Electronic Supplementary Material (ESI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2024

Supplementary Information

Active Learning of Neural Network Potentials for Rare Events

Gang Seob Jung[1†], Jong Youl Choi[2] and Sangkeun Matthew Lee[2]

[1]Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[2]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

†Email: jungg@ornl.gov

Supplementary Note 1

We tested the effect of the data augmentation by adding random noise to the coordinates of atoms in the sampled configurations. The inherited data selection strategy with seven models shows the best performance regarding the activation energy's accuracy and convergence. We utilized the same conditions but without data augmentation. Figure S8 shows the result. As expected, the error of the activation energy does not go below 1 kcal/mol. Also, the convergence is not good. There is an effect from the absolute number of the newly added data number. However, NNP has no chance to learn whether the newly sampled configuration is energetically stable. The snapshot in Figure S8c (iv) shows that the broken parts have an unphysical configuration, e.g., a cluster of hydrogen and isolated carbons without any bonds. The results may imply that the challenge in training SchNet may come from the sparse configurations to reconstruct continuous PES.

Supplementary Note 2

We tested them and investigated the behaviors of the SchNet. First, we checked whether TorchANI shows similar behaviors with the configurations sampled from the SchNet and vice versa. Interestingly, SchNet does not show such unphysical behaviors with the configurations sampled from SMD through $NNP_{TorchANI}$ as shown in Figure S9. Also, TorchANI works fine. Then, we double-checked whether the training conditions of SchNet can train such deformed configurations. We utilized combinations of training and validation sets that allow high accuracy of the activation energy to train SchNet. One is from the 9th AL iteration of the random selection strategy with 3 models, and the other is from the 22nd AL iteration of the inherited selection strategy with 7 models. Both datasets are working well, as shown in Figure S10. Therefore, we hypothesized that configurations sampled through SchNet cause the challenge of training during the AL iterations. We closely checked the configurations sampled from SMD simulation through $NNP_{SchNet\ and}$ found a unique difference from DFTB and TorchANI. Two carbons (CH3) at the edge rotate just before the bond breaks with SchNet, which does not occur with DFTB and TorchANI. This was consistently observed when the SchNet was trained from datasets not sampled from the SchNet (Figure 4 and Figure S9). The exact mechanism of how SchNet produces such configurations needs to be explored further in future studies. However, this finding may provide an opportunity to improve the performance of the SchNet.

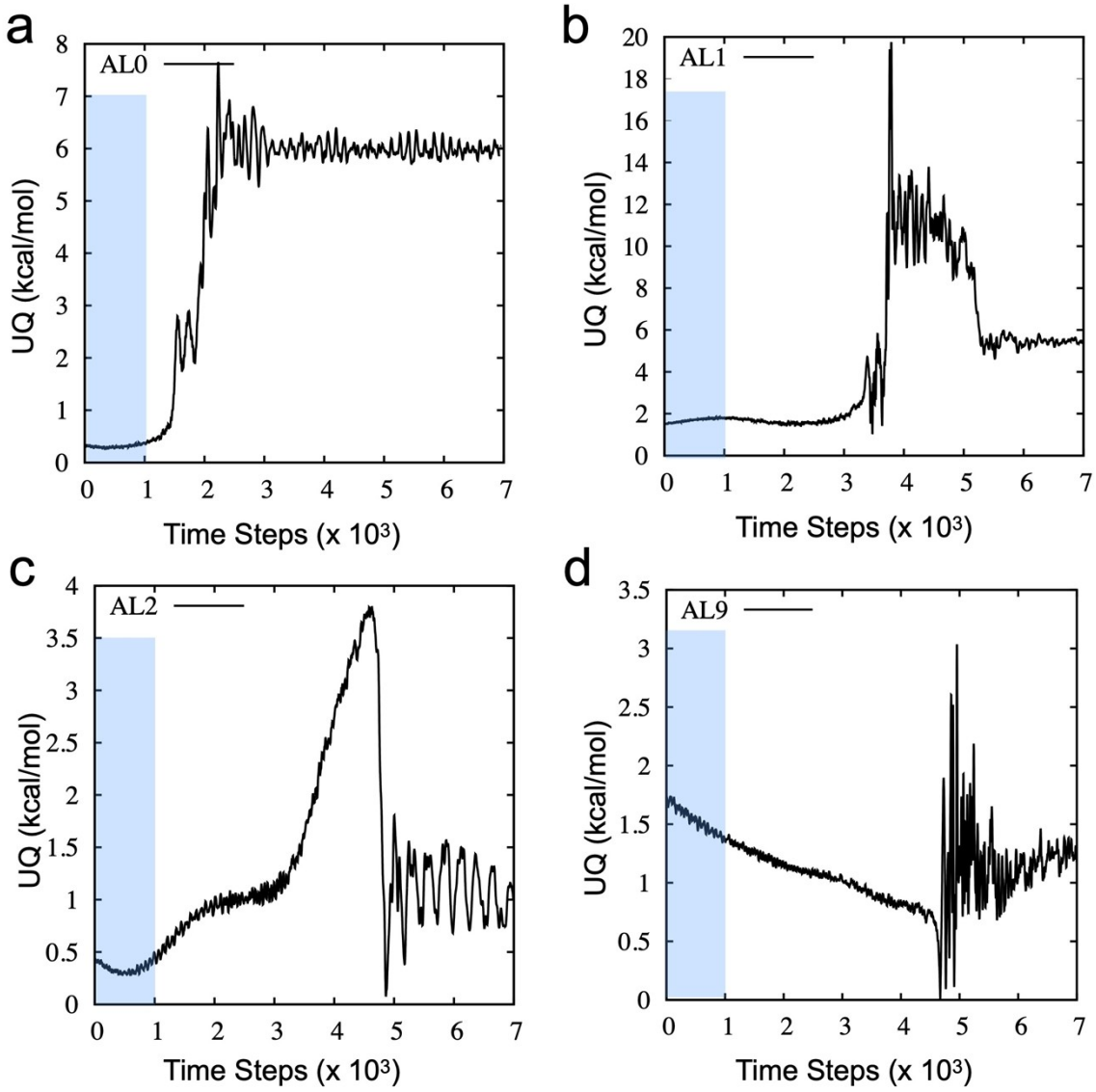

Figure S1 The examples of UQ values (TorchANI model) obtained from different active learning stages: AL0 (a), AL1 (b), AL2 (c), and AL9 (d) of random selection with 3 models. The graded area (blue) is where we gather 1000 UQ values from the ensemble model. Then we estimated the mean (μ_{UQ}) and standard deviation (σ_{UQ}). As observed in the examples, UQ values can vary even in the same sampling region (0~1x10^3 steps). Therefore, we consider the value of $\mu_{UQ} + 3\sigma_{UQ}$ as a criterion to decide whether the sampled configurations are new or not.

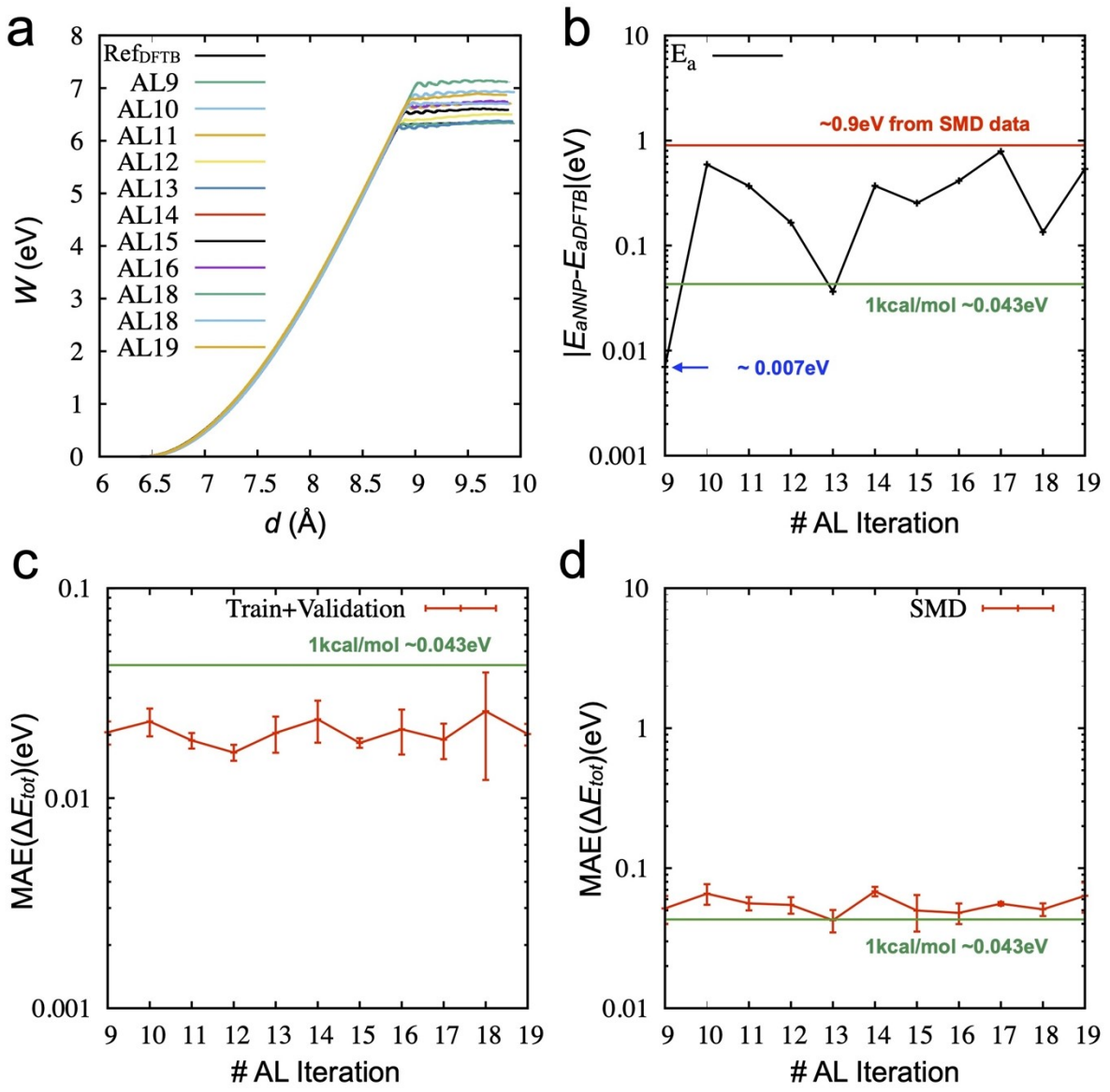

Figure S2 Active learning (AL) results from 9th to 19th iteration based on the TorchANI models: (**a**) the work function of the best model from three models with the random selection strategy. (**b**) Absolute error ($|E_{a\mathrm{NNP}} - E_{a\mathrm{DFTB}}|$) of the best model trained and selected at each AL iteration (**c**) MAE of the relative energy during the active learning based on the train/validation data at each iteration (**d**) MAE of the relative energy during the active learning based on the unseen SMD data (6,500 points).

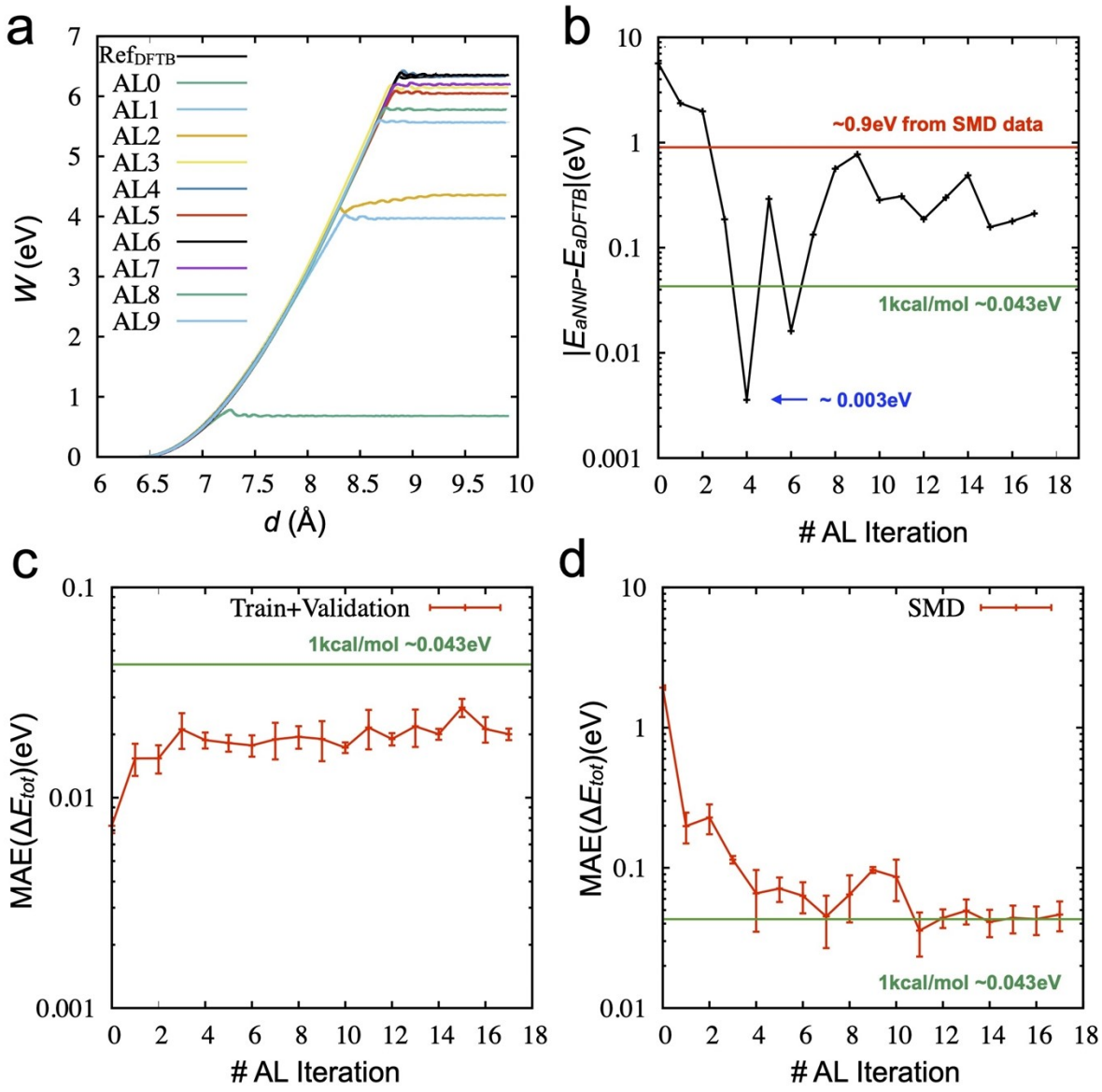

Figure S3 Active learning (AL) results based on the TorchANI models: (**a**) the work function of the best model from three models with the inherited selection strategy. (**b**) Absolute error ($|E_{a\mathrm{NNP}} - E_{a\mathrm{DFTB}}|$) of the best model trained and selected at each AL iteration (**c**) MAE of the relative energy during the active learning based on the train/validation data at each iteration (**d**) MAE of the relative energy during the active learning based on the unseen SMD data (6,500 points).

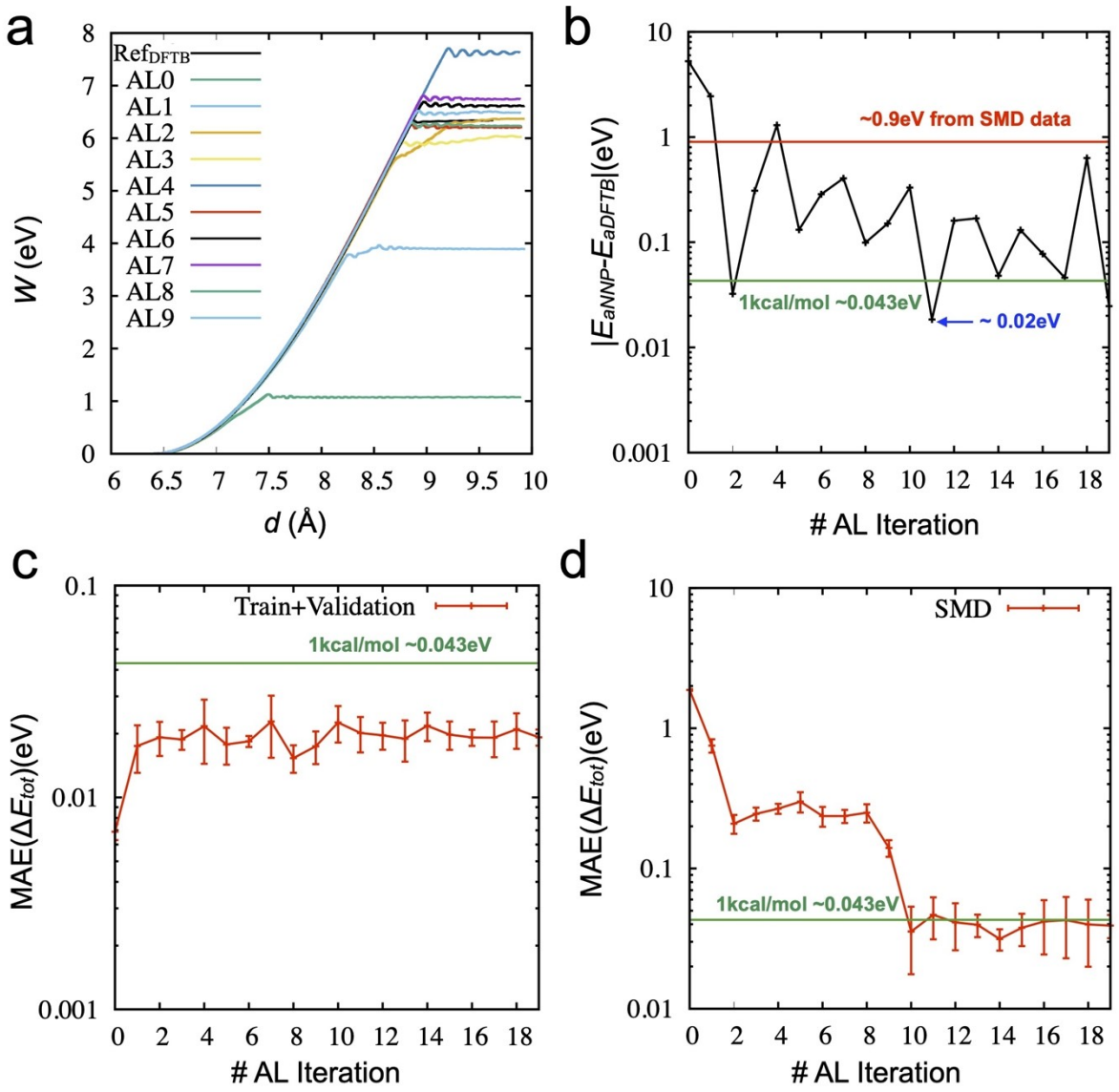

Figure S4 Active learning (AL) results based on the TorchANI models: (**a**) the work function of the best model from five models with the inherited selection strategy. (**b**) Absolute error ($|E_{a\mathrm{NNP}} - E_{a\mathrm{DFTB}}|$) of the best model trained and selected at each AL iteration (**c**) MAE of the relative energy during the active learning based on the train/validation data at each iteration (**d**) MAE of the relative energy during the active learning based on the unseen SMD data (6,500 points).

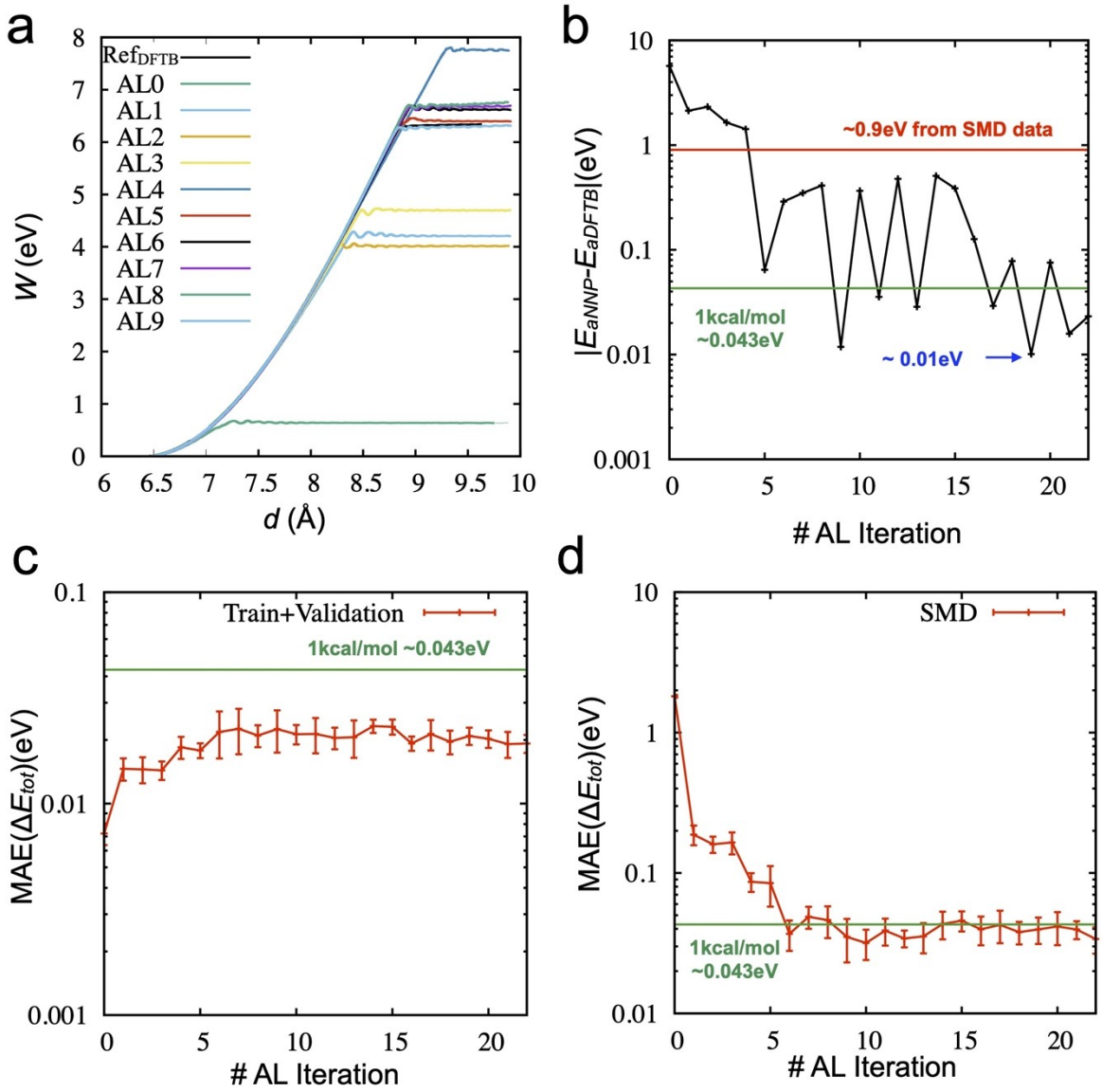

Figure S5 Active learning (AL) results based on the TorchANI models: (**a**) the work function of the best model from seven models with the inherited selection strategy. (**b**) Absolute error ($|E_{a\mathrm{NNP}} - E_{a\mathrm{DFTB}}|$) of the best model trained and selected at each AL iteration (**c**) MAE of the relative energy during the active learning based on the train/validation data at each iteration (**d**) MAE of the relative energy during the active learning based on the unseen SMD data (6,500 points).

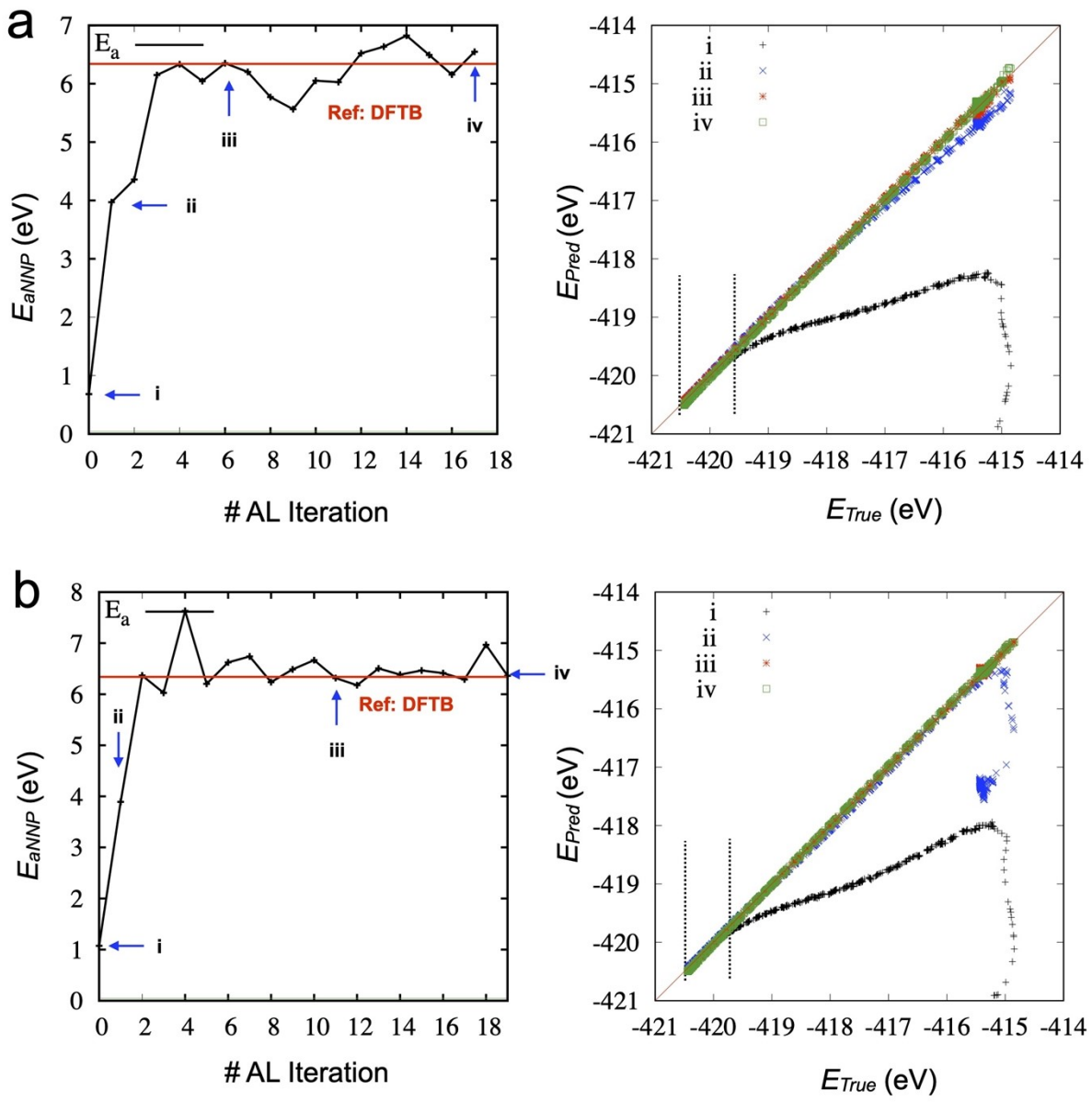

Figure S6 (left) Activation energy obtained from trained TorchANI-based SMD simulations during the active learning. (right) Scattering plots of unseen data points (6,500 points) from DFTB-based SMD simulation. The results from inherited data selection strategy. (**a**) three models, (**b**) five models.

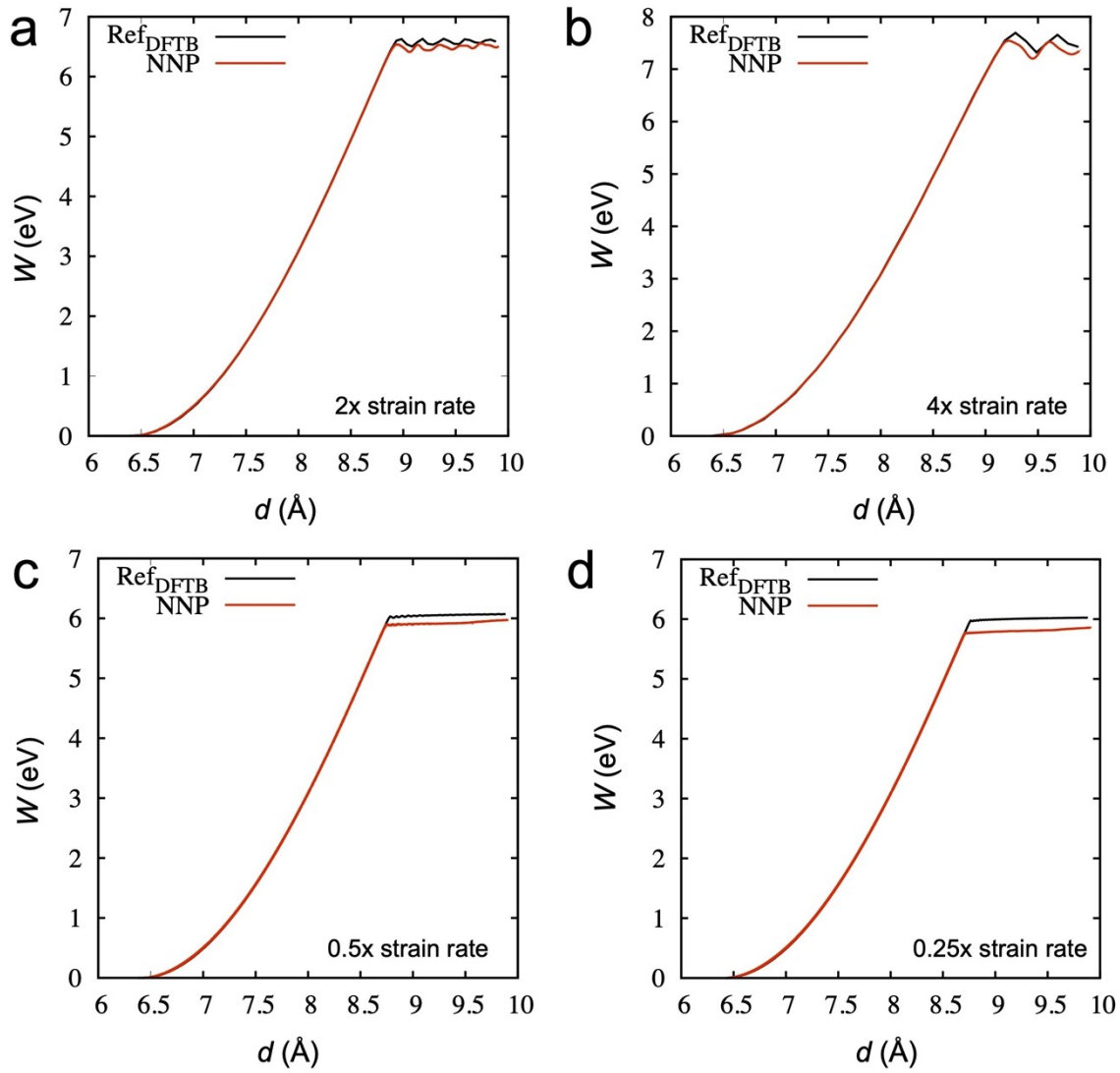

Figure S7 Loading rate effects on the estimation of activation energy based on the TorchANI models (**a**) 2 Å/ps, (**b**) 4 Å/ps, (**c**) 0.5 Å/ps, and (**d**) 0.25 Å/ps. In all cases, the error becomes larger than 1kcal/mol. However, the NNP still could capture the trend of strain rate effects, e.g., higher energy and longer breaking distance for a higher strain rate.

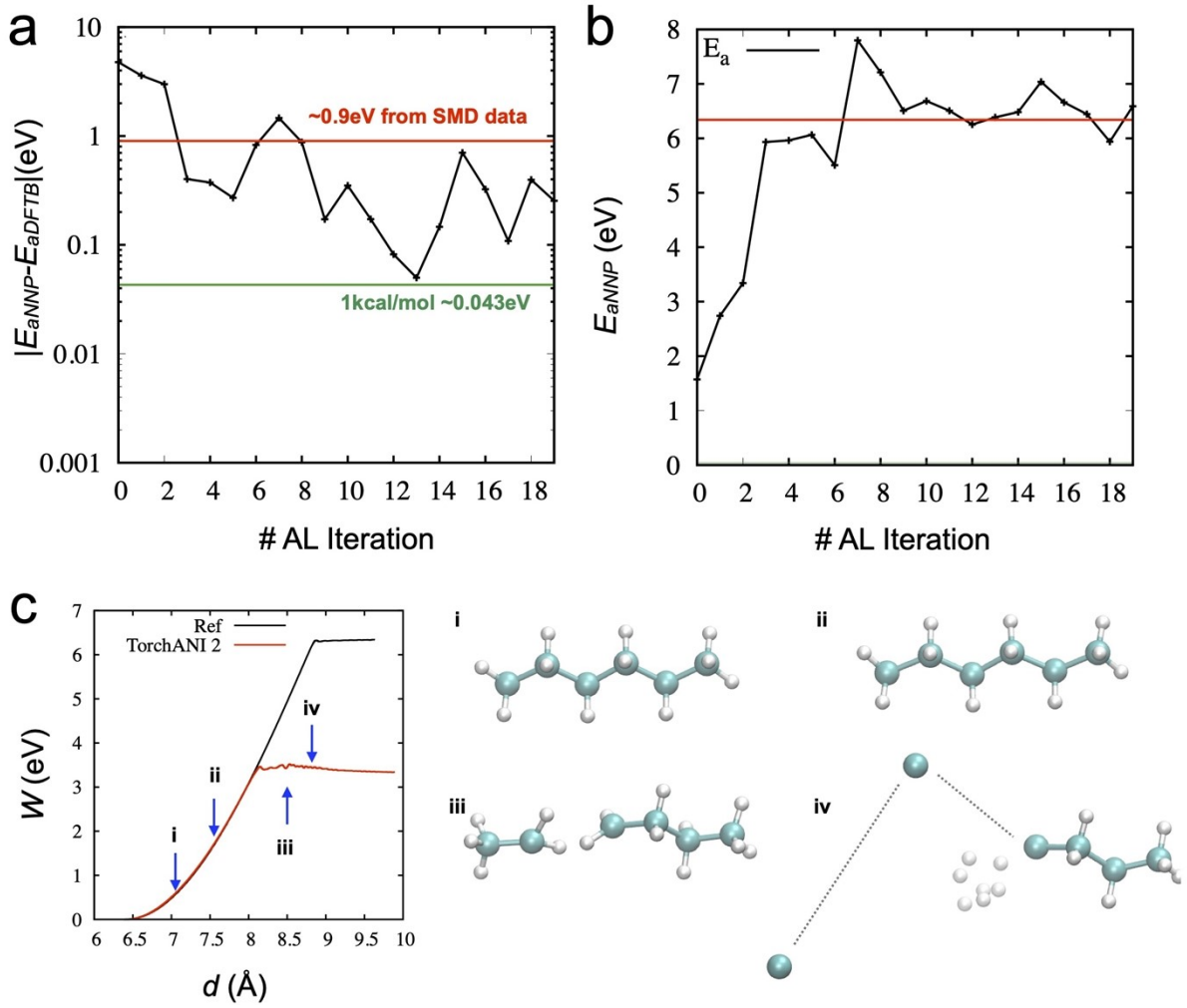

Figure S8 Active learning (AL) results based on the TorchANI models without data augmentation (seven models in the ensemble and the inherited strategy). (**a**) Absolute error ($|E_{a\mathrm{NNP}} - E_{a\mathrm{DFTB}}|$) of the best model at each AL iteration. (**b**) Activation energy obtained from the TorchANI-based SMD simulations during the active learning. (**c**) SMD simulation from the model trained and selected from the 2nd active learning with ANI model. Since there is no augmentation from newly sampled configurations, training structures of broken parts is challenging. As a result, the unphysical configuration is sampled as the snapshot (iv) in pane c.

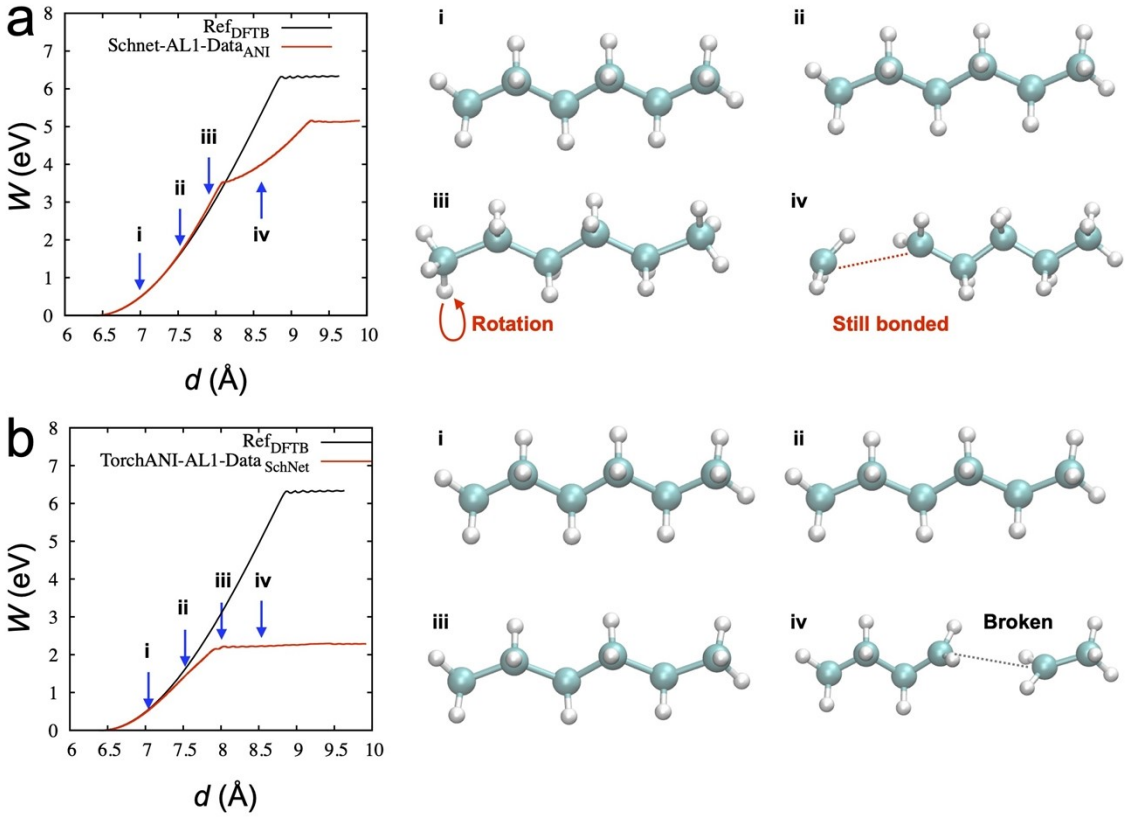

Figure S9 (**a**) Results of SMD simulation from the SchNet model trained from the configuration sampled by TorchANI model at the first AL iteration (AL1) (**b**) Results of SMD simulation from the TorchANI model trained from the configuration sampled by SchNet model at the first AL iteration (AL1). SchNet does not show unphysical configurations like Figure 4b. However, we still observed the rotation of carbons at the edge near the failure point.

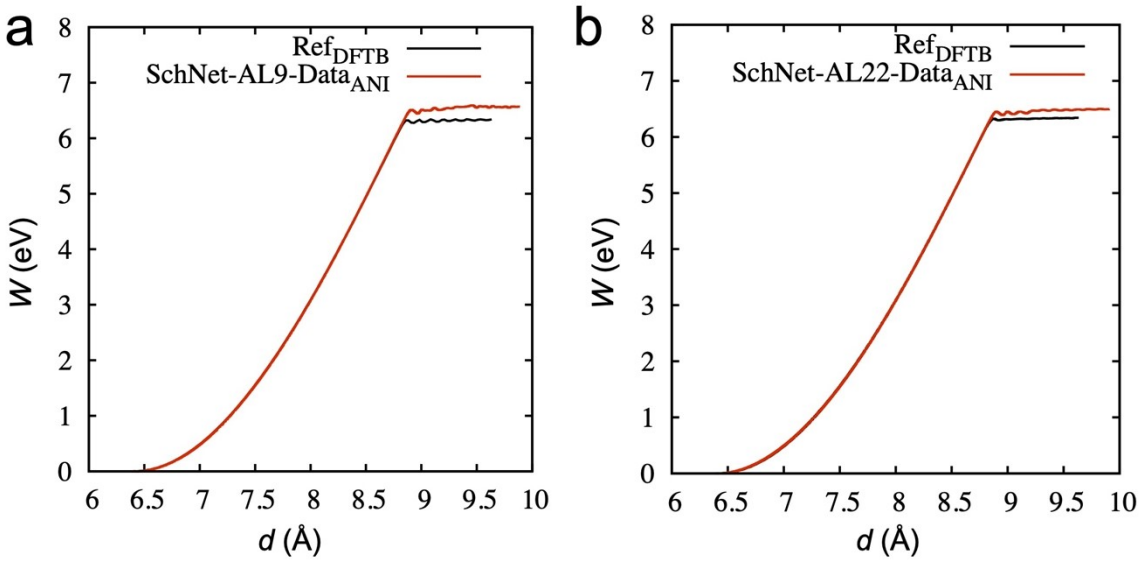

Figure S10 (**a**) Results of SMD simulation from the SchNet model trained from the configuration sampled by TorchANI model at the 9th AL iteration (AL9) with the random selection (**b**) Results of SMD simulation from the SchNet model trained from the configuration sampled by TorchANI model at the 22th AL iteration (AL22) with the inherited selection.