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

© Copyright by KU Leuven

Without written permission of the promotors 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

Contribution statement

Summary

List of abbreviations

Contents

1 Introduction **1**
- 1.1 Role of phosphates in biological systems 1
- 1.2 Enzymes involved in phosphate hydrolysis 1
- 1.3 Reaction mechanism 1
- 1.4 Research goals 1

2 Theory **2**
- 2.1 A brief introduction to statistical mechanics 3
  - 2.1.1 Classical forcefields and molecular dynamics 3
  - 2.1.2 The canonical ensemble and free energy calculations 3
  - 2.1.3 Enhanced sampling techniques 3
- 2.2 Density functional theory 3
  - 2.2.1 The Kohn-Sham approach 3
  - 2.2.2 Generalised gradient approximation and PBE functional 3
  - 2.2.3 *Ab initio* molecular dynamics and GPW method 3
- 2.3 Extended tight binding 3
- 2.4 Neural network potentials 3
  - 2.4.1 Deep neural networks 3
  - 2.4.2 Invariance and equivariance 3
  - 2.4.3 Behler-Parrinello neural network potentials 3
  - 2.4.4 Equivariant neural network potentials 3

3 Computational Details **4**
- 3.1 Training dataset generation 4
  - 3.1.1 System preparation 4
  - 3.1.2 Initial equilibration using the classical forcefields 5
  - 3.1.3 Collective variables 6
  - 3.1.4 GFN1-xTB based exploration of the configuration space 8
  - 3.1.5 Data labeling 9
  - 3.1.6 Iterative training of the neural network potential 9

3.2 Production runs at different temperatures 9
3.3 Validation of the transition states 9
3.4 Lifetime of the transition states 9
3.5 Data analysis and visualisation 9

4 Results and Discussion 10

5 Conclusions 11

Bibliography 12

A Supplementary information 14

Chapter 1

Introduction

1.1 Role of phosphates in biological systems

1.2 Enzymes involved in phosphate hydrolysis

1.3 Reaction mechanism

1.4 Research goals

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 Enhanced sampling techniques

Metadynamics and its well-tempered flavour

Kinetics from metadynamics

2.2 Density functional theory

2.2.1 The Kohn-Sham approach

2.2.2 Generalised gradient approximation and PBE functional

2.2.3 *Ab initio* molecular dynamics and GPW method

2.3 Extended tight binding

2.4 Neural network potentials

2.4.1 Deep neural networks

Multilayer perceptron

Graph neural networks

Message passing neural networks

2.4.2 Invariance and equivariance

2.4.3 Behler-Parrinello neural network potentials

Chapter 3

Computational Details

This chapter provides the details of the computational methods used in this work. The first section describes the generation of the training dataset, including the preparation of the system, initial equilibration using the molecular mechanics, exploration of the configuration space at the xTB level, further data labeling, and iterative training of the neural network potential. The second section discusses the production runs at different temperatures using the fitted neural network potential. The third section describes the workflow of validating the transition states obtained from the simulations following the partial Hessian formalism. Finally, the fourth section presents the data analysis and visualisation techniques employed to interpret the results.

3.1 Training dataset generation

3.1.1 System preparation

The systems were prepared using the CHARMM-GUI webserver's functionality [1]. In particular, the Multicomponent Assembler interface [2] was utilised.

As a first step, the singly protonated and deprotonated forms of the methyl diphosphate were parametrised in CGenFF [3], i.e. CHARMM General Forcefield. These states of the methyl diphosphate were chosen based on the fact that pyrophosphoric (diphosphoric) acid has the following dissociation constants [4]:

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

Thus, at the physiological pH of 7.4 this acid exists as an equillibrium between the doubly and singly protonated forms. As an assumption, the methyl group can be considered as a proton, therefore we condsidered the methyl diphosphate molecule to exist as a mixture of the singly (MeHDP) and deprotonated (MeDP) forms at the physiological pH.

After succesfully parametrising the molecules, the system was solvated in a cubic box of TIP3 water molecules together with the sodium counterions Na^+ to neutralise the charge. The final system composition can be seen in Table 3.1.

3.1.2 Initial equilibration using the classical forcefields

The equilibration of the system was performed following the standard protocol generated by the CHARMM-GUI webserver [1]. The system was first energy minimised using the steepest descent algorithm for 5000 steps.

Subsequently the system was equillibrated in the NVT (constant number of particles, volume, and temperature) ensemble for 5 ns. During the minimisation and NVT equilibration, the heavy atoms of the solute were restrained using a harmonic potential with a force constant of 400 kJ mol^{-1} nm^{-2}.

As a last step, the system was equilibrated in the NPT (constant number of particles, pressure, and temperature) ensemble for 45 ns. Throughout the whole protocol, the temperature was set to 300 K and the pressure was set to 1 bar. To ensure the constant temperature and pressure, the system was coupled to a ν-rescale thermostat with a coupling constant of 1 ps and an isotropic *c*-rescale barostat with a coupling constant of 5 ps. During the NPT run, the cut-off for the non-bonded interactions was set to 0.6 nm and the long-range electrostatics were treated using the Particle Mesh Ewald (PME) method.

All simulations were conducted in GROMACS 2021.4 [5] using the CHARMM36m forcefield [6] and the leap-frog integration method with a time step of 1 fs. All hydrogen involving bonds were constrained using the LINCS algorithm. The final dimensions of the box for all further calculations were obtained after the NPT run and are shown in Table 3.1. The last frame of the NPT runs was used as a starting point in all further calculations unless stated otherwise.

Table 3.1: System composition and simulation box details. ^{1}The final dimensions were obtained after the NPT run using the CHARMM36m forcefield.

System	Equillibrated box dimensions^1 (Å)	No. of water molecules	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

In order to properly sample the reaction space, we used two types of collective variables (CVs) to bias the system, e.g. distances and coordination numbers (CNs). The coordination number is defined as the smooth coordination number 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 the atoms i and j from the groups A and B, d_0 is a shift in the distance where CN starts to decay, r_0 is a scaling parameter or a characteristic decay length that defines how fast the CN drops off with distance, and n and m are the positive integers that control the steepness of the function. Normally, $m > n$ and the purpose of these parameters is to control how quickly the neighbours stop contributing as the distance increases. Therefore, the CN_{ij} ranges from ≈ 1 when $r_{ij} \leq r_0$ and ≈ 0 when $r_{ij} \geq r_0$. The CVs used to bias the systems in this work are shown in Figure 3.1. The corresponding parameters to describe each collective variable are as follows:

- The distance between the β-phosphorus and the oxygen atom connecting it to the rest of the molecule (CV_1, d($O_{remaining}$ - $P_{leaving}$),
- The coordination number of all oxygens around the β-phosphorus atom (CV_2, CN($P_{leaving}$ - O_{all}): d_0 = 0, r_0 = 2.1 Å, n = 8, m = 16,
- The coordination number of all hydrogens excluding the methyl ones around the oxygen atoms bound to the β-phosphorus atom (CV_3, CN($O_{leaving}$ - H_{all}): d_0 = 0, r_0 = 1.4 Å, n = 6, m = 12.

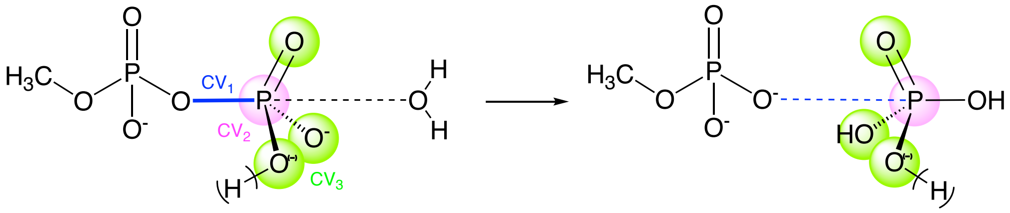
Besides the above mentioned CVs that were biased during the simulations, we also used the following collective variables to track the number of H_3O^+ and OH^- in the solution:

- The number of H_3O^+ ($n_{H_3O^+}$): TODO,
- The number of OH^- (n_{OH^-}): TODO.

To prevent the system from exploring the unphysical regions of the potential energy surface, we additionally applied quadratic (harmonic-like) wall potentials to softly restrain the accessible regions of some of the degrees of freedom.

The aforementioned wall potentials have the following form:

$$\text{for UPPER WALLS: } \sum_i k_i \left(\frac{CV_i - a_i + o_i}{s_i}\right)^{e_i} \tag{3.2}$$



CV_1 = d($O_{remaining}$ - $P_{leaving}$), CV_2 = CN($P_{leaving}$ - O_{all}), CV_3 = CN($O_{leaving}$ - H_{all})

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

$$\text{for LOWER WALLS: } \sum_i k_i \left| \frac{CV_i - a_i - o_i}{s_i} \right|^{e_i} \tag{3.3}$$

where CV_i is the value of the collective variable, k_i is the force constant (wall strength) that controls how strong the wall pushes back, a_i is the position of the wall, o_i is the offset to shift the wall slightly, s_i is the rescaling factor, and e_i is the exponent that determines the sharpness of the wall. When $e_i = 2$ the wall is harmonic-like.

Thus this potential is equal to zero while the CV is within the boundaries but once the CV tries to go beyond the boundaries, the potential increases and penalises the system.

The wall potentials were applied to the following collective variables:

- CV_1, d($O_{remaining}$ - $P_{leaving}$): k = 500 kcal mol^{-1} Å^{-2}, a = 5.0 Å(upper wall), o = 0 Å, s = 1 Å, e = 2),
- TODO (CONSIDER TRANSFORMING INTO A TABLE)

All CV related computations were carried out using the CP2K 2023.1 in-built tools [7] or PLUMED 2.9.3 [8]. It's important to note that the number of the CVs used in the simulations varies depending on the part of the overall workflow. In the next sections we will clearly mention the respective collective variables used in each part of the workflow.

3.1.4 GFN1-xTB based exploration of the configuration space

To generate the first set of the configurations for the training dataset, the system was subjected to molecular dynamics simulations using the semi-empirical GFN1-xTB [9] level of theory. GFN1-xTB gives a good first approximation of the potential energy surface and is computationally efficient hence making it suitable for relatively long MD simulations of big systems.

Each system was first equilibrated for 5 ps in the NVT ensemble at 300 K to relax the structures at the GFN1-xTB level of theory with a D3 dispersion correction [10]. Afterwards, we performed 50 ps long well-tempered metadynamics (WTMD) simulations in the NVT ensemble as well. In the latter simulation, the biasing potential was applied to force the system to explore the configuration space outside of the reactants basin. The biasing potential was added along two collective variables (CV): 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 around the β-phosphorus atom (CV_2).

All calculations were performed using the CP2K 2023.1 package [7]. The temperature was controlled by means of the velocity-rescaling thermostat [11] with a time constant of 50 fs for the equillibration and 100 fs for the WTMD run. The SCF convergence was set to 10^{-5} a.u. The biasing potential was applied 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. Lastly, the time step for the integration was set to 0.5 fs. Throughout all simulations, the periodic boundary conditions (PBC) were applied in all directions.

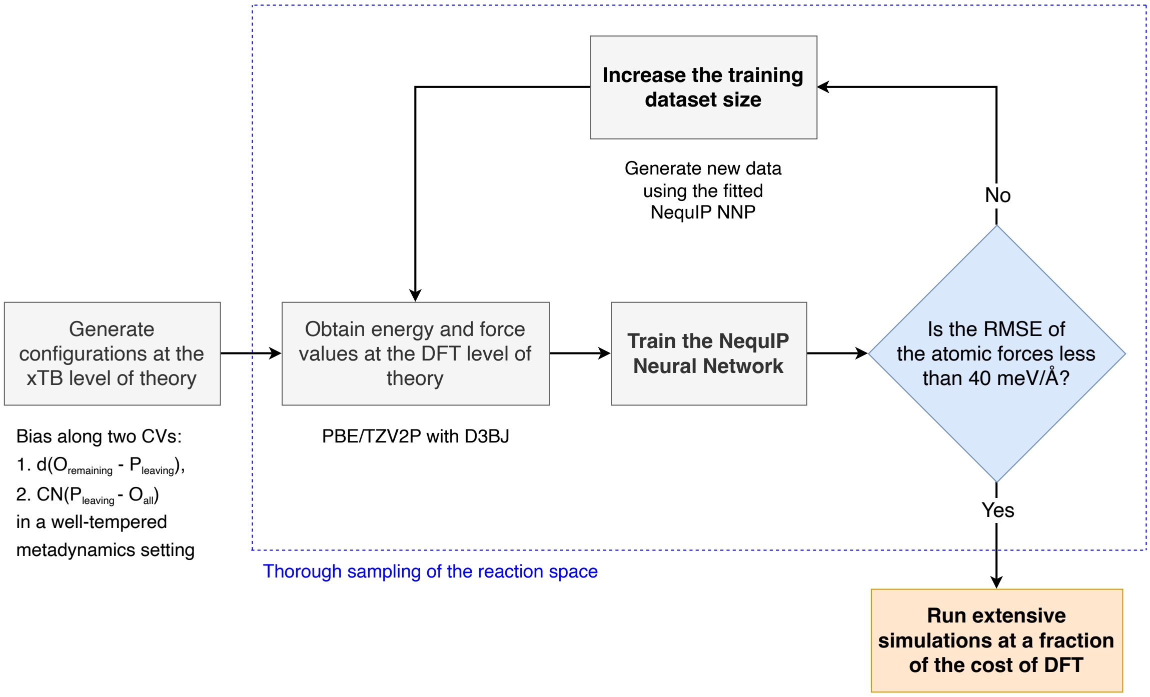

Increase the training dataset size
Generate new data using the fitted NequIP NNP
No
Generate configurations at the xTB level of theory
Obtain energy and force values at the DFT level of theory
Train the NequIP Neural Network
Is the RMSE of the atomic forces less than 40 meV/Å?
Bias along two CVs:
1. d(O$_{remaining}$ - P$_{leaving}$),
2. CN(P$_{leaving}$ - O$_{all}$)
in a well-tempered metadynamics setting
PBE/TZV2P with D3BJ
Yes
Thorough sampling of the reaction space
Run extensive simulations at a fraction of the cost of DFT

Figure 3.2: The iterative training of the NequIP neural network potential.

3.1.5 Data labeling

3.1.6 Iterative training of the neural network potential

First round

Second round

Third round

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

Chapter 5

Conclusions

Bibliography

[1] 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). URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20945`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.20945.

[2] 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). URL `https://www.nature.com/articles/s41467-024-49700-4`. Publisher: Nature Publishing Group.

[3] 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). URL `https://onlinelibrary.wiley.com/doi/10.1002/jcc.24829`.

[4] Haynes, W. M. *CRC Handbook of Chemistry and Physics* (CRC Press, 2016). Google-Books-ID: VVezDAAAQBAJ.

[5] Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015). URL `https://www.sciencedirect.com/science/article/pii/S2352711015000059`.

[6] Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* **14**, 71–73 (2017). URL `https://www.nature.com/articles/nmeth.4067`. Publisher: Nature Publishing Group.

[7] 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). URL `https://doi.org/10.1063/5.0007045`.

[8] 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). URL `https://www.sciencedirect.com/science/article/pii/S0010465513003196`.

[9] 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). URL `https://doi.org/10.1021/acs.jctc.7b00118`. Publisher: American Chemical Society.

[10] 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). URL `https://doi.org/10.1063/1.3382344`.

[11] Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**, 014101 (2007). URL `https://doi.org/10.1063/1.2408420`.

Appendix A

Supplementary information

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