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 classical force fields 5
  - 3.1.3 Collective variables 6
  - 3.1.4 GFN1-xTB based exploration of the configuration space 7
  - 3.1.5 Data labeling 8
  - 3.1.6 Iterative training of the neural network potential 9

3.2 Production runs at different temperatures 12
3.3 Validation of the transition states 12
3.4 Lifetime of the transition states 12
3.5 Data analysis and visualisation 12

4 Results and Discussion 13

5 Conclusions 14

Bibliography 15

A Supplementary information 18

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 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 [1], specifically the Multicomponent Assembler interface [2].

As a first step, the singly protonated and deprotonated forms of methyl diphosphate were parameterised using CGenFF [3], i.e., the CHARMM General Force Field. These protonation states were chosen based on the dissociation constants of pyrophosphoric (diphosphoric) acid [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 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 [1]. 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 [5] with a coupling constant of 1 ps, and pressure was regulated using an isotropic *c*-rescale barostat [6] 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 [7] with the CHARMM36m force field [8]. 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:

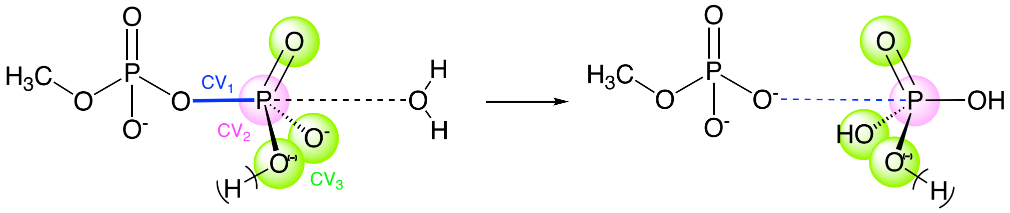
- Number of H_3O^+ ($n_{H_3O^+}$): TODO,
- Number of OH^- (n_{OH^-}): TODO.

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:

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

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



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.

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.

These wall potentials were applied to the following CVs:

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

All CV-related computations were performed using the built-in tools of CP2K 2023.1 [9] or PLUMED 2.9.3 [10]. 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.

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 [11] 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 [12]. Following equilibration, we performed 50 ps of well-tempered metadynamics (WTMD) [13] 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 [9]. Temperature control was achieved using the ν-rescale thermostat [5], 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 [14], along with the D3 dispersion correction and the Becke-Johnson damping function [12, 15]. In all calculations, the Goedecker-Teter-Hutter (GTH) pseudopotentials [16, 17] 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 [18] of the CP2K 2023.1 package [9]. 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 [19]. 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 [20], 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 1×10^{-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 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

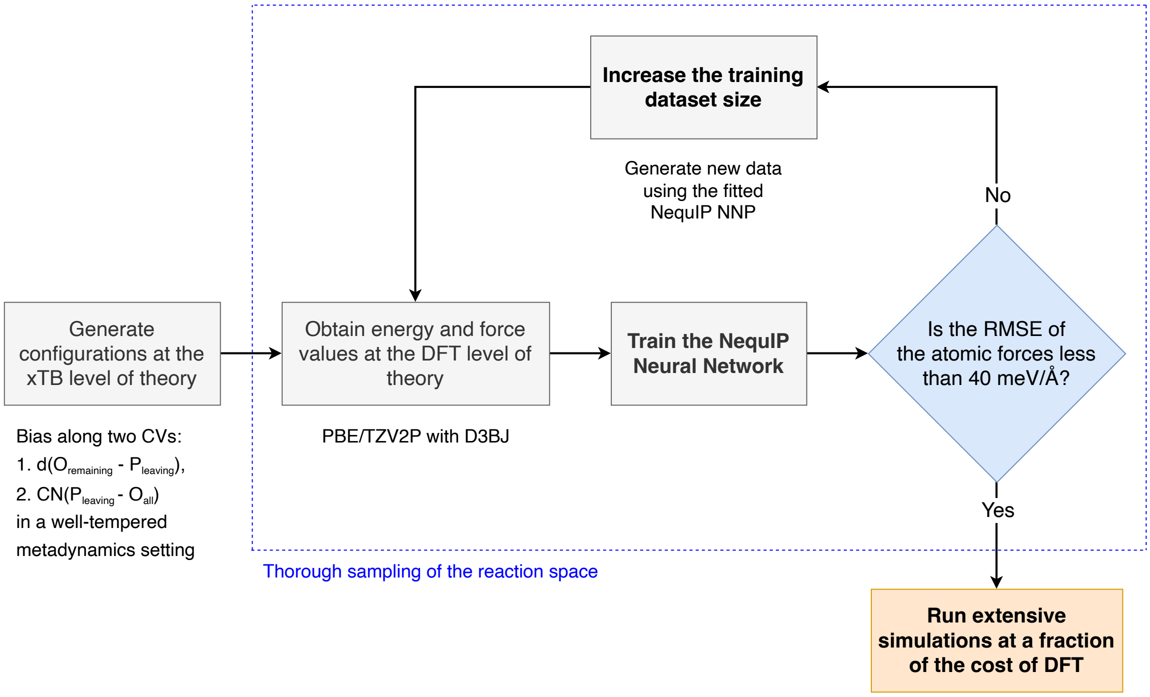

Increase the training dataset size
Generate new data using the fitted NequIP NNP
No
Generate configurations at the xTB level of theory
Obtain energy and force values at the DFT level of theory
Train the NequIP Neural Network
Is the RMSE of the atomic forces less than 40 meV/Å?
Bias along two CVs:
1. 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: Iterative training of the NequIP neural network potential.

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 $k = \max\left(10, \min\left(\left\lfloor \frac{N}{50} \right\rfloor, \left\lfloor \frac{n_{\mathrm{samples}}}{10} \right\rfloor\right)\right)$, 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 [21] compiled with PLUMED 2.9.3 [10] and `pair_nequip` [22] 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 [23, 24] 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_2 in order to favour either a dissociative or associative mechanism of the reaction. 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

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).

[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).

[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).

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

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

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

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

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

[9] 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).

[10] 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).

[11] 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).

[12] 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).

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

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

[15] 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).

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

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

[18] 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).

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

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

[21] 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).

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

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

[24] 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. $\Delta E(450\,\text{Ry}) = E(450\,\text{Ry}) - E(400\,\text{Ry})$. 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.)**	**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 $n_{\text{samples}}$
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