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

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]:

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

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 [5] with a coupling constant of 1 ps and an isotropic *c*-rescale barostat [6] 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. In all steps, the periodic boundary conditions (PBC) were applied in all directions.

All simulations were conducted in GROMACS 2021.4 [7] using the CHARMM36m forcefield [8] 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

Table 3.1: System composition and simulation box details.

System	Equillibrated box dimensions (Å^3)	No. of H_2O	No. of Na^+
MeDP	15.877 × 15.877 × 15.877	119	3
MeHDP	15.901 × 15.901 × 15.901	124	2

calculations unless stated otherwise.

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:

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

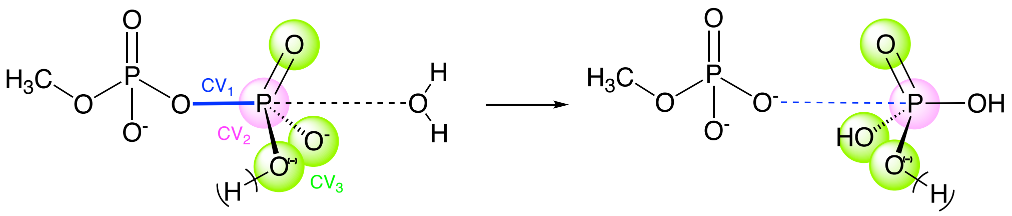
where r_{ij} is the distance between the atoms i and j from the groups A and B, d_0 is a shift in the distance where CN starts to decay, r_0 is a scaling parameter or a characteristic decay length that defines how fast the CN drops off with distance, and n and m are the positive integers that control 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$_{\text{remaining}}$ - P$_{\text{leaving}}$),
- The coordination number of all oxygens around the β-phosphorus atom (CV$_2$, CN(P$_{\text{leaving}}$ - O$_{\text{all}}$)): 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$_{\text{leaving}}$ - H$_{\text{all}}$)): 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.



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

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

The aforementioned wall potentials have the following form:

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

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

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:

- $\mathrm{CV_1}$, $\mathrm{d(O_{remaining} - P_{leaving})}$: 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 either the CP2K 2023.1 in-built tools [9] or PLUMED 2.9.3 [10]. It's important to note that the number of the CVs and the restraints 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 and wall potentials used in each part.

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 [11] 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 [12]. Afterwards, we performed 50 ps long well-tempered metadynamics (WTMD) [13] 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 [9]. The temperature was controlled by means of the ν-rescale thermostat [5] with a time constant of 50 fs for the equillibration and 100 fs for the WTMD run. The self-consistent field (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 were applied in all directions.

3.1.5 Data labeling

All data points were labelled by means of running the single-point calculations in order to obtain the energy and force values. The single-point calculations were performed using the Perdew–Burke–Ernzerhof (PBE) exchange-correlation functional [14] 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 together with the triple-ζ valence basis set with two polarisation functions (TZV2P).

The single-point calculations were carried out using the Gaussian Plane Wave (GPW) method implemented in QUICKSTEP module [18] of the CP2K 2023.1 package [9]. The SCF convergence was set to 10^{-6} a.u. A plane-wave cutoff of 800 Ry for the total density and a cutoff of 60 Ry for the Kohn-Sham orbitals were utilised.

The above mentioned cutoffs were determined based on the convergence test performed on one of the configurations as described in [19]. The error in total energy less than 10^{-8} a.u. for the convergence test has been considered as sufficient. The convergence test was performed using the following parameters: the cutoff for the total density was varied from 400 to 1500 Ry and the cutoff for the Kohn-Sham orbitals was varied 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. Speaking of 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 modeling consisted of two layers with 64 hidden neurons. Self-connections were enabled, and the average number of neighbors 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 ϵ = 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 test purposes for both systems (MeDP and MeHDP)

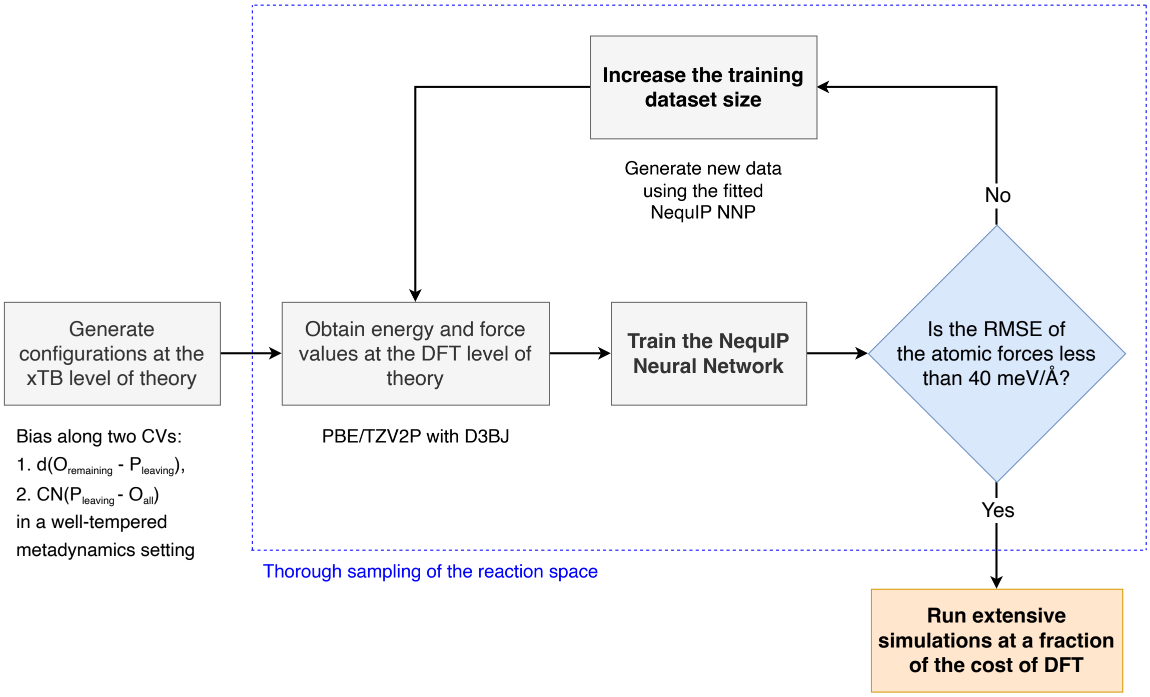

Increase the training dataset size
Generate new data using the fitted NequIP NNP
No
Generate configurations at the xTB level of theory
Obtain energy and force values at the DFT level of theory
Train the NequIP Neural Network
Is the RMSE of the atomic forces less than 40 meV/Å?
Bias along two CVs:
1.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

combined together. It was obtained within the three rounds of iterative training. In each round of training the model was retrained on a bigger dataset. The data obtained from each round will be discussed in the following sections.

Selection of configurations for training and testing

The important part of the iterative training is the selection of the 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 COLVAR 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 serves 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 are 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 a training and test dataset that closely mirrors 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 1.

First round

In the first round of training of the neural network potential, the model was trained on a small dataset consisting of 4,000 configurations. The configurations were obtained from the very first 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 the CV_1 and CV_2. No restraints were applied to the system. The training was performed 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 the second round of exploration of the configuration space driven by the NNP obtained after the first round of training.

To run the simulations with the NNP, 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 Nosé-Hoover thermostat [23, 24] with a time constant of 50 fs. The biasing potential was applied to the CV_1 and CV_2 every 50 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. The time step for the integration was set to 0.5 fs.

The restraints were applied to CV_2 in order to favour dissociative or associative mechanism of the reaction. The training was performed using the same hyperparameters as in the first round.

Third round

In the last round of training, the model was trained on a dataset consisting of 12,000 configurations. The additional configurations were obtained from the 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 in the same setup as in the second round. The only difference is that the temperature in this round was set to 320 and 340 K to explore the configuration space at higher temperatures. No restraints were applied to the system. The training was performed using the same hyperparameters as in the first round. The final dataset can be seen 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. $\Delta E(450\,\mathrm{Ry}) = E(450\,\mathrm{Ry}) - E(400\,\mathrm{Ry})$. 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 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