3 (Feb. 2025). e2360542. ISSN: 0026-8976. DOI: 10.1080/00268976.2024.2360542.
- [40] Frank Jensen. 'Atomic orbital basis sets'. en. In: *WIREs Computational Molecular Science* 3.3 (2013). Pp. 273–295. ISSN: 1759-0884. DOI: 10.1002/wcms.1123.
- [41] Denis Mario Maag. 'Enhanced sampling methods applied to chemical reactions in proteins'. de. PhD thesis. Karlsruher Institut für Technologie (KIT), July 2022. DOI: 10.5445/IR/1000146741.
- [42] RCSB Protein Data Bank. *RCSB PDB: Homepage*. en-US.
- [43] Helen M. Berman et al. 'The Protein Data Bank'. In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.235.
- [44] *H++ (web-based computational prediction of protonation states and pK of ionizable groups in macromolecules)*.
- [45] John C. Gordon et al. 'H++: a server for estimating pKa's and adding missing hydrogens to macromolecules'. In: *Nucleic Acids Research* 33.suppl_2 (July 2005), W368–W371. ISSN: 0305-1048. DOI: 10.1093/nar/gki464.
- [46] Jonathan Myers et al. 'A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules'. en. In: *Proteins: Structure, Function, and Bioinformatics* 63.4 (2006). Pp. 928–938. ISSN: 1097-0134. DOI: 10.1002/prot.20922.
- [47] Ramu Anandakrishnan, Boris Aguilar and Alexey V. Onufriev. 'H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations'. In: *Nucleic Acids Research* 40.W1 (July 2012), W537–W541. ISSN: 0305-1048. DOI: 10.1093/nar/gks375.
- [48] Mats H. M. Olsson et al. 'PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions'. In: *Journal of Chemical Theory and Computation* 7.2 (Feb. 2011), pp. 525–537. ISSN: 1549-9618. DOI: 10.1021/ct100578z.
- [49] Chresten R. Søndergaard et al. 'Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values'. In: *Journal of Chemical Theory and Computation* 7.7 (July 2011), pp. 2284–2295. ISSN: 1549-9618. DOI: 10.1021/ct200133y.
- [50] Koji Okuwaki et al. 'Geometry Optimization Using the Frozen Domain and Partial Dimer Approaches in the Fragment Molecular Orbital Method: Implementation, Benchmark, and Applications to Protein Ligand-Binding Sites'. In: *Journal of Chemical Information and Modeling* 64.24 (Dec. 2024), pp. 9449–9458. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.4c01280.
- [51] Loup Verlet. 'Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules'. In: *Physical Review* 159.1 (July 1967), pp. 98–103. DOI: 10.1103/PhysRev.159.98.
- [52] William C. Swope et al. 'A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters'. In: *The Journal of Chemical Physics* 76.1 (Jan. 1982), pp. 637–649. ISSN: 0021-9606. DOI: 10.1063/1.442716.
- [53] M. P. Allen and Dominic J. Tildesley. *Computer simulation of liquids*. eng. Oxford: Clarendon press, 1994. ISBN: 978-0-19-855645-9.
- [54] H. J. C. Berendsen et al. 'Molecular dynamics with coupling to an external bath'. In: *The Journal of Chemical Physics* 81.8 (Oct. 1984), pp. 3684–3690. ISSN: 0021-9606. DOI: 10.1063/1.448118.
- [55] Giovanni Bussi, Davide Donadio and Michele Parrinello. 'Canonical sampling through velocity rescaling'. In: *The Journal of Chemical Physics* 126.1 (Jan. 2007), p. 014101. ISSN: 0021-9606. DOI: 10.1063/1.2408420.
- [56] Stephen C. Harvey, Robert K.-Z. Tan and Thomas E. Cheatham III. 'The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition'. en. In: *Journal of Computational Chemistry* 19.7 (1998). Pp. 726–740. ISSN: 1096-987X. DOI: 10.1002/(SICI)1096-987X(199805)19:7<726::AID-JCC4>3.0.CO;2-S.
- [57] Shuichi Nosé. 'A unified formulation of the constant temperature molecular dynamics methods'. In: *The Journal of Chemical Physics* 81.1 (July 1984), pp. 511–519. ISSN: 0021-9606. DOI: 10.1063/1.447334.

- [58] William G. Hoover. 'Canonical dynamics: Equilibrium phase-space distributions'. In: *Physical Review A* 31.3 (Mar. 1985), pp. 1695–1697. DOI: 10.1103/PhysRevA.31.1695.
- [59] Yuqing Lin et al. 'Application of Berendsen barostat in dissipative particle dynamics for nonequilibrium dynamic simulation'. In: *The Journal of Chemical Physics* 146.12 (Mar. 2017), p. 124108. ISSN: 0021-9606. DOI: 10.1063/1.4978807.
- [60] M. Parrinello and A. Rahman. 'Polymorphic transitions in single crystals: A new molecular dynamics method'. In: *Journal of Applied Physics* 52.12 (Dec. 1981), pp. 7182–7190. ISSN: 0021-8979. DOI: 10.1063/1.328693.
- [61] Qia Ke et al. 'Effects of thermostats/barostats on physical properties of liquids by molecular dynamics simulations'. In: *Journal of Molecular Liquids* 365 (Nov. 2022), p. 120116. ISSN: 0167-7322. DOI: 10.1016/j.molliq.2022.120116.
- [62] Martin Karplus. 'Development of Multiscale Models for Complex Chemical Systems: From H+H2 to Biomolecules (Nobel Lecture)'. In: *Angewandte Chemie International Edition* 53.38 (2014). Pp. 9992–10005. ISSN: 1521-3773. DOI: 10.1002/anie.201403924.
- [63] Michael Levitt. 'Birth and future of multiscale modeling for macromolecular systems (Nobel Lecture)'. eng. In: *Angewandte Chemie (International Ed. in English)* 53.38 (Sept. 2014), pp. 10006–10018. ISSN: 1521-3773. DOI: 10.1002/anie.201403691.
- [64] Arieh Warshel. 'Multiscale Modeling of Biological Functions: From Enzymes to Molecular Machines (Nobel Lecture)'. en. In: *Angewandte Chemie International Edition* 53.38 (2014). Pp. 10020–10031. ISSN: 1521-3773. DOI: 10.1002/anie.201403689.
- [65] Gerrit Groenhof. 'Introduction to QM/MM Simulations'. en. In: *Biomolecular Simulations*. Ed. by Luca Monticelli and Emppu Salonen. Vol. 924. Totowa, NJ: Humana Press, 2013, pp. 43–66. ISBN: 978-1-62703-016-8 978-1-62703-017-5. DOI: 10.1007/978-1-62703-017-5_3.
- [66] Thomas P. Fay, Nicolas Ferré and Miquel Huix-Rotllant. 'Efficient Polarizable QM/MM Using the Direct Reaction Field Hamiltonian with Electrostatic Potential Fitted Multipole Operators'. In: *Journal of Chemical Theory and Computation* 21.1 (Jan. 2025), pp. 183–201. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.4c01219.
- [67] Martin J. Field, Paul A. Bash and Martin Karplus. 'A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations'. en. In: *Journal of Computational Chemistry* 11.6 (1990). Pp. 700–733. ISSN: 1096-987X. DOI: 10.1002/jcc.540110605.
- [68] Tamás Veszprémi and Miklós Fehér. *Quantum Chemistry*. en. Boston, MA: Springer US, 1999. ISBN: 978-1-4613-6879-3 978-1-4615-4189-9. DOI: 10.1007/978-1-4615-4189-9.
- [69] Qiang Cui and Marcus Elstner. 'Density functional tight binding: values of semi-empirical methods in an ab initio era'. en. In: *Physical Chemistry Chemical Physics* 16.28 (June 2014), pp. 14368–14377. ISSN: 1463-9084. DOI: 10.1039/C4CP00908H.
- [70] M. Born and R. Oppenheimer. 'Zur Quantentheorie der Molekeln'. en. In: *Annalen der Physik* 389.20 (1927). Pp. 457–484. ISSN: 1521-3889. DOI: 10.1002/andp.19273892002.
- [71] Pietro d'Avenia, Liliane Maia and Gaetano Siciliano. 'Hartree-Fock type systems: Existence of ground states and asymptotic behavior'. In: *Journal of Differential Equations* 335 (Oct. 2022), pp. 580–614. ISSN: 0022-0396. DOI: 10.1016/j.jde.2022.07.012.
- [72] Clémence Corminboeuf, Fabien Tran and Jacques Weber. 'The role of density functional theory in chemistry: Some historical landmarks and applications to zeolites'. In: *Journal of Molecular Structure: THEOCHEM*. A Collection of Papers Dedicated to Professor Annick Goursot on the Occasion of her 60th Birthday 762.1 (Apr. 2006), pp. 1–7. ISSN: 0166-1280. DOI: 10.1016/j.theochem.2005.07.036.
- [73] R. O. Jones. 'Density functional theory: Its origins, rise to prominence, and future'. In: *Reviews of Modern Physics* 87.3 (Aug. 2015), pp. 897–923. DOI: 10.1103/RevModPhys.87.897.
- [74] D. R. Hartree. 'The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I. Theory and Methods'. en. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.1 (Jan. 1928), pp. 89–110. ISSN: 1469-8064, 0305-0041. DOI: 10.1017/S0305004100011919.
- [75] V. Fock. 'Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems'. de. In: *Zeitschrift für Physik* 61.1 (Jan. 1930), pp. 126–148. ISSN: 0044-3328. DOI: 10.1007/BF01340294.

- [76] V. Fock. ‘„Selfconsistent field“ mit Austausch für Natrium’. de. In: *Zeitschrift für Physik* 62.11 (Nov. 1930), pp. 795–805. ISSN: 0044-3328. DOI: 10.1007/BF01330439.
- [77] Pablo Echenique and J. L. Alonso. ‘A mathematical and computational review of Hartree–Fock SCF methods in quantum chemistry’. In: *Molecular Physics* 105.23-24 (Dec. 2007). Pp. 3057–3098. ISSN: 0026-8976. DOI: 10.1080/00268970701757875.
- [78] W. Pauli. ‘Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren’. de. In: *Zeitschrift für Physik* 31.1 (Feb. 1925), pp. 765–783. ISSN: 0044-3328. DOI: 10.1007/BF02980631.
- [79] ‘Das Mehrelektronenproblem der Molekülphysik und Quantenchemie’. de. In: *Molekülphysik und Quantenchemie: Einführung in die experimentellen und theoretischen Grundlagen*. Ed. by Hermann Haken and Hans Christoph Wolf. Berlin, Heidelberg: Springer, 2006, pp. 129–151. ISBN: 978-3-540-30315-2. DOI: 10.1007/3-540-30315-4_7.
- [80] J. C. Slater. ‘The Theory of Complex Spectra’. In: *Physical Review* 34.10 (1929), pp. 1293–1322. DOI: 10.1103/PhysRev.34.1293.
- [81] Gerd Wedler and Hans-Joachim Freund. *Lehr- und Arbeitsbuch physikalische Chemie*. ger. Siebte, wesentlich überarbeitete und erweiterte Auflage. Weinheim: Wiley-VCH, 2018. ISBN: 978-3-527-82727-5 978-3-527-82728-2.
- [82] Attila Szabo and Neil S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. en. Courier Corporation, June 2012. ISBN: 978-0-486-13459-8.
- [83] Per-Olov Löwdin. ‘Quantum Theory of Many-Particle Systems. III. Extension of the Hartree-Fock Scheme to Include Degenerate Systems and Correlation Effects’. In: *Physical Review* 97.6 (Mar. 1955), pp. 1509–1520. DOI: 10.1103/PhysRev.97.1509.
- [84] Takao Tsuneda. *Density Functional Theory in Quantum Chemistry*. eng. 1st ed. ISBN: 978-4-431-54825-6.
- [85] Marcus Elstner. *Physikalische Chemie II: Quantenmechanik und Spektroskopie*. ger. 1. Aufl. 2021. Berlin, Heidelberg: Springer Berlin Heidelberg, 2021. ISBN: 978-3-662-61461-7 978-3-662-61462-4. DOI: 10.1007/978-3-662-61462-4.
- [86] M. Ya Amusia, A. Z. Msezane and V. R. Shaginyan. ‘Density Functional Theory versus the Hartree–Fock Method: Comparative Assessment’. en. In: *Physica Scripta* 68.6 (Jan. 2003), p. C133. ISSN: 1402-4896. DOI: 10.1238/Physica.Regular.068aC0133.
- [87] P. Hohenberg and W. Kohn. ‘Inhomogeneous Electron Gas’. In: *Physical Review* 136.3B (Nov. 1964), B864–B871. DOI: 10.1103/PhysRev.136.B864.
- [88] P. Geerlings, F. De Proft and W. Langenaeker. ‘Conceptual Density Functional Theory’. In: *Chemical Reviews* 103.5 (May 2003), pp. 1793–1874. ISSN: 0009-2665. DOI: 10.1021/cr990029p.
- [89] Wolfram Koch and Max C. Holthausen. *A Chemist’s Guide to Density Functional Theory*. eng. Second edition. Weinheim: Wiley-VCH, Jan. 2015. ISBN: 978-3-527-30372-4.
- [90] W. Kohn and L. J. Sham. ‘Self-Consistent Equations Including Exchange and Correlation Effects’. In: *Physical Review* 140.4A (Nov. 1965), A1133–A1138. DOI: 10.1103/PhysRev.140.A1133.
- [91] Walter John Moore and Wolfgang Paterno. *Grundlagen der Physikalischen Chemie*. ger. Berlin: Walter de Gruyter, 2011. ISBN: 978-3-11-009941-6 978-3-11-085462-6. DOI: 10.1515/9783110854626.
- [92] Kieron Burke and Lucas O. Wagner. ‘DFT in a nutshell’. en. In: *International Journal of Quantum Chemistry* 113.2 (2013). Pp. 96–101. ISSN: 1097-461X. DOI: 10.1002/qua.24259.
- [93] Narbe Mardirossian and Martin Head-Gordon. ‘Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals’. In: *Molecular Physics* 115.19 (Oct. 2017), pp. 2315–2372. ISSN: 0026-8976. DOI: 10.1080/00268976.2017.1333644.
- [94] Ahmed Badreldin, Othmane Bouhali and Ahmed Abdel-Wahab. ‘Complimentary Computational Cues for Water Electrocatalysis: A DFT and ML Perspective’. en. In: *Advanced Functional Materials* 34.12 (2024). P. 2312425. ISSN: 1616-3028. DOI: 10.1002/adfm.202312425.
- [95] John P. Perdew and Wang Yue. ‘Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation’. In: *Physical Review B* 33.12 (June 1986), pp. 8800–8802. DOI: 10.1103/PhysRevB.33.8800.

- [96] John P. Perdew, Kieron Burke and Matthias Ernzerhof. 'Generalized Gradient Approximation Made Simple'. In: *Physical Review Letters* 77.18 (Oct. 1996), pp. 3865–3868. DOI: 10.1103/PhysRevLett.77.3865.
- [97] Chengteh Lee, Weitao Yang and Robert G. Parr. 'Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density'. In: *Physical Review B* 37.2 (Jan. 1988), pp. 785–789. DOI: 10.1103/PhysRevB.37.785.
- [98] A. D. Becke. 'Density-functional exchange-energy approximation with correct asymptotic behavior'. In: *Physical Review A* 38.6 (Sept. 1988), pp. 3098–3100. DOI: 10.1103/PhysRevA.38.3098.
- [99] Axel D. Becke. 'Density-functional thermochemistry. III. The role of exact exchange'. In: *The Journal of Chemical Physics* 98.7 (Apr. 1993), pp. 5648–5652. ISSN: 0021-9606. DOI: 10.1063/1.464913.
- [100] M Marsman et al. 'Hybrid functionals applied to extended systems'. en. In: *Journal of Physics: Condensed Matter* 20.6 (Jan. 2008), p. 064201. ISSN: 0953-8984. DOI: 10.1088/0953-8984/20/6/064201.
- [101] Narbe Mardirossian and Martin Head-Gordon. 'ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation'. In: *The Journal of Chemical Physics* 144.21 (June 2016), p. 214110. ISSN: 0021-9606. DOI: 10.1063/1.4952647.
- [102] W. J. Hehre, R. F. Stewart and J. A. Pople. 'Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals'. In: *The Journal of Chemical Physics* 51.6 (Sept. 1969), pp. 2657–2664. ISSN: 0021-9606. DOI: 10.1063/1.1672392.
- [103] W. J. Hehre et al. 'Self-Consistent Molecular Orbital Methods. IV. Use of Gaussian Expansions of Slater-Type Orbitals. Extension to Second-Row Molecules'. In: *The Journal of Chemical Physics* 52.5 (Mar. 1970), pp. 2769–2773. ISSN: 0021-9606. DOI: 10.1063/1.1673374.
- [104] R. Ditchfield, W. J. Hehre and J. A. Pople. 'Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules'. In: *The Journal of Chemical Physics* 54.2 (Jan. 1971), pp. 724–728. ISSN: 0021-9606. DOI: 10.1063/1.1674902.
- [105] W. J. Hehre, R. Ditchfield and J. A. Pople. 'Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules'. In: *The Journal of Chemical Physics* 56.5 (Mar. 1972), pp. 2257–2261. ISSN: 0021-9606. DOI: 10.1063/1.1677527.
- [106] Warren J. Hehre and William A. Lathan. 'Self-Consistent Molecular Orbital Methods. XIV. An Extended Gaussian-Type Basis for Molecular Orbital Studies of Organic Molecules. Inclusion of Second Row Elements'. In: *The Journal of Chemical Physics* 56.11 (June 1972), pp. 5255–5257. ISSN: 0021-9606. DOI: 10.1063/1.1677028.
- [107] Florian Weigend and Reinhart Ahlrichs. 'Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy'. en. In: *Physical Chemistry Chemical Physics* 7.18 (Aug. 2005), pp. 3297–3305. ISSN: 1463-9084. DOI: 10.1039/B508541A.
- [108] Samuel J. Pitman et al. 'Benchmarking Basis Sets for Density Functional Theory Thermochemistry Calculations: Why Unpolarized Basis Sets and the Polarized 6-311G Family Should Be Avoided'. In: *The Journal of Physical Chemistry A* 127.48 (Dec. 2023), pp. 10295–10306. ISSN: 1089-5639. DOI: 10.1021/acs.jpca.3c05573.
- [109] Jingjing Zheng, Xuefei Xu and Donald G. Truhlar. 'Minimally augmented Karlsruhe basis sets'. en. In: *Theoretical Chemistry Accounts* 128.3 (Feb. 2011), pp. 295–305. ISSN: 1432-2234. DOI: 10.1007/s00214-010-0846-z.
- [110] W. Kohn. 'Nobel Lecture: Electronic structure of matter—wave functions and density functionals'. In: *Reviews of Modern Physics* 71.5 (Oct. 1999), pp. 1253–1266. DOI: 10.1103/RevModPhys.71.1253.
- [111] John A. Pople. 'Nobel Lecture: Quantum chemical models'. In: *Reviews of Modern Physics* 71.5 (1999), pp. 1267–1274. DOI: 10.1103/RevModPhys.71.1267.
- [112] Roberto Car et al. 'First-Principles Molecular Dynamics'. en. In: *Handbook of Materials Modeling: Methods*. Ed. by Sidney Yip. Dordrecht: Springer Netherlands, 2005, pp. 59–76. ISBN: 978-1-4020-3286-8. DOI: 10.1007/978-1-4020-3286-8_5.
- [113] Clarivate Analytics. *Web of Science Core Collection*.
- [114] G. Seifert, D. Porezag and Th. Frauenheim. 'Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme'. en. In: *International Journal of Quantum Chemistry* 58.2 (1996). Pp. 185–192. ISSN: 1097-461X. DOI: 10.1002/(SICI)1097-461X(1996)58:2<185::AID-QUA7>3.0.CO;2-U.

- [115] D. Porezag et al. 'Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon'. In: *Physical Review B* 51.19 (May 1995), pp. 12947–12957. DOI: 10.1103/PhysRevB.51.12947.
- [116] M. Elstner et al. 'Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties'. In: *Physical Review B* 58.11 (Sept. 1998), pp. 7260–7268. DOI: 10.1103/PhysRevB.58.7260.
- [117] Michael Gaus, Qiang Cui and Marcus Elstner. 'DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB)'. In: *Journal of Chemical Theory and Computation* 7.4 (Apr. 2011), pp. 931–948. ISSN: 1549-9618. DOI: 10.1021/ct100684s.
- [118] M. Elstner. 'The SCC-DFTB method and its application to biological systems'. en. In: *Theoretical Chemistry Accounts* 116.1 (Aug. 2006), pp. 316–325. ISSN: 1432-2234. DOI: 10.1007/s00214-005-0066-0.
- [119] Marcus Elstner and Gotthard Seifert. 'Density functional tight binding'. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 372.2011 (Mar. 2014), p. 20120483. ISSN: 1364-503X. DOI: 10.1098/rsta.2012.0483.
- [120] Qiang Cui et al. 'A QM/MM Implementation of the Self-Consistent Charge Density Functional Tight Binding (SCC-DFTB) Method'. In: *The Journal of Physical Chemistry B* 105.2 (Jan. 2001), pp. 569–585. ISSN: 1520-6106. DOI: 10.1021/jp0029109.
- [121] Michael Gaus et al. 'Automatized Parametrization of SCC-DFTB Repulsive Potentials: Application to Hydrocarbons'. In: *The Journal of Physical Chemistry A* 113.43 (Oct. 2009), pp. 11866–11881. ISSN: 1089-5639. DOI: 10.1021/jp902973m.
- [122] Pekka Koskinen and Ville Mäkinen. 'Density-functional tight-binding for beginners'. In: *Computational Materials Science* 47.1 (Nov. 2009), pp. 237–253. ISSN: 0927-0256. DOI: 10.1016/j.commatsci.2009.07.013.
- [123] Haiyan Liu et al. 'Quantum mechanics simulation of protein dynamics on long timescale'. fr. In: *Proteins: Structure, Function, and Bioinformatics* 44.4 (2001). Pp. 484–489. ISSN: 1097-0134. DOI: 10.1002/prot.1114.
- [124] Alessandro Barducci, Massimiliano Bonomi and Michele Parrinello. 'Metadynamics'. en. In: *WIREs Computational Molecular Science* 1.5 (2011). Pp. 826–843. ISSN: 1759-0884. DOI: 10.1002/wcms.31.
- [125] Marcus Elstner, Qiang Cui and Maja Gruden. *Introduction to Statistical Thermodynamics: A Molecular Perspective*. en. Physical Chemistry in Action. Cham: Springer International Publishing, 2024. ISBN: 978-3-031-54993-9 978-3-031-54994-6. DOI: 10.1007/978-3-031-54994-6.
- [126] Jérôme Hénin et al. 'Enhanced sampling methods for molecular dynamics simulations'. In: *Living Journal of Computational Molecular Science* 4.1 (2022). ISSN: 25756524. DOI: 10.33011/livecoms.4.1.1583.
- [127] Marcus Elstner and Tomáš Kubař. 'Moleküldynamik-Simulationen'. Lecture. Karlsruher Institut für Technologie (KIT), June 2021.
- [128] Nils Hansen. 'Advances in Free-Energy Calculations Based on Classical Molecular Dynamics Simulations'. Habilitation Thesis. ETH Zürich, 2014.
- [129] Ming Chen. 'Collective variable-based enhanced sampling and machine learning'. en. In: *The European Physical Journal B* 94.10 (Oct. 2021), p. 211. ISSN: 1434-6036. DOI: 10.1140/epjb/s10051-021-00220-w.
- [130] Mohammad M. Sultan and Vijay S. Pande. 'Automated design of collective variables using supervised machine learning'. In: *The Journal of Chemical Physics* 149.9 (Sept. 2018), p. 094106. ISSN: 0021-9606. DOI: 10.1063/1.5029972.
- [131] Luigi Bonati et al. 'A unified framework for machine learning collective variables for enhanced sampling simulations: mlcolvar'. In: *The Journal of Chemical Physics* 159.1 (July 2023), p. 014801. ISSN: 0021-9606. DOI: 10.1063/5.0156343.
- [132] Haohao Fu et al. 'Collective Variable-Based Enhanced Sampling: From Human Learning to Machine Learning'. In: *The Journal of Physical Chemistry Letters* 15.6 (Feb. 2024), pp. 1774–1783. DOI: 10.1021/acs.jpclett.3c03542.
- [133] Hythem Sidky, Wei Chen and Andrew L. Ferguson. 'Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation'. In: *Molecular Physics* 118.5 (Mar. 2020). e1737742. ISSN: 0026-8976. DOI: 10.1080/00268976.2020.1737742.

- [134] Ayan Majumder and John E. Straub. 'Machine Learning Derived Collective Variables for the Study of Protein Homodimerization in Membrane'. In: *Journal of Chemical Theory and Computation* 20.13 (July 2024), pp. 5774–5783. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.4c00454.
- [135] Rangsiman Ketkaew and Sandra Lubner. 'DeepCV: A Deep Learning Framework for Blind Search of Collective Variables in Expanded Configurational Space'. In: *Journal of Chemical Information and Modeling* 62.24 (Dec. 2022), pp. 6352–6364. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.2c00883.
- [136] Kai Zhu et al. 'Enhanced Sampling in the Age of Machine Learning: Algorithms and Applications'. In: *Chemical Reviews* (Oct. 2025). ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.5c00700.
- [137] Yuji Sugita and Yuko Okamoto. 'Replica-exchange molecular dynamics method for protein folding'. In: *Chemical Physics Letters* 314.1 (Nov. 1999), pp. 141–151. ISSN: 0009-2614. DOI: 10.1016/S0009-2614(99)01123-9.
- [138] Ruxi Qi et al. 'Replica Exchange Molecular Dynamics: A Practical Application Protocol with Solutions to Common Problems and a Peptide Aggregation and Self-Assembly Example'. In: *Methods in molecular biology (Clifton, N.J.)* 1777 (2018), pp. 101–119. ISSN: 1064-3745. DOI: 10.1007/978-1-4939-7811-3_5.
- [139] Donald Hamelberg, John Mongan and J. Andrew McCammon. 'Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules'. In: *The Journal of Chemical Physics* 120.24 (June 2004), pp. 11919–11929. ISSN: 0021-9606. DOI: 10.1063/1.1755656.
- [140] Yi Wang et al. 'Implementation of Accelerated Molecular Dynamics in NAMD'. In: *Computational science & discovery* 4.1 (2011), p. 015002. ISSN: 1749-4699. DOI: 10.1088/1749-4699/4/1/015002.
- [141] Soumav Nath. 'Thermodynamic and conformational analysis of monomeric A β isoforms binding with A β (1-42) fibril - insights into the secondary nucleation of amyloidosis'. PhD thesis. May 2022. DOI: 10.13140/RG.2.2.28745.93288.
- [142] Yinglong Miao, Sara E. Nichols and J. Andrew McCammon. 'Free energy landscape of G-protein coupled receptors, explored by accelerated molecular dynamics'. en. In: *Physical Chemistry Chemical Physics* 16.14 (Mar. 2014), pp. 6398–6406. ISSN: 1463-9084. DOI: 10.1039/C3CP53962H.
- [143] Qinghua Liao. 'Enhanced sampling and free energy calculations for protein simulations'. en-US. In: *Progress in Molecular Biology and Translational Science*. Vol. 170. Academic Press, Jan. 2020, pp. 177–213. DOI: 10.1016/bs.pmbts.2020.01.006.
- [144] Rafael C. Bernardi, Marcelo C. R. Melo and Klaus Schulten. 'Enhanced sampling techniques in molecular dynamics simulations of biological systems'. In: *Biochimica et Biophysica Acta (BBA) - General Subjects*. Recent developments of molecular dynamics 1850.5 (May 2015), pp. 872–877. ISSN: 0304-4165. DOI: 10.1016/j.bbagen.2014.10.019.
- [145] Yi Isaac Yang et al. 'Enhanced sampling in molecular dynamics'. In: *The Journal of Chemical Physics* 151.7 (Aug. 2019), p. 070902. ISSN: 0021-9606. DOI: 10.1063/1.5109531.
- [146] Glenn M. Torrie and John P. Valleau. 'Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid'. In: *Chemical Physics Letters* 28.4 (Oct. 1974), pp. 578–581. ISSN: 0009-2614. DOI: 10.1016/0009-2614(74)80109-0.
- [147] G. M. Torrie and J. P. Valleau. 'Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling'. In: *Journal of Computational Physics* 23.2 (Feb. 1977), pp. 187–199. ISSN: 0021-9991. DOI: 10.1016/0021-9991(77)90121-8.
- [148] Johannes Kästner. 'Umbrella sampling'. en. In: *WIREs Computational Molecular Science* 1.6 (2011). Pp. 932–942. ISSN: 1759-0884. DOI: 10.1002/wcms.66.
- [149] Alessandro Laio and Michele Parrinello. 'Escaping free-energy minima'. In: *Proceedings of the National Academy of Sciences* 99.20 (Oct. 2002), pp. 12562–12566. DOI: 10.1073/pnas.202427399.
- [150] Dalibor Trapl and Wojtech Spiwok. 'Analysis of the Results of Metadynamics Simulations by metadynminer and metadynminer3d'. en. In: *The R Journal* 14.3 (Dec. 2022), pp. 46–58. ISSN: 2073-4859. DOI: 10.32614/RJ-2022-057.
- [151] Giovanni Bussi and Davide Branduardi. 'Free-Energy Calculations with Metadynamics: Theory and Practice'. en. In: *Reviews in Computational Chemistry*. Ed. by Abby L. Parrill and Kenny B. Lipkowitz. 1st ed. Wiley, Apr. 2015, pp. 1–49. ISBN: 978-1-118-40777-6 978-1-118-88988-6. DOI: 10.1002/9781118889886.ch1.

- [152] Alessandro Barducci, Giovanni Bussi and Michele Parrinello. ‘Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method’. In: *Physical Review Letters* 100.2 (Jan. 2008), p. 020603. DOI: 10.1103/PhysRevLett.100.020603.
- [153] Rui Sun et al. ‘Transition-Tempered Metadynamics Is a Promising Tool for Studying the Permeation of Drug-like Molecules through Membranes’. In: *Journal of Chemical Theory and Computation* 12.10 (Oct. 2016), pp. 5157–5169. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.6b00206.
- [154] Paolo Raiteri et al. ‘Efficient Reconstruction of Complex Free Energy Landscapes by Multiple Walkers Metadynamics’. In: *The Journal of Physical Chemistry B* 110.8 (Mar. 2006), pp. 3533–3539. ISSN: 1520-6106. DOI: 10.1021/jp054359r.
- [155] ‘Nobel Prize in Physics 2024’. en-US. In: *NobelPrize.org* (2024). DOI: <https://www.nobelprize.org/prizes/physics/2024/summary/>.
- [156] ‘Nobel Prize in Chemistry 2024’. en-US. In: *NobelPrize.org* (2024). DOI: <https://www.nobelprize.org/prizes/chemistry/2024/summary/>.
- [157] Raghunathan Ramakrishnan et al. ‘Big Data Meets Quantum Chemistry Approximations: The Δ -Machine Learning Approach’. In: *Journal of Chemical Theory and Computation* 11.5 (May 2015), pp. 2087–2096. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.5b00099.
- [158] Lin Shen, Jingheng Wu and Weitao Yang. ‘Multiscale Quantum Mechanics/Molecular Mechanics Simulations with Neural Networks’. In: *Journal of Chemical Theory and Computation* 12.10 (Oct. 2016), pp. 4934–4946. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.6b00663.
- [159] Junmian Zhu et al. ‘Artificial neural network correction for density-functional tight-binding molecular dynamics simulations’. en. In: *MRS Communications* 9.3 (Sept. 2019), pp. 867–873. ISSN: 2159-6867. DOI: 10.1557/mrc.2019.80.
- [160] Claudia L. Gómez-Flores et al. ‘Accurate Free Energies for Complex Condensed-Phase Reactions Using an Artificial Neural Network Corrected DFTB/MM Methodology’. In: *Journal of Chemical Theory and Computation* 18.2 (Feb. 2022), pp. 1213–1226. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.1c00811.
- [161] Julian Böser et al. ‘Reduction pathway of glutaredoxin 1 investigated with QM/MM molecular dynamics using a neural network correction’. In: *The Journal of Chemical Physics* 157.15 (Oct. 2022), p. 154104. ISSN: 0021-9606. DOI: 10.1063/5.0123089.
- [162] Lennard Bösel, Moritz Thürlmann and Sereina Riniker. ‘Machine Learning in QM/MM Molecular Dynamics Simulations of Condensed-Phase Systems’. In: *Journal of Chemical Theory and Computation* 17.5 (May 2021), pp. 2641–2658. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.0c01112.
- [163] Leonardo Medrano Sandonas et al. ‘Advancing density functional tight-binding method for large organic molecules through equivariant neural networks’. en. In: *Physical Chemistry Chemical Physics* (Jan. 2026). ISSN: 1463-9084. DOI: 10.1039/D6CP00038J.
- [164] Jinzhe Zeng et al. ‘Development of Range-Corrected Deep Learning Potentials for Fast, Accurate Quantum Mechanical/Molecular Mechanical Simulations of Chemical Reactions in Solution’. In: *Journal of Chemical Theory and Computation* 17.11 (Nov. 2021), pp. 6993–7009. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.1c00201.
- [165] Jörg Behler. ‘Four Generations of High-Dimensional Neural Network Potentials’. In: *Chemical Reviews* 121.16 (Aug. 2021), pp. 10037–10072. ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.0c00868.
- [166] Jörg Behler and Michele Parrinello. ‘Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces’. In: *Physical Review Letters* 98.14 (Apr. 2007), p. 146401. DOI: 10.1103/PhysRevLett.98.146401.
- [167] Jörg Behler. ‘Atom-centered symmetry functions for constructing high-dimensional neural network potentials’. In: *The Journal of Chemical Physics* 134.7 (Feb. 2011), p. 074106. ISSN: 0021-9606. DOI: 10.1063/1.3553717.
- [168] Andreas Singraber, Jörg Behler and Christoph Dellago. ‘Library-Based LAMMPS Implementation of High-Dimensional Neural Network Potentials’. In: *Journal of Chemical Theory and Computation* 15.3 (Mar. 2019), pp. 1827–1840. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.8b00770.
- [169] K. V. Jovan Jose, Nongnuch Artrith and Jörg Behler. ‘Construction of high-dimensional neural network potentials using environment-dependent atom pairs’. In: *The Journal of Chemical Physics* 136.19 (May 2012), p. 194111. ISSN: 0021-9606. DOI: 10.1063/1.4712397.

- [170] Zachary L. Glick et al. 'AP-Net: An atomic-pairwise neural network for smooth and transferable interaction potentials'. In: *The Journal of Chemical Physics* 153.4 (July 2020), p. 044112. ISSN: 0021-9606. DOI: 10.1063/1.50011521.
- [171] Kun Yao et al. 'The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics'. en. In: *Chemical Science* 9.8 (Feb. 2018), pp. 2261–2269. ISSN: 2041-6539. DOI: 10.1039/C7SC04934J.
- [172] M. Gastegger et al. 'wACSF—Weighted atom-centered symmetry functions as descriptors in machine learning potentials'. In: *The Journal of Chemical Physics* 148.24 (Mar. 2018), p. 241709. ISSN: 0021-9606. DOI: 10.1063/1.5019667.
- [173] Yunxing Zuo et al. 'Performance and Cost Assessment of Machine Learning Interatomic Potentials'. In: *The Journal of Physical Chemistry A* 124.4 (Jan. 2020), pp. 731–745. ISSN: 1089-5639. DOI: 10.1021/acs.jpca.9b08723.
- [174] Christian Carl Schmidt. 'Machine Learning for Multiscale Simulation of Complex Molecular Systems: Phosphorylation Reactions and Excited States Energy Surfaces'. PhD Thesis. Karlsruher Institut für Technologie (KIT), Dec. 2025. DOI: 10.5445/IR/1000189441.
- [175] <https://cdn.ps.emap.com/wp-content/uploads/sites/3/2024/05/Fig-2.-Major-roles-of-phosphate-in-the-body.png> (Accessed: 1 Jan. 2026).
- [176] Tony Hunter. 'Why nature chose phosphate to modify proteins'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1602 (Sept. 2012), pp. 2513–2516. ISSN: 0962-8436. DOI: 10.1098/rstb.2012.0013.
- [177] Christian Fernández-García, Adam J. Coggins and Matthew W. Powner. 'A Chemist's Perspective on the Role of Phosphorus at the Origins of Life'. In: *Life* 7.3 (July 2017), p. 31. ISSN: 2075-1729. DOI: 10.3390/life7030031.
- [178] Philip Cohen. 'The regulation of protein function by multisite phosphorylation – a 25 year update'. In: *Trends in Biochemical Sciences* 25.12 (Dec. 2000), pp. 596–601. ISSN: 0968-0004. DOI: 10.1016/S0968-0004(00)01712-6.
- [179] Richard Wolfenden and Mark J. Snider. 'The Depth of Chemical Time and the Power of Enzymes as Catalysts'. In: *Accounts of Chemical Research* 34.12 (Dec. 2001), pp. 938–945. ISSN: 0001-4842. DOI: 10.1021/ar000058i.
- [180] F. H. Westheimer. 'Why nature chose phosphates'. In: *Science* 235.4793 (Mar. 1987), pp. 1173–1178. DOI: 10.1126/science.2434996.
- [181] Kyle W. Knouse et al. 'Nature Chose Phosphates and Chemists Should Too: How Emerging P(V) Methods Can Augment Existing Strategies'. In: *ACS Central Science* 7.9 (Sept. 2021), pp. 1473–1485. ISSN: 2374-7943. DOI: 10.1021/acscentsci.1c00487.
- [182] Shina C. L. Kamerlin et al. 'Why nature really chose phosphate'. en. In: *Quarterly Reviews of Biophysics* 46.1 (Feb. 2013), pp. 1–132. ISSN: 0033-5835, 1469-8994. DOI: 10.1017/S0033583512000157.
- [183] Edmond H. Fischer. 'Protein Phosphorylation and Cellular Regulation II (Nobel Lecture)'. In: *Angewandte Chemie International Edition in English* 32.8 (1993). Pp. 1130–1137. ISSN: 1521-3773. DOI: 10.1002/anie.199311301.
- [184] Edwin G. Krebs. 'Protein Phosphorylation and Cellular Regulation I (Nobel Lecture)'. In: *Angewandte Chemie International Edition in English* 32.8 (1993). Pp. 1122–1129. ISSN: 1521-3773. DOI: 10.1002/anie.199311221.
- [185] Paul D. Boyer. 'Energy, Life, and ATP (Nobel Lecture)'. en. In: *Angewandte Chemie International Edition* 37.17 (1998). Pp. 2296–2307. ISSN: 1521-3773. DOI: 10.1002/(SICI)1521-3773(19980918)37:17<2296::AID-ANIE2296>3.0.CO;2-W.
- [186] John E. Walker. 'ATP Synthesis by Rotary Catalysis (Nobel lecture)'. en. In: *Angewandte Chemie International Edition* 37.17 (1998). Pp. 2308–2319. ISSN: 1521-3773. DOI: 10.1002/(SICI)1521-3773(19980918)37:17<2308::AID-ANIE2308>3.0.CO;2-W.
- [187] Ann M Stock, Victoria L Robinson and Paul N Goudreau. 'Two-component signal transduction'. en. In: *Annual review of biochemistry* 69.1 (2000), pp. 183–215.
- [188] Janusz J. Petkowski, William Bains and Sara Seager. 'Natural Products Containing 'Rare' Organophosphorus Functional Groups'. In: *Molecules* 24.5 (Feb. 2019), p. 866. ISSN: 1420-3049. DOI: 10.3390/molecules24050866.

- [189] Dénes Berta et al. 'Mechanism-Based Redesign of GAP to Activate Oncogenic Ras'. In: *Journal of the American Chemical Society* 145.37 (Sept. 2023), pp. 20302–20310. ISSN: 0002-7863. DOI: 10.1021/jacs.3c04330.
- [190] Rodrigo Recabarren et al. 'Mechanistic insights into the phosphoryl transfer reaction in cyclin-dependent kinase 2: A QM/MM study'. en. In: *PLOS ONE* 14.9 (Apr. 2019), e0215793. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0215793.
- [191] Juliana Murillo-López et al. 'Studying the phosphoryl transfer mechanism of the E. coli phosphofructokinase-2: from X-ray structure to quantum mechanics/molecular mechanics simulations'. en. In: *Chemical Science* 10.10 (2019), pp. 2882–2892. DOI: 10.1039/C9SC00094A.
- [192] Michael Gaus et al. 'Parameterization of DFTB3/3OB for Sulfur and Phosphorus for Chemical and Biological Applications'. In: *Journal of Chemical Theory and Computation* 10.4 (Apr. 2014), pp. 1518–1537. ISSN: 1549-9618. DOI: 10.1021/ct401002w.
- [193] Marcus D. Hanwell et al. 'Avogadro: an advanced semantic chemical editor, visualization, and analysis platform'. en. In: *Journal of Cheminformatics* 4.1 (Aug. 2012), p. 17. ISSN: 1758-2946. DOI: 10.1186/1758-2946-4-17.
- [194] The Avogadro Team. *Avogadro: an open-source molecular builder and visualization tool. Version 1.2.0*. en.
- [195] Michael Gaus, Albrecht Goetz and Marcus Elstner. 'Parametrization and Benchmark of DFTB3 for Organic Molecules'. In: *Journal of Chemical Theory and Computation* 9.1 (Jan. 2013), pp. 338–354. ISSN: 1549-9618. DOI: 10.1021/ct300849w.
- [196] Xiya Lu et al. 'Parametrization of DFTB3/3OB for Magnesium and Zinc for Chemical and Biological Applications'. In: *The Journal of Physical Chemistry B* 119.3 (Jan. 2015), pp. 1062–1082. ISSN: 1520-6106. DOI: 10.1021/jp506557r.
- [197] William L. Jorgensen et al. 'Comparison of simple potential functions for simulating liquid water'. In: *The Journal of Chemical Physics* 79.2 (July 1983), pp. 926–935. ISSN: 0021-9606. DOI: 10.1063/1.445869.
- [198] Viktor Hornak et al. 'Comparison of multiple Amber force fields and development of improved protein backbone parameters'. en. In: *Proteins: Structure, Function, and Bioinformatics* 65.3 (2006). Pp. 712–725. ISSN: 1097-0134. DOI: 10.1002/prot.21123.
- [199] Kresten Lindorff-Larsen et al. 'Improved side-chain torsion potentials for the Amber ff99SB protein force field'. en. In: *Proteins: Structure, Function, and Bioinformatics* 78.8 (2010). Pp. 1950–1958. ISSN: 1097-0134. DOI: 10.1002/prot.22711.
- [200] H. J. C. Berendsen, D. van der Spoel and R. van Drunen. 'GROMACS: A message-passing parallel molecular dynamics implementation'. In: *Computer Physics Communications* 91.1 (Sept. 1995), pp. 43–56. ISSN: 0010-4655. DOI: 10.1016/0010-4655(95)00042-E.
- [201] David Van Der Spoel et al. 'GROMACS: Fast, flexible, and free'. en. In: *Journal of Computational Chemistry* 26.16 (2005). Pp. 1701–1718. ISSN: 1096-987X. DOI: 10.1002/jcc.20291.
- [202] Sander Pronk et al. 'GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit'. In: *Bioinformatics* 29.7 (Apr. 2013), pp. 845–854. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt055.
- [203] Mark James Abraham et al. 'GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers'. In: *SoftwareX* 1-2 (Sept. 2015), pp. 19–25. ISSN: 2352-7110. DOI: 10.1016/j.softx.2015.06.001.
- [204] Mark Abraham et al. 'GROMACS 2025.4 Manual'. In: (Nov. 2025). DOI: 10.5281/zenodo.17671776.
- [205] B. Hourahine et al. 'DFTB+, a software package for efficient approximate density functional theory based atomistic simulations'. In: *The Journal of Chemical Physics* 152.12 (Mar. 2020), p. 124101. ISSN: 0021-9606. DOI: 10.1063/1.5143190.
- [206] B. Hourahine et al. 'Recent Developments in DFTB+, a Software Package for Efficient Atomistic Quantum Mechanical Simulations'. In: *The Journal of Physical Chemistry A* 129.24 (June 2025), pp. 5373–5390. ISSN: 1089-5639. DOI: 10.1021/acs.jpca.5c01146.
- [207] Tomáš Kubař. <https://github.com/tomaskubar/gromacs-dftbplus>. Nov. 2025.
- [208] Tomáš Kubař. <https://github.com/tomaskubar/dftbplus>. Oct. 2025.

- [209] Tom Darden, Darrin York and Lee Pedersen. 'Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems'. In: *The Journal of Chemical Physics* 98.12 (June 1993), pp. 10089–10092. ISSN: 0021-9606. DOI: 10.1063/1.464397.
- [210] Massimiliano Bonomi et al. 'Promoting transparency and reproducibility in enhanced molecular simulations'. en. In: *Nature Methods* 16.8 (Aug. 2019), pp. 670–673. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0506-8.
- [211] Gareth A. Tribello et al. 'PLUMED 2: New feathers for an old bird'. In: *Computer Physics Communications* 185.2 (Feb. 2014), pp. 604–613. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2013.09.018.
- [212] Frank Neese. 'The ORCA program system'. en. In: *WIREs Computational Molecular Science* 2.1 (2012). Pp. 73–78. ISSN: 1759-0884. DOI: 10.1002/wcms.81.
- [213] Frank Neese. 'Software update: The ORCA program system—Version 5.0'. en. In: *WIREs Computational Molecular Science* 12.5 (2022). e1606. ISSN: 1759-0884. DOI: 10.1002/wcms.1606.
- [214] TensorFlow Developers. *TensorFlow*. Aug. 2025. DOI: 10.5281/zenodo.16852354.
- [215] Martín Abadi et al. 'TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems'. en. In: (2015).
- [216] Martín Abadi et al. 'TensorFlow: A System for Large-Scale Machine Learning'. en. In: (Nov. 2016).
- [217] François and others Chollet. *Keras*. 2015.
- [218] Alessandro Laio and Francesco L Gervasio. 'Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science'. en. In: *Reports on Progress in Physics* 71.12 (Nov. 2008), p. 126601. ISSN: 0034-4885. DOI: 10.1088/0034-4885/71/12/126601.
- [219] Yanier Crespo et al. 'Metadynamics convergence law in a multidimensional system'. en. In: *Physical Review E* 81.5 (May 2010), p. 055701. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.81.055701.
- [220] Mayukh Kansari. 'Improvements of the density-functional tight-binding parameters and their application in histidine kinase'. eng. PhD Thesis. Karlsruher Institut für Technologie (KIT), July 2022. DOI: 10.5445/IR/1000155207.
- [221] B. Hourahine et al. 'Self-interaction and strong correlation in DFTB'. eng. In: *The Journal of Physical Chemistry. A* 111.26 (July 2007), pp. 5671–5677. ISSN: 1089-5639. DOI: 10.1021/jp070173b.
- [222] Aron J. Cohen, Paula Mori-Sánchez and Weitao Yang. 'Insights into Current Limitations of Density Functional Theory'. In: *Science* 321.5890 (Aug. 2008), pp. 792–794. DOI: 10.1126/science.1158722.
- [223] Mathias Rapacioli et al. 'Modeling Charge Resonance in Cationic Molecular Clusters: Combining DFT-Tight Binding with Configuration Interaction'. In: *Journal of Chemical Theory and Computation* 7.1 (Jan. 2011), pp. 44–55. ISSN: 1549-9618. DOI: 10.1021/ct100412f.
- [224] Thomas A. Niehaus and Fabio Della Sala. 'Range separated functionals in the density functional based tight-binding method: Formalism'. en. In: *physica status solidi (b)* 249.2 (2012). Pp. 237–244. ISSN: 1521-3951. DOI: 10.1002/pssb.201100694.
- [225] Ariel E. Mechaly et al. 'Structural Coupling between Autokinase and Phosphotransferase Reactions in a Bacterial Histidine Kinase'. In: *Structure* 25.6 (June 2017), 939–944.e3. ISSN: 0969-2126. DOI: 10.1016/j.str.2017.04.011.
- [226] Katharina Spies. 'Multiscale Simulations of Proton Coupled Electron Transfer (PCET) in Biological Systems'. PhD Thesis. Karlsruher Institut für Technologie (KIT), Oct. 2025. DOI: 10.5445/IR/1000189149.
- [227] Timothy J. Giese, Jinzhe Zeng and Darrin M. York. 'Transferability of MACE Graph Neural Network for Range Corrected Δ -Machine Learning Potential QM/MM Applications'. In: *The Journal of Physical Chemistry B* 129.22 (June 2025), pp. 5477–5490. ISSN: 1520-6106. DOI: 10.1021/acs.jpcc.5c02006.
- [228] Naveen Michaud-Agrawal et al. 'MDAnalysis: A toolkit for the analysis of molecular dynamics simulations'. en. In: *Journal of Computational Chemistry* 32.10 (2011). Pp. 2319–2327. ISSN: 1096-987X. DOI: 10.1002/jcc.21787.
- [229] Richard J. Gowers et al. 'MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations'. en. In: *SciPy 2016* (June 2016). DOI: 10.25080/Majora-629e541a-00e.

- [230] Tino Krell et al. 'Bacterial Sensor Kinases: Diversity in the Recognition of Environmental Signals'. en. In: *Annual Review of Microbiology* 64. Volume 64, 2010 (Oct. 2010), pp. 539–559. ISSN: 0066-4227, 1545-3251. DOI: 10.1146/annurev.micro.112408.134054.
- [231] Christopher P. Zschiedrich, Victoria Keidel and Hendrik Szurmant. 'Molecular Mechanisms of Two-Component Signal Transduction'. In: *Journal of Molecular Biology*. Molecular Basis of Signal Transduction 428.19 (Sept. 2016), pp. 3752–3775. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2016.08.003.
- [232] Yasuhiro Gotoh et al. 'Two-component signal transduction as potential drug targets in pathogenic bacteria'. In: *Current Opinion in Microbiology*. Cell regulation 13.2 (Apr. 2010), pp. 232–239. ISSN: 1369-5274. DOI: 10.1016/j.mib.2010.01.008.
- [233] Sandeep Tiwari et al. 'Two-Component Signal Transduction Systems of Pathogenic Bacteria As Targets for Anti-microbial Therapy: An Overview'. English. In: *Frontiers in Microbiology* 8 (Oct. 2017). ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.01878.
- [234] Madhumati Sevvana et al. 'A Ligand-Induced Switch in the Periplasmic Domain of Sensor Histidine Kinase CitA'. In: *Journal of Molecular Biology* 377.2 (Mar. 2008), pp. 512–523. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2008.01.024.
- [235] Daniela Albanesi et al. 'Structural plasticity and catalysis regulation of a thermosensor histidine kinase'. In: *Proceedings of the National Academy of Sciences* 106.38 (Sept. 2009), pp. 16185–16190. DOI: 10.1073/pnas.0906699106.
- [236] Gottfried Unden, Stephanie Nilkens and Mareike Singenstreu. 'Bacterial sensor kinases using Fe–S cluster binding PAS or GAF domains for O₂ sensing'. en. In: *Dalton Transactions* 42.9 (2013), pp. 3082–3087. DOI: 10.1039/C2DT32089D.
- [237] Manasi P. Bhate et al. 'Signal Transduction in Histidine Kinases: Insights from New Structures'. English. In: *Structure* 23.6 (June 2015), pp. 981–994. ISSN: 0969-2126. DOI: 10.1016/j.str.2015.04.002.
- [238] P. V. Attwood et al. 'Focus on phosphohistidine'. en. In: *Amino Acids* 32.1 (Jan. 2007), pp. 145–156. ISSN: 1438-2199. DOI: 10.1007/s00726-006-0443-6.
- [239] Tony Hunter. 'A journey from phosphotyrosine to phosphohistidine and beyond'. English. In: *Molecular Cell* 82.12 (June 2022), pp. 2190–2200. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2022.05.007.
- [240] Jung-Min Kee and Tom W. Muir. 'Chasing Phosphohistidine, an Elusive Sibling in the Phosphoamino Acid Family'. In: *ACS Chemical Biology* 7.1 (Jan. 2012), pp. 44–51. ISSN: 1554-8929. DOI: 10.1021/cb200445w.
- [241] Eric F. Pettersen et al. 'UCSF Chimera—A visualization system for exploratory research and analysis'. en. In: *Journal of Computational Chemistry* 25.13 (2004). Pp. 1605–1612. ISSN: 1096-987X. DOI: 10.1002/jcc.20084.
- [242] Berk Hess et al. 'LINCS: A linear constraint solver for molecular simulations'. en. In: *Journal of Computational Chemistry* 18.12 (1997). Pp. 1463–1472. ISSN: 1096-987X. DOI: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.
- [243] Tomáš Kubař, Kai Welke and Gerrit Groenhof. 'New QM/MM implementation of the DFTB3 method in the gromacs package'. en. In: *Journal of Computational Chemistry* 36.26 (2015). Pp. 1978–1989. ISSN: 1096-987X. DOI: 10.1002/jcc.24029.
- [244] William Humphrey, Andrew Dalke and Klaus Schulten. 'VMD: Visual molecular dynamics'. In: *Journal of Molecular Graphics* 14.1 (Feb. 1996), pp. 33–38. ISSN: 0263-7855. DOI: 10.1016/0263-7855(96)00018-5.
- [245] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, Feb. 2009. ISBN: 978-1-4414-1269-0.
- [246] Charles R. Harris et al. 'Array programming with NumPy'. en. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2.
- [247] The pandas development team. *pandas-dev/pandas: Pandas*. Jan. 2026. DOI: 10.5281/zenodo.18328522.
- [248] Wes McKinney. 'Data Structures for Statistical Computing in Python'. en. In: *SciPy 2010* (May 2010). DOI: 10.25080/Majora-92bf1922-00a.
- [249] John D. Hunter. 'Matplotlib: A 2D Graphics Environment'. In: *Computing in Science & Engineering* 9.3 (May 2007), pp. 90–95. ISSN: 1558-366X. DOI: 10.1109/MCSE.2007.55.

- [250] Igor Leontyev and Alexei Stuchebrukhov. 'Accounting for electronic polarization in non-polarizable force fields'. en. In: *Physical Chemistry Chemical Physics* 13.7 (Feb. 2011), pp. 2613–2626. ISSN: 1463-9084. DOI: 10.1039/C0CP01971B.
- [251] Brian J. Kirby and Pavel Jungwirth. 'Charge Scaling Manifesto: A Way of Reconciling the Inherently Macroscopic and Microscopic Natures of Molecular Simulations'. In: *The Journal of Physical Chemistry Letters* 10.23 (Dec. 2019), pp. 7531–7536. DOI: 10.1021/acs.jpclett.9b02652.
- [252] Franco Marsico et al. 'Multiscale approach to the activation and phosphotransfer mechanism of CpxA histidine kinase reveals a tight coupling between conformational and chemical steps'. In: *Biochemical and Biophysical Research Communications. Multiscale Modeling* 498.2 (Mar. 2018), pp. 305–312. ISSN: 0006-291X. DOI: 10.1016/j.bbrc.2017.09.039.
- [253] Valerie A. Clausen et al. 'Biochemical Characterization of the First Essential Two-Component Signal Transduction System from *Staphylococcus aureus* and *Streptococcus pneumoniae*'. In: *Journal of Molecular Microbiology and Biotechnology* 5.4 (July 2003), pp. 252–260. ISSN: 1464-1801. DOI: 10.1159/000071077.
- [254] Alberto Marina et al. 'Structural and Mutational Analysis of the PhoQ Histidine Kinase Catalytic Domain: INSIGHT INTO THE REACTION MECHANISM*'. In: *Journal of Biological Chemistry* 276.44 (Nov. 2001), pp. 41182–41190. ISSN: 0021-9258. DOI: 10.1074/jbc.M106080200.
- [255] Reha Celikel et al. 'ATP forms a stable complex with the essential histidine kinase WalK (YycG) domain'. In: *Acta Crystallographica Section D: Biological Crystallography* 68.Pt 7 (July 2012), pp. 839–845. ISSN: 0907-4449. DOI: 10.1107/S090744491201373X.
- [256] Joachim Weber et al. 'Mg²⁺ Coordination in Catalytic Sites of F1-ATPase'. In: *Biochemistry* 37.2 (Jan. 1998), pp. 608–614. ISSN: 0006-2960. DOI: 10.1021/bi972370e.
- [257] Derek R. Case, Jon Zubieta and Robert P. Doyle. 'The Coordination Chemistry of Bio-Relevant Ligands and Their Magnesium Complexes'. en. In: *Molecules* 25.14 (Jan. 2020), p. 3172. ISSN: 1420-3049. DOI: 10.3390/molecules25143172.
- [258] Guanhua Hou et al. 'A Modified QM/MM Hamiltonian with the Self-Consistent-Charge Density-Functional-Tight-Binding Theory for Highly Charged QM Regions'. In: *Journal of Chemical Theory and Computation* 8.11 (Nov. 2012), pp. 4293–4304. ISSN: 1549-9618. DOI: 10.1021/ct300649f.
- [259] Michael J. McLeish and George L. Kenyon. 'Relating Structure to Mechanism in Creatine Kinase'. In: *Critical Reviews in Biochemistry and Molecular Biology* 40.1 (Jan. 2005). Pp. 1–20. ISSN: 1040-9238. DOI: 10.1080/10409230590918577.
- [260] T. Wallimann et al. 'Intracellular compartmentation, structure and function of creatine kinase isoenzymes in tissues with high and fluctuating energy demands: the 'phosphocreatine circuit' for cellular energy homeostasis'. en. In: *Biochemical Journal* 281.Pt 1 (Jan. 1992), p. 21. DOI: 10.1042/bj2810021.
- [261] Werner Müller-Esterl. *Biochemie*. de. Berlin, Heidelberg: Springer, 2018. ISBN: 978-3-662-54850-9 978-3-662-54851-6. DOI: 10.1007/978-3-662-54851-6.
- [262] Jeffrey F. Ohren et al. 'Structural asymmetry and intersubunit communication in muscle creatine kinase'. en. In: *Acta Crystallographica Section D Biological Crystallography* 63.3 (Mar. 2007), pp. 381–389. ISSN: 0907-4449. DOI: 10.1107/S0907444906056204.
- [263] Lucien Bettendorff and Pierre Wins. 'Energy-Rich Molecules and Group Transfer Potentials in Energetic Coupling Reactions'. In: *Molecules* 31.2 (Jan. 2026), p. 242. ISSN: 1420-3049. DOI: 10.3390/molecules31020242.
- [264] S. M. Mühlebach et al. 'Sequence homology and structure predictions of the creatine kinase isoenzymes'. en. In: *Molecular and Cellular Biochemistry* 133.1 (Apr. 1994), pp. 245–262. ISSN: 1573-4919. DOI: 10.1007/BF01267958.
- [265] Uwe Schlattner, Malgorzata Tokarska-Schlattner and Theo Wallimann. 'Mitochondrial creatine kinase in human health and disease'. In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. Mitochondria in Diseases and Therapeutics* 1762.2 (Feb. 2006), pp. 164–180. ISSN: 0925-4439. DOI: 10.1016/j.bbadis.2005.09.004.
- [266] Lucas Guimarães-Ferreira. 'Role of the phosphocreatine system on energetic homeostasis in skeletal and cardiac muscles'. In: *Einstein* 12.1 (2014), pp. 126–131. ISSN: 1679-4508. DOI: 10.1590/S1679-45082014RB2741.

- [267] Sang Young Kim et al. 'Cardiac biomarkers and detection methods for myocardial infarction'. en. In: *Molecular & Cellular Toxicology* 18.4 (Oct. 2022), pp. 443–455. ISSN: 2092-8467. DOI: 10.1007/s13273-022-00287-1.
- [268] Marianne F. Baird et al. 'Creatine-Kinase- and Exercise-Related Muscle Damage Implications for Muscle Performance and Recovery'. en. In: *Journal of Nutrition and Metabolism* 2012.1 (2012). P. 960363. ISSN: 2090-0732. DOI: 10.1155/2012/960363.
- [269] Andrea Barp et al. 'Circulating Biomarkers in Neuromuscular Disorders: What Is Known, What Is New'. en. In: *Biomolecules* 11.8 (Aug. 2021), p. 1246. ISSN: 2218-273X. DOI: 10.3390/biom11081246.
- [270] Martin Kornacker et al. 'Hodgkin disease-derived cell lines expressing ubiquitous mitochondrial creatine kinase show growth inhibition by cyclocreatine treatment independent of apoptosis'. en. In: *International Journal of Cancer* 94.4 (2001). Pp. 513–519. ISSN: 1097-0215. DOI: 10.1002/ijc.1502.
- [271] Kiran Kurmi et al. 'Tyrosine Phosphorylation of Mitochondrial Creatine Kinase 1 Enhances a Druggable Tumor Energy Shuttle Pathway'. In: *Cell Metabolism* 28.6 (Dec. 2018), 833–847.e8. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2018.08.008.
- [272] Xiao-Long Qian et al. 'Overexpression of ubiquitous mitochondrial creatine kinase (uMtCK) accelerates tumor growth by inhibiting apoptosis of breast cancer cells and is associated with a poor prognosis in breast cancer patients'. In: *Biochemical and Biophysical Research Communications* 427.1 (Oct. 2012), pp. 60–66. ISSN: 0006-291X. DOI: 10.1016/j.bbrc.2012.08.147.
- [273] G. Lowe and B. S. Sproat. 'Evidence for an associative mechanism in the phosphoryl transfer step catalyzed by rabbit muscle creatine kinase.' In: *Journal of Biological Chemistry* 255.9 (May 1980), pp. 3944–3951. ISSN: 0021-9258. DOI: 10.1016/S0021-9258(19)85617-X.
- [274] Pan-Fen Wang et al. 'An Unusually Low pKa for Cys282 in the Active Site of Human Muscle Creatine Kinase'. In: *Biochemistry* 40.39 (Oct. 2001), pp. 11698–11705. ISSN: 0006-2960. DOI: 10.1021/bi011208f.
- [275] Hai-Meng Zhou and Chen-Lu Tsou. 'The presence of reactive SH groups in the enzymatically active dicyano derivative of creatine kinase'. In: *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* 911.2 (Jan. 1987), pp. 136–143. ISSN: 0167-4838. DOI: 10.1016/0167-4838(87)90002-1.
- [276] Hong Wu, Qi-Zhi Yao and Chen-Lu Tsou. 'Creatine kinase is modified by 2-chloromercuri-4-nitrophenol at the active site thiols with complete inactivation'. In: *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* 997.1 (July 1989), pp. 78–82. ISSN: 0167-4838. DOI: 10.1016/0167-4838(89)90137-4.
- [277] M. Lijun Lin et al. 'Determination of the catalytic site of creatine kinase by site-directed mutagenesis'. In: *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* 1206.1 (May 1994), pp. 97–104. ISSN: 0167-4838. DOI: 10.1016/0167-4838(94)90077-9.
- [278] Rolf Furter, Elizabeth M. Furter-Graves and Theo Wallimann. 'Creatine kinase: The reactive cysteine is required for synergism but is nonessential for catalysis'. In: *Biochemistry* 32.27 (July 1993), pp. 7022–7029. ISSN: 0006-2960. DOI: 10.1021/bi00078a030.
- [279] Michael Eder et al. 'Crystal structure of human ubiquitous mitochondrial creatine kinase'. en. In: *Proteins: Structure, Function, and Bioinformatics* 39.3 (2000). Pp. 216–225. ISSN: 1097-0134. DOI: 10.1002/(SICI)1097-0134(20000515)39:3<216::AID-PROT40>3.0.CO;2-#.
- [280] Nicolas Guex and Manuel C. Peitsch. 'SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling'. en. In: *ELECTROPHORESIS* 18.15 (1997). Pp. 2714–2723. ISSN: 1522-2683. DOI: 10.1002/elps.1150181505.
- [281] Junmei Wang et al. 'Development and testing of a general amber force field'. en. In: *Journal of Computational Chemistry* 25.9 (2004). Pp. 1157–1174. ISSN: 1096-987X. DOI: 10.1002/jcc.20035.
- [282] Kristin L. Meagher, Luke T. Redman and Heather A. Carlson. 'Development of polyphosphate parameters for use with the AMBER force field'. en. In: *Journal of Computational Chemistry* 24.9 (2003). Pp. 1016–1025. ISSN: 1096-987X. DOI: 10.1002/jcc.10262.
- [283] P. H. König et al. 'Toward Theoretical Analysis of Long-Range Proton Transfer Kinetics in Biomolecular Pumps'. In: *The Journal of Physical Chemistry A* 110.2 (Jan. 2006), pp. 548–563. ISSN: 1089-5639. DOI: 10.1021/jp052328q.

- [284] Pan-Fen Wang et al. 'Loop Movement and Catalysis in Creatine Kinase'. en. In: *IUBMB Life* 57.4-5 (2005). Pp. 355–362. ISSN: 1521-6551. DOI: 10.1080/15216540500091999.
- [285] Charles L. Perrin and Jennifer B. Nielson. "'Strong" hydrogen bonds in chemistry and biology'. en. In: *Annual Review of Physical Chemistry* 48. Volume 48, 1997 (Oct. 1997), pp. 511–544. ISSN: 0066-426X, 1545-1593. DOI: 10.1146/annurev.physchem.48.1.511.
- [286] Demian Riccardi et al. "'Proton Holes" in Long-Range Proton Transfer Reactions in Solution and Enzymes: A Theoretical Analysis'. In: *Journal of the American Chemical Society* 128.50 (Dec. 2006), pp. 16302–16311. ISSN: 0002-7863. DOI: 10.1021/ja065451j.
- [287] Denis Maag et al. 'O to bR transition in bacteriorhodopsin occurs through a proton hole mechanism'. In: *Proceedings of the National Academy of Sciences* 118.39 (Sept. 2021), e2024803118. DOI: 10.1073/pnas.2024803118.
- [288] Pan-Fen Wang et al. 'Exploring the Role of the Active Site Cysteine in Human Muscle Creatine Kinase'. In: *Biochemistry* 45.38 (Sept. 2006), pp. 11464–11472. ISSN: 0006-2960. DOI: 10.1021/bi0607002.
- [289] Ana Maria Teixeira and Grasiely F Borges. 'Creatine Kinase: structure and function'. en. In: *Brazilian Journal of Biomotricity* 6.2 (Jan. 2012), pp. 53–65.
- [290] Thomas K. Harris and George J. Turner. 'Structural Basis of Perturbed pKa Values of Catalytic Groups in Enzyme Active Sites'. en. In: *IUBMB Life* 53.2 (2002). Pp. 85–98. ISSN: 1521-6551. DOI: 10.1080/15216540211468.

List of Publications

- [1] Mayukh Kansari, Lena Eichinger and Tomáš Kubař. ‘Extended-sampling QM/MM simulation of biochemical reactions involving P–N bonds’. In: *Physical Chemistry Chemical Physics* 25.14 (2023), pp. 9824–9836. DOI: 10.1039/D2CP05890A.