Combining Machine Learning Approaches and Accurate Ab Initio Enhanced Sampling Methods for Prebiotic Chemical Reactions in Solution

Timothée Devergne, Théo Magrino, Fabio Pietrucci, A. Marco Saitta

Combining machine learning approaches and accurate ab initio enhanced sampling methods for prebiotic chemical reactions in solution

Timothée Devergne,* Théo Magrino, Fabio Pietrucci, and A. Marco Saitta

Sorbonne Université, UMR CNRS 7590,

Muséum National d' Histoire Naturelle, Institut de Recherche pour le Développement,

Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie,

75252 Paris, France

E-mail: timothee.devergne@sorbonne-universite.fr

Abstract

The study of the thermodynamics, kinetics, and microscopic mechanisms of chemical reactions in solution requires the use of advanced free-energy methods for predictions to be quantitative. This task is however a formidable one for atomistic simulation methods, as the cost of quantum-based *ab initio* approaches, to obtain statistically meaningful samplings of the relevant chemical spaces and networks, becomes exceedingly heavy. In this work, we critically assess the optimal structure and minimal size of an *ab initio* training set able to lead to accurate free energy profiles sampled with neural network potentials. The results allow to propose an *ab initio* protocol where the *ad hoc* inclusion of a machine-learning (ML)-based task can significantly increase the computational efficiency, while keeping the *ab initio* accuracy and, at the same time, avoiding some of the notorious extrapolation risks in typical atomistic ML approaches. We focus on two representative, and computationally challenging, reaction steps of the

classic Strecker-cyanohydrin mechanism for glycine synthesis in water solution, where the main precursors are formaldehyde and hydrogen cyanide. We demonstrate that indistinguishable *ab initio*-quality results are obtained, thanks to the ML-subprotocol, at about one order of magnitude less of computational load.

Introduction

Prebiotic chemistry and *ab initio* molecular dynamics

The study of chemical reactions in solution by means of *ab initio* molecular dynamics (AIMD) is a full branch of computational chemistry, which is of particular interest for prebiotic chemistry, the chemistry of the origins of life.[1] Many hypotheses on how bio molecules formed from elementary bricks exist. One of them is that life building blocks were brought to earth from the interstellar medium: they were either formed in comets[2,3] or in interstellar ice via UV radiation.[4,5] Other works show that elementary biomolecules were formed on earth in a primordial soup, hydrothermal submarine vents[6] or mineral surfaces.[7–9] All these possible prebiotic synthesis environments imply very different conditions to investigate. This is where AIMD comes into play: it allows one to calculate interatomic forces from the quantum-based potential energy surface (PES) of a given systems, in order to simulate the motion of atoms in organic molecules and biomolecules, while explicitly including the non-trivial presence of the chemical environment, *i.e.* the solvent and/or the mineral surfaces. In order to be predictive and quantitative, it needs to be used in conjunction with enhanced sampling methods such as metadynamics,[10,11] umbrella sampling (US)[12] or transition path sampling.[13] This is necessary to overcome the energy barriers of reactions and thus the time limitations of computer simulations in order to explore the possible pathways in a given environment.

In our group, some successful results were obtained in the past few years using these methods in different contexts. The famous Miller-Urey experiment[14] was computationally reproduced[15] and the mechanism of formation of glycine explained through a novel pathway, with formamide as a crucial intermediate. The decomposition pathways of amino acids in the supposed hydrothermal conditions of meteoritic parent bodies was studied,[16] also to obtain clues about a possible origin of their chirality. Recently, we have thoroughly studied the classic Strecker-cyanohydrin synthesis of glycine,[17] generally considered the main synthesis pathway of glycine, the simplest amino acid, in the primordial Earth and universe.

Our goal was both to reproduce this relatively complex textbook reaction, consisting of seven elementary chemical steps, and to complete the sparse existing experimental data on thermodynamics by accurate *ab initio* results.

To this end, as in our previous studies, we calculated the free energy surfaces and barriers for each intermediate chemical step using an AIMD-based enhanced-sampling protocol the exploration of the chemical pathways, the identification of the transition states (TS), the definition of precise RCs, and the accurate calculation of free-energy landscapes via umbrella sampling (US), which in turn requires the calculation of a large number of *ab initio* forces.[17,18] This latter part, although necessary to ensure the statistical convergence of the FES to within 1-2 kcal/mol, is the computationally most expensive one, and thus the bottleneck significantly limiting the extent of potential applications, and the possibility to upscale the protocol to more realistic prebiotic environments, possibly including mineral surfaces.[19] In the past few years, some attempts to overcome this obstacle in computational prebiotic chemistry were carried out exploiting density functional tight-binding approaches.[20] More recently, machine learning approaches to describe the local PES along chemical processes in solution have emerged as a very appealing opportunity.[21]

Machine learning potentials with *ab initio* accuracy

In the past few years, a number of machine learning methods and frameworks have been developed to tackle the problem of obtaining *ab initio*-level PES at a reasonable computational cost.[22–28] Although these methods have significant differences, the basic principle is common: using AIMD trajectories, a model is trained and then used to infer *ab initio* quality PES, and thus perform *ab initio* quality molecular dynamics simulations for a much lesser computational cost.

This approach can be used to access larger system sizes,[29,30] and/or to perform simulations longer than the time reachable using traditional AIMD simulations, in order to observe interesting physical transformations and/or improve statistical sampling.[31]

However, two important questions still need to be fully addressed in ML-based molecular dynamics: how to build optimal training sets, and how to critically assess the quality of machine-learning potentials in chemical reactions, a difficult setting where the system is led to explore very energetic configurations far from the geometries of the metastable minima.

An interesting tool in this respect is the "neural-network committee" method.[32] It consists in training several NNPs on the same training set but with different random seeds. In this way, for the same configuration, the NNPs will give different results. The standard deviation of the predictions of the different members of the committee on some observable is used to assess the reliability of the average prediction. Indeed a good agreement between the different NNP means the configuration is close-enough to the training set for the NNP to be accurate, while a higher value means that the prediction cannot be trusted.

This technique can be used to build a training set using an iterative training[33,34] procedure. First, a small set of NNPs is trained with existing AIMD data; then, a MD simulation is run using one of the NNP of this set. After this step, configurations that display a standard deviation over the set of NNP on the prediction of some observable above a certain threshold, are recomputed using single point *ab initio* calculations and added to the training set. This is repeated until no more configurations are evaluated as mis-predicted.

The iterative learning framework has led to thorough studies of systems with *ab initio* quality at reduced computational cost.[35,36] More recently, the committee method has been also used to quantify the error on an observable computed using NNP-based molecular dynamics simulations,[37] as well as a way to iteratively select configurations from an AIMD trajectory to build optimal training sets.[38] A recent in-depth study of committee methods[39] has shown that in iterative schemes randomly selecting additional configurations to be evaluated at DFT level to improve the training set (random sampling) is equivalent to a selection based on committee disagreement beyond a threshold, but that the latter has to be carefully calibrated.

In this work we adopt NNP-driven enhanced-sampling molecular dynamics simulations

to study chemical reactions in solution. This is a challenging goal as it requires the NNP to explore high-energy configurations far from equilibrium, with highly distorted chemical bond geometries. In particular, for a given $A \rightarrow B$ reaction, we aim at extensively sampling along the RC connecting the two basins, with the aim of reconstructing the accurate free energy landscape through US simulations. To this end, it is crucial to train NNPs capable to yield locally-accurate and well-behaved PES throughout the relevant reaction space. Although enhanced sampling has been combined with machine-learning potentials in a few recent studies,[21,37,40] including a combination of US with NNP,[41] the critical assessment and systematic use of NNP for chemical reactions in solution are still lacking. In the same spirit as this work, but using a different angle of approach, the $\Delta - ML$ idea has already been used along with enhanced sampling in a previous study to obtain high accuracy free energy for a more affordable computational cost.[42]

In the present study we present benchmarks and construction principles for training sets. We carefully assess the total computational cost of the training and data production trajectories, with the goal of limiting the total amount of *ab initio* calculations without losing accuracy. We also introduce a simple approach to ensure long stable trajectories with high NNP-committee agreement: at variance with a previous method where the error is evaluated from deviations between DFT and NNP predictions,[43] our simple scheme avoids the burden of additional *ab initio* calculations.

We apply the new scheme to two reaction steps of the Strecker-cyanohydrin synthesis of glycine in water, previously studied at the DFT level.[17] Our results indicate that a surprisingly reduced amount of suitably-chosen *ab initio* samples is sufficient to train a NNP potential able to sample the full reaction coordinate space between minima, leading to accurate free-energy profiles and barriers at a significantly reduced cost compared to purely DFT simulations.

Methods

Neural network potentials

In this paper, we use the Behler-Parrinello neural networks approach,[24] in which the potential energy of the system is decomposed as the sum of individual atomic contributions as in equation 1, where N_{atoms} is the number of atoms and E_i is the atomic contribution:

$$E = \sum_{i=1}^{N_{atoms}} E_i \tag{1}$$

Each atomic contribution is computed using a neural network, and to ensure permutation invariance, the contributions corresponding to atoms of the same species are computed using the same NN.

Here, we use the deepmd/deepPot-Se code[26,27] based on this architecture. In this framework, a first NN is used to compute symmetry-invariant descriptors for each atom, based on the coordinates of its neighboring atoms within a cutoff sphere. Since this is not the scope of this study, the hyperparameters of the models were not optimized and were taken from other studies.[21] Here, to compute the descriptors we chose a cutoff radius of 6.0 Å and a smoothing cutoff of 1.5 Å. Then, these symmetry-preserving descriptors are given as input to the energy neural networks. The optimization process of the neural networks weights is performed using the Adam optimizer[44] with a learning rate going from 10^{-3} to 5.10^{-8} during 10^6 iterations on a defined loss decreasing by 5% every 5000 iterations.

Our target system is quite heterogeneous, including 251 atoms in a 13.4 Å-side cubic box under periodic boundary conditions, with the following composition: 2 C, 2 N, 6 H and 1 O atom in the reactive molecules, solvated with 81 water explicit water molecules, whose H and O atoms are indistinct from the ones in the reactive part of the system, and treated exactly on the same footing by the NN. As a consequence, the water molecules would comparatively acquire an overwhelming weight, with respect to the important reactive subsystem, during

training. To overcome this potential difficulty, we defined a loss function (equation 2) capable to take into account this heterogeneity:

$$L(\mathbf{w}) = \frac{1}{|B|} \sum_{l \in B} \left[p_E |E_l - E_l^{\mathbf{w}}|^2 + p_f \frac{1}{N_{elem}} \sum_{i=1}^{N_{elem}} \frac{N_{atoms}}{n_i} |\mathbf{F}_l - \mathbf{F}_l^{\mathbf{w}}|^2 \right] \tag{2}$$

,

where n_i is the number of atoms of type i in the system, N_{elem} is the number of different elements in the system, N_{atoms} is the total number of atoms in the system, E_i and F_i denote the DFT energies and forces of the training set, while $E_l^{\mathbf{w}}$ and $F_l^{\mathbf{w}}$ are the forces computed by the NNP, and B is the batch size (i.e., the number of trajectory frames).

This choice of the loss function allows weighting equally each element type via the error on forces. The weight p_f progressively increase from 1 to 10 while the opposite happens for p_e during the training, according to the protocol implemented in the deepmd-kit package.[26]

The neural network used to compute the descriptors was made with 3 layers of 25, 50 and 100 nodes while the energies networks were made with three layers of 240 nodes each. All the AIMD configurations come from a previous study of the group.[17] We used LAMMPS[45] for all the NNP-based molecular dynamics simulations, employing hydrogen atoms instead of deuterium in the original study to have a quicker, more realistic dynamics. All the enhanced sampling simulations performed with neural networks were carried out using the open-source, community-developed PLUMED library,[46] version 2.5.0[47] .

Free energy calculations

Enhanced sampling approaches are necessary to overcome the timescale limitations of molecular dynamics in the simulation of high-barrier rare events like chemical reactions.[48] The technique employed in this study, umbrella sampling, focuses the sampling in the desired regions of configuration space exploiting an external bias potential as a function of a collective

variable[12,49] For the latter, we employ the path collective variables[50] (path CVs) defined as

$$\begin{cases} s(t) = \dfrac{1}{N-1}\left(\dfrac{\sum_{\alpha=1}^{N}\alpha\exp(-\lambda D[x(t),X_\alpha])}{\sum_{\alpha=1}^{N}\exp(-\lambda D[x(t),X_\alpha])}-1\right) \\ z(t) = \dfrac{-1}{\lambda}\log\sum_{\alpha=1}^{N}\exp(-\lambda D[x(t),X_\alpha]) \end{cases} . \tag{3}$$

We introduced a scaling factor[17] so that $s \in [0,1]$. The first variable s measures the progress of a reaction on a path given by N configurations X_α. The second variable z measures the deviation of the configuration at time t from the given path. The metric D used in this study has been extensively used in previous works[16,18,51] and is based on the comparison of coordination patterns:

$$D[X_1,X_2] = \sum_i \sum_\sigma (C_{i\sigma}(X_1) - C_{i\sigma}(X_2))^2 \tag{4}$$

$C_{i\sigma}$ is defined in equation 5 for atom i of atom type α with respect to atom type σ as

$$C_{i\sigma} = \sum_{j\in\sigma} \frac{1-\left[\dfrac{r_{ij}}{r_0^{\alpha\sigma}}\right]^m}{1-\left[\dfrac{r_{ij}}{r_0^{\alpha\sigma}}\right]^n} \tag{5}$$

where the term within the sum is a smooth function going from 1 to 0 when the interatomic distance becomes larger than a typical distance $r_0^{\alpha\sigma}$. This distance was set to 1.8 Å for heavy atom-heavy atom pairs and to 1.4 Å for heavy atom-hydrogen pairs. We chose $m = 8$ and $n = 14$. The exact definition for the reference structures, explained in our previous work,[17] exploits committor-analysis trajectories performed after a preliminary metadynamics simulation[10] (using a pathCVs defined with only the reactants and the products). 10 regularly-spaced reference structures are selected along the latter trajectories using an in-house algorithm[17] based on the nudge elastic band approach.[52] Finally, two more references are added at the start and at the end of the path by linearly extrapolating the coordination numbers of the last and first segments to avoid metastable states to appear as spikes in the

free-energy landscape: the result are pathCVs s, z based on 12 reference structures that we use throughout this work.

We adopt a set of one-dimensional, quadratic umbrella sampling potentials applied on the s variable and centred at positions s_j expressed as:

$$V_{bias,j}(s) = \frac{k}{2}(s - s_j)^2 \tag{6}$$

Windows are equally spaced by $s_{j+1} - s_j = \Delta s$ based on $k = k_B T/(\Delta s/2.5)^2$ in order to have sufficient overlap between two windows (see Ref.[17]). The s path-CV measures the progress along a given reaction pathway, the sampling is therefore performed on that coordinate. The z path-CV measures the deviation from the pathway and helps to detect possible anomalies in the sampling. Once a specific reaction pathway is determined and targeted, US is performed on it, in order to determine the corresponding free-energy profile. Due to the intrinsically high-energy, unstable character of configurations explored close to the barrier top, in US simulations of chemical reactions it is occasionally observed that the system can deviate from the targeted pathway to explore a different one. To focus the sampling on the reaction mechanism under study, a restraining potential is sometimes applied along z, as in the case of the step $(1) \rightarrow (2')$ of reference.[17] We compute the free energy surface, according to:

$$F(s) = -k_B T ln \langle \rho(s) \rangle \tag{7}$$

where $\langle \rho(s) \rangle$ is the unbiased distribution function of s, globally estimated over the whole set of windows using the weighted histogram analysis method[53] implemented in Grossfield's code.[54] We use a convergence criterion of 10^{-7} kcal/mol and 150 bins in s space. We estimate the statistical uncertainty as the deviation between the free energy profiles computed using the third and fourth quarters of each trajectory.

The obtained free energy profile is then used to compute the activation barrier by taking the free energy difference between the highest point in the profile and the reactants free

energy; the free energy difference between reactants and products can also be computed. The value of the activation barriers can then be compared to experimental results as it was done in reference 17. A correction term can also be added to this free energy difference as it was done in reference.[55] However, since the scope of the present work is to compare NNP results to the AIMD ones of the original study,[17] this additional term was not in this work.

From a ML point of view, the splitting of the RC space into independent US windows simplifies the selection of different data sets to form the training set, and the evaluation of the performance of the NNP.

Results and discussion

We first consider the initial step of the Strecker-cyanohydrin synthesis of glycine, particularly emblematic in prebiotic chemistry studies:[14,15]

$$HCN + COH_2 + NH_3 \rightarrow CH_2OH(NH_3)^+ + CN^-$$

as illustrated in Fig. 1. The CVs regarding this reaction will be noted with a subscript $(1) \rightarrow (2')$.[17] The aim is to assess the behavior of a NNP-system, trained on a minimal amount of AIMD US trajectories, carefully selected along an optimized reaction pathway. In the last section we will cross-validate the protocol on another, more complex, chemical reaction step of the Strecker mechanism.

Detecting the frontiers of accurate predictions in a neural network potential

The performance of a NNP is customarily measured via the error with respect to DFT on a training set and a testing set. However, this is not sufficient with respect to our goal, i.e., accurately computing the full free energy landscape. In this respect, it is important

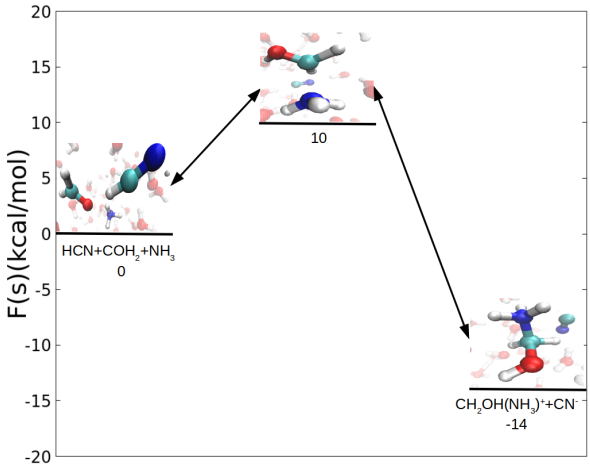


Figure 1: Free energy diagram of the first step of the Strecker-cyanohydrin synthesis of glycine obtained in our previous work.[17] The error is of the order of the kcal/mol and was estimated using block averages. The reactants, products, and transition state configurations are also reported on the graph.

to assess the capacity to generate long, stable and accurate MD trajectories. We base our assessment on the committee approach,[32] including a time-dependent metric similarly to a recent study.[43]

As shown in figure 2, when, during a simulation, the NNP-generated trajectory exits the "safe" region of configuration space, where forces are accurately predicted, the value of the z pathCV significantly increases, indicating a large deviation from the reference transition pathway (see Methods section) and suggesting the likelihood of unphysical configurations. This is confirmed by the analysis of the corresponding structural properties of the system. An inspection of the C-O pair correlation functions $g(r)$ (figure 2 c)), before and after this jump, reveals that an unphysical short-distance peak has appeared, not present in the *ab initio* data. This confirms that the NNP sampled an unphysical region and is trapped in it, as the predicted energy decreases (figure 2 b), a spurious stable configuration. On the other hand, this transition corresponds to a sudden jump in the standard deviation in energy predictions among committee members as well as in the maximum standard deviation on

the predicted atomic forces, defined in the following way:

$$\sigma_{max} = \max_{j \in [1, N_{atoms}]} \sqrt{\left[\sum_{i=1}^{4} \| \mathbf{F}_i^j - \mathbf{F}_{avg}^j \|^2 \right]} \tag{8}$$

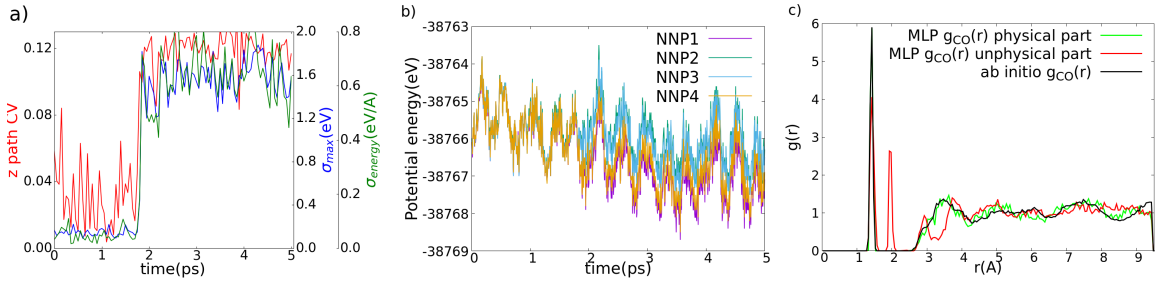


Figure 2: Panel **a)**: left (red line): time evolution of z path CV; right: time evolution of the maximum standard deviation on the prediction of forces(σ_{max}, blue line) and standard deviation on the prediction of energies(σ_{energy}, green line). Panel **b)**: time evolution of the potential energy predicted by the four neural networks. Panel **c)**: Carbon-Oxygen radial distribution function for the AIMD simulation (black line) and for the NNP simulation, before (green line) and after (red line) the unphysical jump in the predictions of energy and forces.

where $\mathbf{F}_i^j$ is the force on atom j computed by NNP i, and $\mathbf{F}_{avg}^j$ is the average over all the neural networks.[21,35,36] We remark that σ_{max} is a more general and accurate indicator of the loss of predictive power than the $g(r)$ or the pathCV z.

This approach allows detecting the frontiers of the configuration space region where the NNP gives accurate prediction. We can thus define a simulation lifetime τ as the time at which σ_{max} surpasses a reliability threshold, simply corresponding to the full simulation time if no pathological behaviors occur. All the trajectory before τ is physically sound, and can be employed to collect statistics. The next logical step is to devise a procedure to maximise τ all along the RC space as a function of the training set composition and extension, in order to carry out reliable US simulations.

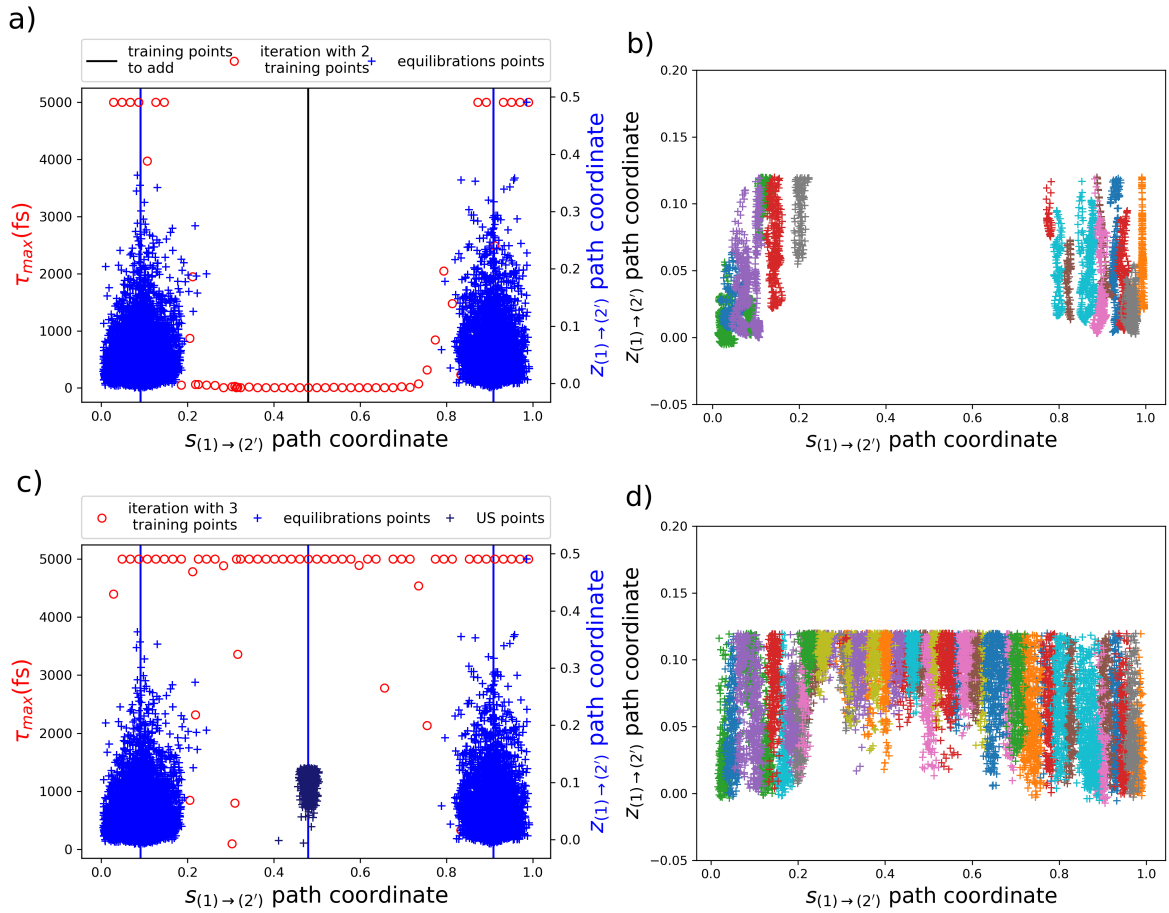


Figure 3: Panel **a)**: maximum lifetime τ_{max} (red dots) in each US window throughout the path CV for a NNP trained only on the AIMD equilibrations of the reactants and the products; the AIMD training points are represented in blue in the (s,z) plane. Panel **b)**: corresponding instantaneous location of the NNP configurations in the (s,z) path CV coordinate plane; panel **c)** maximum lifetime τ_{max} (red dots) in each US window throughout the path CV for a NNP trained only on the AIMD equilibrations of the reactants, the products, and a transition state US window; the AIMD training points are represented in blue in the (s,z) plane. Panel **d)**: corresponding instantaneous location of the NNP configurations in the (s,z) path CV coordinate plane

Generating training sets suited for free energy calculations

In this section, we critically assess the effect of the composition and size of the training set on the error of the free energy profile recontruced with the NNP potential. The aim is to retain *ab initio* accuracy while minimizing the amount of DFT calculations necessary to train the potential. Following the approach employed in our previous full-*ab initio* work[17] on the reaction considered here, we identify the following algorithm:

- Perform a preliminary DFT-based metadynamics simulation[10] employing pathCVs built upon the reactants and products as the only references:[18] this allows a prejudice-free exploration of a reactive pathway (quick, without the need to converge a free-energy estimate), to be refined through committor analysis and leading to the definition of improved pathCVs based on multiple reference structures (see Methods and Ref.[17] for details).

- Generate one AIMD US trajectory (of about 15 ps) in the transition state window along the optimized pathCV (defined at the previous step), and include it in the training set along with AIMD equilibration trajectories (of about 15 ps) of the reactants and the products. These three trajectories represent *ab initio* "milestones" of the RC to train and test the NNPs.

- Train four models on the same training set, with different random seeds. Define a range of US windows densely spanning the full RC range, and for each window perform 50 short (about 5 ps) NNP-based US simulations, with the same starting point but different random initial velocities taken from the Boltzmann distribution.

- Plot the maximum lifetime τ_{max} over each US window (see figure 3): if in some RC region τ_{max} is smaller than the autocorrelation time of the pathCV, i.e., if it is impossible to generate uncorrelated samples, it is necessary to generate an additional AIMD US trajectory in that region (since the NNP is locally unreliable) to be added to the training set.

- Repeat the two last steps until a satisfactory NNP is obtained, *i.e.* with an acceptable lifetime τ across the full RC space, hence capable of producing a dense sampling and a converged US free energy landscape.

The procedure is illustrated in figure 3, showing the training configurations for each iteration (about 600 structures saved every 20 fs) and the resulting NNP samples in the

(s, z) plane and lifetime for every US window. Clearly, adding training points according to our scheme allows to progressively increase the simulation lifetime and the overlap between NNP US simulations. In the next two sections, we discuss the error of the NNP on a testing set, and the accuracy of the NNP free energy landscape with respect to the *ab initio* one.

Error estimate on the benchmark reaction

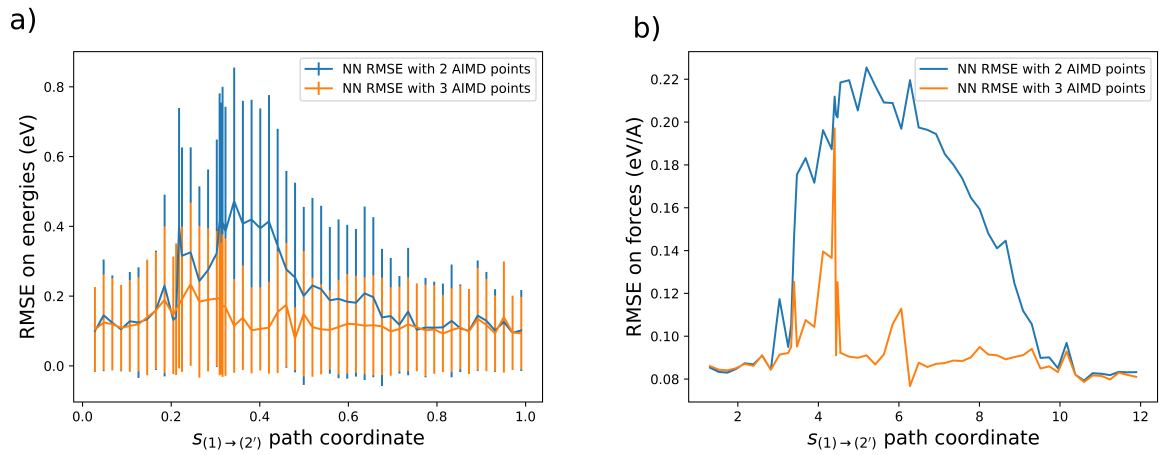


Figure 4: Panel **a)**: Root mean squared error (RMSE) of the NNP-estimated energy along the chemical path for a NNP trained only on the AIMD equilibrations of the reactants and the products (blue line) and for a NNP trained on those configurations and on the US transition state window (orange line). The test set is built only with configurations that are not present in any training set. Panel **b)**: same as panel **a)** but the RMSE was computed on forces instead of energies

Before proceeding on the actual calculation of the free energies, it is important to verify that the NNP displays a rather uniform error over the s-coordinate. To this aim, we compute the root mean square error (RMSE) of the NNP-predicted energies and forces on a test set built with configurations from *ab initio* simulations. Those configurations were never included in any of the training sets. The results for the case of a training set including only reactants and products (2-AIMD points) are compared with the case including also one transition-state-like US window in the trainng set (3-AIMD points) in figure 4. Clearly, adding the third trajectory to the training set improves the error in the central region of RC space and leads to a uniform RMSE.

Although this error-evaluation step gives us an idea on the quality of the NNP produced using our new procedure, it cannot be applied on a system that has not been extensively studied with AIMD.

Calculation of the NNP free energy surface for the benchmark reaction

We now assess the accuracy of the free energy landscape reconstructed from the NNP. Figure 5 shows that it is possible to obtain a first-principle quality free energy surface using, in the training set, only one US AIMD simulations and the two end-point AIMD equilibration simulations of the reactants and the products, and to recover the full FES mostly via NNP-US windows. This is a major gain of computational time, which could allow to study, at the same level of *ab initio* accuracy, larger and more complex systems.

The full original *ab initio* FES in Ref.[17] was obtained using 55 US AIMD windows; our procedure significantly reduces, by more than an order of magnitude, the number of *ab initio* MD simulations to be carried out. From the computational point of view, the training of 1 neural network takes approximately 24 GPUh; while one 5-ps simulation takes about 0.05 GPUh. Hence, for the full study of this test reaction, we needed a total of 161 GPUh, to which one needs to add the 40k CPUh used to build the *ab initio* training set. The fact that all the calculations were run on GPU makes non-trivial the comparison with respect to the CPU simulations. However, the whole *ab initio* study of this reaction in Ref.[17] took around 700k CPUh.

In figure 5, the free energy was computed using short simulations of 5 ps per window. In the next section, we propose a method to generate longer, but still stable, NNP trajectories.

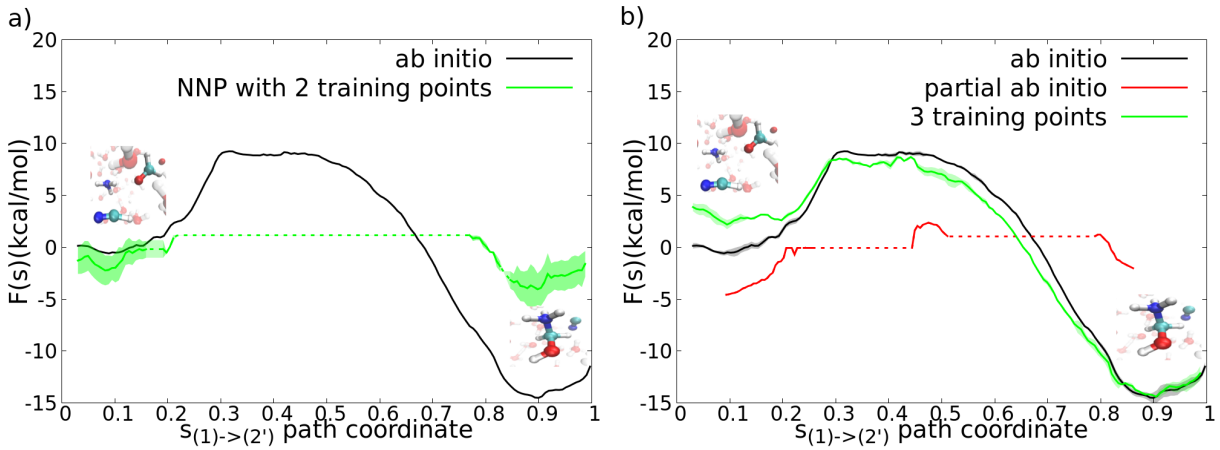


Figure 5: Panel **a)**: US-obtained free energy profiles using full AIMD (black line), and using a NNP trained on AIMD equilibration of reactants and products (green line). The shaded zones correspond to the estimated statistical errors. Panel **b)**: US-obtained free energy profiles using full AIMD (black line), and using a NNP trained on AIMD equilibration of reactants, products, and the transition state US windows (green line). The red line represents the free energy which would be obtained using the three AIMD US windows (reactants, products, transition state). The reactants and products endpoint-configurations are also reported on the graphs.

Generating stable NNP trajectories

After properly training our NNP, our aim is to be able to generate long and stable MD simulations. However, it is known that, at some point, the NNP-system will eventually

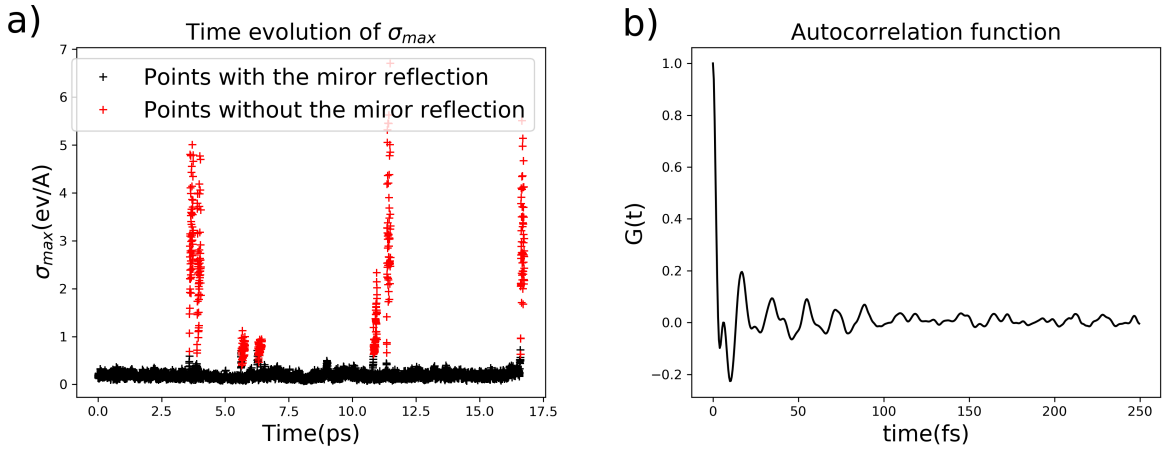


Figure 6: Panel **a)**: Time evolution of the maximum standard deviation on the prediction of forces (σ_{max}) during a NNP umbrella sampling simulation. We report with black crosses its instantaneous values before the the unphysical behavior sets in, and with red crosses its behavior after that. Panel **b)**: NNP-autocorrelation function of the s reaction coordinate computed in the same NNP-umbrella sampling window.

explore untrained regions of the configuration space, where energies and forces will necessarily be extrapolated, likely failing to describe the correct system anymore. To overcome this common drawback, one possibility to restrain the sampling to low σ_{max} regions is to add a quadratic potential on σ_{max}, as it was proposed in reference.[32] In this work, instead, we find it more effective to completely avoid high-σ_{max} values, which would correspond to an infinite spring constant with respect to the above-mentioned reference. Therefore, we can define such a "mirror reflection operation" as follows:

We identify two regions in the configuration space sampled by each US window: the region where he NNP behaves properly, *i.e.* where the maximum standard deviation on the prediction of forces σ_{max} is low, and the region where significant deviations, hence unphysical behavior, are observed.

Whenever σ_{max} exceeds a given threshold, we define for each atom α the following vector $\mathbf{G}_\alpha$, the value of this vector in the k direction is:

$$G_{\alpha k} = \frac{1}{n_{comm}} \sum_{i=1}^{n_{comm}} (F_{\alpha k}^{i} - F_{\alpha k}^{avg})^2 \tag{9}$$

With n_{com} the number of neural networks in the committee ($n_{comm} = 4$ in this study). This quantity provides the normal vector of a frontier between the reliable, "interpolation region" of the NNP, and a NNP-unknown "extrapolation region". We can therefore mirror-reflect the trajectory of the atoms on this hypersurface when σ_{max} displays a pathological jump to high values.

In order to do so, and force the NNP-system to remain in the interpolation region, the simulation is stopped a t_{step} number of timesteps before the instability, and is restarted after imposing the following transformation on the velocities:

$$\mathbf{v}_\alpha^{new} = \mathbf{v}_\alpha^{old} - \frac{2\mathbf{v}_\alpha^{old}.\mathbf{G}_\alpha}{\|\mathbf{G}_\alpha\|^2}\mathbf{G}_\alpha \tag{10}$$

where $\mathbf{v}_\alpha^{old}$ is the velocity of atom α at the moment in which we mirror-reflect the trajectory,

while **G** is evaluated at the moment in which where σ_{max} passes the instability threshold.

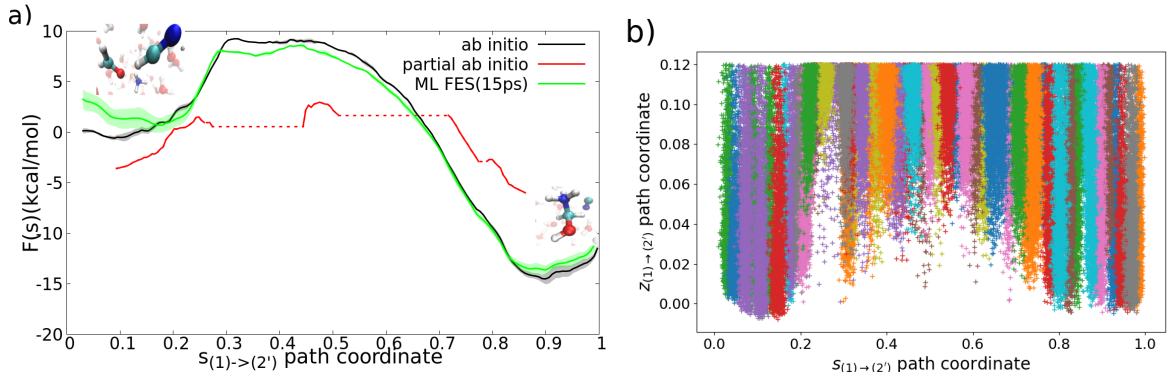


Figure 7: Panel **a)**: US-obtained free energy profiles using full AIMD (black line), and using NNP trained on AIMD equilibration of reactants, products and transition state US windows when the velocity transformation is applied (green line). The red line represents the free energy which would be obtained using only the three AIMD US windows (reactants, products, transition state). The reactants and products endpoint-configurations are also reported on the graph. Panel **b)**: corresponding instantaneous location of the NNP configurations in the (s,z) path CV coordinate plane. As in the original study[17] z path CV was confined to within 0.12 via a semiparabolic wall.

This operation preserves the kinetic energy of the system, without pathological effects on the US procedure. We consider that it is a useful and legitimate complement to our NNP-based free-energy calculation protocol whenever the typical lifetime τ is significantly longer than the auto-correlation time of the CV, so that uncorrelated samples can be collected between successive reflections. In the opposite case, it is advisable to improve the NNP via a larger training set before employing it to perform statistical sampling.

To illustrate the choice of the parameters of the protocol, the time evolution of σ_{max} during a US simulation is reported in figure 6 along with the NNP-autocorrelation function of the CV. The threshold of σ_{max} is chosen as 0.6eV/A, while choosing the number of timesteps at which the reflection is performed before the instability between 200-500 gives similar results. Clearly, the typical time lapse between two reflections (ranging here from 0.3 to 5 ps) is much larger than the autocorrelation time ($\sim$ 0.1 ps). Reflections therefore appear not harmful from a statistical viewpoint, and allow to carry out long, stable US simulations localized in the region of configuration space where the NNP is reliable, leading

to free-energy estimations retaining *ab initio* accuracy (Figure 7 and table 3). The profile obtained in figure 7 lies within the typical 2 kcal/mol uncertainty of the PBE functional for this kind of studies, as well as the predicted free energies, shown in table 3.

Table 1: Activation barrier ($\Delta F^{\dagger}$) and free energy difference between reactants and products($\Delta F_{(1)\to(2')}$) obtained in the *ab initio* study along with the ones obtained in this work

	Ab initio study	machine learning study
$\Delta F^{\dagger}_{(1)\to(2')}$ (kcal/mol)	10	8
$\Delta F_{(1)\to(2')}$ (kcal/mol)	-14	-14.5 ± 1

Application of the protocol to a more complex reaction

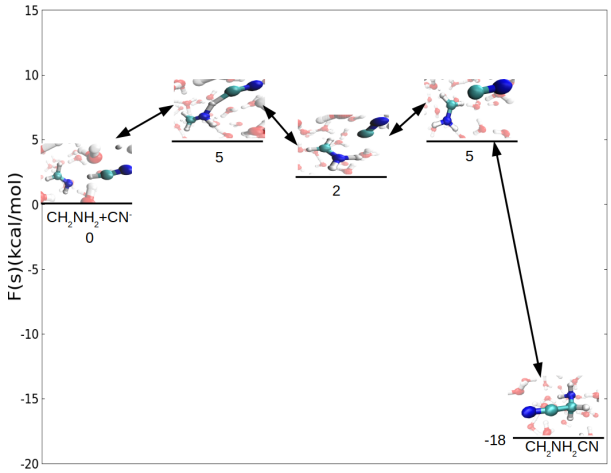


Figure 8: Free energy diagram of the reaction (3) $\to$ (4) of the Strecker-cyanohydrin mechanism.[17] The error is of the order of the kcal/mol and was estimated using block average.[17] The relevant configurations are also reported on the graph.

The aim of this part of our work is to challenge the method previously devised to reproduce a complex FES. To this end, we chose another intermediate step of the Strecker-cyanohydrin synthesis, presenting a more complex free energy landscape, consisting of a two-step process, as shown in figure 8: a hydrogen bond breaking between the imine nitrogen, a hydrogen atom and the cyanide, followed by the addition of the cyanide to the imine.

This process displays a small barrier of 5 kcal/mol, a small drop of 3 kcal/mol towards a metastable state, another small barrier of 3 kcal/mol with respect to the latter step, followed by a large 23 kcal/mol drop to the products, as obtained from our original AIMD FES in Ref.[17] by using 44 US windows. Such small free-energy barriers are likely more difficult to be quantitatively and qualitatively described by a NNP, than in the previous reaction. We underline that, although the stoichiometry of the chemical species is the same as in the previous case, the reactants and the products are different, and thus the NNPs are generated from scratch, independently from the previous case.

Results of US simulations employing a series of three NNPs with different training sets are presented in figure 9. We progressively enlarged the training set as previously described, by using at first 1, then 3 and finally 6 AIMD US windows, besides the reactants and products ones (respectively 3, 5 and 8 training points) .

The NNP FES obtained starting from the different training sets are shown in figure 10. Judging from the accuracy of the FES, NNP training is optimal when 6 *AIMD* US windows are used, which is reasonable considering the fact that this reaction consists of two successive steps. Instead, unsurprisingly, the FES obtained with only 3 training points compares very poorly with the *ab initio* one.

Adding AIMD US windows to the training set leads to a progressive increase of the accuracy of the NNP FES, until a satisfactory agreement with the benchmark AIMD one is achieved. As in the case of the previous reaction, the cumulative duration of the AIMD trajectories necessary for training the NNP is one order of magnitude shorter than the total duration required for computing an accurate AIMD FES, despite the fact that the present reaction is clearly more challenging to be reproduced in its fine details. The free energy profile obtained using the velocity transformation technique is shown in figure 11. We report in Table 2 the CPU time comparison between a full ab initio FES calculation and a NNP-based one. As we have no guarantee that our NNPs are transferable to another reaction, we chose to add the CPU time needed to perform the training simulations in the comparison

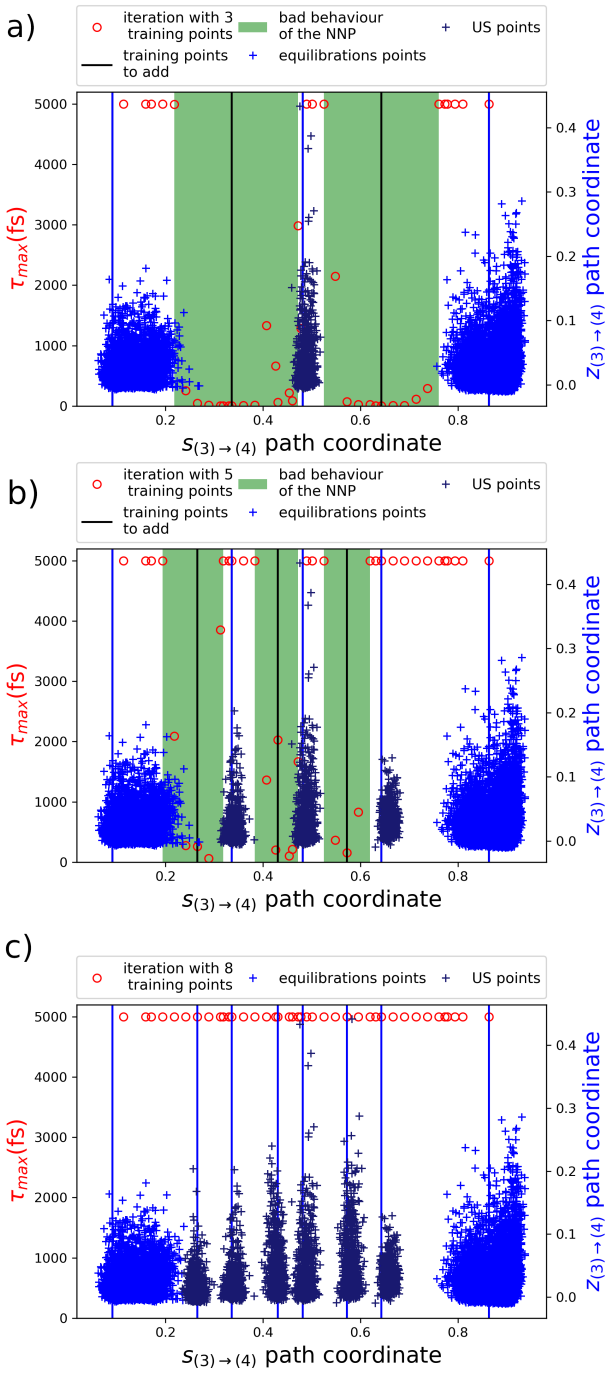


Figure 9: Illustration of the iterative procedure followed to build a converged NNP for reaction (3) $\rightarrow$ (4). Panel **a)**: maximum lifetime τ_{max} (red dots) in each US window throughout the path CV for a NNP trained only on the AIMD equilibrations of the reactants, the products, and the transition state; the AIMD training points are represented with blue crosses in the (s,z) plane. Panel **b)**: same as **a)** but adding two additional AIMD US windows in the NNP training set. Panel **c)**: same as **b)** but adding three additional AIMD US windows in the NNP training set.

with the *ab initio* study

Table 2: Summary and comparison of the computational times for the (3) → (4) reaction between the pure AIMD protocol and the combined AIMD-ML one.

	Ab initio study	machine learning study
training time	0	$24h(GPU) \times 4 \times 3$ iterations
US ab initio time	$15 \times 44 = 660ps$	$15 \times 6 = 90ps$
CPU/GPU simulation time	540k CPU.h	100k CPU.h + 200 GPU.h
Total CPU/GPU time	540kCPU.h	488 GPU.h + 100k CPU.h

Table 3: Activation barrier ($\Delta F^{\dagger}_{(3)\to(4)}$) and free energy difference between reactants and products($\Delta F_{(3)\to(4)}$ obtained in the *ab initio* study along with the ones obtained in this work

	Ab initio study	machine learning study
$\Delta F^{\dagger}_{(3)\to(4)}$ (kcal/mol)	3	3.5 ± 1
$\Delta F_{(3)\to(4)}$ (kcal/mol)	-20	-20

Conclusions

In this work, we perform a critical assessment of different procedures to build a NNP training set starting from AIMD US trajectories of chemical reactions in solution, exploiting two reactions along the Strecker pathway[17] as test cases. The systematic comparison of AIMD training sets increasingly more dense along the reaction coordinate clearly indicates a threshold for achieving NNP free energy profiles of ab initio accuracy.

We also provide an approach to exploit the disagreement between predictions of equally-trained NNPs not only as diagnostics but also to render robust and stable the long-time dynamics, hence to achieve satisfactory statistical sampling, a crucial advantage for free-energy reconstruction.

Our investigation suggests a new protocol for the accurate, ab initio quality, NNP calculation of FES of chemical reactions in solution: the computational load of pure and costly AIMD simulations would be limited to preliminary metadynamics exploration with mechanism-agnostic general-purpose CVs,[18] followed by committor analysis to define mechanism-

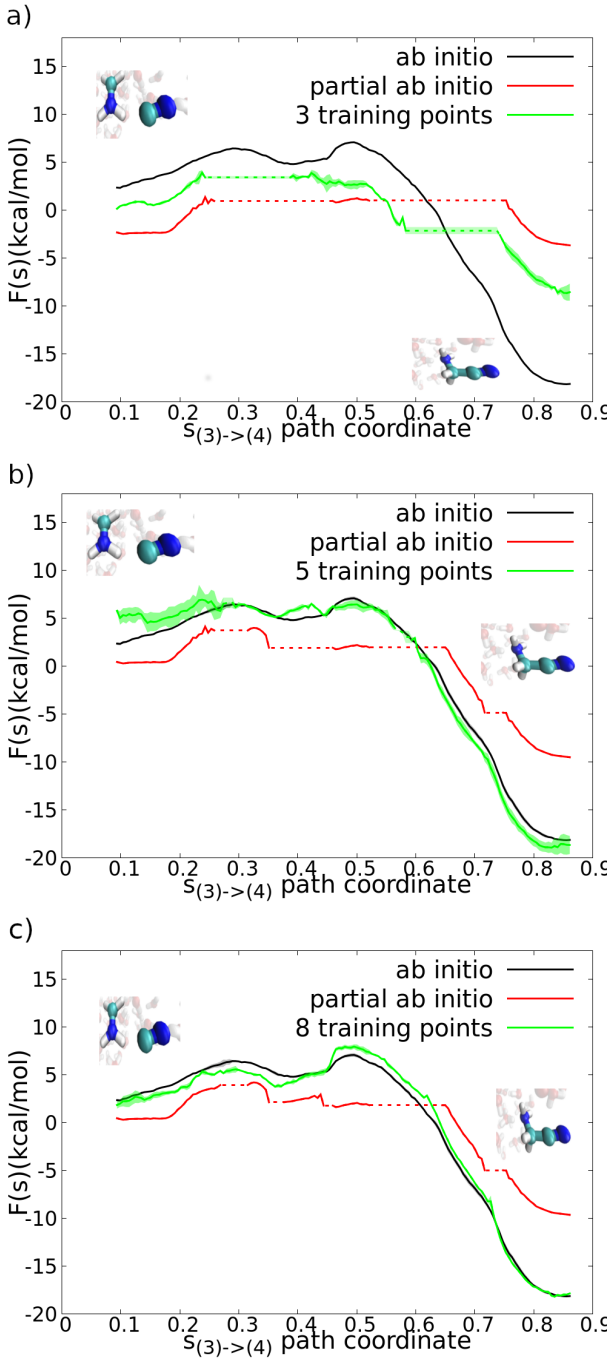


Figure 10: Panel **a)**: US-obtained free energy profiles using full AIMD (black line), and NNP trained on AIMD equilibration of reactants, products and transition state windows (green line). The red line represents the free energy which would be obtained using only the three AIMD US windows (reactants, products, transition state). The shaded zones correspond to the estimated statistical errors. Panel **b)**: same as **a)** but using the five training points in panel **b)** of figure 9. Panel **c)**: same as **b)** but using the eight training points in panel **c)** of figure 9. The reactants and products endpoint-configurations are also reported on the graphs.

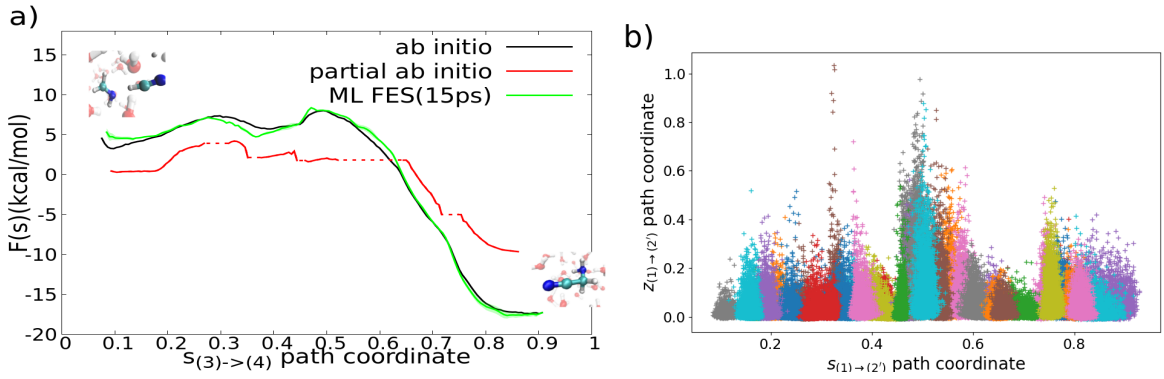


Figure 11: Panel **a)** free energy profile obtained as in panel **c)** of figure 10 but applying our velocity transformation (equation (10)). Panel **b)** corresponding instantaneous location of the NNP configurations in the (s,z) path CV coordinate plane.

specific CVs,[17] and finishing with a limited number of AIMD US trajectories along the reaction path, used for training a NNP.

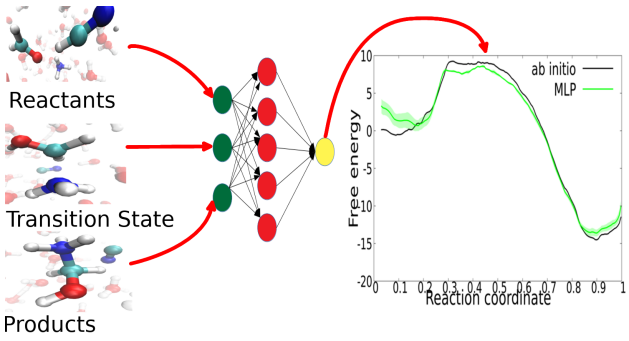
As tested on two important chemical steps (with very different FES) of the classic Strecker-cyanohydrin reaction for the synthesis of amino-acids in solution, the proposed protocol for optimal training set construction and NNP trajectory stabilization allows reproducing with excellent agreement the benchmark ab initio FES for a fraction of the computational effort.

We expect our approach to be easily generalizable to a range chemical reactions in solution, allowing a limited and incremental use of costly AIMD calculations only if and when needed. In perspective, this controlled and efficient scheme will help to exploit NNPs to overcome a significant computational bottleneck in the accurate calculation of free-energy profiles in solution chemistry.

Acknowledgement

This work was granted access to the HPC ressources of IDRIS (Institut du Développement et des Ressources en Informatique Scientifique) under the allocations 2020-A0080901387 and 2021-A0100910143 attributed by GENCI (Grand Equipement National de Calcul Intensif).

For table of contents only



References

(1) Pérez-Villa, A.; Pietrucci, F.; Saitta, A. M. Prebiotic chemistry and origins of life research with atomistic computer simulations. *Phys. Life Rev.* **2020**, *34-35*, 105–135.

(2) Napier, W. M.; Wickramasinghe, J. T.; Wickramasinghe, N. C. The origin of life in comets. *Int. J. Astrobiol.* **2007**, *6*, 321–323.

(3) Bar-Nun, A.; Bar-Nun, N.; Bauer, S. H.; Sagan, C. Shock Synthesis of Amino Acids in Simulated Primitive Environments. *Science* **1970**, *168*, 470–472.

(4) Burton, A. S.; Stern, J. C.; Elsila, J. E.; Glavin, D. P.; Dworkin, J. P. Understanding prebiotic chemistry through the analysis of extraterrestrial amino acids and nucleobases in meteorites. *Chem. Soc. Rev.* **2012**, *41*, 5459–5472.

(5) Marcellus, P. d.; Meinert, C.; Myrgorodska, I.; Nahon, L.; Buhse, T.; d'Hendecourt, L. L. S.; Meierhenrich, U. J. Aldehydes and sugars from evolved precometary ice analogs: Importance of ices in astrochemical and prebiotic evolution. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 965–970.

(6) Martin, W.; Baross, J.; Kelley, D.; Russell, M. J. Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* **2008**, *6*, 805–814.

(7) Wächtershäuser, G. Groundworks for an evolutionary biochemistry: The iron-sulphur world. *Prog. Biophys. Mol.* **1992**, *58*, 85–201.

(8) Brack, A. In *Developments in Clay Science*; Bergaya, F., Lagaly, G., Eds.; Handbook of Clay Science; Elsevier, 2013; Vol. 5; Chapter 7.4, pp 507–521.

(9) Ferris, J. P. Mineral Catalysis and Prebiotic Synthesis: Montmorillonite-Catalyzed Formation of RNA. *Elements* **2005**, *1*, 145–149.

(10) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562–12566.

(11) Laio, A.; Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* **2008**, *71*, 126601.

(12) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

(13) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

(14) Miller, S. L. A Production of Amino Acids Under Possible Primitive Earth Conditions. *Science* **1953**, *117*, 528–529.

(15) Saitta, A. M.; Saija, F. Miller experiments in atomistic computer simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 13768–13773.

(16) Pietrucci, F.; Aponte, J. C.; Starr, R.; Pérez-Villa, A.; Elsila, J. E.; Dworkin, J. P.; Saitta, A. M. Hydrothermal Decomposition of Amino Acids and Origins of Prebiotic Meteoritic Organic Compounds. *ACS Earth Space Chem.* **2018**, *2*, 588–598.

(17) Magrino, T.; Pietrucci, F.; Saitta, A. M. Step by Step Strecker Amino Acid Synthesis from Ab Initio Prebiotic Chemistry. *J. Phys. Chem. Lett.* **2021**, *12*, 2630–2637.

(18) Pietrucci, F.; Saitta, A. M. Formamide reaction network in gas phase and solution via a unified theoretical approach: Toward a reconciliation of different prebiotic scenarios. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 15030–15035.

(19) Brigiano, F. S.; Gierada, M.; Tielens, F.; Pietrucci, F. Mechanism and Free-Energy Landscape of Peptide Bond Formation at the Silica–Water Interface. *ACS Catalysis* **2022**, *12*, 2821–2830.

(20) Kroonblawd, M. P.; Pietrucci, F.; Saitta, A. M.; Goldman, N. Generating Converged Accurate Free Energy Surfaces for Chemical Reactions with a Force-Matched Semiempirical Model. *J. Chem. Theory Comput.* **2018**, *14*, 2207–2218.

(21) Yang, M.; Bonati, L.; Polino, D.; Parrinello, M. Using metadynamics to build neural network potentials for reactive events: the case of urea decomposition in water. *Catal.* **2022**, *387*, 143–149.

(22) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(23) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(24) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(25) Lindsey, R. K.; Fried, L. E.; Goldman, N. ChIMES: A Force Matched Potential with Explicit Three-Body Interactions for Molten Carbon. *J. Chem. Theory Comput.* **2017**, *13*, 6222–6229.

(26) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178–184.

(27) Zhang, L.; Han, J.; Wang, H.; Saidi, W. A.; Car, R.; E, W. End-to-end Symmetry Preserving Inter-atomic Potential Energy Model for Finite and Extended Systems. *Proceedings of the 32nd International Conference on Neural Information Processing System* **2018**, 4441–4451.

(28) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **2020**, *11*, 5223.

(29) Lu, D.; Wang, H.; Chen, M.; Lin, L.; Car, R.; E, W.; Jia, W.; Zhang, L. 86 PFLOPS Deep Potential Molecular Dynamics simulation of 100 million atoms with ab initio accuracy. *Comput. Phys. Commun.* **2021**, *259*, 107624.

(30) Morawietz, T.; Singraber, A.; Dellago, C.; Behler, J. How van der Waals interactions determine the unique properties of water. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 8368–8373.

(31) Gerrits, N.; Shakouri, K.; Behler, J.; Kroes, G.-J. Accurate Probabilities for Highly Activated Reaction of Polyatomic Molecules on Surfaces Using a High-Dimensional Neural Network Potential: CHD3 + Cu(111). *J. Phys. Chem. Lett.* **2019**, *10*, 1763–1768.

(32) Schran, C.; Brezina, K.; Marsalek, O. Committee neural network potentials control generalization errors and enable active learning. *J. Chem. Phys.* **2020**, *153*, 104105.

(33) Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.* **2015**, *115*, 1032–1050.

(34) Zhang, Y.; Wang, H.; Chen, W.; Zeng, J.; Zhang, L.; Wang, H.; E, W. DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Comput. Phys. Commun.* **2020**, *253*, 107206.

(35) Xu, N.; Shi, Y.; He, Y.; Shao, Q. A Deep-Learning Potential for Crystalline and Amorphous Li–Si Alloys. *J. Phys. Chem. C* **2020**, *124*, 16278–16288.

(36) Zhang, L.; Wang, H.; Car, R.; E, W. Phase Diagram of a Deep Potential Water Model. *Phys. Rev. Lett.* **2021**, *126*, 236001.

(37) Imbalzano, G.; Zhuang, Y.; Kapil, V.; Rossi, K.; Engel, E. A.; Grasselli, F.; Ceriotti, M. Uncertainty estimation for molecular dynamics and sampling. *J. Chem. Phys.* **2021**, *154*, 074102.

(38) Schran, C.; Thiemann, F. L.; Rowe, P.; Müller, E. A.; Marsalek, O.; Michaelides, A. Machine learning potentials for complex aqueous systems made simple. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, e2110077118.

(39) Kahle, L.; Zipoli, F. On the Quality of Uncertainty Estimates from Neural Network Potential Ensembles. *Phys. Rev. E* **2022**, *105*, 015311.

(40) Xu, J.; Cao, X.-M.; Hu, P. Accelerating Metadynamics-Based Free-Energy Calculations with Adaptive Machine Learning Potentials. *J. Chem. Theory Comput.* **2021**, *17*, 4465–4476.

(41) Pan, X.; Yang, J.; Van, R.; Epifanovsky, E.; Ho, J.; Huang, J.; Pu, J.; Mei, Y.; Nam, K.; Shao, Y. Machine-Learning-Assisted Free Energy Simulation of Solution-Phase and Enzyme Reactions. *J. Chem. Theory Comput.* **2021**, *17*, 5745–5758.

(42) Bučko, T.; Gešvandtnerová, M.; Rocca, D. Ab Initio Calculations of Free Energy of Activation at Multiple Electronic Structure Levels Made Affordable: An Effective Combination of Perturbation Theory and Machine Learning. *Journal of Chemical Theory and Computation* **2020**, *16*, 6049–6060, PMID: 32786917.

(43) Young, T. A.; Johnston-Wood, T.; Deringer, V. L.; Duarte, F. A transferable active-learning strategy for reactive molecular force fields. *Chem. Sci.* **2021**, *12*, 10944–10955.

(44) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* **2017**,

(45) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

(46) Bonomi, M. et al. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **2019**, *16*, 670–673.

(47) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

(48) Pietrucci, F. Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead. *Rev. Phys.* **2017**, *2*, 32–45.

(49) Roux, B. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.* **1995**, *91*, 275–282.

(50) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *J. Chem. Phys.* **2007**, *126*, 054103.

(51) Pérez-Villa, A.; Saitta, A. M.; Georgelin, T.; Lambert, J.-F.; Guyot, F.; Maurel, M.-C.; Pietrucci, F. Synthesis of RNA Nucleotides in Plausible Prebiotic Conditions from ab Initio Computer Simulations. *J. Phys. Chem. Lett.* **2018**, *9*, 4981–4987.

(52) Jónsson, H.; Mills, G.; Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. **1998**,

(53) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. Multidimensional free-energy calculations using the weighted histogram analysis method. *Journal of Computational Chemistry* **1995**, *16*, 1339–1350.

(54) Grossfield, A. WHAM: the weighted histogram analysis method version 2.0.10.2. `http://membrane.urmc.rochester.edu/?page_id=126`.

(55) Bučko, T.; Chibani, S.; Paul, J.-F.; Cantrel, L.; Badawi, M. Dissociative iodomethane adsorption on Ag-MOR and the formation of AgI clusters: an ab initio molecular dynamics study. *Phys. Chem. Chem. Phys.* **2017**, *19*, 27530–27543.