KU LEUVEN

FACULTY OF SCIENCE

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

Foreword

Contribution statement

Summary

List of abbreviations

Contents

Chapter 1

Introduction

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℃ 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 β and γ 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 ($\Delta G^0 = -30.5$ kJ mol^{-1}), and under cellular conditions, the actual free energy release is often considerably greater. Specifically, the experimentally obtained ΔG values are approximately -59 to -53.5 kJ mol^{-1} in the liver and about -61.7 to -59.5 kJ mol^{-1} in the heart [3].

Beyond ATP, inorganic polyphosphate (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, 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

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
Nicotine 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].

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 (Δp) 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 essence, 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 adenosine triphosphate (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, 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 inorganic polyphosphate (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

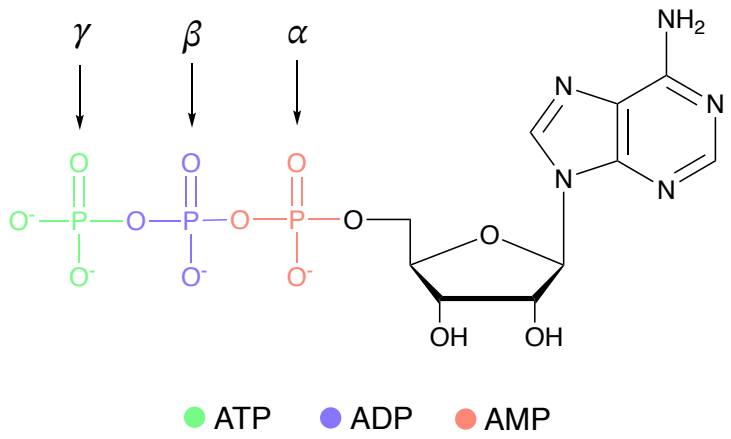


Figure 1.1: Chemical structures of the AMP, ADP, and ATP molecules with the phosphates marked as α, β, and γ, respectively.

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 β subunits, which interact with the α 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 γ, δ, and ϵ 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 (Δp). This reverse function is especially important under conditions of low Δp, 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 β and γ in Figure 1.1) follows the principles of the binding change mechanism [10]. The rotation of the asymmetric γ subunit induces sequential conformational changes in the three β 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 that releases ADP and P_i [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 polyphosphate (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 immunocaptured F_0F_1-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_0F_1-ATP synthase inhibitor. Medium- and long-chain polyP molecules, made of 60 and 130

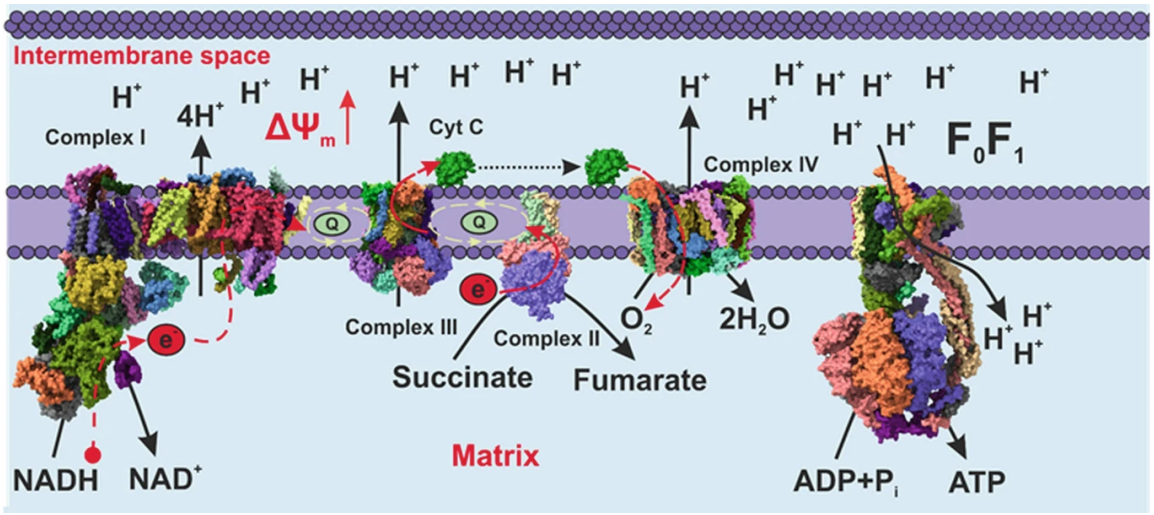


Figure 1.2: Electron transport chain coupled with oxidative phosphorylation in mitochondria. Reproduced from [11].

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_1 domain. This suggests that polyP could serve as an alternative energy source for the F_0F_1 complex, potentially helping to maintain mitochondrial membrane potential when ATP levels are compromised [11, 12].

There is growing confidence that the F_1-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_1-ATPase in particular—is both fascinating and complex. The F_1-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

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 (MeMP) 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 (A_N + D_N, where A_N stands for nucleophilic addition and D_N for nucleophilic departure) or concerted (A_ND_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 (A_N + 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 occurs to one of the non-bridging oxygens. The reaction proceeds through a compact pentacoordinated transition state with a trigonal bipyramidal geometry, followed by the elimination of the leaving group in a subsequent transition state.

Regarding the associative/concerted mechanism (A_ND_N), it proceeds in a manner

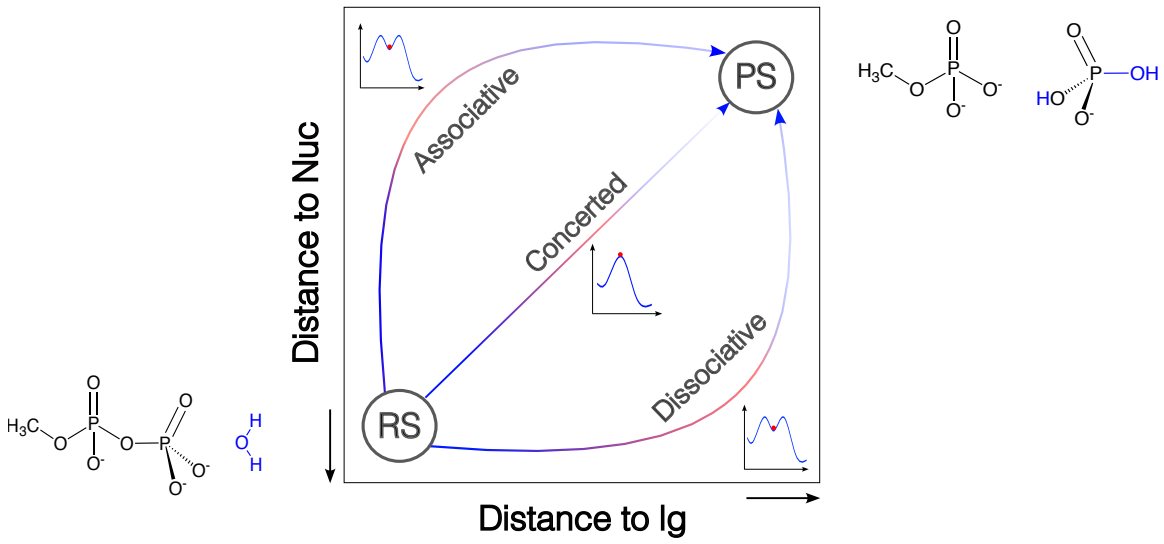


Figure 1.3: More O'Ferrall-Jencks (MFJ) plot of the possible reaction mechanisms for phosphate hydrolysis. The plot shows the free energy as a function of two reaction coordinates: the distance between phosphorus and the nucleophile (Nuc), and the distance between the leaving group (lg) and the phosphorus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

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 (MeHMP), the calculated barrier height $\Delta G^{\ddagger}_{calc}$ is approximately 6-7 kcal/mol lower than that of the methyl phosphate dianion (MeMP). A similar effect was observed when OH^- acted as a nucleophile instead of a water molecule [16] (40 vs 47 kcal/mol, respectively).

For the associative mechanism, calculated barrier heights $\Delta G^{\ddagger}_{calc}$ 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_ND_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

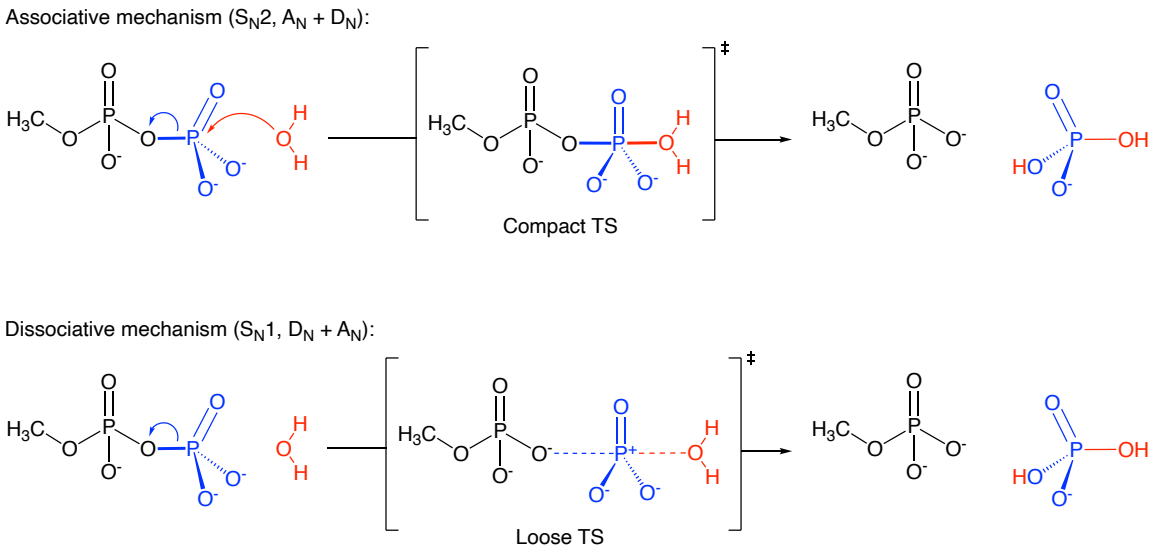


Figure 1.4: Associative and dissociative reaction mechanisms. For the transition states, only one of the two is shown. The nucleophile (Nuc) is shown in red, the leaving group (lg) in black, and the phosphoryl group in yellow.

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. Nevertheless, the dissociative mechanism is unlikely to take place, or at least it has not been observed.

1.3.2 Diphosphates

1.3.3 Triphosphates

In summary, computational investigations reveal a nuanced picture of phosphate hydrolysis. The preferred mechanism (associative/concerted/dissociative) and proton transfer route (1W/2W/substrate-assisted) depend significantly on the specific substrate (leaving group pKa), its protonation state, the presence of metal ions like Mg^{2+}, and the solvation model used. Near-zero activation entropies do not uniquely identify dissociative pathways. QM/MM free energy methods are essential for navigating the complex potential energy surfaces and accurately determining reaction barriers and mechanisms.

1.4 Research goals

Table 1.2: Summary of computational and experimental studies on phosphate hydrolysis. In case of barrier height, the rate-limiting step is given.

System	Method	Level of theory	Mechanism	$\Delta G^{\ddagger}$ (kcal/mol)	Ref.
$MeMP^{2-}$ + H_2O	DFT	B3LYP/6-311+G** and COSMO	Associative Concerted	47.2 44.5	[15]
$MeMP^{2-}$ + H_2O	DFT	B3LYP/6-311++G** and COSMO	Associative	47	[16]
$MeMP^{2-}$ + 3 H_2O	DFT	M06-2X/6-311+G** and SMD	Associative Concerted	$\approx$ 36 $\approx$ 44	[17]
$MeMP^{2-}$ + 4 H_2O	DFT	M06-2X/6-311+G**	Associative	$\approx 40.8 \pm 1.9$	[18]
$MeHMP^{-}$ + 4 H_2O	DFT	M06-2X/6-311+G**	Associative	$\approx 33.7 \pm 1.7$	[18]
$MeDP^{3-}$ + 2 H_2O	DFT	B3LYP/6-311++G** and PCM	Associative Concerted	34.64 35.24	[19]
$MeDP^{3-}$ + H_2O	DFT	B3LYP/6-311+G** and COSMO	Associative Dissociative	34.8 30.3	[15]
$MeDP^{3-}$ + H_2O	DFT	B3LYP/6-311++G** and COSMO	Associative Dissociative	38 34	[16]
$MeHDP^{2-}$ + H_2O	DFT	B3LYP/6-311++G** and COSMO	Associative Dissociative	34 31	[16]
ATP^{4-} + Mg^{2+} + 4163 H_2O + counterions	QM/MM, NEB	B3LYP/6-311++G** and MM	Concerted	32.5	[20]
$MeDP^{3-}$ + Mg^{2+} + 5 H_2O	QM/MM, FEP (EVB)	B3LYP/6-311++G** and MM	Associative Concerted Dissociative	35 34 35	[16]
ATP^{4-} + Mg^{2+} + 1800 H_2O + counterions	QM/MM, QM = CPMD	BLYP/PW with Troullier-Martins pseudopotentials and MM	Associative Dissociative	36.2 33.4	[21]
$MeTP^{4-}$ + Mg^{2+} + 54 H_2O	CPMD	PBE/PW with Troullier-Martins pseudopotentials	Associative Flexible Dissociative	39.1 35.1 36.6	[22]
$MeTP^{4-}$ + Mg^{2+} + 113 H_2O	BOMD, metadynamics	BLYP/TZV2P with GTH pseudopotentials	Associative (acidic solution) Concerted (neutral solution)	29-30 29	[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^{2-} (or ATP^{3-})	Exp. at 25 °C	–	–	27.5	[2]
$ATPH^{3-}$ (or ATP^{4-})	Exp. at 70 °C	pH=6.69-7.66	–	24.34-24.78	[24]
$dADPH^{2-}$ (or $dADP^{3-}$)	Exp. at 70 °C	pH=6.82	–	24.25	[25]
$dATPH^{3-}$ (or $dATP^{4-}$)	Exp. at 70 °C	pH=7.00	–	24.50	[25]
$MgATPH^{-}$ (or $MgATP^{2-}$)	Exp. at 70 °C	pH=6.59-7.63	–	24.59-24.64	[24]
$CaATPH^{-}$ (or $CaATP^{2-}$)	Exp. at 70 °C	pH=6.67-7.01	–	25.71-25.72	[24]

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

System	Mechanism	TS_1		Intermediate		TS_2		Ref.
		d(P-O_{Nuc})	**d(P-O_{lg})**	**d(P-O_{Nuc})**	**d(P-O_{lg})**	**d(P-O_{Nuc})**	**d(P-O_{lg})**	
$MeMP^{2-}$	Associative (A_ND_N)	2.0	1.8	–	–	–	–	[16]
	Associative (A_ND_N)	1.9	2.15	–	–	–	–	[15]
	Associative ($A_N + D_N$)	2.16	1.71	1.84	1.78	1.71	2.24	[17]
	Associative ($A_N + D_N$)	2.08	1.78	1.99	1.80	1.77	2.52	[18]
	Concerted (A_ND_N)	2.75	2.75	–	–	–	–	[15]
	Concerted (A_ND_N)	2.06	2.61	–	–	–	–	[17]
$MeHMP^{-}$	Associative ($A_N + D_N$)	2.26	1.66	1.76	1.77	1.68	2.25	[18]

Chapter 2

Theory

2.1 A brief introduction to statistical mechanics

2.1.1 Classical forcefields and molecular dynamics

2.1.2 The canonical ensemble and free energy calculations

2.1.3 Free energy techniques

2.2 Transition state theory

2.3 Density functional theory

2.3.1 The Kohn-Sham approach

2.3.2 Generalised gradient approximation and PBE functional

2.3.3 *Ab initio* molecular dynamics and GPW method

2.4 Extended tight binding

2.5 Neural network potentials

2.5.1 Deep neural networks

Multilayer perceptron

Graph neural networks

Message passing neural networks

2.5.2 Invariance and equivariance

2.5.3 Behler-Parrinello neural network potentials

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 various temperatures using the fitted neural network potential. The third section describes the workflow for validating the transition states obtained from the simulations, based on the partial Hessian formalism. Finally, the fourth 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 [26], specifically the Multicomponent Assembler interface [27].

As a first step, the singly protonated and deprotonated forms of methyl diphosphate were parameterised using CGenFF [28], i.e., the CHARMM General Force Field. These protonation states were chosen based on the dissociation constants of pyrophosphoric (diphosphoric) acid [29]:

$$\begin{aligned} H_4P_2O_7 &\rightleftharpoons [H_3P_2O_7]^- + H^+, \quad pK_a = 0.91 \\ [H_3P_2O_7]^- &\rightleftharpoons [H_2P_2O_7]^{2-} + H^+, \quad pK_a = 2.10 \\ [H_2P_2O_7]^{2-} &\rightleftharpoons [HP_2O_7]^{3-} + H^+, \quad pK_a = 6.70 \\ [HP_2O_7]^{3-} &\rightleftharpoons [P_2O_7]^{4-} + H^+, \quad pK_a = 9.32 \end{aligned}$$

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 singly protonated (MeHDP) and deprotonated (MeDP) forms under physiological conditions.

Following successful parameterisation, the system was solvated in a cubic box of TIP3P water molecules, with sodium counterions (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 [26]. Initially, energy minimisation was conducted using the steepest descent algorithm for 5,000 steps.

This was followed by equilibration in the NVT (constant number of particles, volume, and temperature) 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 kJ mol^{-1} nm^{-2}.

Subsequently, the system was equilibrated in the NPT (constant number of particles, pressure, and temperature) 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 ν-rescale thermostat [30] with a coupling constant of 1 ps, and pressure was regulated using an isotropic *c*-rescale barostat [31] 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. Periodic boundary conditions (PBC) were applied in all directions throughout the simulation.

All simulations were carried out using GROMACS 2021.4 [32] with CHARMM36m force field [33]. The leap-frog integrator was employed with a time step of 1 fs. All hydrogen-involving bonds were constrained using the LINCS algorithm. The final 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	Equilibrated box dimensions ($Å^3$)	No. of H_2O	No. of Na^+
MeDP	$15.877 \times 15.877 \times 15.877$	119	3
MeHDP	$15.901 \times 15.901 \times 15.901$	124	2

3.1.3 Collective variables

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

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

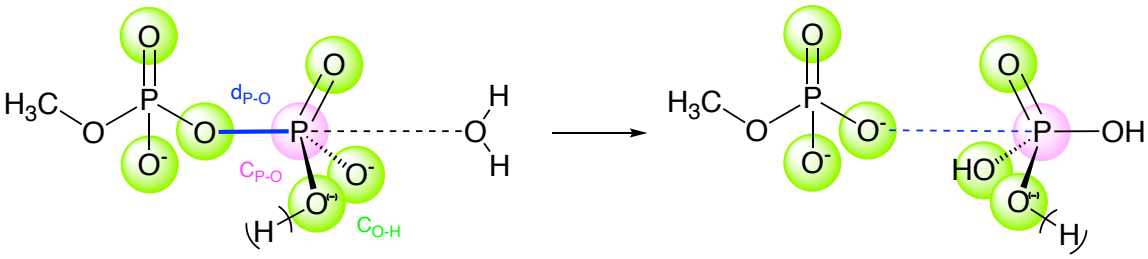
where r_{ij} is the distance between atoms i and j from groups A and B, d_0 is the distance at which the CN begins to decay, r_0 is a characteristic decay length, and n and m are integers that control the steepness of the decay. Typically, $m > n$, ensuring a smooth transition of CN_{ij} from approximately 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 β-phosphorus and the bridging oxygen (CV_1, d($O_{remaining}$ - $P_{leaving}$)),
- Coordination number of all oxygen atoms surrounding the β-phosphorus (CV_2, CN($P_{leaving}$ - O_{all})): $d_0 = 0$, $r_0 = 2.1$ Å, $n = 8$, $m = 16$,
- Coordination number of non-methyl hydrogen atoms around the oxygen atoms bonded to the β-phosphorus (CV_3, CN($O_{leaving}$ - H_{all})): $d_0 = 0$, $r_0 = 1.4$ Å, $n = 6$, $m = 12$.

Additionally, the following CVs were monitored to estimate the number of H_3O^+ and OH^- species in solution:

- Number of H_3O^+ ($n_{H_3O^+}$): TODO,



$d_{P\text{-}O}$ = d(P - O_{lg}), $C_{P\text{-}O}$ = CN(P - O_{all}), $C_{O\text{-}H}$ = CN(O - H_{all})

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

- Number of OH^- (n_{OH^-}): TODO.

To avoid sampling 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:

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

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

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 [34] or PLUMED 2.9.3 [35]. 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. [1]Different values for the walls were used depending on the system MeDP/MeHDP.

CV	Lower wall	Upper wall	Force constant (kcal mol^{-1} Å^{-2})
$d_{P\text{-}O}$	–	5	500
$C_{O\text{-}H_{all}}$	– / 1.2[1]	1.2 / 2.2[1]	1000
$C_{P\text{-}O_{metaphosphate}}$	2.6	–	2000
$C_{P\text{-}O_{water}}$	–	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 [36] level of theory. GFN1-xTB provides a good first approximation of the potential energy surface and is computationally efficient, thus making it suitable for relatively long MD simulations of large systems.

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 D3 dispersion correction [37]. Following equilibration, we performed 50 ps of well-tempered metadynamics (WTMD) [38] 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 collective variables (CVs): the distance between the β-phosphorus and the oxygen atom connecting it to the rest of the molecule (CV_1), and the coordination number of all oxygens surrounding the β-phosphorus atom (CV_2).

All calculations were carried out using the CP2K 2023.1 package [34]. Temperature control was achieved using the ν-rescale thermostat [30], with a time constant of 50 fs during equilibration and 100 fs during the WTMD 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^{-1} 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 all simulations, periodic boundary conditions 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 (PBE) exchange-correlation functional [39], along with the D3 dispersion correction and the Becke-Johnson damping function [37, 40]. In all calculations, the Goedecker-Teter-Hutter (GTH) pseudopotentials [41, 42] were used to represent the core electrons, in combination with the triple-ζ valence basis set with two polarisation functions (TZV2P).

The single-point calculations were performed using the Gaussian Plane Wave (GPW) method implemented in the QUICKSTEP module [43] of the CP2K 2023.1 package [34]. 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 [44]. 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.

3.1.6 Iterative training of the neural network potential

We trained a neural network potential using the NequIP framework [45], 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 ℓ = 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 β_1 = 0.9, β_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 1×10^{-5}, or if it exceeded 1×10^{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.

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 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 1,200 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

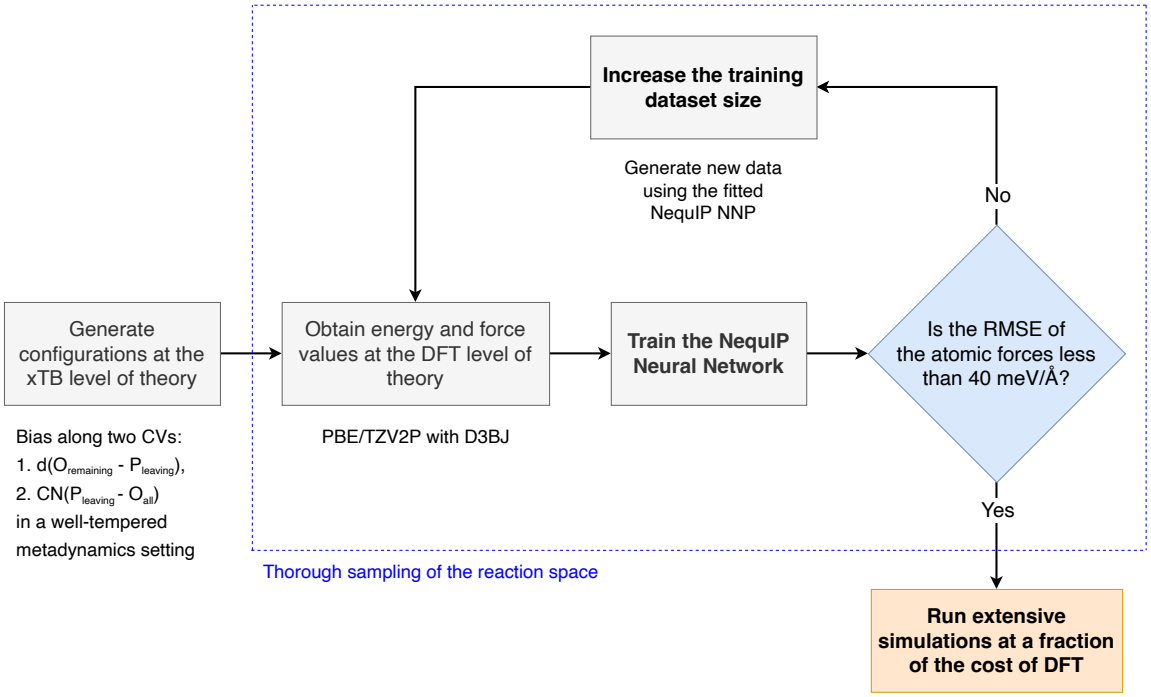


Figure 3.2: Iterative training of the NequIP neural network potential.

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 $d(\mathrm{O}_{\mathrm{remaining}} - \mathrm{P}_{\mathrm{leaving}})$ and the coordination number $\mathrm{CN}(\mathrm{P}_{\mathrm{leaving}} - \mathrm{O}_{\mathrm{all}})$.

The two CVs were combined into a two-dimensional feature space $\mathbf{X} = (d, \mathrm{CN})$, 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 =

$\max\left(10,\ \min\left(\left\lfloor \frac{N}{50} \right\rfloor, \left\lfloor \frac{n_{\text{samples}}}{10} \right\rfloor\right)\right)$, 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 neural network potential, 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 CV_1 and CV_2, 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 neural network potential (NNP) obtained after the first round of training.

To run the simulations with the NNP, the LAMMPS package [46] compiled with PLUMED 2.9.3 [35] and `pair_nequip` [47] was used. The simulations were performed for 100 ps in the NVT ensemble at 300 K. The temperature was controlled by a Nosé–Hoover thermostat [48, 49] with a time constant of 50 fs. The biasing potential was

applied to CV_1 and CV_2 every 50 fs, using a Gaussian hill height of 2 kcal mol^{-1} 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 CV_1 and CV_2 in order to favour either a dissociative or associative mechanism of the reaction and sample more configurations from the transition state 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. 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 different temperatures

3.3 Validation of the transition states

3.4 Lifetime of the transition states

3.5 Data analysis and visualisation

Chapter 4

Results and Discussion

4.1 Accuracy of the neural network potential

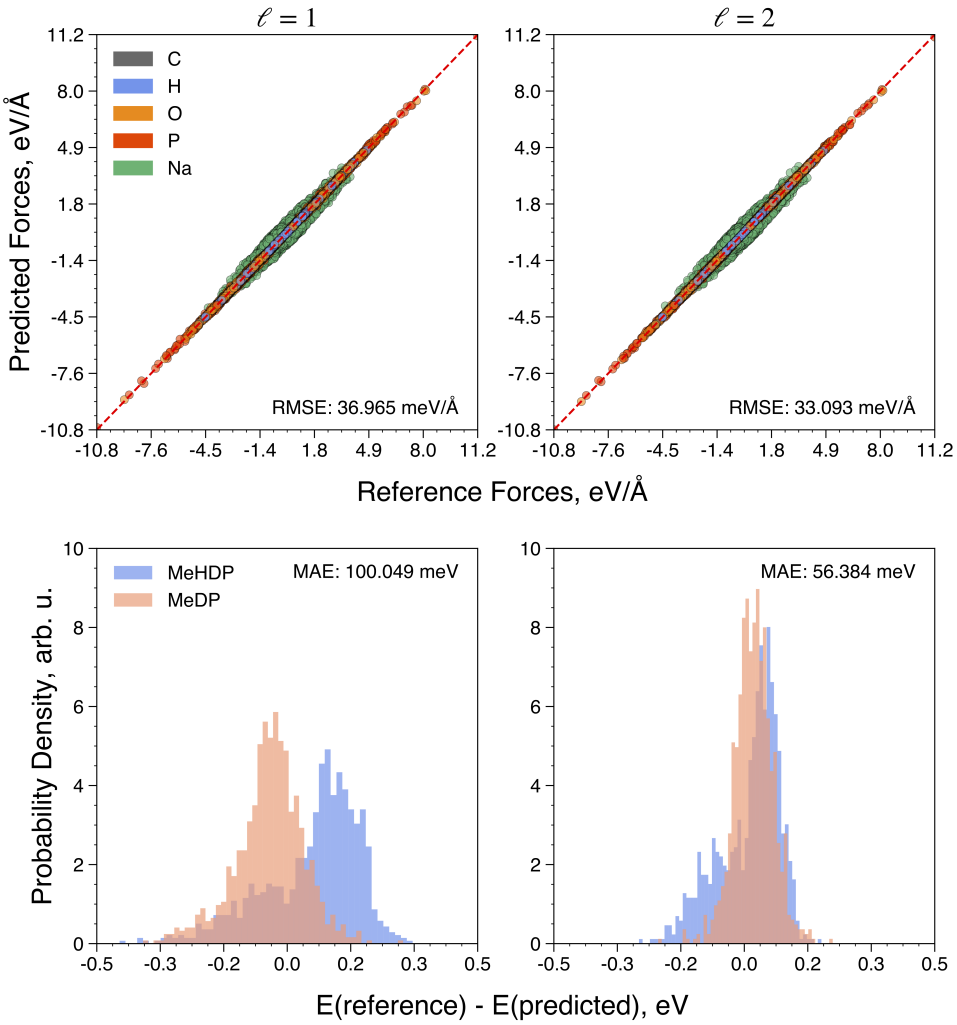


Figure 4.1: Accuracy of the neural network potential trained on 12,000 data points. The left panel shows the errors in the forces and energy for the tensor rank $\ell = 1$ and the right panel shows the errors for $\ell = 2$. The errors are calculated as the difference between the neural network potential and the reference DFT values. For the histograms, the number of bins was set to 50.

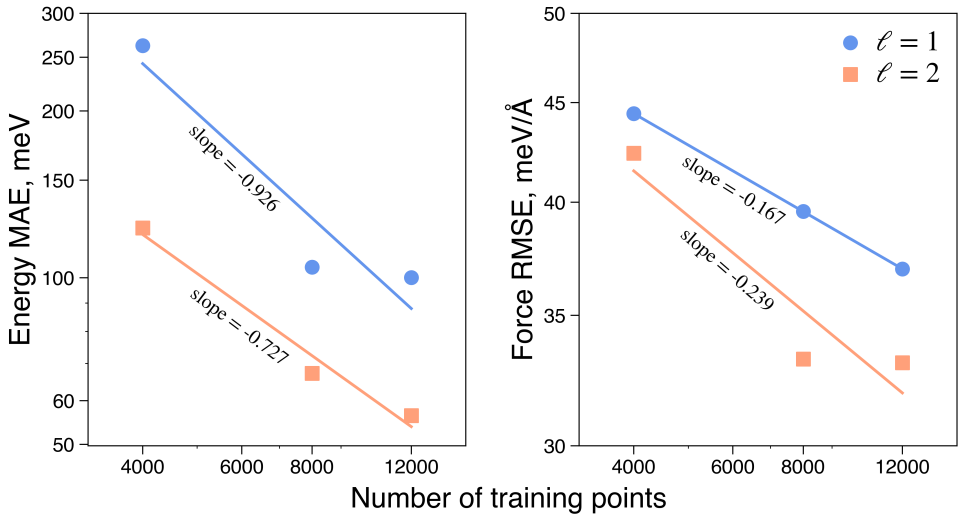


Figure 4.2: Log-log plot of the errors in the energy and forces for the neural network potential with the respect to the training dataset size. In all cases, the errors were calculated on the final test set of 1,200 data points.

Chapter 5

Conclusions

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] 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).

[21] 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).

[22] Akola, J. & Jones, R. O. ATP Hydrolysis in Water - A Density Functional Study. *The Journal of Physical Chemistry B* **107**, 11774–11783 (2003).

[23] 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).

[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*ı*-Deoxyadenosine and Adenosine 5*ı*-Triphosphates and 5*ı*-Diphosphates. *Phosphorus and Sulfur and the Related Elements* **13**, 249–257 (1982).

[26] 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).

[27] 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).

[28] 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).

[29] Haynes, W. M. *CRC Handbook of Chemistry and Physics* (CRC Press, 2016).

[30] Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**, 014101 (2007).

[31] Bernetti, M. & Bussi, G. Pressure control using stochastic cell rescaling. *The Journal of Chemical Physics* **153**, 114107 (2020).

[32] Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

[33] Huang, J. *et al.* CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14**, 71–73 (2017).

[34] 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).

[35] 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).

[36] 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 = 1–86). *Journal of Chemical Theory and Computation* **13**, 1989–2009 (2017).

[37] 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).

[38] Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters* **100**, 020603 (2008).

[39] Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **77**, 3865–3868 (1996).

[40] 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).

[41] Goedecker, S., Teter, M. & Hutter, J. Separable dual-space Gaussian pseudopotentials. *Physical Review B* **54**, 1703–1710 (1996).

[42] Hartwigsen, C., Goedecker, S. & Hutter, J. Relativistic separable dual-space Gaussian pseudopotentials from H to Rn. *Physical Review B* **58**, 3641–3662 (1998).

[43] 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).

[44] CP2K_Developers. How to Converge the CUTOFF and REL_CUTOFF. https://manual.cp2k.org/trunk/methods/dft/cutoff.html.

[45] Batzner, S. *et al.* E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications* **13**, 2453 (2022).

[46] 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).

[47] Mir-group/pair_nequip. https://github.com/mir-group/pair_nequip.

[48] Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics* **81**, 511–519 (1984).

[49] Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* **31**, 1695–1697 (1985).

Appendix A

Supplementary information

Table A.1: The plane-wave cutoff convergence test for DFT calculations. The calculation of ΔE involves subtracting the previous energy, e.g. $\Delta E(450\,\mathrm{Ry}) = E(450\,\mathrm{Ry}) - E(400\,\mathrm{Ry})$. When the cutoff ≥ 800 and the rel cutoff ≥ 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.)**	**ΔE (a.u.)**
400	60	-2352.6355962810	–
450	60	-2352.6262868887	9.31×10^{-3}
500	60	-2352.6262867349	1.54×10^{-7}
550	60	-2352.6254866602	8.00×10^{-4}
600	60	-2352.6243443853	1.14×10^{-3}
650	60	-2352.6242425582	1.02×10^{-4}
700	60	-2352.6224669798	1.78×10^{-3}
750	60	-2352.6209571227	1.51×10^{-3}
800	60	-2352.6212901605	-3.33×10^{-4}
850	60	-2352.6212901727	-1.22×10^{-8}
900	60	-2352.6212901873	-1.46×10^{-8}
950	60	-2352.6213082173	-1.80×10^{-5}
1000	60	-2352.6208957304	4.12×10^{-4}
10	800	-2354.4562984779	–
20	800	-2352.6775968461	1.78
30	800	-2352.6281701514	4.94×10^{-2}
40	800	-2352.6213637375	6.81×10^{-3}
50	800	-2352.6212892865	7.45×10^{-5}
60	800	-2352.6212901605	-8.74×10^{-7}
70	800	-2352.6212901729	-1.24×10^{-8}
80	800	-2352.6212901739	-1.00×10^{-9}
90	800	-2352.6212901739	0.00
100	800	-2352.6212901739	0.00

Algorithm A.1 Density-aware sampling of configurations

Require: Feature matrix $\mathbf{X} \in \mathbb{R}^{N\times 2}$ of N configurations, number of samples n_{samples}
Ensure: List of selected configuration indices

Determine number of clusters:

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

Apply K-means clustering to $\mathbf{X}$ with k clusters
Initialize empty list for sampled cofiguration indices $S \leftarrow [\,]$
for each cluster C_i, $i = 1$ to k **do**
    $n_i \leftarrow \max\left(1, \left\lfloor \frac{|C_i|}{N} \cdot n_{\text{samples}} \right\rfloor\right)$
    Select n_i random configurations from C_i with fixed random seed
    Append selected indices to S
end for
return S

Table A.2: Composition of the full dataset used for training and testing. Well-tempered metadynamics settings used to run the simulations: [1]GFN1-xTB for energies and forces, gaussian height = 2 kcal/mol, spawning frequency = 25 fs, bias factor = 30 and [2]NNP for energies and forces, gaussian height = 2 kcal/mol, spawning frequency = 50 fs, bias factor = 30.

System	Temperature (K)	Simulation length (ps)	Train/Val	Test
MeDP[1]	300	50 ps	2000	150
MeDP[2]	300	100 ps	2000	150
MeDP[2]	320	500 ps	1000	150
MeDP[2]	340	500 ps	1000	150
MeHDP[1]	300	50 ps	2000	150
MeHDP[2]	300	100 ps	2000	150
MeHDP[2]	320	500 ps	1000	150
MeHDP[2]	340	500 ps	1000	150
Total			**12000**	**1200**

Quantum Chemistry and Physical Chemistry
Celestijnenlaan 200F bus 2404
3001 LEUVEN, BELGIË
tel. + 32 16 37 21 98
jeremy.harvey@kuleuven.be
www.kuleuven.be