

# *Ab Initio* Molecular Dynamics Simulations of Phosphate Hydrolysis Using Neural Network Potentials

**Albert MAKHMUDOV**

Supervisor: Prof. J. Harvey  
KU Leuven

Thesis presented in  
fulfillment of the requirements  
for the degree of Master of Science  
in Theoretical Chemistry and Computational Modelling

Academic year 2024-2025

**© Copyright by KU Leuven**

Without written permission of the promtors and the authors it is forbidden to reproduce or adapt in any form or by any means any part of this publication. Requests for obtaining the right to reproduce or utilize parts of this publication should be addressed to KU Leuven, Faculteit Wetenschappen, Celestijnenlaan 200H bus 2100, 3001 Leuven (Heverlee), telephone +32 16 32 14 01.

A written permission of the promotor is also required to use the methods, products, schematics and programs described in this work for industrial or commercial use, and for submitting this publication in scientific contests.

This thesis is an exam document that obtained no further correction of possible errors after the defense. Referring to this thesis in papers and analogous documents is only allowed after written consent of the supervisor(s), mentioned on the title page.

# Foreword

The TCCM master's programme - and this thesis project in particular - provided me with the opportunity to open a completely new chapter of my life. Not only did it allow me to study at a top-notch university, but, more importantly, it gave me the chance to be surrounded by brilliant people who inspired me to think outside the box and to learn many new things, not merely in a professional sense, but also on a personal level.

I'm wholeheartedly grateful to my supervisor, Prof. Jeremy Harvey, who agreed to take me on from the time of my internship in the first year and continued to support me throughout my master's thesis. His constant guidance and encouragement made this project possible. However, most significantly, our discussions and brainstorming sessions had the greatest impact on me. They helped me to think critically about my research topic and to consider it from different perspectives.

I'd like to thank my family in Uzbekistan for always being there for me and supporting me within the course of my studies. Their love and support helped me get through all the ups and downs of living in a completely new environment. Большое спасибо моей семье: маме, бабушке, папе и младшему брату за веру в меня. Люблю вас всех и крепко обнимаю! Без вашей поддержки у меня бы точно не получилось начать новую жизнь в совершенно другой части мира.

Student life is not only about studying, but also about making friends and having fun. In this regard, I would like to thank those I hold dear: my friends, all group and division members, as well as the TCCM cohort, for making these two years an unforgettable experience. Special *dank u wel, muchas gracias, merci beaucoup, and muito obrigado* go to everyone for the coffee breaks we shared - they were not only a great way to unwind, but also a source of eureka moments and ideas.

Last but not least, I would like to express my gratitude to my former supervisors and colleagues in Uzbekistan, who motivated me to embark on an academic journey and to pursue a master's degree abroad. For that, I would like to thank Dr. Artyom Baev, Dr. Anvar Sariev, Prof. Zukhra Kadirova, and Prof. Shahlo Daminova.

This project would not have been possible without the computational resources provided by KU Leuven and the Erasmus+ scholarship from the EMJMD TCCM.

# Contribution statement

Albert Makhmudov proposed revisiting phosphate hydrolysis using a more advanced computational setup. Prof. Jeremy Harvey suggested the investigation of a system containing methyl diphosphate and proposed the use of a neural network potential to describe the reaction mechanism. The overall workflow was designed together by Albert Makhmudov and Prof. Jeremy Harvey, who also jointly analysed the results. All calculations, figures, and visualisations were carried out by Albert Makhmudov. The thesis was written by Albert Makhmudov with feedback and corrections from Prof. Jeremy Harvey. Ehsan Moravveji (VSC, KU Leuven), Hans Vansweevelt (KU Leuven), and Anders Johansson (Harvard) helped with the installing and compiling the software on the high-performance computing clusters.

# **Summary**

# List of abbreviations

**ADP** adenosine diphosphate

**AIMD** *ab initio* molecular dynamics

**ATP** adenosine triphosphate

**BOMD** Born-Oppenheimer molecular dynamics

**CN** coordination number

**CSVR** canonical sampling through velocity-rescaling

**CV** collective variable

**DFT** density functional theory

**DFTB** density functional tight-binding

**DNA** deoxyribonucleic acid

**FES** free energy surface

**GGA** generalised gradient approximation

**GNN** graph neural network

**GPW** Gaussian plane wave method

**GTH** Goedecker-Teter-Hutter pseudopotentials

**Ig** leaving group

**MD** molecular dynamics

**MAE** mean absolute error

**MeDP** methyl diphosphate trianion

**MeHDP** methyl diphosphate dianion

**MeHMP** methyl phosphate monoanion

**MeMP** methyl phosphate dianion

**MeTP** methyl triphosphate tetraniion

**MFEP** minimum free energy path

**MSE** mean squared error

**NNP** neural network potential

**NPT** constant number of particles, pressure and temperature

**Nuc** nucleophile

**NVT** constant number of particles, volume and temperature

**PBC** periodic boundary conditions

**PBE** Perdew-Burke-Ernzerhof exchange-correlation functional

**P<sub>i</sub>** inorganic phosphate

**PME** Particle Mesh Ewald

**polyP** polyphosphate

**PT** proton transfer

**RDF** radial distribution function

**RMSE** root mean square error

**RNA** ribonucleic acid

**RRHO** Rigid-Rotor Harmonic-Oscillator

**SCF** self-consistent field

**TS** transition state

**TZV2P** triple- $\zeta$  valence basis set with two polarisation functions

**WTMetaD** well-tempered metadynamics

**xTB** extended tight-binding

# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>   | <b>1</b>  |
| 1.1      | Role of phosphates in biological systems . . . . .              | 1         |
| 1.2      | Enzymes involved in phosphate hydrolysis . . . . .              | 3         |
| 1.3      | Reaction mechanism: phosphate monoesters . . . . .              | 6         |
| 1.3.1    | Phosphates . . . . .  | 6         |
| 1.3.2    | Diphosphates . . . . .  | 8         |
| 1.3.3    | Triphosphates . . . . .   | 12        |
| 1.4      | Research goals . . . . .  | 12        |
| <b>2</b> | <b>Theory</b>   | <b>14</b> |
| 2.1      | A brief introduction to statistical mechanics . . . . .         | 14        |
| 2.1.1    | Partition functions . . . . .                                   | 14        |
| 2.1.2    | Macroscopic properties and thermodynamic functions . . . . .    | 16        |
| 2.1.3    | The canonical ensemble . . . . .                                | 17        |
| 2.1.4    | Classical forcefields and molecular dynamics . . . . .          | 18        |
| 2.1.5    | Enhanced sampling techniques . . . . .                          | 19        |
| 2.2      | The density functional theory tourist . . . . .                 | 21        |
| 2.2.1    | The Kohn-Sham approach . . . . .                                | 21        |
| 2.2.2    | Generalised gradient approximation and PBE functional . . . . . | 23        |
| 2.2.3    | <i>Ab initio</i> molecular dynamics . . . . .                   | 24        |
| 2.3      | Extended tight binding . . . . .                                | 24        |
| 2.4      | Neural network potentials . . . . .                             | 25        |
| 2.4.1    | Graph neural networks . . . . .                                 | 25        |
| 2.4.2    | Invariance and equivariance . . . . .                           | 27        |
| 2.4.3    | Equivariant graph neural networks . . . . .                     | 28        |
| <b>3</b> | <b>Computational details</b>                                    | <b>32</b> |
| 3.1      | Training dataset generation . . . . .                           | 32        |
| 3.1.1    | System preparation . . . . .                                    | 32        |
| 3.1.2    | Initial equilibration using classical force fields . . . . .    | 33        |

|                                    |  |           |
|------------------------------------|--|-----------|
| 3.1.3                              | Collective variables . . . . .                                     | 34        |
| 3.1.4                              | GFN1-xTB based exploration of the configuration space . . . . .    | 36        |
| 3.1.5                              | Data labeling . . . . .  | 36        |
| 3.1.6                              | Iterative training of the neural network potential . . . . .       | 37        |
| 3.2                                | Production runs at ambient temperature . . . . .                   | 41        |
| 3.3                                | Data analysis and visualisation . . . . .                          | 41        |
| <b>4</b>                           | <b>Results and discussion</b>                                      | <b>42</b> |
| 4.1                                | Final dataset composition . . . . .                                | 42        |
| 4.2                                | Accuracy and performance of the neural network potential . . . . . | 43        |
| 4.3                                | Stability of the production runs . . . . .                         | 47        |
| 4.4                                | Radial distribution function of water . . . . .                    | 48        |
| 4.5                                | Convergence of the free energy profiles . . . . .                  | 49        |
| 4.6                                | Evolution of the collective variables over time . . . . .          | 51        |
| 4.7                                | Reaction mechanism for methyl diphosphate trianion . . . . .       | 52        |
| 4.7.1                              | Minimum free energy path . . . . .                                 | 52        |
| 4.7.2                              | Proton transfer mechanism . . . . .                                | 52        |
| 4.8                                | Reaction mechanism for methyl diphosphate dianion . . . . .        | 53        |
| 4.8.1                              | Minimum free energy path . . . . .                                 | 53        |
| 4.8.2                              | Proton transfer mechanism . . . . .                                | 53        |
| <b>5</b>                           | <b>Conclusions and outlook</b>                                     | <b>54</b> |
| <b>Bibliography</b>                |  | <b>55</b> |
| <b>A Supplementary information</b> |  | <b>61</b> |

# Chapter 1

## Introduction

This chapter provides an overview of the role of phosphates in biological systems, the enzymes involved in phosphate hydrolysis, and a detailed explanation of the reaction mechanisms associated with these processes - topics that have puzzled researchers for a long time. The discussion begins with the fundamental importance of phosphates in life, particularly in energy transfer and storage. This is followed by a brief overview of the enzymes that catalyse phosphate hydrolysis and their implications for cellular function. Finally, the chapter explores the reaction mechanisms of phosphate hydrolysis, highlighting key studies and findings in this area.

### 1.1 Role of phosphates in biological systems

Phosphates are among the fundamental building blocks that play a central role in life on Earth. They form the basis for both the storage and transfer of genetic information, as well as the flow of metabolic energy within biological systems. The ubiquitous nature of phosphate esters and anhydrides - such as those found in deoxyribonucleic acid (DNA), ribonucleic acid (RNA), adenosine triphosphate (ATP), and polyphosphate (polyP) - highlights their fundamental importance [1]. Some of the phosphates found in biological systems and their respective functions are summarised in Table 1.1.

A key characteristic enabling these roles is the ability of phosphoric acid to link molecular units while retaining an ionisable group. This inherent negative charge at physiological pH serves a dual purpose: it helps to retain these molecules within cellular boundaries defined by lipid membranes, and more importantly, it confers kinetic stability upon phosphate esters and anhydrides by electrostatically repelling nucleophilic attack, particularly from water [1]. For instance, the half-time for hydrolysis at 25 °C for a phosphomonoester monoanion (P-O) is about 90 years; however, for a phosphodiester anion (P-O), this number increases dramatically to approximately 16 million years

[2]. This stability is crucial for maintaining the integrity of genetic material but can be readily overcome by enzymatic catalysis when there is a metabolic demand.

Phosphates are involved in numerous processes in living systems, such as cell signalling and sensation, regulation of metabolism, blood coagulation, and bone formation [3, 4]. Their role is perhaps most evident in cellular energetics, where ATP functions as the universal energy currency. The energy derived from nutrients such as glucose is captured and stored within the high-energy phosphoanhydride bonds linking the phosphate groups of ATP. This energy is released upon hydrolysis of the terminal phosphoanhydride bond (P-O bond between  $\beta$  and  $\gamma$  in Figure 1.1), typically yielding adenosine diphosphate (ADP) and inorganic phosphate ( $P_i$ ). The cleavage of this bond provides the thermodynamic driving force for the majority of cellular processes, including biosynthesis, active transport, and mechanical work such as muscle contraction. The standard free energy change for ATP hydrolysis is substantial ( $\Delta G^0 = -30.5 \text{ kJ mol}^{-1}$ ), and under cellular conditions, the actual free energy release is often considerably greater. Specifically, the experimentally obtained  $\Delta G$  values are approximately  $-59$  to  $-53.5 \text{ kJ mol}^{-1}$  in the liver and about  $-61.7$  to  $-59.5 \text{ kJ mol}^{-1}$  in the heart [3].

Beyond ATP, polyP - a linear polymer of orthophosphate residues linked by similar high-energy phosphoanhydride bonds - represents another significant phosphate-based energy storage found across all domains of life, including mammalian cells. However, in mammalian cells, the concentration of polyP is significantly lower compared to that in microorganisms. While its roles in mammals are still being fully elucidated,

| Phosphate                         | Biological role                     |
|-----------------------------------|-------------------------------------|
| DNA/RNA                           | Genetic material                    |
| ADP/ATP                           | Intracellular energy transfer       |
| cAMP                              | Cellular signalling                 |
| Polyphosphate                     | Energy storage, Cellular signalling |
| Creatine phosphate                | Intracellular energy transfer       |
| Phosphoenolpyruvate               | Metabolism                          |
| Pyridoxal phosphate               | Coenzyme                            |
| Nicotinamide adenine dinucleotide | Calcium signalling                  |
| Fructose 1,6-diphosphate          | Metabolism                          |
| Glucose-6-phosphate               | Metabolism                          |
| Isopentenyl pyrophosphate         | Metabolism                          |
| Ribose-6-phosphate                | Metabolism                          |
| Glycerol 3-phosphate              | Metabolism                          |
| Dihydroxyacetone phosphate        | Calvin cycle, metabolism            |
| Inositol phosphates               | Cellular signalling                 |

Table 1.1: Examples of biologically relevant phosphates and their roles. Reproduced and adapted from [5].

dated, polyP metabolism is intrinsically linked to the cellular energy status. Mitochondrial polyP levels fluctuate with respiratory activity and appear to depend on  $F_0F_1$ -ATP synthase function, suggesting a role in mitochondrial bioenergetics, potentially acting as an energy reservoir [6].

The efficient transfer of energy stored in phosphate bonds from sites of production (e.g., mitochondria) to sites of utilisation (e.g., ATPases involved in muscle contraction or ion transport) is crucial. Simple diffusion of ATP is often insufficient due to the complexity of intracellular architecture and the potential for large concentration gradients to arise, which would be thermodynamically inefficient. Instead, cells employ phosphotransfer networks, utilising enzymes such as creatine kinase and adenylate kinase, which catalyse phosphoryl exchange reactions. These networks act as 'phosphoryl wires', facilitating the efficient conduction of high-energy phosphoryl groups and energetic signals throughout the cell with minimal energy dissipation or accumulation of inhibitory products such as ADP. The existence of these networks underscores the dynamic and highly organised nature of cellular energy management, where phosphates - mainly in the form of ATP - serve as the key energy carriers [7].

The synthesis of ATP occurs primarily through oxidative phosphorylation in mitochondria, a process tightly coupled to the electron transport chain, which establishes a proton-motive force ( $\Delta p$ ) across the inner mitochondrial membrane. This electrochemical potential energy is used by the molecular machine ATP synthase. Interestingly, the principal energy input required by ATP synthase is not for the chemical formation of the phosphoanhydride bond itself, but rather for the conformational changes necessary to release the newly synthesised, tightly bound ATP molecule from the enzyme's catalytic site. This 'binding change mechanism' involves the cooperative, sequential action of the enzyme's multiple catalytic sites, driven by proton flow. The hydrolysis of ATP to ADP and  $P_i$  is catalysed by a variety of enzymes, including ATPases and possibly  $F_1$ -ATPase, which are frequently coupled to other cellular processes [8].

In summary, the unique chemical properties of phosphates - their ability to form stable esters and energy-rich anhydrides, along with their negative charge - combined with the evolution of sophisticated enzymatic machinery for their synthesis, transfer, and hydrolysis, have secured their vital role in virtually all life processes.

## 1.2 Enzymes involved in phosphate hydrolysis

The hydrolysis of high-energy phosphoanhydride bonds, particularly the terminal bond in ATP, is a cornerstone of cellular bioenergetics. While numerous enzymes utilise ATP hydrolysis, the  $F_0F_1$ -ATP synthase complex, primarily known for synthesising ATP,

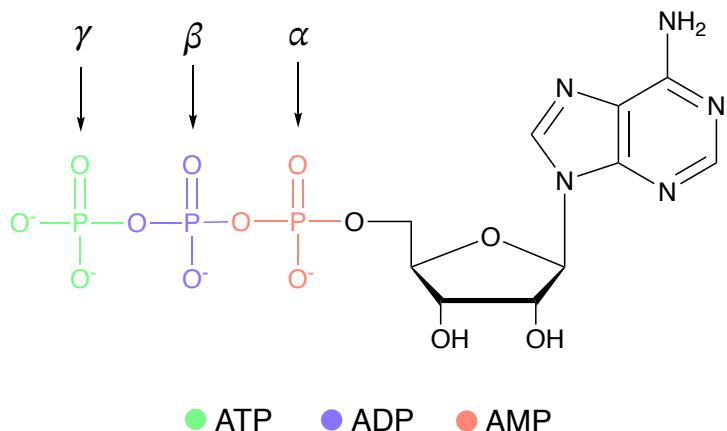


Figure 1.1: Chemical structures of the AMP, ADP, and ATP molecules with the phosphates marked as  $\alpha$ ,  $\beta$ , and  $\gamma$ , respectively.

also shows potential ATP hydrolytic activity, particularly through its  $F_1$  component ( $F_1$ -ATPase). This enzyme complex, therefore, plays a dual role in managing the cell's primary energy currency [8–10]. Furthermore, recent evidence suggests that this complex may also participate in the metabolism, including the hydrolysis, of polyP in mammalian cells [11, 12].

The  $F_0F_1$ -ATP synthase is a molecular motor embedded in the mitochondrial membrane. It consists of two major domains: the  $F_1$  domain, which carries the catalytic sites, and the  $F_0$  domain, which is embedded within the membrane. These domains are connected by a central rotor stalk and a peripheral stator stalk [10, 13, 14]. The activity of this enzyme is coupled with the electron-transport chain, as illustrated in Figure 1.2.

The  $F_1$  domain ( $\alpha_3\beta_3\gamma\delta\epsilon$  stoichiometry) extends into the mitochondrial matrix. It has a globular shape as can be seen in Figure 1.2. The catalytic sites for ATP synthesis and hydrolysis are located on the three  $\beta$  subunits, which interact with the  $\alpha$  subunits. When functioning in reverse, the  $F_1$  domain acts as an  $F_1$ -ATPase, hydrolysing ATP. This hydrolysis drives the counterclockwise rotation (as viewed from the membrane) of the central stalk, composed of the  $\gamma$ ,  $\delta$ , and  $\epsilon$  subunits [8, 10, 13]. If coupled to the  $F_0$  domain, this rotation actively pumps protons from the matrix, thereby generating or maintaining the proton-motive force ( $\Delta p$ ). This reverse function is especially important under conditions of low  $\Delta p$ , where it helps prevent its complete dissipation at the expense of cellular ATP and possibly polyP [9, 10, 12].

The mechanism of ATP hydrolysis (cleavage of the P-O bond between  $\beta$  and  $\gamma$  in Figure 1.1) follows the principles of the binding change mechanism [10]. The rotation of the asymmetric  $\gamma$  subunit induces sequential conformational changes in the three  $\beta$  subunits, cycling them through states analogous to those in synthesis: an 'open' state that binds ATP, a 'tight' state that facilitates hydrolysis, and a subsequent 'open' state

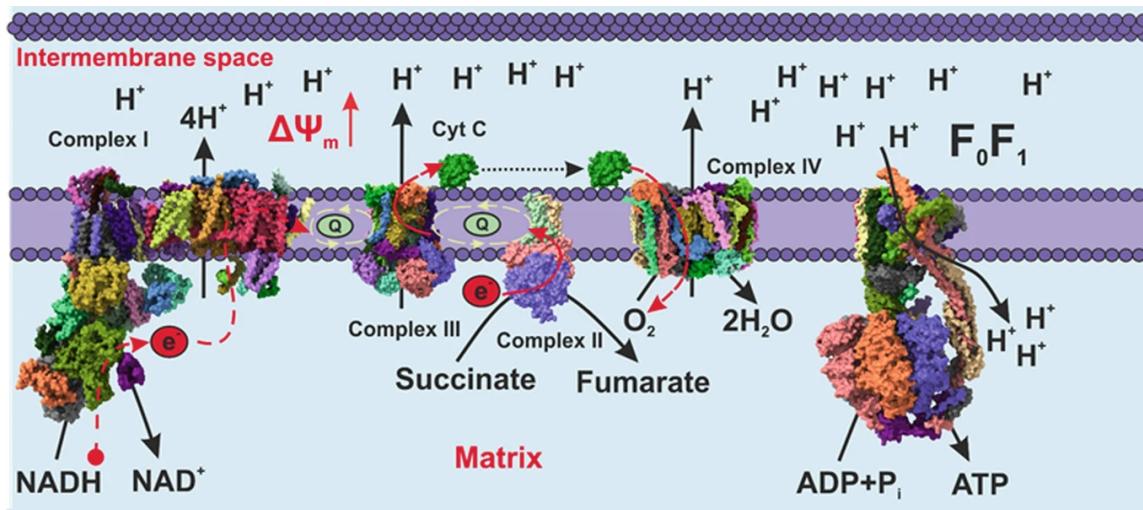


Figure 1.2: Electron transport chain coupled with oxidative phosphorylation in mitochondria. This figure was taken from [11].

that releases ADP and P<sub>i</sub> [8, 13]. The hydrolysis of each ATP molecule is associated with a 120° rotation of the central stalk, which occurs in substeps [10].

While the metabolism of inorganic polyP is well-characterised in microorganisms via specific kinases (PPK) and phosphatases (PPX), the enzymes responsible for its turnover in mammalian cells remain largely unknown. Recent studies using immuno-captured F<sub>0</sub>F<sub>1</sub>-ATPase have demonstrated that the enzyme complex can hydrolyse polyP. This polyP hydrolysis appears to drive the enzyme's proton-pumping activity, akin to ATP hydrolysis, and is sensitive to oligomycin, a specific F<sub>0</sub>F<sub>1</sub>-ATP synthase inhibitor. Medium- and long-chain polyP molecules, made of 60 and 130 orthophosphate units, respectively, seem to be effective substrates for this hydrolytic activity. Docking simulations support the feasibility of polyP binding to the nucleotide-binding sites within the F<sub>1</sub> domain. This suggests that polyP could serve as an alternative energy source for the F<sub>0</sub>F<sub>1</sub> complex, potentially helping to maintain mitochondrial membrane potential when ATP levels are compromised [11, 12].

There is growing confidence that the F<sub>1</sub>-ATPase could act not only as an ATP hydrolase but also potentially as a polyP hydrolase. However, other enzymes contribute to phosphate metabolism as well. In the context of polyP, mammalian enzymes such as alkaline phosphatase (ALP) have demonstrated exopolyphosphatase activity, capable of degrading polyP chains of various lengths [11].

The world of enzymes - and phosphate hydrolysis by F<sub>1</sub>-ATPase in particular - is both fascinating and complex. The F<sub>1</sub>-ATPase is a molecular machine capable of hydrolysing ATP and polyP, yet the precise mechanism of hydrolysis remains not well understood. In order to address this gap, it is necessary to investigate the fundamental reaction mechanisms of phosphate hydrolysis, beginning with the simplest phosphate

esters in less complex environments such as bulk water.

## 1.3 Reaction mechanism: phosphate monoesters

Computational and experimental studies have provided significant insights into the mechanisms of phosphate hydrolysis reactions. Various systems and methodologies have been employed to explore the details of these fundamental biological processes. The debate often centres on whether the reaction proceeds via an associative mechanism (bond formation precedes bond breaking) or a dissociative mechanism (bond breaking precedes bond formation), and the nature of the proton transfer.

### 1.3.1 Phosphates

Starting from the simplest possible system, it has been shown that the hydrolysis of methyl phosphate dianion ( $\text{MeMP}^{\cdot\cdot-}$ ) in water can proceed via either associative or concerted mechanisms [5, 15–17]. A schematic representation of these mechanisms is presented in Figure 1.3, which illustrates the More O’Ferrall-Jencks (MFJ) diagram. The MFJ plot is a useful two-dimensional graphical representation of multidimensional free energy surfaces.

The associative mechanism may proceed in two ways: stepwise ( $\text{A}_N + \text{D}_N$ , where  $\text{A}_N$  stands for nucleophilic addition and  $\text{D}_N$  for nucleophilic departure) or concerted ( $\text{A}_N\text{D}_N$ ). The stepwise mechanism involves two transition states and an intermediate. In contrast, the concerted mechanism proceeds through a single transition state without the formation of intermediates [17].

In the case of the associative/stepwise mechanism ( $\text{A}_N + \text{D}_N$ , Figure 1.4), the nucleophile (Nuc) approaches the phosphorus atom while the leaving group (lg) is still attached. Upon the nucleophile’s approach, a concerted proton transfer (PT) occurs to one of the non-bridging oxygens. The reaction proceeds through a compact pentaco-ordinated transition state with a trigonal bipyramidal geometry, followed by a compact intermediate and the elimination of the leaving group in a subsequent transition state.

Regarding the associative/concerted mechanism ( $\text{A}_N\text{D}_N$ ), it proceeds in a manner quite similar to the first step of the associative/stepwise pathway. The reaction also involves a compact transition state in which bond formation and bond cleavage occur simultaneously.

It has been shown that the protonation state of methyl phosphate lowers the overall barrier height of the rate-limiting step; however, it does not alter the reaction mechanism [18]. For the methyl phosphate monoanion ( $\text{MeHMP}^{\cdot+}$ ), the calculated barrier

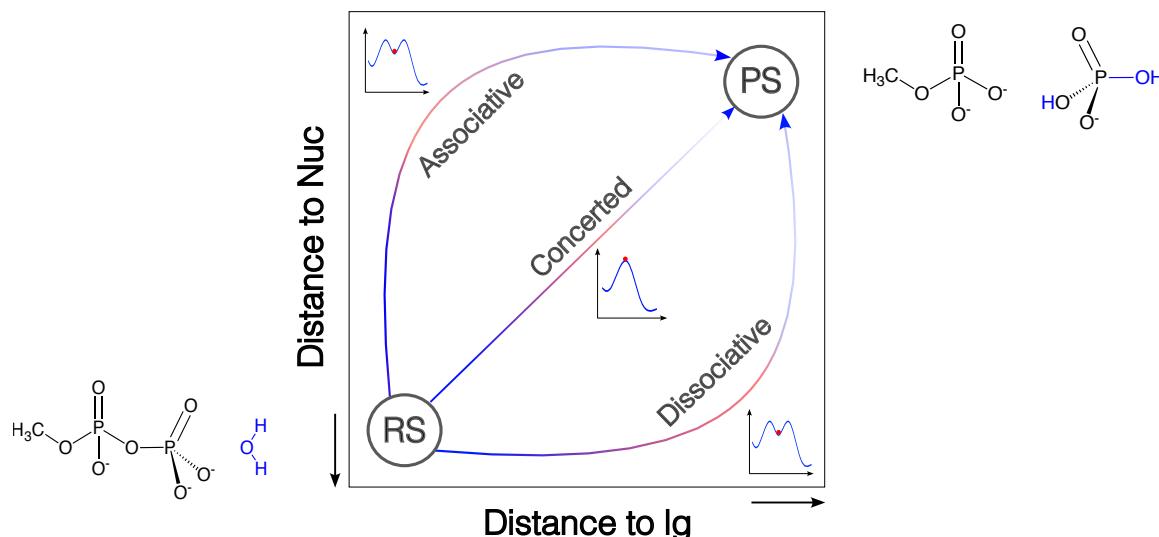


Figure 1.3: More O'Ferrall-Jencks (MFJ) plot of the possible reaction mechanisms for phosphate hydrolysis. The plot shows the free energy as a function of two reaction coordinates: the distance between phosphorus and the nucleophile (Nuc), and the distance between the leaving group (lg) and the phosphorus atom. RS stands for reactant state, PS for product state.

height  $\Delta G_{\text{calc}}^\ddagger$  is approximately 6-7 kcal/mol lower than that of the MeMP. A similar effect was observed when  $\text{OH}^-$  acted as a nucleophile instead of a water molecule [16] (40 vs 47 kcal/mol, respectively). The latter fact arises a question about the proton-transfer in this reaction, for instance, whether it could happen in a concerted or a step-wise manner, in which the PT happens in the pre-equilibration phase.

For the associative mechanism, calculated barrier heights  $\Delta G_{\text{calc}}^\ddagger$  lie in the range of 33.7-47.2 kcal/mol, while experimental values obtained at 25 °C range between 30.6 and 44.3 kcal/mol, depending on the protonation state. Detailed information about the calculated and experimentally determined barrier heights can be found in Table 1.2. Corresponding data on transition state structures and intermediates is summarised in Table 1.3.

The concerted mechanism ( $A_N D_N$ ) is characterised by a single transition state where the nucleophile approaches the phosphorus atom while the leaving group remains attached. The reaction proceeds via a compact pentacoordinated transition state with a trigonal bipyramidal geometry, which is more expansive compared to that of the associative mechanism (Table 1.3). In this transition state, the distance between the phosphorus atom and the nucleophile is approximately 2.06-2.75 Å, while the distance between the leaving group and the phosphorus atom is around 2.61-2.75 Å. The barrier heights for the concerted mechanism are 44.5 and 47 kcal/mol (Table 1.2).

As can be observed, it is rather difficult to clearly distinguish between the associative and concerted mechanisms, and it appears that both may occur in bulk water. Even by looking at the activation entropies of both reaction pathways, it's clear that the

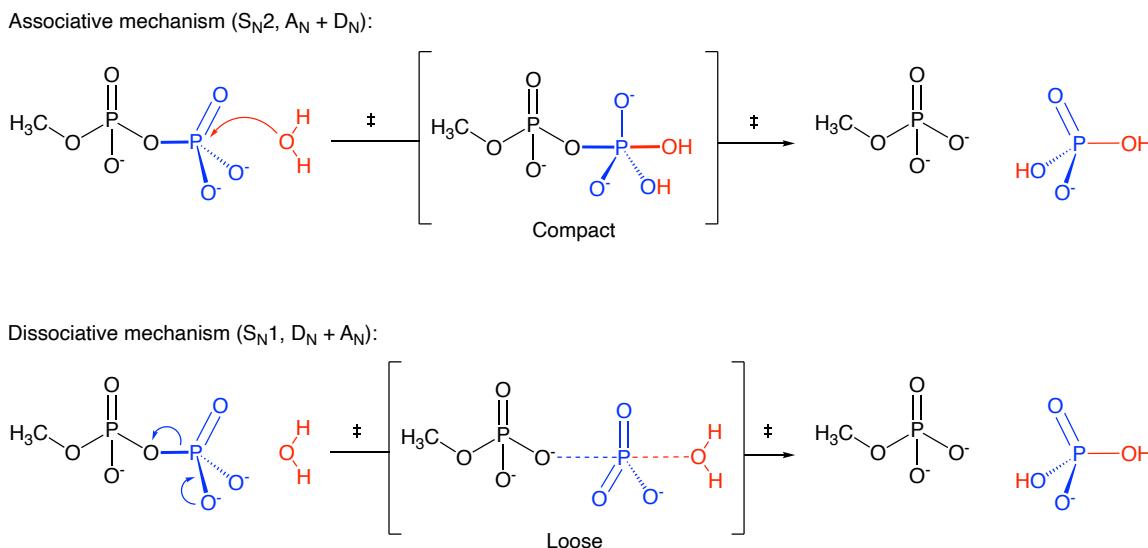


Figure 1.4: Associative and dissociative stepwise reaction mechanisms. The nucleophile (Nuc) is shown in red, the leaving group (lg) in black, and the phosphoryl group in blue.

corresponding values are similarly small: 0.7 and -1.6 kcal/mol for the associative and concerted mechanisms, respectively [17]. Nevertheless, the dissociative mechanism is unlikely to take place, or at least it has not been observed.

### 1.3.2 Diphosphates

Moving on to more complex systems, the hydrolysis of pyrophosphates (diphosphates) (e.g. methyl diphosphate trianion (MeDP)), which are the main focus of this thesis, has also been thoroughly studied. It has been shown that the reaction mechanism can proceed through either associative or concerted pathways, just as in the case of methyl phosphate. In contrast, the associative mechanism has been demonstrated to be solely concerted. However, there is a twist to this story: the dissociative mechanism has also been observed, in both concerted and stepwise forms [15, 16, 19]. A schematic representation of the mechanisms can be found in Figures 1.3 and 1.4.

In the associative/concerted mechanism, the reaction undergoes the same steps as discussed in Subsection 1.3.1. The transition state has a similarly compact geometry: the distance between the phosphorus atom and the nucleophile is approximately 2.03–2.26 Å, and the distance between the leaving group and the phosphorus atom is around 1.82–2.4 Å, as shown in Table 1.3. However, the calculated barrier heights are slightly lower in comparison to methyl phosphate, ranging from 34 to 38 kcal/mol (Table 1.2).

The concerted pathway is characterised by the same general mechanism as discussed in Subsection 1.3.1. The transition state is more expansive than in the associative mechanism, with the distance between the phosphorus atom and the nucleophile

Table 1.2: Summary of computational and experimental studies on phosphate hydrolysis. In the case of calculated  $\Delta G^\ddagger$ , the rate-limiting step is given. <sup>1</sup>The values were calculated using the transition state theory (TST).

| System  | Method                | Level of theory  | Mechanism                                | $\Delta G^\ddagger$ (kcal/mol) | Ref. |
|---|-----------------------|--|--|--------------------------------|------|
| MeMP <sup>2-</sup> + H <sub>2</sub> O   | DFT                   | B3LYP/6-311+G** and COSMO                              | Associative<br>Concerted                 | 47.2<br>44.5                   | [15] |
| MeMP <sup>2-</sup> + H <sub>2</sub> O   | DFT                   | B3LYP/6-311++G** and COSMO                             | Associative                              | 47                             | [16] |
| MeMP <sup>2-</sup> + 3 H <sub>2</sub> O                                       | DFT                   | M06-2X/6-311+G** and SMD                               | Associative<br>Concerted                 | $\approx$ 36<br>$\approx$ 44   | [17] |
| MeMP <sup>2-</sup> + 4 H <sub>2</sub> O                                       | DFT                   | M06-2X/6-311+G**                                       | Associative                              | $\approx$ 40.8 $\pm$ 1.9       | [18] |
| MeHMP <sup>-</sup> + 4 H <sub>2</sub> O                                       | DFT                   | M06-2X/6-311+G**                                       | Associative                              | $\approx$ 33.7 $\pm$ 1.7       | [18] |
| MeDP <sup>3-</sup> + 2 H <sub>2</sub> O                                       | DFT                   | B3LYP/6-311++G** and PCM                               | Associative<br>Dissociative              | 34.64<br>35.24                 | [19] |
| MeDP <sup>3-</sup> + H <sub>2</sub> O   | DFT                   | B3LYP/6-311+G** and COSMO                              | Associative<br>Dissociative              | 34.8<br>30.3                   | [15] |
| MeDP <sup>3-</sup> + H <sub>2</sub> O   | DFT                   | B3LYP/6-311++G** and COSMO                             | Associative<br>Concerted                 | 38<br>34                       | [16] |
| MeHDP <sup>2-</sup> + H <sub>2</sub> O  | DFT                   | B3LYP/6-311++G** and COSMO                             | Associative<br>Concerted                 | 34<br>31                       | [16] |
| MeDP <sup>3-</sup> + Mg <sup>2+</sup> + 5 H <sub>2</sub> O                    | QM/MM,<br>FEP (EVB)   | B3LYP/6-311++G** and MM                                | Associative<br>Concerted<br>Dissociative | 35<br>34<br>35                 | [16] |
| MeTP <sup>4-</sup> + Mg <sup>2+</sup> + 54 H <sub>2</sub> O                   | CPMD                  | PBE/PW with Troullier-Martins pseudopotentials         | Associative<br>Concerted<br>Dissociative | 39.1<br>35.1<br>36.6           | [20] |
| MeTP <sup>4-</sup> + Mg <sup>2+</sup> + 113 H <sub>2</sub> O                  | BOMD,<br>metadynamics | BLYP/TZV2P with GTH pseudopotentials                   | Associative<br>Concerted                 | 29<br>29-30                    | [21] |
| ATP <sup>4-</sup> + Mg <sup>2+</sup> + 4163 H <sub>2</sub> O<br>+ counterions | QM/MM, NEB            | B3LYP/6-311++G** and MM                                | Concerted                                | 32.5                           | [22] |
| ATP <sup>4-</sup> + Mg <sup>2+</sup> + 1800 H <sub>2</sub> O<br>+ counterions | QM/MM, QM =<br>CPMD   | BLYP/PW with Troullier-Martins pseudopotentials and MM | Associative<br>Dissociative              | 36.2<br>33.4                   | [23] |
| Methyl phosphate dianion  | Exp. at 25°C          | –  | –  | 44.3                           | [2]  |
| Methyl phosphate monoanion  | Exp. at 25°C          | –  | –  | 30.6                           | [2]  |
| Pyrophosphate trianion  | Exp. at 25°C          | –  | –  | 29.2                           | [2]  |
| Pyrophosphate dianion   | Exp. at 25°C          | –  | –  | 27.7                           | [2]  |
| ADP <sup>2-</sup> (or ATP <sup>3-</sup> )                                     | Exp. at 25°C          | –  | –  | 27.5                           | [2]  |
| ATPH <sup>3-</sup> (or ATP <sup>4-</sup> )                                    | Exp. at 70°C          | pH=6.69-7.66   | –  | 24.34-24.78 <sup>1</sup>       | [24] |
| dADPH <sup>2-</sup> (or dADP <sup>3-</sup> )                                  | Exp. at 70°C          | pH=6.82  | –  | 24.25 <sup>1</sup>             | [25] |
| dATPH <sup>3-</sup> (or dATP <sup>4-</sup> )                                  | Exp. at 70°C          | pH=7.00  | –  | 24.50 <sup>1</sup>             | [25] |
| MgATPH <sup>-</sup> (or MgATP <sup>2-</sup> )                                 | Exp. at 70°C          | pH=6.59-7.63   | –  | 24.59-24.64 <sup>1</sup>       | [24] |
| CaATPH <sup>-</sup> (or CaATP <sup>2-</sup> )                                 | Exp. at 70°C          | pH=6.67-7.01   | –  | 25.71-25.72 <sup>1</sup>       | [24] |

being approximately 2.26–2.5 Å, while the distance between the leaving group and the phosphorus atom is around 2.7 Å (Table 1.3). The barrier heights for the concerted mechanism are 31 and 34 kcal/mol (Table 1.2), which is notably lower than in the case of methyl phosphate.

In general, it can be noted that the barrier height is strongly dependent on the  $pK_a$  value of the leaving group. The lower the  $pK_a$ , the lower the  $\Delta G^\ddagger$  ( $pK_a(\text{CH}_3\text{O}^-) = 15.5$  vs  $pK_a(\text{CH}_3\text{PO}_4^{2-}) = 6.3$ ). Not only does a lower  $pK_a$  reduce the barrier height, but it also favours the concerted and dissociative mechanisms [16].

The dissociative mechanism can proceed via both concerted and stepwise routes. The dissociative/concerted pathway is quite similar to the general concerted mechanism. The main difference lies in the synchrony of the transition state: while the general concerted mechanism has a more synchronous transition state, the dissociative/concerted one features a greater distance between the phosphorus atom and the leaving group.

On the other hand, the dissociative/stepwise pathway ( $D_N + A_N$ ) is characterised by the departure of the leaving group from the phosphorus atom before the nucleophile approaches. Thus, there is clearly no bond remaining between the phosphorus and the leaving group. Following the departure of the leaving group, a planar metaphosphate  $\text{PO}_3^-$  is formed, as illustrated in Figure 1.4. The transition state is more expansive compared to that of the associative mechanism, with the distance between the phosphorus atom and the nucleophile being 2.7 Å, and the distance between the leaving group and the phosphorus atom being 3.4 Å (Table 1.3). Consequently, after  $\text{TS}_1$ , the system reaches an intermediate step in which the nucleophile is positioned closer to the metaphosphate, followed by an attack and bond formation in  $\text{TS}_2$ .

The barrier heights for the dissociative mechanism lie in the range of 30.3–35.24 kcal/mol (Table 1.2), which is lower than those for any of the previously mentioned mechanisms. Comparing the calculated and experimentally obtained  $\Delta G^\ddagger$  clearly indicates that the dissociative mechanism is more favourable than the associative and concerted ones. The  $\Delta G_{\text{exp}}^\ddagger$  values for the pyrophosphate trianion and dianion are 29.2 and 27.7 kcal/mol, respectively. The influence of one-water (1W) or two-water (2W) mechanisms has also been explored [19], but the overall barriers remain similar.

The dissociative mechanism is further favoured by the presence of metal ions, such as  $\text{Mg}^{2+}$ , as well as in cases where MeDP is protonated, i.e. methyl diphosphate dianion (MeHDP). Interestingly, in the latter case, the proton always transfers to the leaving group en route to the product state [16].

Even though computational studies suggest that  $\text{Mg}^{2+}$  promotes the dissociative mechanism, experimental data do not support this hypothesis [24], since the  $\Delta G_{\text{exp}}^\ddagger$  obtained at 70 °C shows little to no difference, at least in the case of adenosine triphos-

Table 1.3: Summary of the distances between the phosphorus atom and the nucleophile as well as the leaving group in the transition states and intermediates. All distances are in Å.

| System                                | Mechanism                                       | TS <sub>1</sub>        |                       | Intermediate           |                       | TS <sub>2</sub>        |                       | Ref. |
|---------------------------------------|---|------------------------|-----------------------|------------------------|-----------------------|------------------------|-----------------------|------|
|                                       |   | d(P-O <sub>Nuc</sub> ) | d(P-O <sub>lg</sub> ) | d(P-O <sub>Nuc</sub> ) | d(P-O <sub>lg</sub> ) | d(P-O <sub>Nuc</sub> ) | d(P-O <sub>lg</sub> ) |      |
| MeMP <sup>2-</sup>                    | Associative (A <sub>N</sub> D <sub>N</sub> )    | 2.0                    | 1.8                   | —                      | —                     | —                      | —                     | [16] |
|                                       | Associative (A <sub>N</sub> D <sub>N</sub> )    | 1.9                    | 2.15                  | —                      | —                     | —                      | —                     | [15] |
|                                       | Associative (A <sub>N</sub> + D <sub>N</sub> )  | 2.16                   | 1.71                  | 1.84                   | 1.78                  | 1.71                   | 2.24                  | [17] |
|                                       | Associative (A <sub>N</sub> + D <sub>N</sub> )  | 2.08                   | 1.78                  | 1.99                   | 1.80                  | 1.77                   | 2.52                  | [18] |
|                                       | Concerted (A <sub>N</sub> D <sub>N</sub> )      | 2.75                   | 2.75                  | —                      | —                     | —                      | —                     | [15] |
|                                       | Concerted (A <sub>N</sub> D <sub>N</sub> )      | 2.06                   | 2.61                  | —                      | —                     | —                      | —                     | [17] |
| MeHMP <sup>-</sup>                    | Associative (A <sub>N</sub> + D <sub>N</sub> )  | 2.26                   | 1.66                  | 1.76                   | 1.77                  | 1.68                   | 2.25                  | [18] |
| MeDP <sup>3-</sup>                    | Associative (A <sub>N</sub> D <sub>N</sub> )    | 2.2                    | 2.0                   | —                      | —                     | —                      | —                     | [15] |
|                                       | Associative (A <sub>N</sub> D <sub>N</sub> )    | 2.03                   | 1.88                  | —                      | —                     | —                      | —                     | [16] |
|                                       | Associative (A <sub>N</sub> D <sub>N</sub> )    | 2.2                    | 2.0                   | —                      | —                     | —                      | —                     | [19] |
|                                       | Concerted (A <sub>N</sub> D <sub>N</sub> )      | 2.5                    | 2.7                   | —                      | —                     | —                      | —                     | [16] |
|                                       | Dissociative (A <sub>N</sub> D <sub>N</sub> )   | 2.8                    | 3.25                  | —                      | —                     | —                      | —                     | [15] |
|                                       | Dissociative (D <sub>N</sub> + A <sub>N</sub> ) | 2.7                    | 3.4                   | 2.0                    | 3.75                  | 1.7                    | 3.75                  | [19] |
| MeDP <sup>3-</sup> + Mg <sup>2+</sup> | Associative (A <sub>N</sub> D <sub>N</sub> )    | 2.1                    | 2.4                   | —                      | —                     | —                      | —                     | [16] |
|                                       | Concerted (A <sub>N</sub> D <sub>N</sub> )      | 2.3                    | 2.7                   | —                      | —                     | —                      | —                     | [16] |
|                                       | Dissociative (A <sub>N</sub> D <sub>N</sub> )   | 2.8                    | 3.4                   | —                      | —                     | —                      | —                     | [16] |
| MeHDP <sup>2-</sup>                   | Associative (A <sub>N</sub> D <sub>N</sub> )    | 2.26                   | 1.82                  | —                      | —                     | —                      | —                     | [16] |
|                                       | Concerted (A <sub>N</sub> D <sub>N</sub> )      | 2.26                   | 2.78                  | —                      | —                     | —                      | —                     | [16] |
| MeTP <sup>4-</sup> + Mg <sup>2+</sup> | Associative (A <sub>N</sub> D <sub>N</sub> )    | 1.9                    | 2.0                   | —                      | —                     | —                      | —                     | [20] |
|                                       | Associative (A <sub>N</sub> + D <sub>N</sub> )  | 2.03                   | 3.11                  | 1.95                   | 3.06                  | 1.66                   | 3.26                  | [21] |
|                                       | Concerted (A <sub>N</sub> D <sub>N</sub> )      | 2.5                    | 2.6                   | —                      | —                     | —                      | —                     | [20] |
|                                       | Concerted (A <sub>N</sub> D <sub>N</sub> )      | 2.28                   | 2.69                  | —                      | —                     | —                      | —                     | [21] |
|                                       | Dissociative (A <sub>N</sub> D <sub>N</sub> )   | 3.6                    | 3.5                   | —                      | —                     | —                      | —                     | [20] |
|                                       | ATP <sup>4-</sup> + Mg <sup>2+</sup>            | 1.9                    | 1.9                   | —                      | —                     | —                      | —                     | [23] |
| ATP <sup>4-</sup> + Mg <sup>2+</sup>  | Concerted (A <sub>N</sub> D <sub>N</sub> )      | 2.8                    | 3.2                   | —                      | —                     | —                      | —                     | [22] |
|                                       | Dissociative (A <sub>N</sub> D <sub>N</sub> )   | 3.5                    | 3.5                   | —                      | —                     | —                      | —                     | [22] |

phate (Table 1.2).

### 1.3.3 Triphosphates

Last but not least, let us consider the hydrolysis of triphosphates. These systems more closely resemble the biological environment, particularly the processes associated with energy metabolism.

The hydrolysis of methyl triphosphate tetranion (MeTP) and ATP has been studied using a range of computational methods. It has been shown that the reaction mechanisms share many similarities with those observed in mono- and diphosphates. Specifically, the mechanism may proceed via associative/concerted and associative/stepwise routes, as well as concerted and dissociative/concerted pathways (Table 1.3).

When comparing the calculated and experimentally obtained  $\Delta G^\ddagger$  values, it is difficult to clearly distinguish between the aforementioned mechanisms. The calculated barrier heights span a range from 29 to 39.1 kcal/mol, whereas experimental data suggest that the barrier height for ATP should be around 27.5 kcal/mol, as shown in Table 1.2. The more complex the system becomes, the more factors one must likely take into account.

In summary, computational investigations reveal a nuanced and peculiar picture of phosphate hydrolysis. The preferred mechanism (associative, concerted, or dissociative) and the proton transfer route (1W, 2W, etc.) depend significantly on specific factors such as the pKa of the leaving group, the protonation state, the presence of metal ions like  $Mg^{2+}$ , and the surrounding solvent environment.

## 1.4 Research goals

To properly study the underlying free energy surface of the reaction mechanism, adequate sampling of the configurational space is crucial. To achieve this, various free energy techniques, such as metadynamics, can be employed - provided that the level of theory is sufficient to describe a system of realistic size while allowing results to be obtained within a reasonable timeframe. This is precisely the goal of the present project.

As a model system, methyl diphosphate hydrolysis in water has been chosen. Methyl diphosphate is a relatively simple phosphate ester that can represent more complex molecules like ADP and ATP. The main focus of this thesis is to investigate the reaction mechanism of MeDP hydrolysis in water using well-tempered metadynamics, and to explore the influence of the protonation state and the solvent environment

on the reaction mechanism. To run the metadynamics simulations, a neural network potential will be trained, thus allowing *ab initio* molecular dynamics simulations to be performed at low computational cost.

Specifically, the following goals have been set for this thesis:

- Compose a comprehensive dataset of molecular configurations covering all steps of the methyl diphosphate hydrolysis reaction mechanism.
- Train the NequIP neural network to fit a neural network potential that serves as an engine for the *ab initio* molecular dynamics simulations.
- Assess the accuracy and performance of the neural network potential with respect to network complexity and the size of the training set.
- Perform extensive well-tempered metadynamics simulations to obtain the free energy surface of the methyl diphosphate hydrolysis reaction mechanism.
- Gain insights into the reaction mechanism, i.e. the kinetics and thermodynamics of the reaction, and compare the results with previously published data.
- Gain insights into the proton transfer mechanism.

Taking the above-mentioned goals into account, the present thesis follows the idea of Paul A. M. Dirac [26], who emphasised the need for approximate practical methods in quantum mechanics to explain complex atomic systems without excessive computational demands:

The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. *It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.*

The use of neural network potentials in this thesis aims to achieve precisely that: to provide a practical and efficient approach to studying the reaction mechanism while maintaining a high level of accuracy.

# Chapter 2

## Theory

This chapter provides a short and condensed introduction to the theoretical concepts of the methods used in this work. It begins with a discussion of statistical mechanics and free energy techniques, followed by an overview of density functional theory and *ab initio* molecular dynamics. Finally, the chapter concludes with a brief introduction to graph neural networks and neural network potentials. The aim is to provide a comprehensive background to the methods used, while the technical details and derivations are left to the literature sources cited throughout the text.

### 2.1 A brief introduction to statistical mechanics

The discussion in this section is mostly based on the “Introduction to Computational Chemistry” textbook written by Jensen [27], “Statistical Mechanics: Theory and Molecular Simulation” by Tuckermann [28], and “Understanding Molecular Simulation: From Algorithms to Applications” by Frenkel and Smit [29] unless stated otherwise.

#### 2.1.1 Partition functions

The development of the field of statistical mechanics has been crucial for the computational chemistry community, as it enables the connection between the jigglings and wiggles of atoms and the properties of much larger systems such as liquids and solids.

Let us begin with the most fundamental concept: the partition function. The partition function is akin to a Swiss army knife in statistical mechanics, meaning it is a versatile tool that makes the connection between microscopic and macroscopic properties in thermodynamics possible. In the simplest case of a single molecule, the partition function  $q$  takes the following form:

$$q = \sum_{i=\text{levels}}^{\infty} g_i e^{-\epsilon_i/kT} \quad (2.1)$$

Here, it is expressed as a sum over all energy levels  $\epsilon_i$  of a molecule (or particle), multiplied by a degeneracy factor  $g_i$  in cases where multiple levels have the same energy. The term  $kT$  represents the Boltzmann factor.

Moving on to a more complex scenario in which the partition function describes multiple molecules, we arrive at the partition function  $Q$  for non-interacting particles, such as those in an ideal gas:

$$Q = q^N \text{ (different particles)} \quad Q = \frac{q^N}{N!} \text{ (identical particles)} \quad (2.2)$$

Here,  $N$  denotes the total number of particles. However, one could argue that if we wish to describe a real system such as bulk water, we must account for interactions between molecules. Consequently, Equation 2.2 must be rewritten:

$$Q = \sum_i^{\infty} e^{-E_i/kT} \quad (2.3)$$

In this case, the partition function  $Q$  includes contributions from all possible energy states  $E_i$  of the system.

Although the concept of the partition function might initially appear abstract, it can be clarified by expressing it in a different form, namely, within the context of the Rigid-Rotor Harmonic-Oscillator (RRHO) approximation, where the electronic, vibrational, and rotational degrees of freedom can be separated. For a single molecule case it would look like:

$$q_{\text{tot}} = q_{\text{trans}} \times q_{\text{rot}} \times q_{\text{vib}} \times q_{\text{elec}} \quad (2.4)$$

Let us now examine each contribution in more detail. From this point onward we will consider polyatomic molecules in the formulation of the partition functions, unless stated otherwise.

The translational partition function  $q_{\text{trans}}$  can be derived from the energy expression for a particle in a one-dimensional box and is given by:

$$q_{\text{trans}} = \left( \frac{2\pi MkT}{h^2} \right)^{3/2} V \quad (2.5)$$

Here,  $M$  is the total molecular mass, and  $V$  is the volume. Turning to the rotational partition function  $q_{\text{rot}}$ , it can be derived from the Schrödinger equation for a diatomic "rigid rotor" and has the following form:

$$q_{\text{rot}} = \frac{8\pi^2 I k T}{h^2 \sigma} \quad (2.6)$$

In this expression,  $I$  denotes the moment of inertia, and  $\sigma$  represents the symmetry factor, i.e. the order of the rotational subgroup within the molecular point group. For polyatomic molecules, writing an exact expression is more complex, but an approximate form can be used:

$$q_{\text{rot}} = \frac{\sqrt{\pi}}{\sigma} \left( \frac{8\pi^2 k T}{h^2} \right)^{3/2} \sqrt{I_1 I_2 I_3} \quad (2.7)$$

For the vibrational partition function  $q_{\text{vib}}$ , it is expressed as a product over the various vibrational modes of a molecule, each with frequency  $\nu_i$ :

$$q_{\text{vib}} = \prod_i \frac{e^{-h\nu_i/2kT}}{1 - e^{-h\nu_i/kT}} \quad (2.8)$$

Lastly, the electronic partition function  $q_{\text{elec}}$  is given as a sum over all electronic states of a molecule, from the ground state to all excited states. However, since the energy difference between the ground state and higher states is usually much greater than  $kT$  at ambient temperatures, the function can typically be approximated by considering only the ground state:

$$q_{\text{elec}} = \sum_{i=0}^{\infty} g_i e^{-\epsilon_i/kT} \approx g_0 e^{-\epsilon_0/kT} \quad (2.9)$$

### 2.1.2 Macroscopic properties and thermodynamic functions

Once the partition function is determined, it provides a direct means of evaluating macroscopic properties. For instance, the internal energy  $U$  and the Helmholtz free energy  $A$  can be calculated from the partition function  $Q$ :

$$U = kT^2 \left( \frac{\partial \ln Q}{\partial T} \right)_V \quad (2.10)$$

$$A = -kT \ln Q \quad (2.11)$$

In addition, other macroscopic properties, such as pressure  $P$  and the heat capacity at constant volume  $C_V$ , can also be expressed in terms of the partition function:

$$P = - \left( \frac{\partial A}{\partial V} \right)_T = kT \left( \frac{\partial \ln Q}{\partial V} \right)_T \quad (2.12)$$

$$C_V = \left( \frac{\partial U}{\partial T} \right)_V = 2kT \left( \frac{\partial \ln Q}{\partial T} \right)_V + kT^2 \left( \frac{\partial^2 \ln Q}{\partial T^2} \right)_V \quad (2.13)$$

Turning to thermodynamic functions, namely enthalpy  $H$ , entropy  $S$ , and Gibbs free energy  $G$ , these can also be derived from the partition function  $Q$ :

$$H = U + PV = kT^2 \left( \frac{\partial \ln Q}{\partial T} \right)_V + kTV \left( \frac{\partial \ln Q}{\partial V} \right)_T \quad (2.14)$$

$$S = \frac{U - A}{T} = kT \left( \frac{\partial \ln Q}{\partial T} \right)_V + k \ln Q \quad (2.15)$$

$$G = H - TS = kTV \left( \frac{\partial \ln Q}{\partial V} \right)_T - kT \ln Q \quad (2.16)$$

This connection between macroscopic observables, thermodynamic functions, and the partition function once again highlights its fundamental importance in statistical thermodynamics.

### 2.1.3 The canonical ensemble

Having established a method to calculate the macroscopic properties of a system we implicitly relied on averaging over a large enough number of states. Therefore one may naturally ask: how can we sample enough configurations to apply the equations described in the previous section under conditions that resemble those in experiments? One such answer is the canonical ensemble.

The canonical ensemble describes a system at constant temperature  $T$ , fixed volume  $V$ , and a fixed number of particles  $N$  (NVT). In this ensemble, the system is in contact with a heat bath, which makes it particularly relevant to most molecular simulations that are describing the experimental conditions, where the temperature is externally controlled while the internal energy of the system is allowed to fluctuate.

Since the energy fluctuates in the canonical ensemble, a logical step is to estimate the magnitude of these fluctuations:

$$\frac{\Delta E}{E} \sim \frac{\sqrt{N}}{N} \sim \frac{1}{\sqrt{N}} \quad (2.17)$$

Here,  $N$  denotes the number of particles, and thus for sufficiently large systems, the relative energy fluctuations become negligible.

The use of the canonical ensemble implicitly assumes that the system is ergodic, meaning that time averages obtained from simulation trajectories are equivalent to ensemble averages over the Boltzmann distribution. This assumption is, for instance, central to molecular dynamics simulations where the canonical ensemble can be sampled.

### 2.1.4 Classical forcefields and molecular dynamics

Bringing all the puzzle pieces together, we can now discuss how to simulate a molecular or atomic system of interest. One widely used approach is molecular dynamics (MD) simulations. The first step involves defining a potential energy function that describes the interactions between atoms. This function, often referred to as a forcefield, is typically parameterised based on experimental data or high-level quantum mechanical calculations.

In classical MD, the evolution of a system of  $N$  atoms is governed by Newton's equations of motion. A commonly used form of the potential energy function is:

$$\begin{aligned} U(\mathbf{r}_1, \dots, \mathbf{r}_N) = & \sum_{\text{bonds}} \frac{1}{2} K_{\text{bond}} (r - r_0)^2 + \sum_{\text{bends}} \frac{1}{2} K_{\text{bend}} (\theta - \theta_0)^2 \\ & + \sum_{\text{tors}} \sum_{n=0}^6 A_n [1 + \cos(C_n \phi + \delta_n)] \\ & + \sum_{i,j \in \text{nb}} \left\{ \left[ 4\epsilon_{ij} \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{r_{ij}} \right\} \end{aligned} \quad (2.18)$$

Here, the total energy is decomposed into bonded interactions (bonds, angles, and torsions) and non-bonded interactions, including Lennard-Jones and Coulombic terms. Once the potential is specified, the force on each atom  $i$  is obtained via:

$$\mathbf{F}_i = -\frac{\partial U}{\partial \mathbf{r}_i} \quad (2.19)$$

To propagate the positions and velocities of atoms in time, numerical integration schemes are employed. Among these, the velocity Verlet algorithm is widely used in perhaps all MD engines. Let us consider the Taylor expansion of the position of particle  $i$  to second order in the time step  $\Delta t$ :

$$\mathbf{r}_i(t + \Delta t) \approx \mathbf{r}_i(t) + \Delta t \mathbf{v}_i(t) + \frac{\Delta t^2}{2m_i} \mathbf{F}_i(t) \quad (2.20)$$

Here,  $\mathbf{F}_i(t)$  is the force on particle  $i$  at time  $t$ , and  $m_i$  is its mass, calculated using Equation 2.19 and  $\mathbf{v}_i(t)$  is its velocity. This expression provides a prediction of the new

position based on the current velocity and force.

We can also consider a backward expansion in time from  $\mathbf{r}_i(t + \Delta t)$  and  $\mathbf{v}_i(t + \Delta t)$ , yielding:

$$\mathbf{r}_i(t) = \mathbf{r}_i(t + \Delta t) - \Delta t \mathbf{v}_i(t + \Delta t) + \frac{\Delta t^2}{2m_i} \mathbf{F}_i(t + \Delta t) \quad (2.21)$$

By substituting Equation 2.20 into Equation 2.21 and solving for  $\mathbf{v}_i(t + \Delta t)$ , we get:

$$\mathbf{v}_i(t + \Delta t) = \mathbf{v}_i(t) + \frac{\Delta t}{2m_i} [\mathbf{F}_i(t) + \mathbf{F}_i(t + \Delta t)] \quad (2.22)$$

Equations 2.20 and 2.22 together form the velocity Verlet integrator. The algorithm proceeds as follows:

1. First, update positions using Equation 2.20.
2. Then, compute new forces  $\mathbf{F}_i(t + \Delta t)$  based on the updated positions.
3. Finally, update velocities using Equation 2.22.

To correctly sample the canonical ensemble, one should consider the use of thermostats to maintain the system temperature. In this work, we focus on two widely used thermostats: the Nosé–Hoover thermostat [30, 31] and the canonical sampling through velocity-rescaling (CSVR) thermostat [32]. In the former, the equations of motion are modified to include a friction term that couples the system to a heat bath, allowing for energy exchange. The CSVR thermostat, on the other hand, uses a velocity rescaling approach to maintain the desired temperature by adjusting particle velocities at each time step.

### 2.1.5 Enhanced sampling techniques

Even though MD simulations are a powerful tool for studying molecular systems, their applicability can be limited due to the presence of energy barriers separating minima in the potential energy landscape. As a result, the system may remain trapped in local minima, leading to insufficient sampling of the relevant configurational space. This issue becomes particularly pronounced in the context of reactive systems, where rare events involve transitions between states separated by high free energy barriers and occur on timescales much longer than typical simulation durations.

To address this challenge, various enhanced sampling techniques have been developed. These methods aim to accelerate the exploration of phase space. In general,

they bias the system along reaction coordinates, or collective variables (CVs), by applying a biasing potential that drives the system towards regions of interest. One such approach is metadynamics [33, 34].

In metadynamics, a biasing (external) potential is added to the system's potential energy surface. This biasing potential takes the following form:

$$V(S(x), t) = w \sum_{t'=\tau_G, 2\tau_G, \dots}^{t' < t} \exp\left(-\frac{(S(x) - s(t'))^2}{2\delta s^2}\right) \quad (2.23)$$

where  $s(t) = S(x(t))$  is the value of the CV at time  $t$ . The height of the Gaussian kernel is denoted by  $w$ ,  $\delta s$  is its width, and  $\tau_G$  is the deposition rate.

The approach used in metadynamics can be explained using the Panama Canal as an analogy as illustrated in Figure 2.1. The idea is to fill the basins of the free energy landscape with a Gaussian potential, which can be thought of as water gradually filling the basins, lifting the system (like a ship in a lock) out of a free energy minimum and helping it traverse to other states.

The assumption in metadynamics is that, after sufficiently long sampling, the biasing potential  $V_G(S, t)$  converges to the negative of the underlying free energy surface:

$$\lim_{t \rightarrow \infty} V(s, t) \sim -F(s) \quad (2.24)$$

Despite the many benefits that metadynamics offers, it is important to note that it has limitations. For example, obtaining a converged free energy surface is not straightforward, especially when multiple CVs are involved. In principle, Gaussian kernels can be deposited indefinitely, making it difficult to assess convergence. To address this issue, the well-tempered variant of metadynamics was developed [35]. In this method, a history-dependent potential is added, which is defined as:

$$V(s, t) = \Delta T \ln\left(1 + \frac{\omega N(s, t)}{\Delta T}\right) \quad (2.25)$$

Here,  $N(s, t)$  is the histogram of  $s$  obtained from a biased simulation,  $\Delta T$  is the biasing temperature, and  $\omega$  has the dimension of an energy rate. The rate at which this potential is modified over time is given by:

$$\dot{V}(s, t) = \frac{\omega \Delta T \delta_{s, s(t)}}{\Delta T + \omega N(s, t)} = \omega e^{-[V(s, t)/\Delta T]} \delta_{s, s(t)} \quad (2.26)$$

The height of the Gaussian kernels used is:

$$w = \omega e^{-[V(s, t)/\Delta T]} \tau_G \quad (2.27)$$

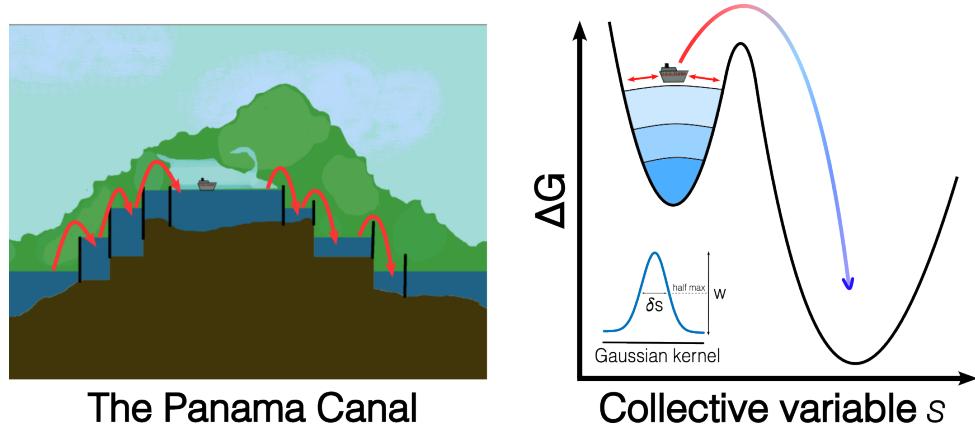


Figure 2.1: The concept of metadynamics.  $w$  stands for the Gaussian kernel height, and  $\delta s$  stands for its width. The Panama Canal cartoon was taken from [36].

where  $\tau_G$  is the deposition rate and  $\omega$  represents the initial bias deposition rate. The Gaussian kernel height is now dependent on the history of the system, allowing for a more controlled exploration of the free energy landscape.

Ultimately, the underlying free energy surface can be estimated using the following equation:

$$\tilde{F}(s, t) = -\frac{T + \Delta T}{\Delta T} V(s, t) = -(T + \Delta T) \ln \left( 1 + \frac{\omega N(s, t)}{\Delta T} \right) \quad (2.28)$$

The advantage of well-tempered metadynamics (WTMetaD) is that it enables more efficient exploration of the free energy landscape, as the biasing potential adapts according to the trajectory's history. Moreover, the convergence can be easily monitored by observing the decay of the Gaussian height, which should approach zero as the system fully explores the relevant phase space.

## 2.2 The density functional theory tourist

The discussion in this section is primarily based on the “Introduction to Computational Chemistry” textbook written by Jensen [27], “Density Functional Theory: a Practical Introduction” by Scholl and Steckel [37], and “A Chemist’s Guide to Density Functional Theory” by Koch and Holthausen [38] unless stated otherwise.

### 2.2.1 The Kohn-Sham approach

In order to describe reactive events in relatively large systems, up to approximately 1,000 atoms, it is necessary to use methods that offer a good balance between accuracy and computational cost. One such method is density functional theory (DFT),

which is based on the Hohenberg-Kohn theorems and the Kohn-Sham equations.

The central idea behind DFT, established by Hohenberg and Kohn, is that the ground state energy of a many-electron system can be expressed as a functional of the electron density. This reformulation reduces the problem from solving a  $3N$ -dimensional wavefunction to working with a 3-dimensional electron density.

The energy functional can be written as:

$$\begin{aligned} E[\rho(\mathbf{r})] &= T_s[\rho] + J[\rho] + E_{\text{ee}}[\rho] + E_{\text{xc}}[\rho] = \\ &= -\frac{1}{2} \sum_i^N \langle \phi_i | \nabla^2 | \phi_i \rangle \\ &\quad + \frac{1}{2} \sum_i^N \sum_j^N \iint |\phi_i(\mathbf{r}_1)|^2 \frac{1}{r_{12}} |\phi_j(\mathbf{r}_2)|^2 d\mathbf{r}_1 d\mathbf{r}_2 \\ &\quad - \sum_i^N \sum_A^M \int \frac{Z_A}{r_{1A}} |\phi_i(\mathbf{r}_1)|^2 d\mathbf{r}_1 \\ &\quad + E_{\text{xc}}[\rho(\mathbf{r})] \end{aligned} \quad (2.29)$$

Here, the first three terms are “known” and represent the kinetic energy of the electrons, the Coulomb interaction between the electrons, and the electron-nucleus interaction, respectively. The final term, the exchange-correlation energy functional, is the unknown component. It contains all the effects that are not straightforward to treat exactly, for instance, the residual part of the kinetic energy and the non-classical electron-electron interactions:

$$E_{\text{xc}}[\rho] \equiv (T[\rho] - T_s[\rho]) + (E_{\text{ee}}[\rho] - J[\rho]) \quad (2.30)$$

The biggest challenge in DFT lies in the formulation of  $E_{\text{xc}}$ . This term is particularly important, as finding the minimum of the total energy functional, as expressed in Equation 2.29, depends on its accurate representation.

To address this, the Kohn-Sham approach introduces a set of single-electron equations that can be solved iteratively to obtain the electron density and the total energy of the system. The Kohn-Sham equations are given by:

$$\left( -\frac{1}{2} \nabla^2 + V_{\text{eff}}(\mathbf{r}) \right) \phi_i = \varepsilon_i \phi_i \quad (2.31)$$

Here,  $V_{\text{eff}}$  takes the form:

$$V_{\text{eff}}(\mathbf{r}) = \int \frac{\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_2 + V_{\text{xc}}(\mathbf{r}) - \sum_A^M \frac{Z_A}{r_{1A}} \quad (2.32)$$

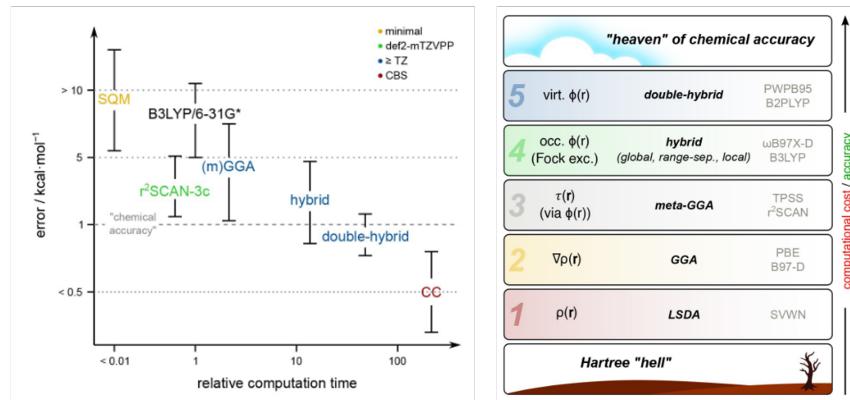


Figure 2.2: Left panel: accuracy of the common quantum chemical methods as a function of the computational cost. Right panel: categorisation of the exchange-correlation functionals according to Perdew’s “Jacob’s ladder”. The figure was reproduced from [39].

The iterative procedure to solve these equations proceeds as follows: first, a trial electron density is defined. Then, the Kohn-Sham equations are solved using this trial density to obtain the single-particle wavefunctions. Next, a new electron density is calculated from the obtained wavefunctions. Finally, the new density is compared with the initial trial density. If the two densities match within a given convergence criterion, the ground state electron density has been found, and the total energy of the system can be computed.

## 2.2.2 Generalised gradient approximation and PBE functional

The field of DFT has opened new avenues for computational chemists and physicists, enabling them to study the properties of materials and molecules, as well as to investigate the reaction pathways of chemical processes. However, the accuracy of DFT calculations is highly dependent on the choice of the exchange-correlation functional.

In this work, we focus on the generalised gradient approximation (GGA) exchange-correlation functionals, which are widely used in DFT calculations and are known to provide results close to chemical accuracy at a relatively low computational cost, as can be seen in Figure 2.2. In fact, the development of GGA functionals marked a turning point in the acceptance of the DFT method by the quantum chemistry community.

The GGA functionals are based on the idea that the exchange-correlation energy can be expressed as a functional of the electron density  $\rho$  and its gradient  $\nabla\rho$ . The general form of a GGA functional is given by:

$$E_{\text{XC}}^{\text{GGA}}[\rho] = \int f(\rho, \nabla\rho) d\mathbf{r} \quad (2.33)$$

The exchange-correlation energy can be explicitly divided into two parts:

$$E_{XC}^{\text{GGA}} = E_X^{\text{GGA}} + E_C^{\text{GGA}} \quad (2.34)$$

One of the most widely used GGA functionals is the Perdew-Burke-Ernzerhof (PBE) functional [40]. Its formulation incorporates 4 parameters in the exchange and correlation parts derived from first principles, making it a truly *ab initio* functional.

### 2.2.3 *Ab initio* molecular dynamics

In one of the previous sections, we touched upon the topic of classical MD simulations. However, classical force fields are unable to simulate bond-breaking and bond-forming processes. Although reactive events can also be studied using static approaches, by calculating the potential energy surface at a given set of coordinates, we believe that incorporating dynamics provides more informative insights and a clearer picture of the reaction mechanism.

To simulate the dynamics of a chemical reaction, one could consider using *ab initio* molecular dynamics (AIMD), and in particular, Born-Oppenheimer molecular dynamics (BOMD). In BOMD, the forces acting on the atoms are calculated at each time step using quantum mechanical methods, such as DFT, while the nuclei are propagated according to classical mechanics. This process can be described using the Lagrangian formalism,  $L$ , which offers an alternative formulation of classical dynamics:

$$L = K - U = \frac{1}{2} \sum_{i=1}^{3N} m_i v_i^2 - E[\phi(\mathbf{r}_1, \dots, \mathbf{r}_{3N})] \quad (2.35)$$

where  $K$  is the kinetic energy,  $U$  represents the potential energy, and  $\phi(\mathbf{r}_1, \dots, \mathbf{r}_{3N})$  is a set of one-electron Kohn-Sham wave functions.

It is important to note that, since nuclear dynamics are treated classically in this framework, zero-point vibrational energy is not accounted for, nor can tunnelling effects be studied.

## 2.3 Extended tight binding

The tight binding methods can be viewed as a simplification, or semi-empirical approximation, of DFT. Essentially, they are based on the same principles but introduce numerous approximations to reduce the computational cost. In 2017, the extended tight-binding (xTB) method was introduced. This development led to a new family of methods, namely GFNn-xTB, where GFN stands for geometries, frequencies, and non-covalent interactions, and  $n$  denotes the version [41].

In density functional tight-binding (DFTB) methods, the total energy of a system is expressed as a Taylor expansion around  $\Delta\rho$  [42]:

$$E[\rho] = E^{(0)}[\rho_0] + E^{(1)}[\rho_0, \delta\rho] + E^{(2)}[\rho_0, (\delta\rho)^2] + E^{(3)}[\rho_0, (\delta\rho)^3] + \dots \quad (2.36)$$

Here,  $\Delta\rho$  represents the difference between the converged  $\rho$  and the reference density  $\rho_0$ . The terms represent zeroth-order (core-core repulsion), first-order (valence electronic energy), second-order (charge corrections), and higher-order terms [27]. GFNn-xTB is heavily based on DFTB3, which includes terms in the energy expansion up to third order. The energy of GFN1-xTB, which is considered in this work, is given by [42]:

$$\begin{aligned} E_{\text{GFN1-xTB}} &= E_{\text{rep}}^{(0)} + E_{\text{disp}}^{(0)} + E_{\text{XB}}^{(0)} + E_{\text{EHT}}^{(1)} + E_{\text{IES+IXC}}^{(2)} + E_{\text{IES+IXC}}^{(3)} \\ &= E_{\text{rep}} + E_{\text{disp}}^{\text{D3}} + E_{\text{XB}}^{\text{GFN1}} + E_{\text{EHT}} + E_{\gamma} + E_{\Gamma}^{\text{GFN1}} \end{aligned} \quad (2.37)$$

where  $E_{\text{rep}}$  is the repulsion energy,  $E_{\text{disp}}^{\text{D3}}$  is the dispersion energy,  $E_{\text{XB}}^{\text{GFN1}}$  is the exchange-bonding energy,  $E_{\text{EHT}}$  is the extended Hückel theory energy, and  $E_{\gamma}$  and  $E_{\Gamma}^{\text{GFN1}}$  correspond to contributions arising from density fluctuations.

It is important to note that the GFNn-xTB methods are parametrised for 86 elements using a partially polarised minimal valence basis set. GFN1-xTB provides a good first approximation of the potential energy surface and is computationally efficient, thus making it suitable for the initial exploration of large molecular systems and running relatively long MD simulations.

## 2.4 Neural network potentials

The discussion in this section is mainly based on the textbook “Deep Learning: Foundations and Concepts” by C. M. Bishop and H. Bishop [43], the excellent review by Duval, Mathis, Joshi, and Schmidt et al. [44], the research paper by Batatia and Batzner et al. [45], and the research paper by Batzner et al. [46], unless stated otherwise.

### 2.4.1 Graph neural networks

One can easily imagine how computationally expensive it would be to run AIMD simulations for a system containing hundreds of atoms. In recent years, the field of machine learning has made significant progress in addressing this challenge. In particular, the development of graph neural networks (GNNs) has enabled the design of neural network potentials (NNPs), which learn the potential energy surface of molecular systems and can replace DFT in molecular dynamics simulations.

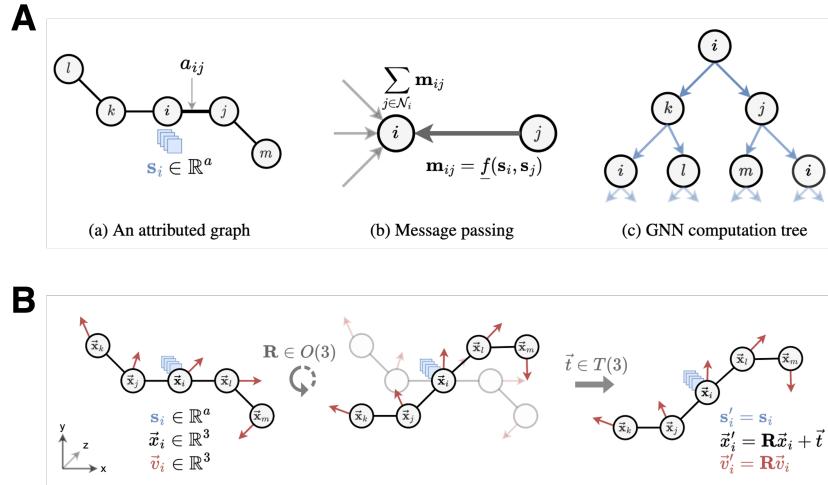


Figure 2.3: The concept of (A) GNNs and (B) geometric GNNs visualised. This figure was reproduced from [44].

A natural question one may ask here: why use GNNs? The answer lies in the fact that GNNs provide a natural way to represent molecular systems. Atoms can be considered as nodes in a graph, with bonds between them represented as edges. This concept is visualised in Figure 2.3 A.

An essential property of any graph  $\mathcal{G} = (\mathbf{A}, \mathbf{S})$  is its adjacency matrix  $\mathbf{A}$ , which encodes the connectivity of the nodes via elements  $a_{ij}$ . This matrix is square with dimensions  $n \times n$ , where  $n$  is the number of nodes in the graph. Each entry is either 0 or 1, indicating the absence or presence of a connection between two nodes.

Because the ordering of nodes in the graph is arbitrary, GNNs are, by construction, permutation symmetric.

In addition to the adjacency matrix, each graph also has a matrix of scalar features  $\mathbf{S}$  associated with its nodes. In the case of molecular systems, these features may include atomic numbers or atom types.

GNNs are neural networks specifically designed to operate on graph-structured data. They learn representations of nodes, edges, or entire graphs by iteratively updating node features based on their neighbours  $\mathcal{N}_i$ . This iterative process is commonly referred to as message passing, illustrated in Figure 2.3 A. Conceptually, the process proceeds as follows:

1. At iteration  $t$ , each node  $i$  receives messages from its neighbours  $\mathcal{N}_i$ :

$$\mathbf{m}_{ij}^{(t)} = \underline{\text{MSG}} \left( \mathbf{s}_i^{(t)}, \mathbf{s}_j^{(t)} \right) \quad (2.38)$$

2. These messages are aggregated to update the node's features:  $\bigoplus_{j \in \mathcal{N}_i} \mathbf{m}_{ij}^{(t)}$ .

3. At iteration  $t + 1$ , the node's representation is updated using the aggregated messages and its current state:

$$\mathbf{s}_i^{(t+1)} = \text{UPD} \left( \mathbf{s}_i^{(t)}, \bigoplus_{j \in \mathcal{N}_i} \mathbf{m}_{ij}^{(t)} \right) \quad (2.39)$$

The MSG and UPD functions form an entire research topic within GNNs, and are typically implemented as neural networks themselves. The aggregation operator  $\bigoplus$  must be permutation-invariant and can take forms such as summation, averaging, etc. After the final iteration, the resulting node representations can be used to predict properties at various levels, whether for the entire graph, such as the total energy of a molecular system, or at the node or edge level.

A particularly important subclass of GNNs is the geometric GNNs, which are designed to handle geometric data in Euclidean space. These are depicted in Figure 2.3 B. Geometric graphs  $\mathcal{G} = (\mathbf{A}, \mathbf{S}, \vec{\mathbf{x}}, \vec{\mathbf{v}})$  contain additional information such as atomic coordinates  $\vec{x}_i$  and other vector features  $\vec{v}_i$ . These vectors may represent quantities such as velocities or forces.

Geometric GNNs are tailored to work with data that has an inherent geometric structure, such as point clouds. Their key advantage is in handling symmetry operations, i.e. that under rotations and translations of the data, scalar features remain invariant, or the same, while vector features transform appropriately. This property is crucial for accurately capturing the underlying physics of molecular systems.

### 2.4.2 Invariance and equivariance

In the context of geometric GNNs, the concepts of invariance and equivariance play a fundamental role in ensuring that models respect the symmetries of the data they are trained on. These properties are especially crucial when working with molecules, where rotations or translations should not change the outcome of the prediction or should change it in a predictable way.

Mathematically, a function  $f$  is said to be invariant to a transformation  $g$  if applying  $g$  to the input  $x$  does not affect the output of the function:

$$f(g \cdot x) = f(x) \quad (2.40)$$

A simple real-world example of invariance is the total energy of a molecule. Typically it stays invariant under rotations and translations of its atomic coordinates.

On the other hand, a function  $f$  is said to be equivariant to a transformation  $g$  if

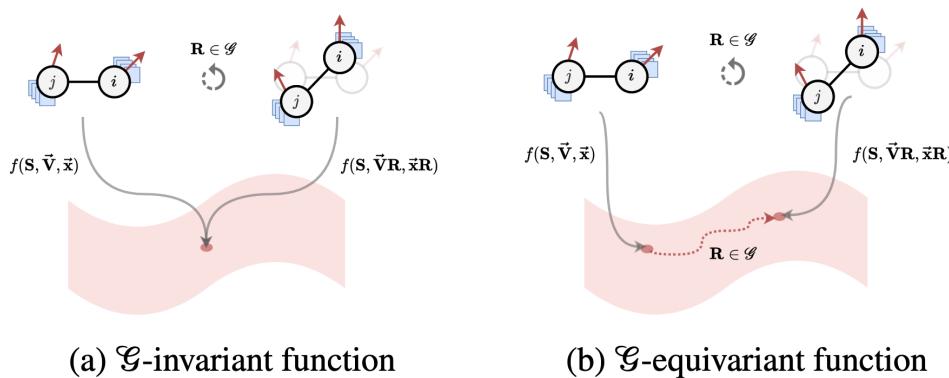


Figure 2.4: Invariance and equivariance. This figure was taken from [44].

applying  $g$  to the input results in the same transformation being applied to the output:

$$f(g \cdot x) = g \cdot f(x) \quad (2.41)$$

A familiar example of equivariance is the behaviour of vector quantities such as velocity or force. If we rotate a coordinate system (or the object within it), the velocity or force vectors will also rotate accordingly. In the context of molecular systems, atomic forces are equivariant under rotations: if the molecule is rotated, the directions of the forces rotate in the same way. Both concepts of invariance and equivariance are illustrated in Figure 2.4.

Designing GNNs that are invariant or equivariant helps reduce the amount of data needed for training. Geometric neural networks, for instance, are explicitly constructed to respect such symmetries, making them particularly well-suited for chemical modelling.

### 2.4.3 Equivariant graph neural networks

In the previous section, we emphasised the importance of implementing invariance and equivariance in GNNs. In general, the field of geometric GNNs applied to computational chemistry problems is still in its early stages, since the first models were introduced around 2018, as shown in Figure 2.5 A.

Nevertheless, the field is rapidly developing, with new architectures being proposed that advance the state of the art. These architectures achieve better performance in learning the intricacies of the potential energy surface while also scaling to systems containing millions of atoms [47].

In this section, we focus on the general pipeline of geometric GNNs, and on the NequIP [46] neural network in particular, which stands for Neural Equivariant Interatomic Potential. The general pipeline of geometric GNNs is shown in Figure 2.5 B.

Broadly speaking, the entire workflow can be divided into three main steps:

1. Create the atomic representations.
2. Learn the embeddings of the atomic representations.
3. Predict the desired output property, such as the total energy or forces.

## Atomic representations

Before learning atomic representations, the network constructs an input graph from the given atomic coordinates, i.e., a point cloud. Atoms are treated as nodes, and edges are formed between atoms that lie within a specified cutoff radius. This radius defines the local environment of each atom and is crucial for ensuring that the network processes only physically meaningful interactions. To ensure a gradual change in interaction strength, a smooth cutoff function is often employed. It has the following form:

$$A_{ij} = \begin{cases} \frac{1}{2} \left( \cos \left( \frac{\pi d_{ij}}{c} \right) + 1 \right) & \text{if } d_{ij} \leq c \\ 0 & \text{otherwise} \end{cases} \quad (2.42)$$

Here,  $d_{ij} = ||\vec{x}_i - \vec{x}_j||$  is the distance between atoms  $i$  and  $j$ , and  $c$  is the cutoff radius. The cutoff function smoothly transitions from 1 to 0 as the distance approaches the cutoff, ensuring that interactions are considered only within the specified range. This function is typically chosen to be continuous and differentiable, such as a cosine or polynomial function, to avoid discontinuities. The choice of cutoff radius is critical, as it determines the size of the local environment considered by the network. A cutoff that is too small may lead to the loss of important long-range interactions, while one that is too large may introduce noise and increase computational cost.

When working with periodic systems, periodic boundary conditionss (PBCs) must be respected during graph construction. Therefore, adjacent unit cells are considered within the cutoff distance to ensure that atomic environments are correctly represented.

Alongside the geometric information, usually encoded as distance-based features, each atom is assigned a type-dependent feature, often implemented as an atomic embedding vector. In addition to pairwise distances, some architectures incorporate angular information or more complex symmetry functions to capture higher-order geometric correlations. Together, these inputs form the basis of the graph on which the neural network operates.

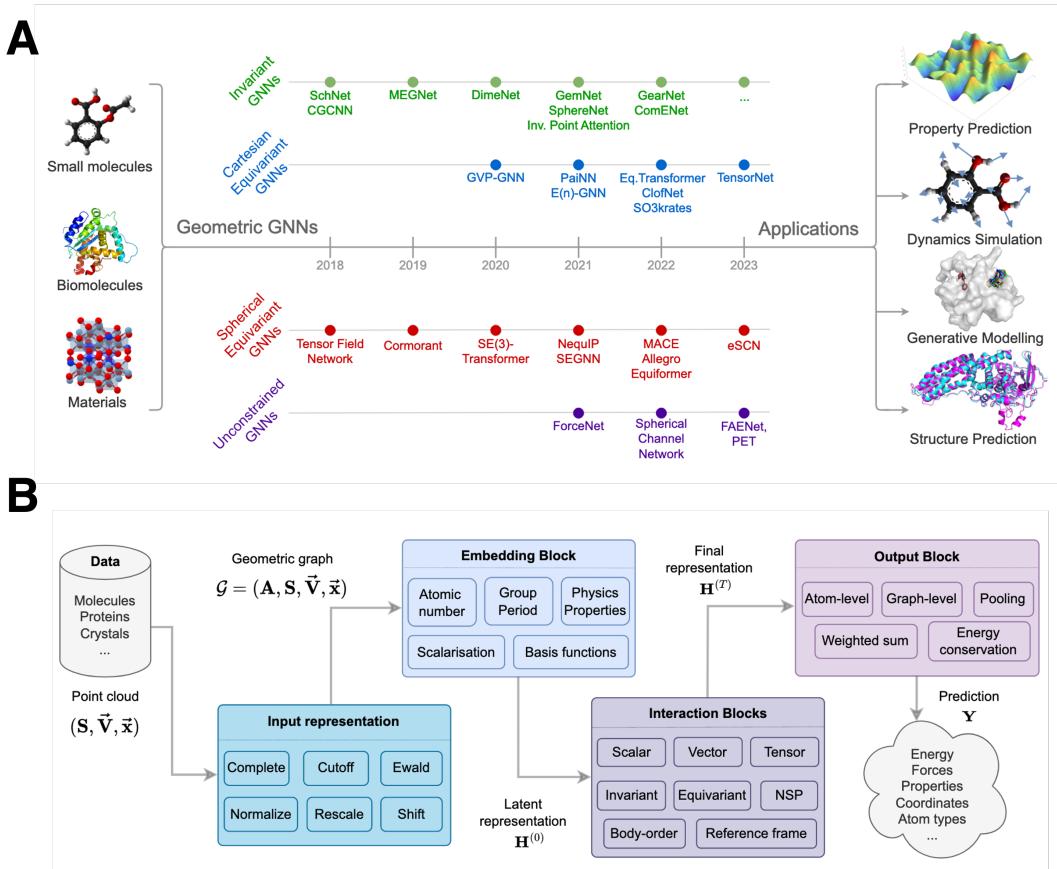


Figure 2.5: (A) The timeline of the development of geometric GNNs aimed at computational chemistry problems. (B) A general architecture of geometric GNNs. This figure was reproduced from [44].

### Embedding and interaction blocks

Once the input features have been defined, the network proceeds through an initial embedding layer. This layer maps the input features into a latent space, which may consist of scalars, vectors, or higher-order tensors  $\ell$ , depending on the network's level of geometric complexity. For instance, in NequIP, the input features are mapped into spherical tensors  $\ell$  that are irreducible representations of  $SO(3)$  thanks to the e3nn library [48], ensuring rotational equivariance, while translational invariance is built in by using relative positional features. In other words, the embedding layer constructs the initial learnable atomic representations that will be refined by the subsequent layers of the network.

The core of the network comprises a series of interaction, or message-passing, blocks. Within each block, nodes aggregate both scalar and vector information from their neighbours, often using edge features such as relative positions or spherical harmonics to guide the update. The aggregated messages are then passed through neural network layers. As a result, each node's feature vector is updated in a way that pre-

serves the geometric structure of the system. By stacking multiple interaction blocks, the network gains the capacity to capture complex many-body interactions. Each layer builds upon the features extracted by the previous one, allowing the model to learn increasingly rich representations of atomic environments.

### Output block

The final stage of a geometric GNN involves translating the atomic representations, refined by the embedding and interaction blocks, into physically meaningful quantities, such as the total energy of the system. This quantity is typically obtained as a sum over per-atom contributions, a strategy that ensures permutation invariance with respect to atomic indexing.

To derive atomic forces, two main approaches can be used: either predicting them directly or utilising automatic differentiation. In the latter case, forces are computed as the negative gradient of the predicted energy with respect to atomic positions using automatic differentiation as implemented in modern deep learning libraries. This approach ensures that the forces are conservative and thus obey the fundamental law of energy conservation. Moreover, since the model is constructed to respect geometric symmetries, the resulting forces are also equivariant under rotations and translations.

Thanks to the incorporating the physics-inspired features and symmetry constraints, the output of the network provides physically meaningful predictions of both total energy and atomic forces.

# Chapter 3

## Computational details

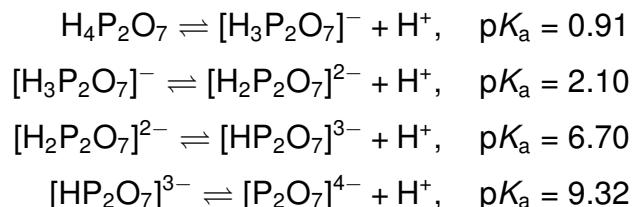
This chapter provides detailed information on the computational methods employed in this work. The first section outlines the generation of the training dataset, including system preparation, initial equilibration using molecular mechanics, exploration of the configuration space at the GFN1-xTB level, further data labelling, and iterative training of the neural network potential. The second section discusses production runs at ambient temperature using the fitted neural network potential. Finally, the third section presents the data analysis and visualisation techniques used to interpret the results.

### 3.1 Training dataset generation

#### 3.1.1 System preparation

The systems were prepared using the functionality of the CHARMM-GUI webserver [49], specifically the Multicomponent Assembler interface [50].

As a first step, methyl diphosphate trianion (MeDP) and methyl diphosphate dianion (MeHDP) were parameterised using CGenFF [51], i.e., the CHARMM General Force Field. These protonation states were chosen based on the dissociation constants of pyrophosphoric (diphosphoric) acid [52]:



Thus, at physiological pH (7.4), this acid exists in equilibrium between the singly and doubly deprotonated forms. Assuming the methyl group behaves similarly to a

proton, the methyl diphosphate molecule was considered to exist as a mixture of the MeHDP and MeDP forms under physiological conditions.

Following successful parameterisation, the system was solvated in a cubic box of TIP3P water molecules, with sodium counterions ( $\text{Na}^+$ ) added to neutralise the system's overall charge. The final system composition is provided in Table 3.1.

### 3.1.2 Initial equilibration using classical force fields

The system equilibration followed the standard protocol generated by the CHARMM-GUI webserver [49]. Initially, energy minimisation was conducted using the steepest descent algorithm for 5,000 steps.

This was followed by equilibration in the constant number of particles, volume and temperature (NVT) ensemble for 5 ns. During both the minimisation and NVT phases, the solute's heavy atoms were restrained using a harmonic potential with a force constant of  $400 \text{ kJ mol}^{-1} \text{ nm}^{-2}$ .

Subsequently, the system was equilibrated in the constant number of particles, pressure and temperature (NPT) ensemble for 45 ns. Throughout this procedure, the temperature and pressure were maintained at 300 K and 1 bar, respectively. Temperature was controlled using a  $\nu$ -rescale thermostat [32] with a coupling constant of 1 ps, and pressure was regulated using an isotropic  $c$ -rescale barostat [53] with a coupling constant of 5 ps. A 0.6 nm cut-off was applied for non-bonded interactions, and long-range electrostatics were treated using the Particle Mesh Ewald (PME) method. PBC were applied in all directions throughout the simulation.

All simulations were carried out using GROMACS 2021.4 [54] with CHARMM36m force field [55]. The leap-frog integrator was employed with a time step of 1 fs. All hydrogen-involving bonds were constrained using the LINCS algorithm. The equilibrated box dimensions used for subsequent simulations were taken from the output of the NPT run and are summarised in Table 3.1. Unless otherwise stated, the last frame of the NPT simulations was used as the starting point for all further calculations.

Table 3.1: System composition and simulation box details.

| System | Final box dimensions ( $\text{\AA}^3$ ) | No. of $\text{H}_2\text{O}$ | No. of $\text{Na}^+$ | No. of atoms |
|--------|---|-----------------------------|----------------------|--------------|
| MeDP   | $15.877 \times 15.877 \times 15.877$    | 119                         | 3                    | 373          |
| MeHDP  | $15.901 \times 15.901 \times 15.901$    | 124                         | 2                    | 388          |

### 3.1.3 Collective variables

To effectively sample the reaction space, two types of CVs were employed to bias the system: distances and coordination numbers (CNs). The CN is defined by the following smooth function:

$$\sum_{i \in A} \sum_{j \in B} CN_{ij} = \frac{1 - \left( \frac{r_{ij} - d_0}{r_0} \right)^n}{1 - \left( \frac{r_{ij} - d_0}{r_0} \right)^m} \quad (3.1)$$

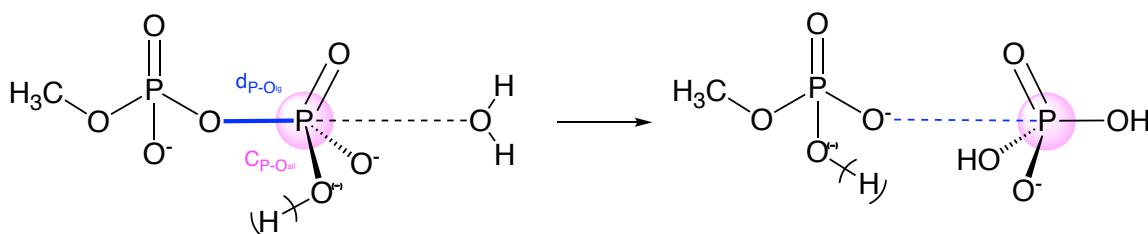
where  $r_{ij}$  is the distance between atoms  $i$  and  $j$  from groups  $A$  and  $B$ ,  $d_0$  is the distance at which the CN begins to decay,  $r_0$  is a characteristic decay length, and  $n$  and  $m$  are integers that control the steepness of the decay. Typically,  $m > n$ , ensuring a smooth transition of  $CN_{ij}$  from approximately 1 to 0 as the distance increases.

The specific CVs used in this work are shown in Figure 3.1, and their corresponding parameters are as follows:

- Distance between the  $\beta$ -phosphorus and the bridging oxygen ( $d_{P-O_{lg}}$ ),
- Coordination number of all water oxygen atoms as well as the bridging oxygen surrounding the  $\beta$ -phosphorus ( $C_{P-O_{all}}$ ):  $d_0 = 0$ ,  $r_0 = 2.1 \text{ \AA}$ ,  $n = 8$ ,  $m = 16$ .

Additionally, the following CVs were monitored to check whether the system was in a reasonable region of the potential energy surface, e.g. to ensure that there is no oxygen exchange between the methyl diphosphate and the water molecules:

- Coordination number of 3 nonbridging phosphoryl oxygens surrounding the  $\beta$ -phosphorus ( $C_{P-O_{nb}}$ ):  $d_0 = 0$ ,  $r_0 = 2.1 \text{ \AA}$ ,  $n = 8$ ,  $m = 16$ ,
- Coordination number of 2 nonbridging oxygens from  $\alpha$ -phosphorus as well as the one forming a bond with the methyl group surrounding the  $\beta$ -phosphorus ( $C_{P-O_{nb-lg}}$ ):  $d_0 = 0$ ,  $r_0 = 2.1 \text{ \AA}$ ,  $n = 8$ ,  $m = 16$ ,



$$d_{P-O_{lg}} = d(P - O_{lg}), C_{P-O_{all}} = CN(P - O_{all})$$

Figure 3.1: The definition of the collective variables (CVs) used in this work.  $d$  stands for distance and CN stands for coordination number.

- Coordination number of water oxygens around the  $\beta$ -phosphorus ( $C_{P-O_w}$ ):  $d_0 = 0$ ,  $r_0 = 2.1 \text{ \AA}$ ,  $n = 8$ ,  $m = 16$ ,
- Coordination number of non-methyl hydrogen atoms around 5 non-bridging and 1 bridging oxygen atoms ( $C_{O-H_{all}}$ ):  $d_0 = 0$ ,  $r_0 = 1.3 \text{ \AA}$ ,  $n = 8$ ,  $m = 16$ .

To avoid sampling of unphysical regions of the potential energy surface, quadratic (harmonic-like) wall potentials were applied to softly constrain certain degrees of freedom. The mathematical form of these wall potentials is given below:

$$\text{For upper walls: } \sum_i k_i \left( \frac{CV_i - a_i + o_i}{s_i} \right)^{e_i} \quad (3.2)$$

$$\text{For lower walls: } \sum_i k_i \left| \frac{CV_i - a_i - o_i}{s_i} \right|^{e_i} \quad (3.3)$$

Here,  $CV_i$  denotes the value of the collective variable,  $k_i$  is the force constant defining the wall's strength,  $a_i$  is the central wall position,  $o_i$  is an offset,  $s_i$  is a scaling factor, and  $e_i$  is the exponent that controls the wall's steepness. When  $e_i = 2$ , the potential acts harmonically.

The wall potentials applied to the CVs during the simulations are summarised in Table 3.2. The parameters for the wall potentials were chosen based on the expected ranges of the CVs. The force constants were set to ensure that the walls were sufficiently strong to prevent unphysical configurations while allowing for reasonable exploration of the configuration space.

All CV-related computations were performed using the built-in tools of CP2K 2023.1 [56] or PLUMED 2.9.3 [57]. It is important to note that the number and type of CVs, as well as the applied restraints, varied depending on the specific stage of the workflow. In the following sections, the relevant collective variables and wall potentials will be specified accordingly.

Table 3.2: The restraints applied to the collective variables during some of the simulations. In all cases,  $o = 0$ ,  $s = 1$ ,  $e = 2$ . <sup>1</sup>During the iterative training/production runs. <sup>2</sup>Different values for the walls were used depending on the system MeDP/MeHDP. Distances are in  $\text{\AA}$  and coordination numbers are unitless.

| CV                | Lower wall | Upper wall             | Force constant (kcal mol <sup>-1</sup> Å <sup>-2</sup> ) |
|-------------------|------------|------------------------|--|
| $d_{P-O_{lg}}$    | —          | 5.0 / 6.0 <sup>1</sup> | 500  |
| $C_{O-H_{all}}$   | —          | 1.3 / 2.5 <sup>2</sup> | 1000   |
| $C_{P-O_{nb}}$    | 2.6        | —                      | 2000   |
| $C_{P-O_{nb-lg}}$ | —          | — / 0.1 <sup>1</sup>   | 2000   |
| $C_{P-O_w}$       | —          | 1.3                    | 2000   |

### 3.1.4 GFN1-xTB based exploration of the configuration space

To generate the initial set of configurations for the training dataset, the system was subjected to molecular dynamics simulations using the semi-empirical GFN1-xTB [41] level of theory.

Each system was first equilibrated for 5 ps in the NVT ensemble at 300 K to allow the structures to relax at the GFN1-xTB level, including a Grimme's D3 dispersion correction [58]. Following equilibration, we performed 50 ps of WTMetaD [35] simulations in the NVT ensemble. In these simulations, a biasing potential was applied to encourage the system to explore regions of the configuration space beyond the reactant basin. This bias was introduced along two CVs:  $d_{P-O_g}$  and  $C_{P-O_{all}}$ . No restraints were applied to the system during this stage.

All calculations were carried out using the CP2K 2023.1 package [56] on CPUs. Temperature control was achieved using the CSVR thermostat [32], with a time constant of 50 fs during equilibration and 100 fs during the WTMetaD simulations. The self-consistent field (SCF) convergence threshold was set to  $10^{-5}$  a.u. The biasing potential was updated every 25 fs, with a Gaussian hill height of 2 kcal mol<sup>-1</sup> and a width of 0.07 for each CV. The bias factor was set to 30. Finally, the integration time step was set to 0.5 fs. Throughout the simulations, PBC were applied in all directions.

### 3.1.5 Data labeling

All data points were labeled by performing single-point calculations to obtain the energy and force values. These single-point calculations were carried out using the Perdew-Burke-Ernzerhof exchange-correlation functional (PBE) [40], along with a Grimme's D3 dispersion correction and the Becke-Johnson damping function [58, 59]. In all calculations, the Goedecker-Teter-Hutter pseudopotentials (GTH) [60, 61] were used to represent the core electrons, in combination with the triple- $\zeta$  valence basis set with two polarisation functions (TZV2P).

The single-point calculations were performed using the Gaussian plane wave method (GPW) implemented in the QUICKSTEP module [62] of the CP2K 2023.1 package [56]. The SCF convergence threshold was set to  $10^{-6}$  a.u. A plane-wave cutoff of 800 Ry was applied for the total density, while a cutoff of 60 Ry was used for the Kohn-Sham orbitals.

The aforementioned cutoffs were determined based on a convergence test performed on one of the configurations, as described in [63]. An error in total energy of less than  $10^{-8}$  a.u. was considered acceptable for the convergence test. The test was conducted by varying the cutoff for the total density from 400 to 1500 Ry, and the cutoff

for the Kohn-Sham orbitals from 10 to 200 Ry. The results of the convergence test are shown in Table A.1.

After the DFT labelling, the data were transformed into the extended XYZ format, which includes atomic positions, forces, total energies, cell dimensions, and whether PBC were used or not. This format was chosen because it is compatible with neural network training, as explained in the next section. The data transformation was carried out using an in-house script, `cp2k2extxyz.py`, which is available in [64].

### 3.1.6 Iterative training of the neural network potential

We trained a NNP using the NequIP framework [46], which implements equivariant message-passing networks for atomistic simulations. Regarding the hyperparameters, a radial cutoff distance of 5.0 Å was chosen to describe the atomic environment of the system.

The equivariant part of the neural network was composed of four interaction layers with a maximum tensor rank of  $\ell = 1$  or  $2$ . Feature parity was enabled to include both even and odd components, and 32 features per irreducible representation were used throughout. Scalar and gating nonlinearities were set to `silu` and `tanh` for even and odd parities, respectively. Eight radial basis functions were employed, in combination with a trainable Bessel basis and a polynomial cutoff of order 6.

The invariant subnetwork for radial interaction modelling consisted of two layers with 64 hidden neurons. Self-connections were enabled, and the average number of neighbours was computed automatically based on the dataset.

Training was performed using the Adam optimizer with the AMSGrad variant enabled, and with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 10^{-8}$ . A starting learning rate of 0.01 was used, and the learning rate was adaptively reduced by a factor of 0.5 upon stagnation of the validation loss (patience = 100 epochs). Early stopping was triggered if the validation loss remained unimproved for 50 epochs, if the loss dropped below  $1 \times 10^{-5}$ , or if it exceeded  $1 \times 10^4$ . The batch size was set to 5. The training was carried out over a period of three days on a single NVIDIA A100 GPU using float64 precision.

The networks were trained using the following mean squared error (MSE) loss function already implemented in the NequIP framework:

$$\mathcal{L} = \lambda_E \|\hat{E} - E\|^2 + \lambda_F \frac{1}{3N} \sum_{i=1}^N \sum_{\alpha=1}^3 \left\| -\frac{\partial \hat{E}}{\partial r_{i,\alpha}} - F_{i,\alpha} \right\|^2 \quad (3.4)$$

Here,  $N$  is the number of atoms,  $\hat{E}$  is the predicted energy,  $E$  is the reference energy,  $F_{i,\alpha}$  is the reference force on atom  $i$  in direction  $\alpha$ , and  $-\frac{\partial \hat{E}}{\partial r_{i,\alpha}}$  is the calculated force by means of autodifferentiating the predicted energy with respect to the atomic

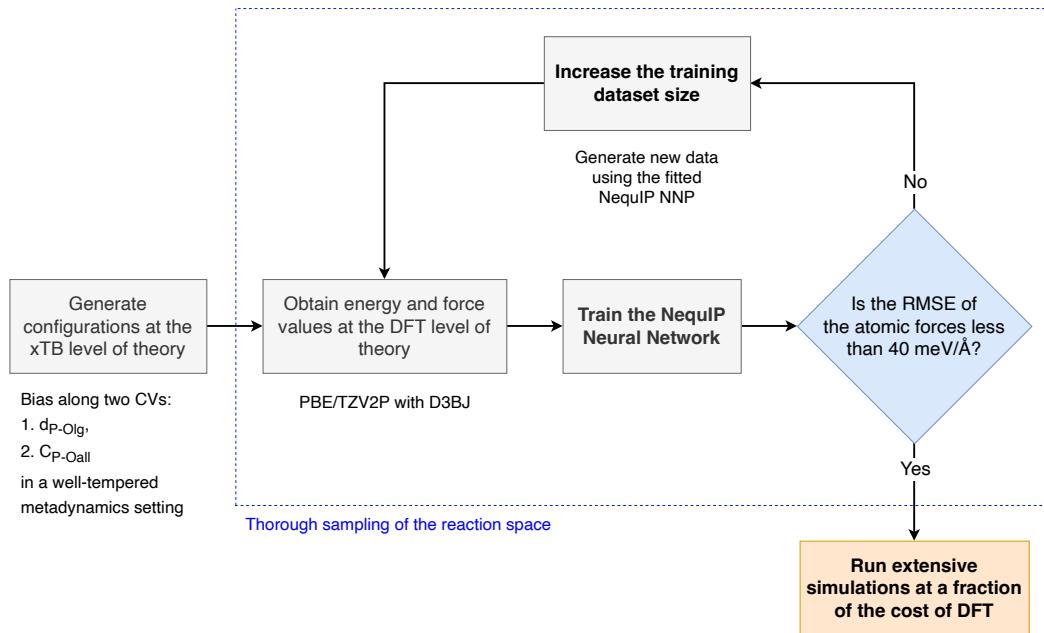


Figure 3.2: Iterative training of the NequIP neural network potential.

position. The hyperparameters  $\lambda_E$  and  $\lambda_F$  by default have a weighting of 1 to  $N_{atoms}^2$ , respectively.

To thoroughly sample the reaction space, the training was performed in an iterative manner, where the model was first trained on a small set of data and then used to generate additional data points. This process was repeated until the model converged, with the root mean square error (RMSE) of the atomic forces being less than 40 meV/Å. The workflow is shown in Figure 3.2.

In the end, the full dataset consisted of 12,000 configurations for training and validation, and 1,800 configurations for testing, for both systems (MeDP and MeHDP) combined. This dataset was obtained within the three rounds of iterative training. In each round of training, the model was retrained on a larger dataset. The data obtained from each round will be discussed in the following sections.

### Selection of configurations for training and testing

An important part of the iterative training process is the selection of configurations that will be used to train the neural network. To construct a representative and diverse dataset for training the neural network potential, configurations were selected from a metadynamics trajectory using a density-aware sampling strategy. The raw data were

extracted from a file generated during the enhanced sampling simulations. Each configuration in this file corresponds to a simulation snapshot, annotated with a time index and two collective variables (CVs): the distance  $d_{P-O_{lg}}$  and the coordination number  $C_{P-O_{all}}$ .

The two CVs were combined into a two-dimensional feature space  $\mathbf{X} = (d, C)$ , which served as the basis for sampling. This feature space often exhibits regions of highly non-uniform data density, due to the biased nature of metadynamics sampling. To account for this, a density-aware sampling method was employed to select configurations for training and testing that maintain good coverage across the feature space.

The selection procedure proceeds as follows:

1. A user-defined number of samples is specified.
2. K-means clustering is applied to the feature space to partition it into a number of clusters,  $k$ . The number of clusters is determined heuristically as  $k = \max(10, \min(\lfloor \frac{N}{50} \rfloor, \lfloor \frac{n_{samples}}{10} \rfloor))$ , where  $N$  is the total number of configurations and  $n_{samples}$  is the desired number of samples.
3. The number of points sampled from each cluster is proportional to its size, ensuring that denser regions do not dominate the dataset. A minimum of one sample is taken from each non-empty cluster.
4. Within each cluster, a fixed number of configurations is randomly selected using a deterministic random seed to ensure reproducibility.
5. After the training set is selected, the remaining configurations are used to construct the test set, following the same density-aware procedure while ensuring no overlap with the training configurations.

This approach results in training and test datasets that closely mirror the overall distribution of the CVs, while ensuring that underrepresented regions of the feature space are adequately sampled. The final output consists of two lists of snapshot indices corresponding to the selected training and test configurations, along with their respective CV values. These snapshots were then extracted from the trajectory files for use in model training and evaluation. The pseudo-code for the density-aware sampling algorithm is provided in Algorithm A.1.

### **First round**

In the first round of training the NNP, the model was trained on a small dataset consisting of 4,000 configurations. These configurations were obtained from the initial

exploration of the configuration space at 300 K using the GFN1-xTB level of theory, as described in Section 3.1.4. The enhanced sampling simulations were biased along  $d_{P-O_{lg}}$  and  $C_{P-O_{all}}$ , and no restraints were applied to the system. The training was carried out using the hyperparameters described in Section 3.1.6.

### Second round

In the second round of training, the model was trained on a larger dataset consisting of 8,000 configurations. The additional configurations were obtained from a second round of exploration of the configuration space, driven by the NNP obtained after the first round of training.

The NNP-driven simulations were run using the LAMMPS package [65] compiled with PLUMED 2.9.3 [57] and pair\_nequip [66] on a single A100 GPU. The simulations were performed for 100 ps in the NVT ensemble at 300 K with the PBC applied in all directions. The temperature was controlled by a Nosé–Hoover thermostat [30, 31] with a time constant of 50 fs. The biasing potential was applied to  $d_{P-O_{lg}}$  and  $C_{P-O_{all}}$  every 50 fs, using a Gaussian hill height of 2 kcal mol<sup>-1</sup> and a width of 0.07 for each CV. The bias factor was set to 30, and the integration time step was 0.5 fs.

Restraints were applied to  $d_{P-O_{lg}}$  and  $C_{P-O_{all}}$  in order to favour either a dissociative or associative mechanism of the reaction and sample more configurations from the transition state (TS) regions. The training was performed using the same hyperparameters as in the first round.

### Third round

In the final round of training, the model was trained on a dataset consisting of 12,000 configurations. These additional configurations were obtained from a third round of exploration of the configuration space, driven by the NNP obtained after the second round of training. The simulations were performed for 500 ps using the same setup as in the second round. The only difference was that the temperature in this round was increased to 320 K and 340 K to explore the configuration space at higher temperatures. The same CVs were biased as in the previous run. No restraints were applied to the system. The training was conducted using the same hyperparameters as in the first round. The final dataset is summarised in Table A.2.

## 3.2 Production runs at ambient temperature

To thoroughly sample the reaction space, the long production runs were performed at ambient temperature (300 K) using the neural network potential obtained in the final round of training. To run the simulations with the NNP, the LAMMPS package [65], compiled with PLUMED 2.9.3 [57] and pair\_nequip [66], was utilised. First, the systems were equilibrated for 75 ps in NVT ensemble, since it was previously shown that water fully relaxes in this timeframe [67]. Then the simulations were carried out for 2 ns in the NVT ensemble, with temperature regulated by a Nosé–Hoover thermostat [30, 31] with a frequency of 50 fs<sup>-1</sup>.

The biasing potential was applied every 50 fs to  $d_{P-O_{lg}}$  and  $C_{P-O_{all}}$  using a Gaussian hill height of 0.5 kcal mol<sup>-1</sup> and a width of 0.07 for each collective variable. Additionally, restraints were imposed on the CVs as mentioned in Table 3.2. The bias factor was set to 30, and the integration time step was maintained at 0.5 fs. All simulations were conducted on a single A100 GPU.

To obtain the free energy profiles of the reactions, the Gaussian kernels applied during the simulations were summed using the `sum_hills` utility provided in the PLUMED 2.9.3 package [57]. Afterwards, the minimum free energy paths (MFEPs) were extracted using the MEPSA 1.4 software [68].

## 3.3 Data analysis and visualisation

Data analysis was performed using Python in Jupyter notebooks [69]. All plots were generated using Matplotlib library [70]. The water oxygen-oxygen  $g_{O-O}(r)$  radial distribution function (RDF) was calculated using the extension in VMD [71] with the bin size of 0.0125 Å using the 2 ns of the production runs.

# Chapter 4

## Results and discussion

### 4.1 Final dataset composition

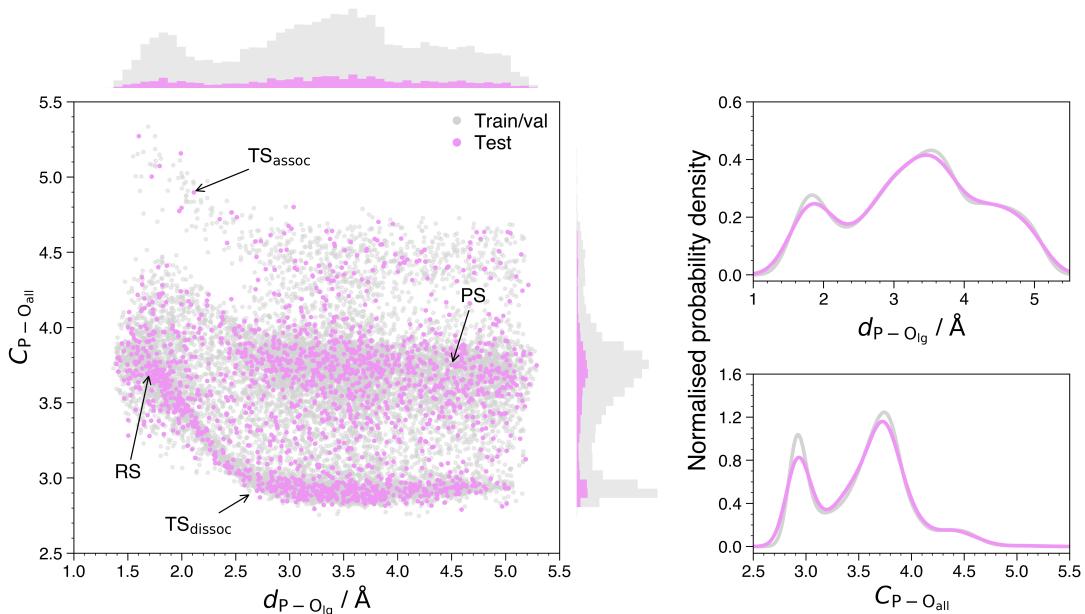


Figure 4.1: Left panel: final dataset composition projected on the two CVs space. RS stands for the reactant state, PS for the product state, and TS for the transition state. 50 bins were used to produce the histograms. Right panel: normalised densities of the two CVs for the training/validation and test sets.

The final dataset was obtained after three iterations of a learning loop, as described in Section 3.1.6. During each iteration, the dataset was expanded by adding points from different regions of the free energy surface (FES). This was achieved by imposing constraints on the CVs. For instance, in the first iteration, sampling primarily targeted the reactant and product basins, while in the second iteration, the focus shifted towards the transition state regions. The final iteration was dedicated to exploring the FES at

elevated temperatures (e.g., 320 and 340 K) in order to enhance the configurational diversity of the dataset. Sampling at higher temperatures generally improves the CV space coverage, as it allows the system to visit higher-energy regions.

The final dataset comprises 12,000 data points for the training and validation sets, and 1,800 points for the test set, as illustrated in the left panel of Figure 4.1.

The reactant basin is well-defined, appearing as a narrow region in the CV space. In contrast, the product basin is broader due to the diffusion of products within the simulation cell.

The sampling quality of the TS regions varies. The dissociative pathway is substantially better represented than the associative one. This difference arises from the fact that the associative path lies higher in energy and is therefore more difficult to access. Nevertheless, it is still represented by a number of points, meaning the fitted potential should be capable of describing it to some extent.

A particularly important aspect of dataset construction was the implementation of density-aware sampling. This approach was adopted for several reasons. First and foremost, the algorithm was used to ensure that the dataset is balanced in terms of the CVs distribution, as shown in the right panel of Figure 4.1. Secondly, it was employed to guarantee sufficient configurational diversity, i.e., the inclusion of points from various physically meaningful regions of the FES, thereby enhancing the potential's ability to generalise. This is crucial, as the neural network potential must be capable of predicting energies and forces for any configuration along the reaction coordinate. Lastly, density-aware sampling was used to ensure that the test set reflects the overall reaction space well, thereby providing a reliable basis for critically evaluating the accuracy and performance of the trained potential.

## 4.2 Accuracy and performance of the neural network potential

At each iteration, the potential was fitted using a NequIP equivariant GNN with a different tensor rank  $\ell$ , namely  $\ell = 1$  and  $\ell = 2$  ( $\ell = 0$  would correspond to an invariant GNN). The reason for using different tensor ranks was to investigate how the network complexity affects the accuracy of the potential. The tensor rank  $\ell$  determines the number of parameters in the network, with higher values leading to more complex node representations.

The final potential was trained on 12,000 data points, and its accuracy is illustrated in Figure 4.2. The left panel shows the errors in the forces and energy for tensor rank  $\ell = 1$ , while the right panel shows the corresponding errors for  $\ell = 2$ . The errors are

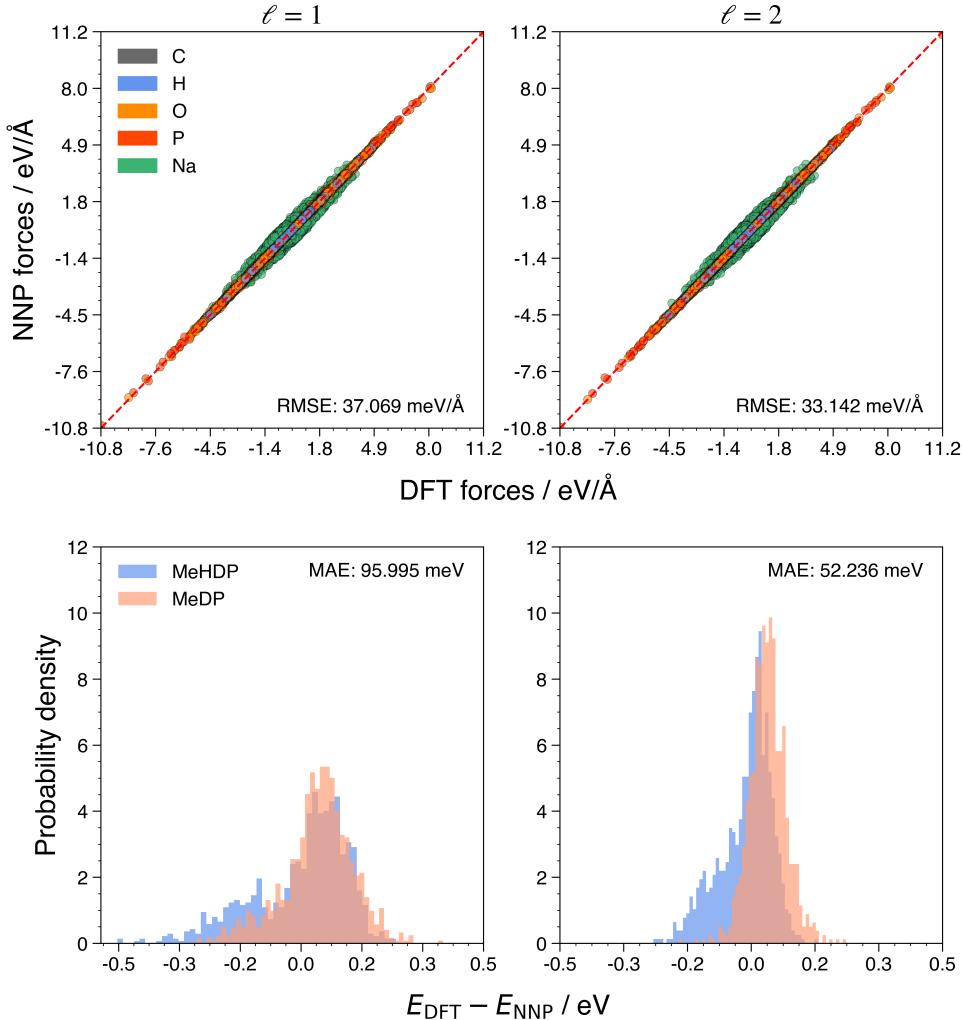


Figure 4.2: Accuracy of the neural network potential trained on 12,000 data points. The left panel shows the errors in the forces and energy for the tensor rank  $\ell = 1$ , and the right panel shows the errors for  $\ell = 2$ . For the histograms, the number of bins was set to 50.

calculated as the difference between the neural network potential and the reference DFT values.

According to current community standards [72], the accuracy of a neural network potential is considered a ‘very good fit’ when the RMSE in the forces lies within the range of 20-40 meV/Å and the mean absolute error (MAE) in the energy is between 1-10 meV/atom. A ‘very accurate fit’ is defined as the RMSE in the forces of approximately 10 meV/Å and the MAE in the energy on the order of 1 meV/atom.

The NNPs obtained in this work fall somewhere in between these two categories. The RMSE in the forces is 37.069 meV/Å for  $\ell = 1$  and 33.142 meV/Å for  $\ell = 2$ , while the MAE in the energy is below 0.3 meV/atom for  $\ell = 1$  and below 0.15 meV/atom for  $\ell = 2$ .

As shown in Figure 4.2, the errors in the forces lie along the diagonal, which rep-

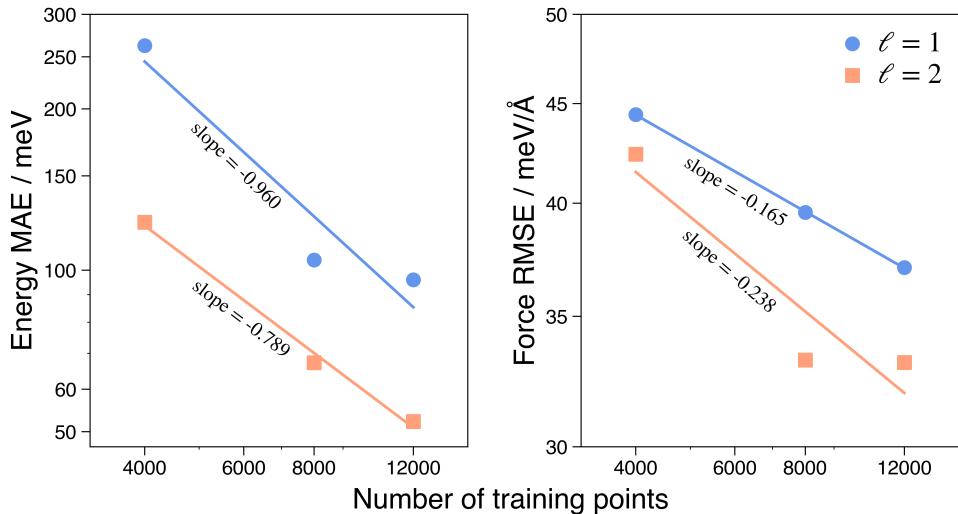


Figure 4.3: Log-log plot of the errors in the energy and forces for the neural network potential with respect to the training dataset size. In all cases, the errors were calculated on the final test set of 1,800 data points.

resents a perfect fit. The only points that are slightly scattered correspond to sodium cations ( $\text{Na}^+$ ) jiggling in solution, not being technically bound to anything. This makes it more challenging for the network to predict the forces on them.

Regarding the energy predictions, the potential with  $\ell = 1$  slightly overestimates them. This can be seen from the tail of the probability density on the left-hand side. The error appears to be systematic, meaning that when the NNP encounters new points close to those in the training data, it would produce a consistent error that may cancel out when evaluating energy differences. In contrast, the potential with  $\ell = 2$  shows normally distributed energy errors, with no significant outliers. The histograms of the errors also demonstrate that the distributions are centred around zero, indicating that the potential does not systematically over- or under-estimate the energies and forces.

Overall, both potentials are very accurate, especially considering that they were trained on a fairly small dataset of 12,000 data points. This fact supports the idea that equivariant GNNs are indeed data-efficient. For instance, a related study [73] investigated phosphoester bond formation between orthophosphate and methanol in bulk water. In that case, to achieve force errors on the order of 50 meV/Å/atom, the authors had to train an invariant neural network, DeePMD [74], on 220,000-400,000 data points.

It is important to note that the potentials are only as accurate as can be assessed by the test set, which contains 1,800 points spanning the entire FES of the reaction. The test set was not used during training nor clashes with the training points, which confirms that the NNP generalises well to unseen data.

The generalisability of the potentials can be further assessed by analysing how the

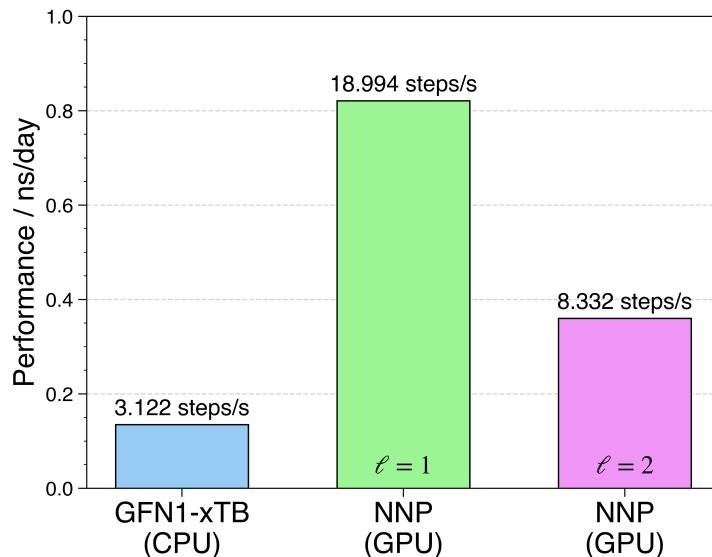


Figure 4.4: Comparison of the performance between the *ab initio* molecular dynamics runs driven by GFN1-xTB and neural network potentials fitted with different tensor ranks. CPU = 2 Intel Xeon Platinum 8468 CPUs (Sapphire Rapids), 48 cores each. GPU = 1 NVIDIA A100 80 GB GPU.

errors in the energy and forces vary with the training dataset size. Figure 4.3 shows a log-log plot of these errors for the NNP at each iteration of the learning loop. The errors were evaluated *a posteriori* on the final test set of 1,800 data points.

By examining the slopes, it becomes evident that the dataset size significantly influences how well the network learns the energies. The errors in the forces, however, are less sensitive to dataset size, which is reasonable given that the forces are not predicted directly by the network, but rather computed as the derivative of the energy with respect to atomic positions. Interestingly, the NNPs began producing sufficiently accurate results after the second training round, i.e., with 8,000 data points. This suggests that a dataset of 12,000 points is indeed sufficient to achieve a very good fit.

The next question to address is which of the two NNPs should be used in subsequent calculations. To answer this, the performance of the potentials was compared in terms of computational time required to run AIMD simulations. The results are presented in Figure 4.4. The AIMD simulations were performed on the MeDP system.

It is clear that the NNP with  $\ell = 1$  is significantly more efficient than both the  $\ell = 2$  NNP and GFN1-xTB. It is important to highlight that the number of trainable parameters in the  $\ell = 1$  NNP is 206,520, whereas for  $\ell = 2$  it is 452,280, making the latter nearly twice as slow in evaluating energies and gradients. Most likely, the performance of the PBE functional would be considerably slower - so much so that its bar would not appear on the plot.

Taking both accuracy and performance into account, the NNP with  $\ell = 1$  was se-

lected for further calculations. It is sufficiently accurate to describe the reaction mechanism, and fast enough to allow for extended AIMD simulations. The NNP with  $\ell = 2$  could be used in future work if an even more accurate potential is required.

### 4.3 Stability of the production runs

Before discussing the reaction mechanism and the free energy profiles, it is important to assess the quality of the fitted NNP in a real world scenario. To do so, we can briefly touch upon the stability of the production runs. The stability of the AIMD simulations was assessed by monitoring temperature and total energy fluctuations over time. The corresponding results are presented in Figure 4.5.

Temperature control was achieved using the Nosé–Hoover thermostat, while the total energy was calculated as the sum of potential and kinetic energies. The production runs were carried out at 300 K, and indeed, the average temperature remained close to this target, with a standard deviation of approximately 10 K. As expected for the NVT ensemble, the total energy showed some fluctuations, yet remained within reasonable bounds in the simulations for both MeDP and MeHDP systems.

It was of particular interest to evaluate the stability of the simulations driven by the NNP. These simulations remained stable, exhibiting neither system explosions nor any

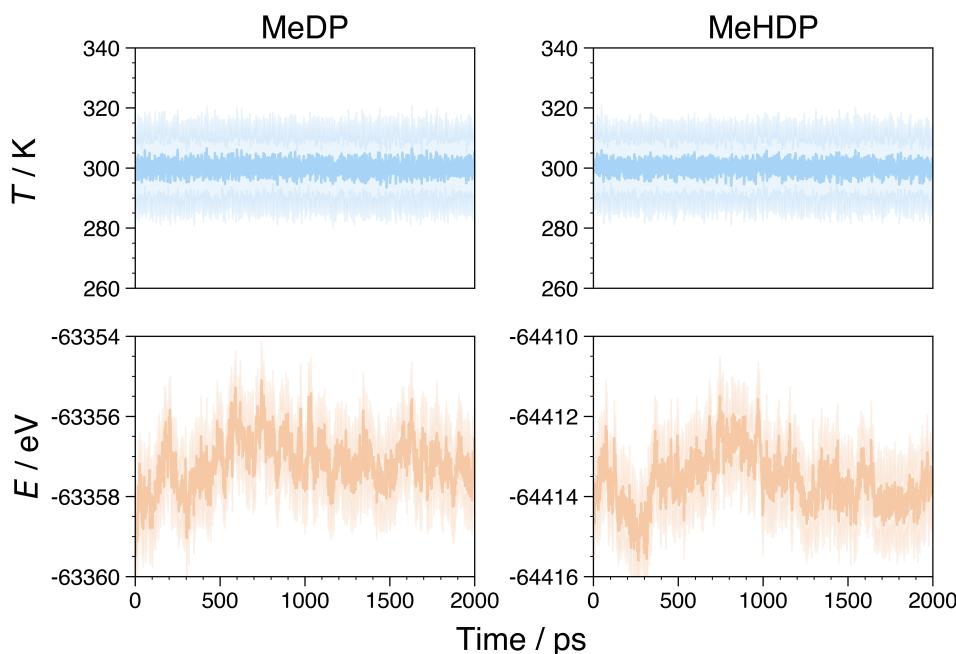


Figure 4.5: Temperature and total energy fluctuations during the production runs at 300 K. The solid line represents the average, while the shaded area indicates the standard deviation. The average and the standard deviation were calculated using a sliding window of 1 ps.

other artefacts over the 4,000,000 time steps corresponding to 2 ns. This indicates the absence of extreme force values that could have caused numerical instabilities. In a recent study [75], the authors compared the stability of AIMD simulations driven by equivariant and invariant NNPs, and found that NequIP was among the most stable potentials tested across various systems. The robustness of the fitted by NequIP NNP employed in this work is consistent with those findings.

## 4.4 Radial distribution function of water

The quality of the NNP can be further assessed by comparing the RDF of water calculated from the production runs with experimental data and previously reported computational findings. The oxygen-oxygen RDF obtained from the MeDP production run is shown in Figure 4.6.

It is evident that the NNP is in qualitative agreement with the experimental data. However, it overestimates the heights of the first two peaks and underestimates the depth of the first minimum. For the NNP, the first maximum occurs at 2.7188 Å with a height of 3.8189, while the first minimum is located at 3.2813 Å with a value of 0.2515. The second maximum appears at 4.4938 Å with a height of 1.5725. There is also a slight shift to the left in the position of the peaks in comparison to the experimental data.

One could argue that the accuracy of the fitted NNP is inherently limited by the quality of the training data. In this study, the dataset was labelled at the PBE-D3(BJ)/TZV2P

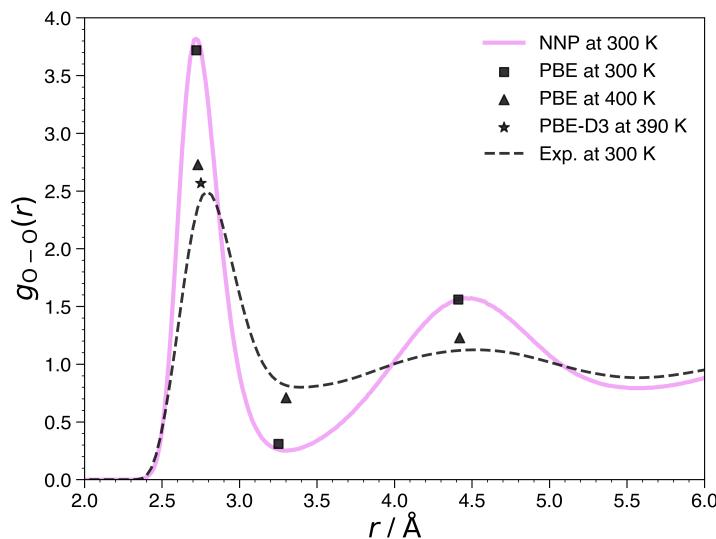


Figure 4.6: Oxygen-oxygen RDF of water calculated from a production run at 300 K. Experimental data was taken from [76]. The PBE and PBE-D3 data were taken from [77] and [78], respectively.

level of theory. Accordingly, the resulting oxygen-oxygen water RDF is consistent with the AIMD study where the RDF was calculated using the PBE functional [77] as can be found in Figure 4.6.

It has previously been shown that, in order to obtain a more accurate representation of bulk water structure, the system temperature should be increased. For instance, [78] reported that the PBE functional with the Grimme's D3 dispersion correction yields a more accurate RDF at 390 K. A similar trend was observed for the PBE functional [77], with better agreement with experimental data achieved at 400 K.

In this work, the production runs were conducted at 300 K. Consequently, the resulting structure and dynamics of water will be less accurately captured than at elevated temperatures. Nevertheless, it remains in good agreement with what would be expected from the PBE functional.

## 4.5 Convergence of the free energy profiles

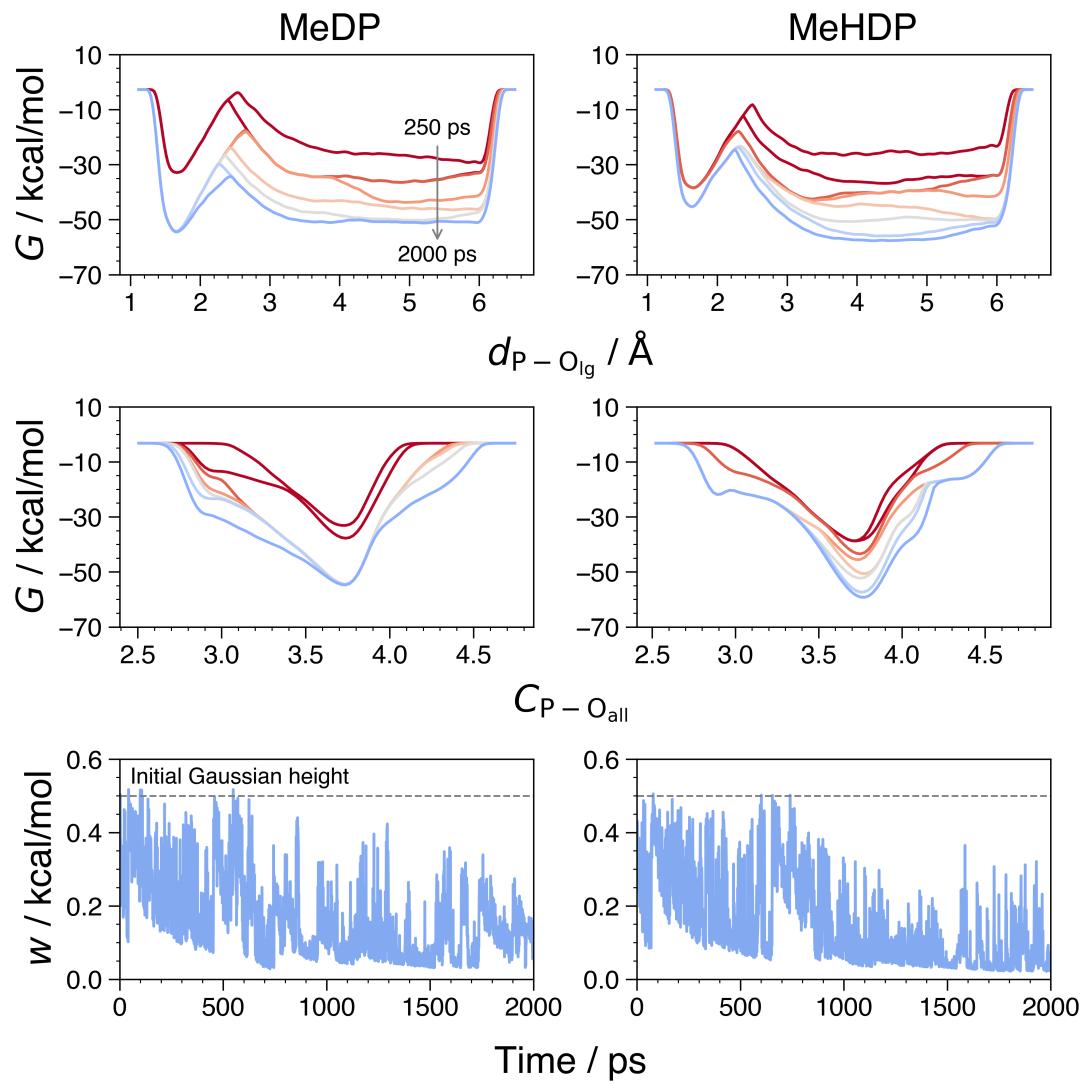


Figure 4.7: FES convergence at 300K

## 4.6 Evolution of the collective variables over time

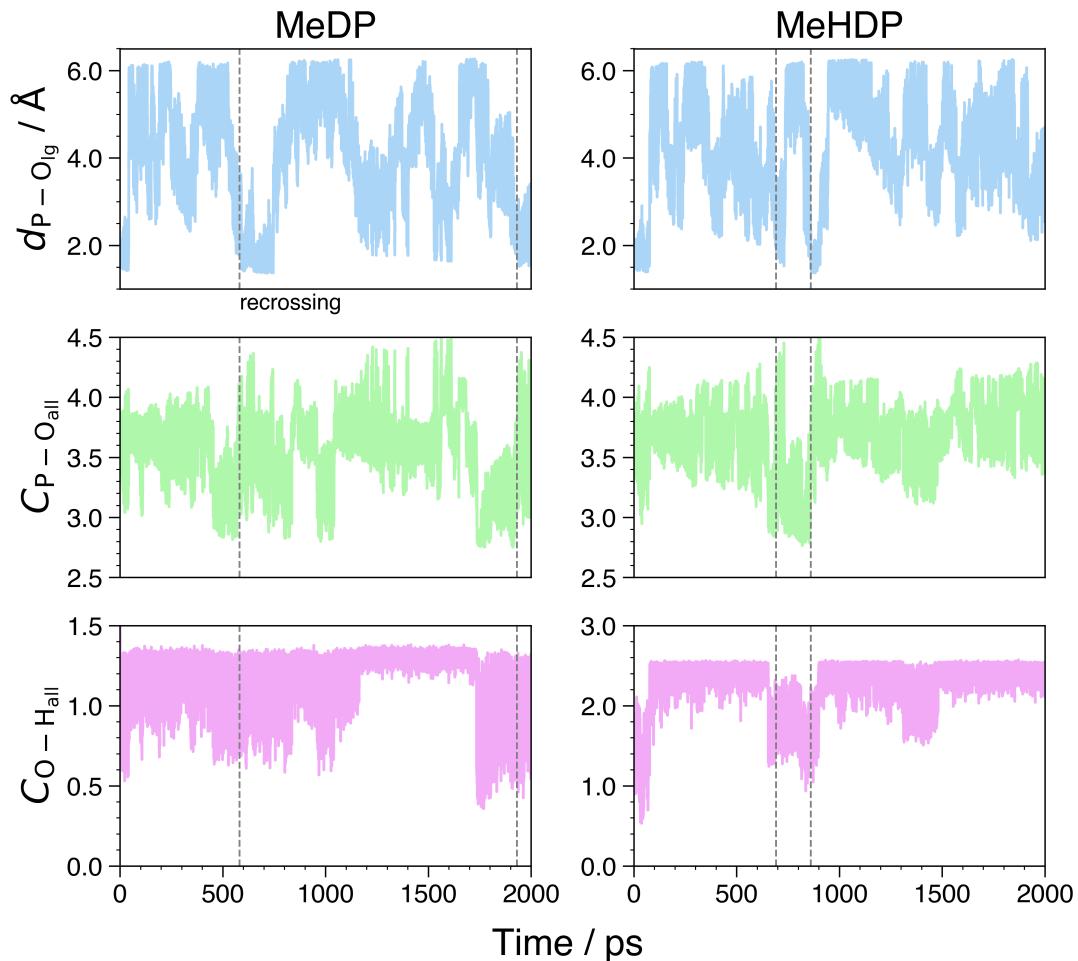


Figure 4.8: CV evolution at 300K

## 4.7 Reaction mechanism for methyl diphosphate trianion

### 4.7.1 Minimum free energy path

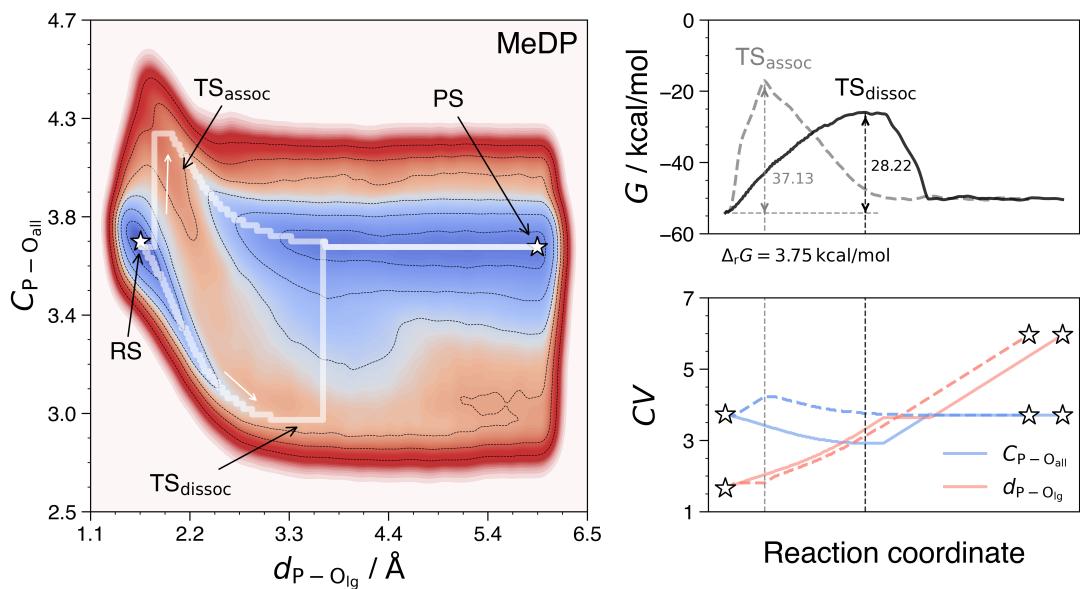


Figure 4.9: FES and MFEP MeDP 300K

### 4.7.2 Proton transfer mechanism

## 4.8 Reaction mechanism for methyl diphosphate dianion

### 4.8.1 Minimum free energy path

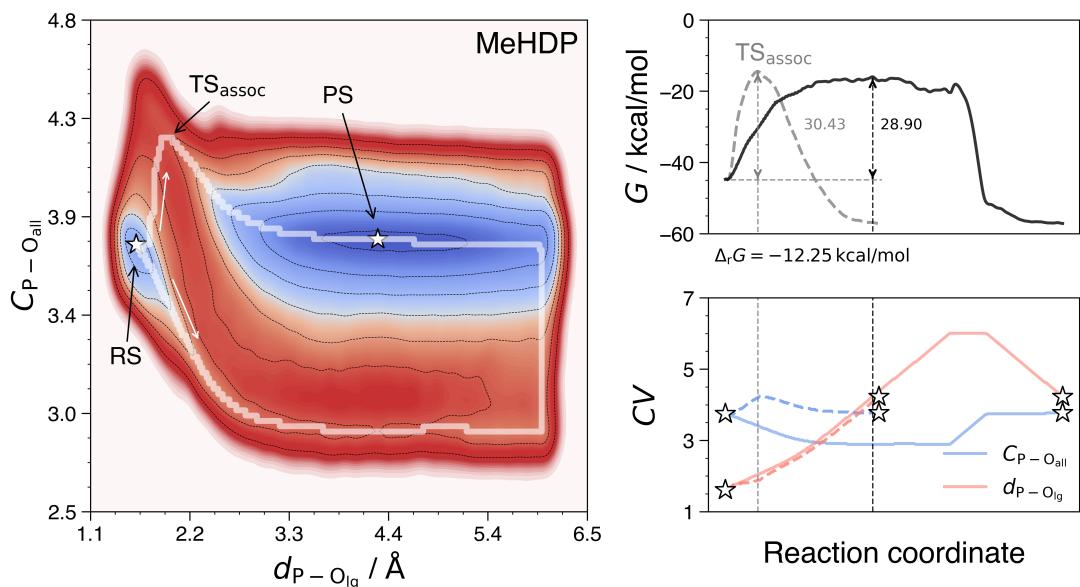


Figure 4.10: FES and MFEP MeHDP 300K

### 4.8.2 Proton transfer mechanism

# **Chapter 5**

## **Conclusions and outlook**

# Bibliography

- [1] Westheimer, F. H. Why Nature Chose Phosphates. *Science* **235**, 1173–1178 (1987).
- [2] Wolfenden, R. Degrees of Difficulty of Water-Consuming Reactions in the Absence of Enzymes. *Chemical Reviews* **106**, 3379–3396 (2006).
- [3] Müller, W. E., Schröder, H. C. & Wang, X. Inorganic Polyphosphates As Storage for and Generator of Metabolic Energy in the Extracellular Matrix. *Chemical Reviews* **119**, 12337–12374 (2019).
- [4] Nebesnaya, K. S. *et al.* Inorganic polyphosphate regulates functions of thymocytes via activation of P2X purinoreceptors. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1868**, 130523 (2024).
- [5] Kamerlin, S. C. L., Sharma, P. K., Prasad, R. B. & Warshel, A. Why nature really chose phosphate. *Quarterly Reviews of Biophysics* **46**, 1–132 (2013).
- [6] Pavlov, E. *et al.* Inorganic Polyphosphate and Energy Metabolism in Mammalian Cells \*. *Journal of Biological Chemistry* **285**, 9420–9428 (2010).
- [7] Dzeja, P. P. & Terzic, A. Phosphotransfer networks and cellular energetics. *Journal of Experimental Biology* **206**, 2039–2047 (2003).
- [8] Boyer, P. D. Energy, Life, and ATP (Nobel Lecture). *Angewandte Chemie International Edition* **37**, 2296–2307 (1998).
- [9] Bonora, M. *et al.* ATP synthesis and storage. *Purinergic Signalling* **8**, 343–357 (2012).
- [10] Walker, J. E. The ATP synthase: The understood, the uncertain and the unknown. *Biochemical Society Transactions* **41**, 1–16 (2013).
- [11] Baev, A. Y. & Abramov, A. Y. Inorganic Polyphosphate and F0F1-ATP Synthase of Mammalian Mitochondria. In Müller, W. E. G., Schröder, H. C., Suess, P. & Wang, X. (eds.) *Inorganic Polyphosphates: From Basic Research to Medical Application*, 1–13 (Springer International Publishing, Cham, 2022).

- [12] Baev, A. Y., Angelova, P. R. & Abramov, A. Y. Inorganic polyphosphate is produced and hydrolyzed in F0F1-ATP synthase of mammalian mitochondria. *Biochemical Journal* **477**, 1515–1524 (2020).
- [13] Walker, J. E. ATP Synthesis by Rotary Catalysis (Nobel lecture). *Angewandte Chemie International Edition* **37**, 2308–2319 (1998).
- [14] Watt, I. N., Montgomery, M. G., Runswick, M. J., Leslie, A. G. W. & Walker, J. E. Bioenergetic cost of making an adenosine triphosphate molecule in animal mitochondria. *Proceedings of the National Academy of Sciences* **107**, 16823–16827 (2010).
- [15] Kamerlin, S. C. L., Florián, J. & Warshel, A. Associative Versus Dissociative Mechanisms of Phosphate Monoester Hydrolysis: On the Interpretation of Activation Entropies. *ChemPhysChem* **9**, 1767–1773 (2008).
- [16] Klähn, M., Rosta, E. & Warshel, A. On the Mechanism of Hydrolysis of Phosphate Monoesters Dianions in Solutions and Proteins. *Journal of the American Chemical Society* **128**, 15310–15323 (2006).
- [17] Duarte, F., Åqvist, J., Williams, N. H. & Kamerlin, S. C. L. Resolving Apparent Conflicts between Theoretical and Experimental Models of Phosphate Monoester Hydrolysis. *Journal of the American Chemical Society* **137**, 1081–1093 (2015).
- [18] Hassan, H. A., Rani, S., Fatima, T., Kiani, F. A. & Fischer, S. Effect of protonation on the mechanism of phosphate monoester hydrolysis and comparison with the hydrolysis of nucleoside triphosphate in biomolecular motors. *Biophysical Chemistry* **230**, 27–35 (2017).
- [19] Prasad, B. R., Plotnikov, N. V. & Warshel, A. Addressing Open Questions about Phosphate Hydrolysis Pathways by Careful Free Energy Mapping. *The Journal of Physical Chemistry B* **117**, 153–163 (2013).
- [20] Akola, J. & Jones, R. O. ATP Hydrolysis in Water - A Density Functional Study. *The Journal of Physical Chemistry B* **107**, 11774–11783 (2003).
- [21] Glaves, R., Mathias, G. & Marx, D. Mechanistic Insights into the Hydrolysis of a Nucleoside Triphosphate Model in Neutral and Acidic Solution. *Journal of the American Chemical Society* **134**, 6995–7000 (2012).
- [22] Wang, C., Huang, W. & Liao, J.-L. QM/MM Investigation of ATP Hydrolysis in Aqueous Solution. *The Journal of Physical Chemistry B* **119**, 3720–3726 (2015).
- [23] Harrison, C. B. & Schulten, K. Quantum and Classical Dynamics Simulations of ATP Hydrolysis in Solution. *Journal of Chemical Theory and Computation* **8**, 2328–2335 (2012).

- [24] Ramirez, F., Marecek, J. F. & Szamosi, J. Magnesium and calcium ion effects on hydrolysis rates of adenosine 5'-triphosphate. *The Journal of Organic Chemistry* **45**, 4748–4752 (1980).
- [25] Ramirez, F., , M., James F. & and Szamosi, J. A Comparative Study of Hydrolysis Rates of 2 $\prime$ -Deoxyadenosine and Adenosine 5 $\prime$ -Triphosphates and 5 $\prime$ -Diphosphates. *Phosphorus and Sulfur and the Related Elements* **13**, 249–257 (1982).
- [26] Dirac, P. A. M. & Fowler, R. H. Quantum mechanics of many-electron systems. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **123**, 714–733 (1997).
- [27] Jensen, F. *Introduction to Computational Chemistry* (John Wiley & Sons, 2017).
- [28] Tuckerman, M. E. & Tuckerman, M. E. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford Graduate Texts (Oxford University Press, Oxford, New York, 2023), second edition edn.
- [29] Frenkel, D. & Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications* (Elsevier, San Diego, 2002), second edition edn.
- [30] Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics* **81**, 511–519 (1984).
- [31] Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* **31**, 1695–1697 (1985).
- [32] Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**, 014101 (2007).
- [33] Laio, A. & Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **99**, 12562–12566 (2002).
- [34] Laio, A. & Gervasio, F. L. Metadynamics: A method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics* **71**, 126601 (2008).
- [35] Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters* **100**, 020603 (2008).
- [36] How the Panama Canal Works. <https://waitbutwhy.com/2014/09/panama-canal-works.html>.
- [37] Sholl, D. S. & Steckel, J. A. *Density Functional Theory: A Practical Introduction* (John Wiley & Sons, 2011).

- [38] Koch, W. & Holthausen, M. C. *A Chemist's Guide to Density Functional Theory* (John Wiley & Sons, 2015).
- [39] Bursch, M., Mewes, J.-M., Hansen, A. & Grimme, S. Best-Practice DFT Protocols for Basic Molecular Computational Chemistry. *Angewandte Chemie International Edition* **61**, e202205735 (2022).
- [40] Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **77**, 3865–3868 (1996).
- [41] Grimme, S., Bannwarth, C. & Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements ( $Z = 1\text{--}86$ ). *Journal of Chemical Theory and Computation* **13**, 1989–2009 (2017).
- [42] Bannwarth, C. *et al.* Extended tight-binding quantum chemistry methods. *WIREs Computational Molecular Science* **11**, e1493 (2021).
- [43] Bishop, C. M. & Bishop, H. *Deep Learning: Foundations and Concepts* (Springer Nature, 2023).
- [44] Duval, A. *et al.* A Hitchhiker's Guide to Geometric GNNs for 3D Atomic Systems (2024). [2312.07511](https://arxiv.org/abs/2312.07511).
- [45] Batatia, I. *et al.* The Design Space of  $E(3)$ -Equivariant Atom-Centered Interatomic Potentials (2022). [2205.06643](https://arxiv.org/abs/2205.06643).
- [46] Batzner, S. *et al.*  $E(3)$ -equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications* **13**, 2453 (2022).
- [47] Musaelian, A. *et al.* Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications* **14**, 579 (2023).
- [48] Geiger, M. & Smidt, T. E3nn: Euclidean Neural Networks (2022). [2207.09453](https://arxiv.org/abs/2207.09453).
- [49] Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* **29**, 1859–1865 (2008).
- [50] Kern, N. R., Lee, J., Choi, Y. K. & Im, W. CHARMM-GUI Multicomponent Assembler for modeling and simulation of complex multicomponent systems. *Nature Communications* **15**, 5459 (2024).
- [51] Kim, S. *et al.* CHARMM-GUI ligand reader and modeler for CHARMM force field generation of small molecules: CHARMM-GUI Ligand Reader and Modeler for CHARMM Force Field Generation of Small Molecules. *Journal of Computational Chemistry* **38**, 1879–1886 (2017).

- [52] Haynes, W. M. *CRC Handbook of Chemistry and Physics* (CRC Press, 2016).
- [53] Bernetti, M. & Bussi, G. Pressure control using stochastic cell rescaling. *The Journal of Chemical Physics* **153**, 114107 (2020).
- [54] Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
- [55] Huang, J. *et al.* CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14**, 71–73 (2017).
- [56] Kühne, T. D. *et al.* CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics* **152**, 194103 (2020).
- [57] Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Computer Physics Communications* **185**, 604–613 (2014).
- [58] Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **132**, 154104 (2010).
- [59] Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry* **32**, 1456–1465 (2011).
- [60] Goedecker, S., Teter, M. & Hutter, J. Separable dual-space Gaussian pseudopotentials. *Physical Review B* **54**, 1703–1710 (1996).
- [61] Hartwigsen, C., Goedecker, S. & Hutter, J. Relativistic separable dual-space Gaussian pseudopotentials from H to Rn. *Physical Review B* **58**, 3641–3662 (1998).
- [62] VandeVondele, J. *et al.* Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Computer Physics Communications* **167**, 103–128 (2005).
- [63] CP2K\_Developers. How to Converge the CUTOFF and REL\_CUTOFF. <https://manual.cp2k.org/trunk/methods/dft/cutoff.html>.
- [64] Makhmudov, A. Cp2k\_to\_extxyz: A script that helps in converting CP2K single-point calculations into an .extxyz file. [https://github.com/almakhmudov/cp2k\\_to\\_extxyz](https://github.com/almakhmudov/cp2k_to_extxyz).
- [65] Thompson, A. P. *et al.* LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications* **271**, 108171 (2022).
- [66] Mir-group/pair\_nequip. [https://github.com/mir-group/pair\\_nequip](https://github.com/mir-group/pair_nequip).

- [67] Morón, M. C., Prada-Gracia, D. & Falo, F. Macro and nano scale modelling of water–water interactions at ambient and low temperature: Relaxation and residence times. *Physical Chemistry Chemical Physics* **18**, 9377–9387 (2016).
- [68] Marcos-Alcalde, I., Setoain, J., Mendieta-Moreno, J. I., Mendieta, J. & Gómez-Puertas, P. MEPSA: Minimum energy pathway analysis for energy landscapes. *Bioinformatics* **31**, 3853–3855 (2015).
- [69] Granger, B. E. & Pérez, F. Jupyter: Thinking and Storytelling With Code and Data. *Computing in Science & Engineering* **23**, 7–14 (2021).
- [70] Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90–95 (2007).
- [71] Humphrey, W., Dalke, A. & Schulter, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33–38 (1996).
- [72] Jacobs, R. *et al.* A practical guide to machine learning interatomic potentials – Status and future. *Current Opinion in Solid State and Materials Science* **35**, 101214 (2025).
- [73] Benayad, Z., David, R. & Stirnemann, G. Prebiotic chemical reactivity in solution with quantum accuracy and microsecond sampling using neural network potentials. *Proceedings of the National Academy of Sciences* **121**, e2322040121 (2024).
- [74] Zeng, J. *et al.* DeePMD-kit v2: A software package for deep potential models. *The Journal of Chemical Physics* **159**, 054801 (2023).
- [75] Fu, X. *et al.* Forces are not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations (2023). [2210.07237](https://arxiv.org/abs/2210.07237).
- [76] Soper, A. K. The Radial Distribution Functions of Water as Derived from Radiation Total Scattering Experiments: Is There Anything We Can Say for Sure? *International Scholarly Research Notices* **2013**, 279463 (2013).
- [77] Pham, T. A., Ogitsu, T., Lau, E. Y. & Schwegler, E. Structure and dynamics of aqueous solutions from PBE-based first-principles molecular dynamics simulations. *The Journal of Chemical Physics* **145**, 154501 (2016).
- [78] Zhou, K., Qian, C. & Liu, Y. Quantifying the Structure of Water and Hydrated Monovalent Ions by Density Functional Theory-Based Molecular Dynamics. *The Journal of Physical Chemistry B* **126**, 10471–10480 (2022).

# Appendix A

## Supplementary information

Table A.1: The plane-wave cutoff convergence test for DFT calculations. The calculation of  $\Delta E$  involves subtracting the previous energy, e.g.  $\Delta E(450 \text{ Ry}) = E(450 \text{ Ry}) - E(400 \text{ Ry})$ . When the cutoff  $\geq 800$  and the rel cutoff  $\geq 60$ , the error in total energy reduces to ca.  $10^{-8}$  a.u. Only part of the results is shown for the sake of clarity.

| Cutoff (Ry) | Rel cutoff (Ry) | Total energy (a.u.) | $\Delta E$ (a.u.)      |
|-------------|-----------------|---------------------|------------------------|
| 400         | 60              | -2352.6355962810    | –                      |
| 450         | 60              | -2352.6262868887    | $9.31 \times 10^{-3}$  |
| 500         | 60              | -2352.6262867349    | $1.54 \times 10^{-7}$  |
| 550         | 60              | -2352.6254866602    | $8.00 \times 10^{-4}$  |
| 600         | 60              | -2352.6243443853    | $1.14 \times 10^{-3}$  |
| 650         | 60              | -2352.6242425582    | $1.02 \times 10^{-4}$  |
| 700         | 60              | -2352.6224669798    | $1.78 \times 10^{-3}$  |
| 750         | 60              | -2352.6209571227    | $1.51 \times 10^{-3}$  |
| 800         | 60              | -2352.6212901605    | $-3.33 \times 10^{-4}$ |
| 850         | 60              | -2352.6212901727    | $-1.22 \times 10^{-8}$ |
| 900         | 60              | -2352.6212901873    | $-1.46 \times 10^{-8}$ |
| 950         | 60              | -2352.6213082173    | $-1.80 \times 10^{-5}$ |
| 1000        | 60              | -2352.6208957304    | $4.12 \times 10^{-4}$  |
| 10          | 800             | -2354.4562984779    | –                      |
| 20          | 800             | -2352.6775968461    | 1.78                   |
| 30          | 800             | -2352.6281701514    | $4.94 \times 10^{-2}$  |
| 40          | 800             | -2352.6213637375    | $6.81 \times 10^{-3}$  |
| 50          | 800             | -2352.6212892865    | $7.45 \times 10^{-5}$  |
| 60          | 800             | -2352.6212901605    | $-8.74 \times 10^{-7}$ |
| 70          | 800             | -2352.6212901729    | $-1.24 \times 10^{-8}$ |
| 80          | 800             | -2352.6212901739    | $-1.00 \times 10^{-9}$ |
| 90          | 800             | -2352.6212901739    | 0.00                   |
| 100         | 800             | -2352.6212901739    | 0.00                   |

**Algorithm A.1** Density-aware sampling of configurations**Input:** Feature matrix  $\mathbf{X} \in \mathbb{R}^{N \times 2}$  of  $N$  configurations, number of samples  $n_{\text{samples}}$ 

1: Determine number of clusters:

$$k \leftarrow \max \left( 10, \min \left( \left\lfloor \frac{N}{50} \right\rfloor, \left\lfloor \frac{n_{\text{samples}}}{10} \right\rfloor \right) \right)$$

2: Apply K-means clustering to  $\mathbf{X}$  with  $k$  clusters3: Initialize empty list for sampled configuration indices  $S \leftarrow []$ 4: **for** each cluster  $C_i$ ,  $i = 1$  to  $k$  **do**

5:    $n_i \leftarrow \max \left( 1, \left\lfloor \frac{|C_i|}{N} \cdot n_{\text{samples}} \right\rfloor \right)$

6:   Select  $n_i$  random configurations from  $C_i$  with fixed random seed7:   Append selected indices to  $S$ 8: **end for**9: **return**  $S$ **Output:** List of selected configuration indices  $S$ 

Table A.2: Composition of the full dataset used for training and testing. Well-tempered meta-dynamics settings used to run the simulations: <sup>1</sup>GPN1-xTB for energies and forces, gaussian height = 2 kcal/mol, spawning frequency = 25 fs<sup>-1</sup>, bias factor = 30 and <sup>2</sup>NNP for energies and forces, gaussian height = 2 kcal/mol, spawning frequency = 50 fs<sup>-1</sup>, bias factor = 30.

| System             | Temperature (K) | Simulation length (ps) | Train/Val    | Test        |
|--------------------|-----------------|------------------------|--------------|-------------|
| MeDP <sup>1</sup>  | 300             | 50 ps                  | 2000         | 150         |
| MeDP <sup>2</sup>  | 300             | 100 ps                 | 2000         | 150         |
| MeDP <sup>2</sup>  | 320             | 500 ps                 | 1000         | 300         |
| MeDP <sup>2</sup>  | 340             | 500 ps                 | 1000         | 300         |
| MeHDP <sup>1</sup> | 300             | 50 ps                  | 2000         | 150         |
| MeHDP <sup>2</sup> | 300             | 100 ps                 | 2000         | 150         |
| MeHDP <sup>2</sup> | 320             | 500 ps                 | 1000         | 300         |
| MeHDP <sup>2</sup> | 340             | 500 ps                 | 1000         | 300         |
| <b>Total</b>       |                 |                        | <b>12000</b> | <b>1800</b> |

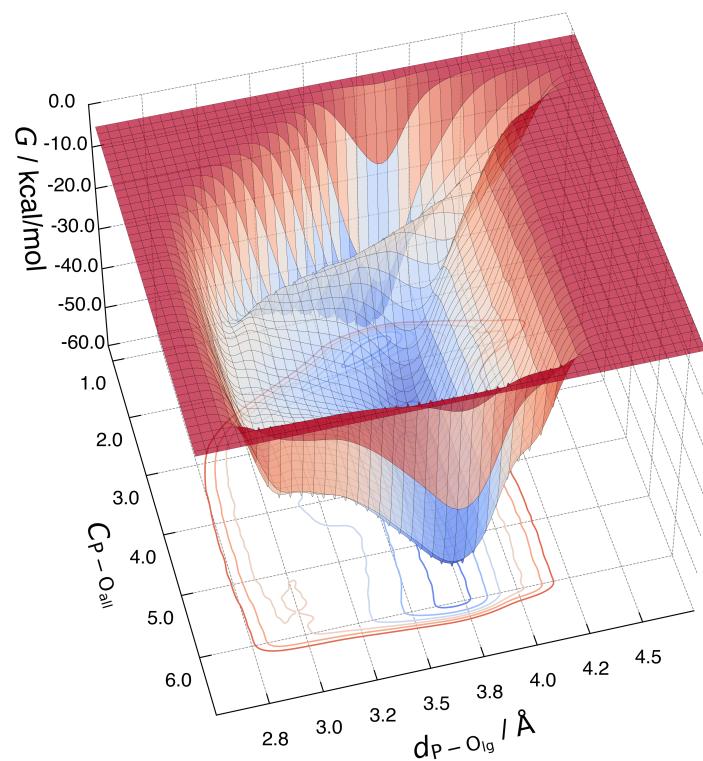


Figure A.1: 3D FES for MeDP at 300 K.

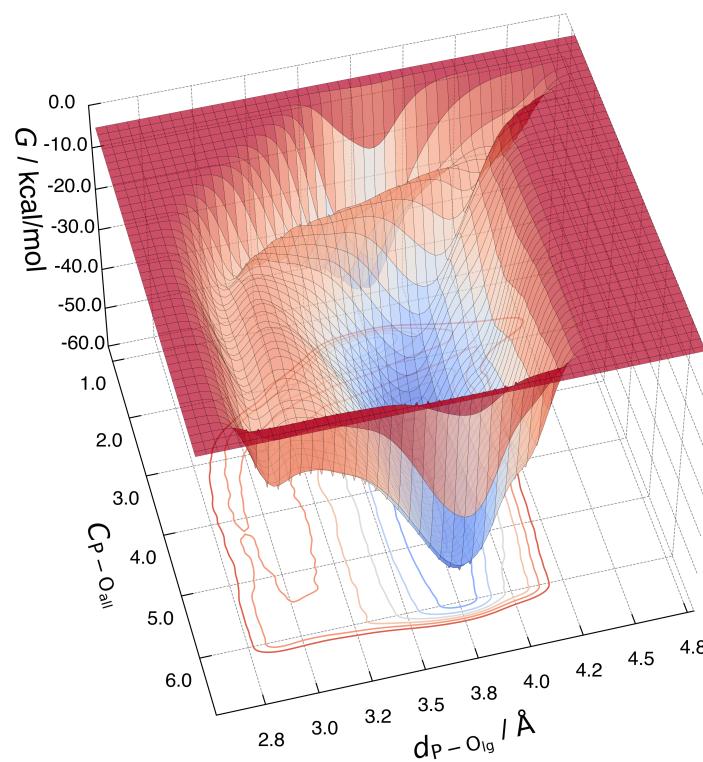


Figure A.2: 3D FES for MeHDP at 300 K.

**Quantum Chemistry and Physical Chemistry**

Celestijnenlaan 200F bus 2404

3001 LEUVEN, BELGIË

tel. + 32 16 37 21 98

[jeremy.harvey@kuleuven.be](mailto:jeremy.harvey@kuleuven.be)

[www.kuleuven.be](http://www.kuleuven.be)

