ADVERSARIAL ATTACKS ON UNCERTAINTY ENABLE ACTIVE LEARNING FOR NEURAL NETWORK POTENTIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural network (NN)-based interatomic potentials provide fast prediction of potential energy surfaces with the accuracy of electronic structure methods. However, NN predictions are only reliable within well-learned training domains, and show volatile behavior when extrapolating. Uncertainty quantification through NN committees identifies domains with low prediction confidence, but thoroughly exploring the configuration space when training NN potentials often requires slow atomistic simulations or exhaustive sampling. Here, we employ adversarial attacks with a differentiable uncertainty metric to sample new molecular geometries and bootstrap NN potentials. In combination with an active learning loop, the extrapolation power of NN potentials is improved beyond the original training data with few additional samples. The framework is demonstrated on the ammonia system, leading to better sampling of kinetic barriers without extensive prior data on the relevant geometries.

1 INTRODUCTION

Neural networks (NNs) have been widely used to fit interatomic potentials with high accuracy and low inference cost (Behler & Parrinello, 2007). Despite their remarkable capacity to interpolate between data points, NNs are known to perform poorly outside of their training domain (Barrett et al., 2018) and may fail catastrophically for rare events. To avoid exhaustive exploration of input space, quantifying model uncertainty becomes key, since it allows distinguishing new inputs that are likely to be informative and worth labeling with *ab initio* simulations, from those close to configurations already represented in the training data. In particular, NN committees have been used as a strategy to quantify epistemic uncertainty (Lakshminarayanan et al., 2017). However, even when uncertainty estimates are available to distinguish informative from uninformative inputs, machine learning (ML)-based potentials still rely on atomistic simulations to generate new trial configurations (Shapeev et al., 2020). With dynamics simulations executed with NN potentials, new outputs are either correlated to the training data or unphysical due to unstable trajectories. Hence, inverting the problem of exploring the configuration space with NN potentials would allow for a more efficient sampling of transition states and dynamic control (Noé et al., 2019).

In this work, we propose an inverse sampling strategy for NN-based atomistic simulations by performing gradient-based optimization of a differentiable uncertainty metric. Building on the concept of adversarial attacks from the ML literature (Szegedy et al., 2013; Goodfellow et al., 2015), new molecular conformations are sampled by backpropagating atomic displacements that maximize the uncertainty of an NN committee while balancing thermodynamic likelihood. These new configurations are then evaluated using atomistic simulations and used to retrain the NNs in an active learning loop. The technique is able to bootstrap training data for NN potentials starting from few configurations, improve their extrapolation power, and efficiently explore the configuration space. This work provides a new method to explore potential energy landscapes without the need for brute-force *ab initio* MD simulations to propose trial configurations.

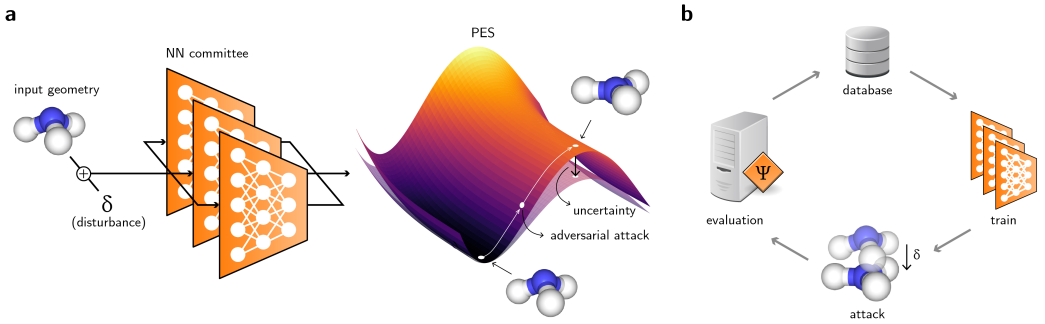

Figure 1: **a**, Schematic diagram of the method. Nuclear coordinates of an input molecule are slightly displaced by δ. Then, a potential energy surface (PES) and its associated uncertainty are calculated with an NN potential committee. By backpropagating an adversarial loss through the NN committee, the disturbance δ can be updated using gradient ascent techniques until the adversarial loss is maximized, thus sampling states that compromise high uncertainty with low energy. **b**, Schematic diagram of the active learning loop used to train the NN potential committee. The evaluation can be performed with classical force fields or electronic structure methods.

2 THEORY

When developing adversarially robust models, the objective is to find the parameters θ that minimize the loss $\mathcal{L}$ subject to a perturbation δ (Tsipras et al., 2018),

$$\min_{\theta} \mathop{\mathbb{E}}_{(X,E,\mathbf{F})\sim\mathcal{D}} \left[\max_{\delta\in\Delta} \mathcal{L}\left(X_\delta, E_\delta, \mathbf{F}_\delta; \theta\right) \right], \tag{1}$$

with Δ the set of allowed perturbations, and X and X_δ the original and perturbed geometries from a dataset $\mathcal{D}$, respectively, with their corresponding energies (E, E_δ) and forces ($\mathbf{F}$, $\mathbf{F}_\delta$). In the ML literature, Δ is often chosen as the set of ℓ_p-bounded perturbations for a given ε, $\Delta = \{\delta \in \mathbb{R} \mid ||\delta||_p \leq \varepsilon\}$.

We propose a method to obtain adversarially-robust NN potentials by combining adversarial attacks, uncertainty quantification, and active learning. In this framework, an adversarial attack maximizes the uncertainty in the property under prediction (Fig. 1a). Then, ground-truth properties are generated for the adversarial examples using density functional theory (DFT) or classical force fields. Finally, the NN committee is retrained on the original dataset and the newly-sampled geometries, restarting the loop (Fig. 1b). New rounds of this active learning can be performed until the test error is sufficiently low or the phase space is explored to a desirable degree.

Within this pipeline, new geometries are sampled by performing an adversarial attack that maximizes an adversarial loss such as

$$\max_{\delta\in\Delta} \mathcal{L}_{\text{adv}}(X, \delta; \theta) = \max_{\delta\in\Delta} \sigma_F^2(X_\delta), \tag{2}$$

where σ_F^2 is the force variance.

In the context of atomistic simulations, the perturbation δ is applied only to the nuclear coordinates, $X_\delta = (\mathbf{Z}, \mathbf{R} + \boldsymbol{\delta})$, $\boldsymbol{\delta} \in \mathbb{R}^{n\times 3}$. The set Δ can be defined by appropriately choosing ε, the maximum p-norm of δ. However, in atomistic simulations, it is often interesting to express these limits in terms of the energy of the states to be sampled, and the sampling temperature. The probability P that a state X_δ with predicted energy $\bar{E}(X_\delta)$ can be sampled, disregarding entropic contributions, is

$$P(X_\delta) = \frac{\exp\left(-\frac{\bar{E}(X_\delta)}{kT}\right)}{\sum_{(X,E,\mathbf{F})\in\mathcal{D}} \exp\left(-\frac{E}{kT}\right)}, \tag{3}$$

with k being the Boltzmann constant and denominator being the partition function of system at a given temperature T constructed from the ground truth data $\mathcal{D}$. Finally, instead of limiting the norm of δ, the adversarial objective can be modified to limit the energy of sampled states by combining Eqs. (2) and (3),

$$\max_\delta \mathcal{L}_{\text{adv}}(X, \delta; \theta) = \max_\delta P(X_\delta)\sigma_F^2(X_\delta). \tag{4}$$

Using automatic differentiation strategies, the value of δ at iteration i can be obtained using gradient ascent techniques,

$$\delta^{(i+1)} = \delta^{(i)} + \alpha_\delta \frac{\partial \mathcal{L}_{\text{adv}}}{\partial \delta}, \tag{5}$$

where α_δ is the learning rate for the adversarial attack.

3 RESULTS

As an example, we bootstrap an NN potential to study the nitrogen inversion in ammonia. An NN committee consisting of SchNet models (Schütt et al., 2018) is first trained with Hessian-displaced geometries data (see Methods). Then, new geometries are sampled by performing an adversarial attack on the ground state geometry, and later evaluated using DFT. After repeating the train-attack-evaluate loop twice, the landscape of conformations is analyzed and compared with random displacements. Fig. 2a shows a UMAP visualization (McInnes et al., 2018) of the conformers, as compared by their similarity using the Smooth Overlap of Atomic Positions (SOAP) representation (Bartók et al., 2013; De et al., 2016). A qualitative analysis of the UMAP plot shows that adversarial attacks from both generations rarely resemble the training set in terms of geometric similarity. In addition, small values of random distortions $\delta \sim \mathcal{U}(-\sigma_\delta, \sigma_\delta)$ for a uniform distribution $\mathcal{U}$ create geometries that are very similar to Hessian-displaced ones. While higher values of σ_δ (e.g. $\sigma_\delta = 0.3$ Å) enable a larger conformational space to be explored, geometries with very high energy end up being sampled (Fig. 2b). On the other hand, energies of adversarially created conformations have a more reasonable upper bound. Fig. 2c compares the degree of distortion of the geometries with respect to their energies. It further confirms that the adversarial strategy navigates the conformational space to find highly distorted states while staying within reasonable energy bounds. Once the adversarially-sampled conformations are used in training, predictions for the energy barrier in the nitrogen inversion improve substantially (Fig. 2d). While the first generation of the NN potential underestimates the energy barrier by about 1 kcal/mol with respect to the DFT reference, the prediction from the second generation has less than 0.25 kcal/mol of error for the inversion barrier. In contrast, predictions from an NN committee trained on randomly-sampled geometries show much higher error. This suggests that adversarial attacks were able to sample geometries similar to the transition state of the nitrogen inversion reaction and accurately interpolate the energy barrier without the need to explicitly add this reaction path into the training set.

The evolution of the phase space of each NN committee is further compared in the projected PES of Fig. 2e (see Methods). Two collective variables (CVs) are defined to represent the phase space of this molecule: the radius of the circumference defined by the three hydrogen atoms (R) and the distance between the nitrogen atom and the plane defined by the three hydrogens (Z) (see also Fig. 3). Fig. 2e shows these CVs normalized by the values found in the ground state geometry. Adversarial attacks expand the configuration space used as train set for NN committees and bring the phase space closer to the ground truth, thus locally lowering the uncertainty of forces. Fig. 2f shows the RMSE between the NN- and DFT-predicted energies for the phase space of Fig. 2e. When energies smaller than 5 kcal/mol are compared, the NN committees from all stages of the active learning loop

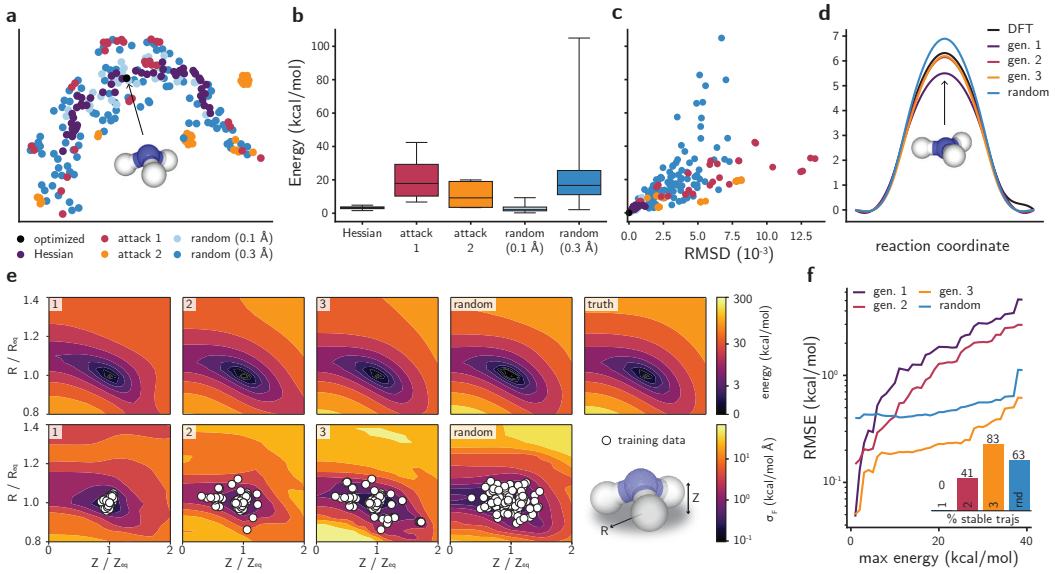

Figure 2: **a**, UMAP plot for the SOAP-based similarity between ammonia geometries. Both axes are on the same scale. **b**, Distribution of DFT energies for conformations sampled with different methods. The horizontal line is the median, the box is the interquartile region and the whiskers span the range of the distribution. **c**, Relationship between DFT energy and root mean square deviation (RMSD) of a geometry with respect to the ground state structure of ammonia. The color scheme follows the legend of **a**. **d**, Energy barrier for the nitrogen inversion calculated with NEB using DFT or using the NN committee. **e**, Evolution of the PES projected onto the CVs (Z, R) for ammonia. The generation of the NN committee is shown in the top left corner of each plot. The scale bar of energies is plotted with the function $\log_{10}(1+E)$, and all energy contour plots have the same levels. Random geometries were generated with $\sigma_\delta = 0.3$ Å (see Methods). **f**, RMSE between the NN and DFT PES for each NN potential when a maximum energy is imposed for the DFT PES. **f, inset**, fraction of stable MD trajectories generated using each NN committee as force field.

display much smaller RMSE than the committee trained on random geometries, probably due to the presence of Hessian-displaced structures in their training set. Moreover, the third generation of NN committees offers the best predictions of energies for structures with DFT energy up to 40 kcal/mol. Finally, the adversarial training yields models capable of performing stable MD simulations. 83% of the trajectories produced by the third generation of adversarially based NN committees are stable, even though the NN-based MD geometries include data points originally not in the training set (Fig. 4). In contrast, only 63% of the trajectories are stable when the NN committee trained on random geometries is used to perform the simulations. This indicates that the adversarial sampling strategy enhances the robustness of NN-based MD simulations by seeking points less likely to be sampled in unbiased MD simulations.

4 Conclusions

In summary, we proposed a new sampling strategy for NN potentials by combining uncertainty quantification, adversarial attacks and active learning. By maximizing the uncertainty of NN predictions through a differentiable metric, new geometries can be sampled without the need for atomistic simulations. This work presents an efficient way to train NN potentials, allowing NN potentials to be bootstrapped with less data without compromising the final accuracy of the models. The method will enable the development of robust NN potentials for increasingly complex or reactive systems.

REFERENCES

David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 511–520, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 5 2013.

Becke and A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6):3098–3100, 9 1988.

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14):146401, 4 2007.

Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical Review Letters*, 97(17):170201, 10 2006.

Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics*, 18(20):13754–13769, 5 2016.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–11, 2015.

Thomas A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519, 4 1996.

Graeme Henkelman and Hannes Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *The Journal of Chemical Physics*, 113(22):9978–9985, 11 2000.

Graeme Henkelman, Blas P. Uberuaga, and Hannes Jónsson. Climbing image nudged elastic band method for finding saddle points and minimum energy paths. *Journal of Chemical Physics*, 113 (22):9901–9904, 12 2000.

Lauri Himanen, Marc O J Jäger, Eiaki V Morooka, Filippo Federici Canova, Yashasvi S Ranawat, David Z Gao, Patrick Rinke, and Adam S Foster. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, Bjørk Hammer, Cory Hargus, Eric D. Hermes, Paul C. Jennings, Peter Bjerre Jensen, James Kermode, John R. Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S. Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W. Jacobsen. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 7 2017.

Hannes Jónsson, Greg Mills, and Karsten W Jacobsen. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*, pp. 385–404. WORLD SCIENTIFIC, 6 1998.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413. Curran Associates, Inc., 2017.

Greg Landrum. RDKit: Open-source cheminformatics, 2006. URL `www.rdkit.org`.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29):861, 2018.

Simone Melchionna, Giovanni Ciccotti, and Brad Lee Holian. Hoover NPT dynamics for systems varying in shape and size. *Molecular Physics*, 78(3):533–544, 2 1993.

Frank Neese. The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(1):73–78, 1 2012.

Frank Neese. Software update: the ORCA program system, version 4.0. *WIREs Computational Molecular Science*, 8(1):e1327, 1 2018.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 9 2019.

John P. Perdew. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Physical Review B*, 33(12):8822–8824, 6 1986.

Kristof Schütt, Huziel Sauceda, P.-J. Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet - A deep learning architecture for molecules and materials. *Journal of Chemical Physics*, 148(24):241722, 6 2018.

Alexander Shapeev, Konstantin Gubaev, Evgenii Tsymbalov, and Evgeny Podryabinkin. Active Learning and Uncertainty Estimation. In Kristof T Schütt, Stefan Chmiela, O Anatole von Lilienfeld, Alexandre Tkatchenko, Koji Tsuda, and Klaus-Robert Müller (eds.), *Machine Learning Meets Quantum Physics*, pp. 309–329. Springer International Publishing, Cham, 2020.

Christian Szegedy, Alexander Toshev, and Dumitru Erhan. Deep Neural Networks for Object Detection. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pp. 2553–2561, Red Hook, NY, USA, 2013. Curran Associates Inc.

Paolo Tosco, Nikolaus Stiefl, and Gregory Landrum. Bringing the MMFF force field to the RDKit: implementation and validation. *Journal of Cheminformatics*, 6(1), 12 2014.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, Aleksander Madry, and Aleksander Mądry. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations*, pp. 161–168, 2018.

Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 9 2005.

A APPENDIX

A.1 METHODS

A.1.1 NEURAL NETWORK POTENTIALS

An NN potential is a hypothesis function h_θ that predicts a real value of energy $\hat{E} = h_\theta(X)$ for a given atomistic configuration X as input. X is generally described by n atoms with atomic numbers $\mathbf{Z} \in \mathbb{Z}_+^n$ and nuclear coordinates $\mathbf{R} \in \mathbb{R}^{n\times 3}$. Atomic forces F_{ij} on atom i and cartesian coordinate j are obtained by differentiating the output energy with respect to the atomic coordinates r_{ij},

$$\hat{F}_{ij} = -\frac{\partial \hat{E}}{\partial r_{ij}}. \tag{6}$$

The parameters θ are trained to minimize the expected loss $\mathcal{L}$ given the distribution of ground truth data $(X, E, \mathbf{F})$ according to the dataset $\mathcal{D}$,

$$\min_\theta \mathop{\mathbb{E}}_{(X,E,\mathbf{F})\sim\mathcal{D}} \left[\mathcal{L}\left(X, E, \mathbf{F}; \theta\right)\right]. \tag{7}$$

During training, the loss $\mathcal{L}$ is usually computed by taking the average mean squared error of the predicted and target properties within a batch of size N,

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left[\alpha_E||E_i - \hat{E}_i||^2 + \alpha_F||\mathbf{F}_i - \hat{\mathbf{F}}_i||^2\right], \tag{8}$$

where α_E and α_F are coefficients indicating the trade-off between energy and force-matching during training (Schütt et al., 2018). The training proceeds using stochastic gradient descent-based techniques.

A.1.2 UNCERTAINTY QUANTIFICATION

To create a differentiable metric of uncertainty, NN commitees are typically implemented by training different h_θ and obtaining a distribution of predictions for each input X. Given M models, the mean, $\bar{E}(X)$ and variance, $\sigma_E^2(X)$ of energy, as well as for forces, $\bar{\mathbf{F}}(X)$ and $\sigma_F^2(X)$, of an NN potential ensemble can be computed as

$$\bar{E}(X) = \frac{1}{M}\sum_{m=1}^{M}\hat{E}^{(m)}(X) \quad ; \quad \sigma_E^2(X) = \frac{1}{M-1}\sum_{m=1}^{M}||\hat{E}^{(m)}(X) - \bar{E}(X)||^2, \tag{9}$$

$$\bar{\mathbf{F}}(X) = \frac{1}{M}\sum_{m=1}^{M}\hat{\mathbf{F}}^{(m)}(X) \quad ; \quad \sigma_F^2(X) = \frac{1}{M-1}\sum_{m=1}^{M}\left[\frac{1}{3n}\sum_{i,j}||\hat{F}_{ij}^{(m)}(X) - \bar{F}_{ij}(X)||^2\right]. \tag{10}$$

Whereas the training objective (7) rewards approaching mean energies or forces to their ground truth values, this is not guaranteed for regions outside of the training set.

A.1.3 ATOMISTIC SIMULATIONS

Initial molecular conformers were generated using RDKit (Landrum, 2006) with the MMFF94 force field (Halgren, 1996; Tosco et al., 2014). DFT structural optimizations and single-point calculations were performed using the BP86-D3/def2-SVP (Becke & Becke, 1988; Perdew, 1986; Weigend & Ahlrichs, 2005) level of theory as implemented in ORCA (Neese, 2012; 2018). NEB calculations (Jónsson et al., 1998; Henkelman et al., 2000; Henkelman & Jónsson, 2000) were performed with 11 images using the FIRE algorithm (Bitzek et al., 2006) as implemented in the Atomic Simulation Environment (Hjorth Larsen et al., 2017). Hessian-displaced geometries were created by randomly

displacing the atoms from their ground state conformation in the direction of normal mode vectors with temperatures between 250 and 750 K. In total, 78 training geometries were used as initial dataset.

For each generation, five NNs with the SchNet architecture (Schütt et al., 2018) were employed. Each model used four convolutions, 256 filters, atom basis of size 256, 32 learnable gaussians and cutoff of 5.0 Å. The models were trained on different splits of the initial dataset (ratios $60:20:20$ for train : validation : test) for 500 epochs, using the Adam optimizer with an initial learning rate of 3×10^{-4} and batch size of 30. A scheduler reduced the learning rate by a factor of 0.5 if 30 epochs passed without improvement in the validation set. The training coefficients α_E and α_F (see Eq. 8) were set to 0.1 and 1, respectively.

Adversarial attacks were initialized by displacing the ground state geometry of ammonia by $\delta \sim \mathcal{N}(0, 0.01\,\text{Å})$ for each coordinate. The resulting attack δ was optimized for 60 iterations using the Adam optimizer with learning rate of 0.01. The normalized temperature kT was set to 20 kcal/mol to ensure that adversarial attacks were not bound by a low sampling temperature, but by the uncertainty in force predictions. 30 adversarial attacks were sampled for each generation.

Random distortions were generated by displacing each coordinate of the ground state geometry of ammonia by a value of $\delta \sim \mathcal{U}(-\sigma_\delta, \sigma_\delta)$. The values of $\sigma_\delta = 0.1$ Å and $\sigma_\delta = 0.3$ Å were adopted. 30 (100) random samples were created for $\sigma_\delta = 0.3$ Å ($\sigma_\delta = 1.0$ Å).

NN-based MD simulations were performed in the NVT ensemble with Nosé-Hoover dynamics (Melchionna et al., 1993), 0.5 fs timesteps, and temperatures of 500, 600, 700, 800, 900, and 1000 K. 100 5 ps-long trajectories were performed for each NN committee and temperature. The ground state geometry of ammonia was used as initial configuration for all MD calculations. Trajectories were considered as unphysical if the distance between hydrogen atoms was closer than 0.80 Å or larger than 2.55 Å, or if the predicted energy was lower than the ground state energy (0 kcal/mol for the reference adopted in this work).

SOAP vectors were created using the DScribe package (Himanen et al., 2020). The cutoff radius was set as 5 Å, with spherical primitive gaussian type orbitals with standard deviation of 1 Å, basis size of 5 functions, and $L_{\max} = 6$. The vectors were averaged over sites before summing the magnetic quantum numbers.

The projected PES shown in Fig. 2e is constructed by evaluating the NN potentials on symmetrical geometries generated for each tuple (Z, R). As such, train points and adversarial attacks are projected onto this space even though the conformers display distortions not captured by the CVs (Z, R) (see Fig. 3). The RMSE between the projected PES of the NN potential and DFT calculations is taken with respect to these symmetrical geometries.

A.2 Supplementary Figures

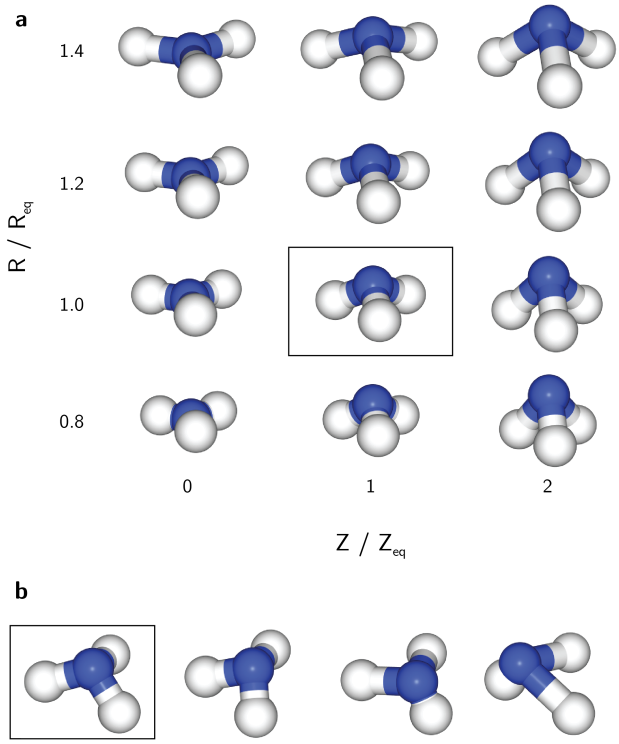

Figure 3: **a**, Example of ammonia geometries created for different values of (Z, R). **b**, Different distortions of the ground state geometry (outlined in black) that have the same values of (Z_{eq}, R_{eq})

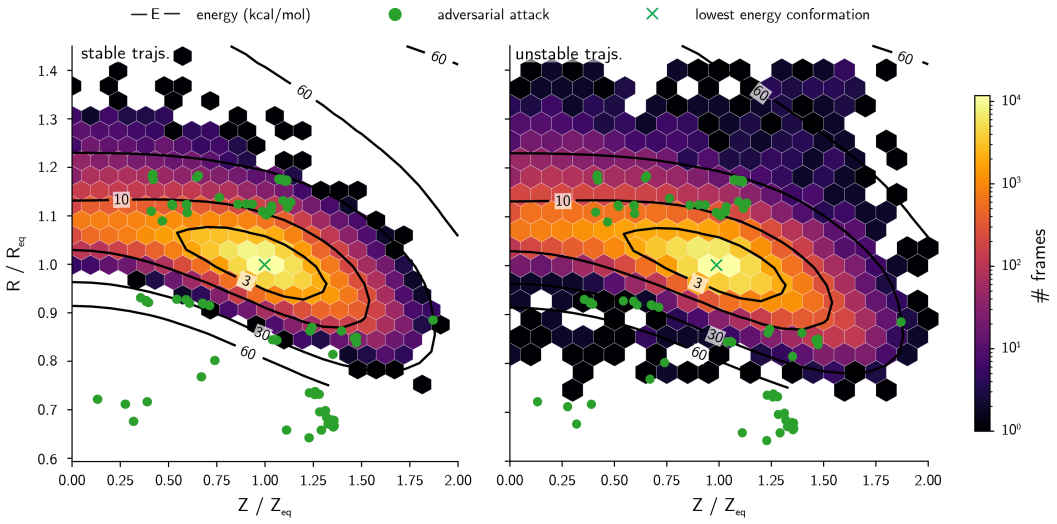

Figure 4: Density of frames for stable (left) and unstable (right) molecular dynamics trajectories obtained with the third generation of NN potentials. Contour lines indicate constant energy levels in the phase space. Points for 100 adversarial attacks performed for the third generation indicate that the adversarial strategy samples points not well explored by the NN potential in MD trajectories. Only a subset of the phase space of unstable trajectories is shown, as several data points fall outside of this region.