A dual-cutoff machine-learned potential for condensed organic systems obtained *via* uncertainty-guided active learning – Supplementary Information

Leonid Kahle,*[a] Benoit Minisini,[a] Tai Bui,[b] Jeremy T. First,[c] Corneliu Buda,[b] Thomas Goldman,[b] and Erich Wimmer[a]

S.1 Choice of DFT functional

Our calculations on polyethylene and methanol showed revPBE-vdw yielded the lowest deviation from experimental data (Figure S1 and S2), with a maximum absolute deviation of 2.2% for the *b* lattice constant of methanol. The mean absolute deviation between the revPBE-vdw functional and experimental results was $1 \pm 0.8\%$ for lattice constants calculated for the six other organic molecular crystals (Figure S3). VASP was used for all calculations with settings as for the training calculations and a k-point spacing of 0.5 $Å^{-1}$.

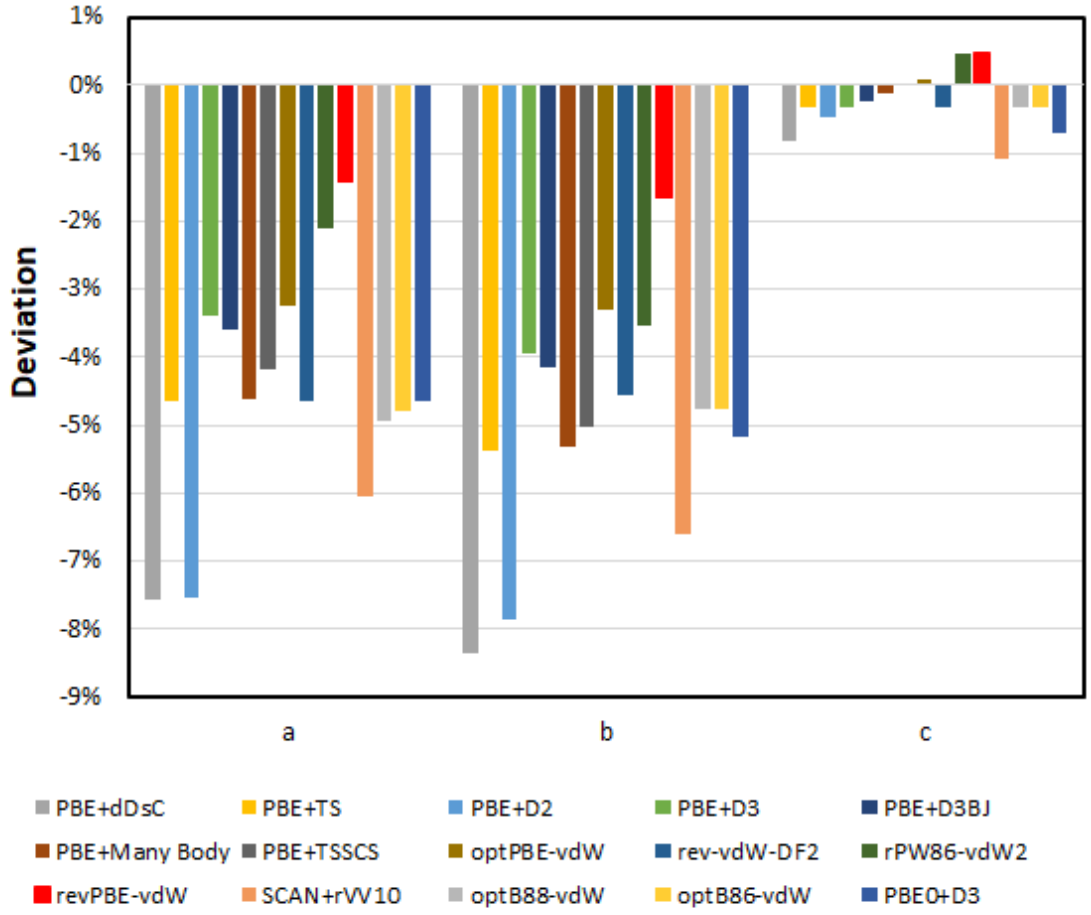

Fig. S1 Deviation between DFT (0 K) and experimental (4 K)? cell parameters (a,b, c) of crystalline polyethylene for different exchange correlation functionals.

[a] *Materials Design SARL, 42 avenue Verdier, 92120 Montrouge, France*
[b] *bp Exploration Operating Co. Ltd, Chertsey Road, Sunbury-on-Thames TW16 7LN, UK*
[c] *bp, Center for High Performance Computing, 225 Westlake Park Blvd, Houston, TX 77079, USA*

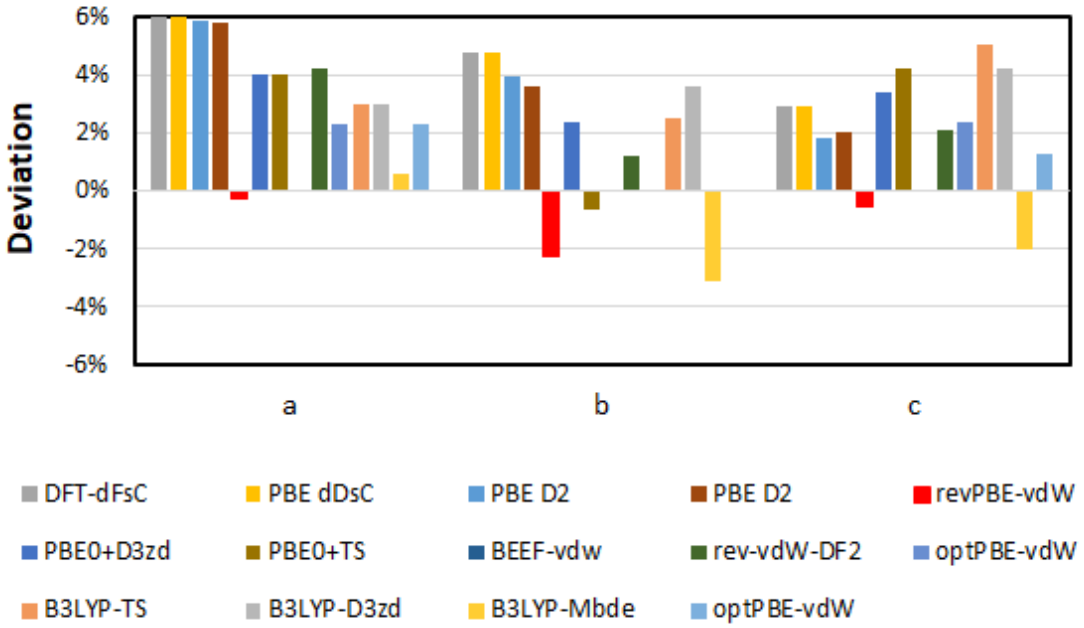

Fig. S2 Deviation between DFT (0 K) and experimental (122 K)[?] cell parameters (a,b, c) of crystalline methanol for different exchange correlation functional.

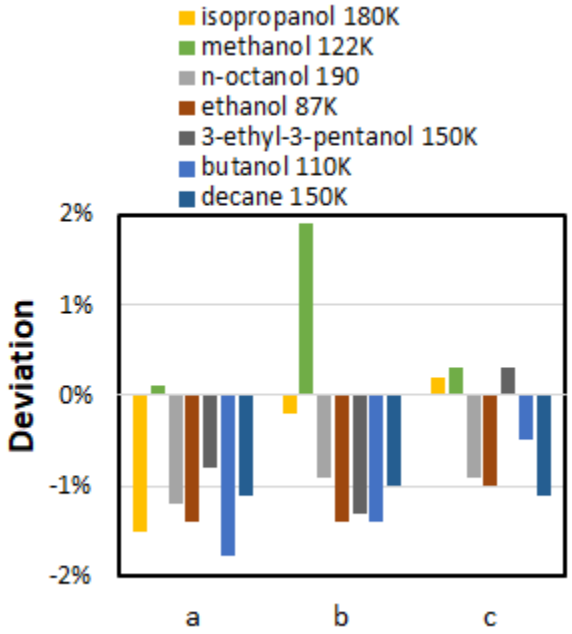

Fig. S3 Deviation between DFT (0 K) and experimental cell parameters (a,b, c) of crystalline organic materials, isopropanol,[?] methanol,[?] n-octanol,[?] ethanol,[?] 3-ethyl-3-pentanol,[?] butanol,[?] and decane.[?]

Step	Systems added	Exploration Technique	Conf. added	Conf. total
0	250 C4ol-8, 50 C4ol-1	Expl. pcff+ and rand. displ.	300	300
1	**70 C4ol-8**, 100 C6ol-6, 100 C8ol-4, 100 C10ol-3	Expl. MLP and pcff+	370	670
2	**19 C4ol-8**, **23 C6ol-6**, **13 C8ol-4**, **40 C10ol-3**	Expl. MLP	95	765
3	**37 C4ol-8, 31 C6ol-6, 26 C8ol-4, 26 C10ol-3**	Expl. MLP	120	885
4	**31 C4ol-8, 18 C6ol-6, 13 C8ol-4, 10 C10ol-3**	Expl. MLP	72	957
5	**60 C4ol-8, 18 C6ol-6, 32 C8ol-4, 73 C10ol-3**	Expl. MLP	183	1140
6	**23 C4ol-8, 7 C6ol-6, 9 C8ol-4, 27 C10ol-3**	Expl. MLP	66	1206
7	**12 C4ol-8, 5 C6ol-6, 5 C8ol-4, 4 C10ol-3, 2 C6-6, 1 C10-3**	Expl. MLP	29	1235
8	**6 C4ol-8, 7 C6ol-6, 3 C8ol-4, 3 C10ol-3, 1 C10-4**, 90 dap-1, 440 dap-3	Expl. MLP and pcff+	550	1785
9	**1 C4ol-8, 1 C6ol-6, 1 C10-4, 26 dap-3**	Expl. MLP	29	1814
10	**4 C4ol-8, 4 dap-3**	Expl. MLP	8	1822
11	*18 C4ol-1*, **5 C4ol-8,** *50 C6ol-1*, *50 C8ol-1*, *50 C10ol-1*, *50 C6-1*, *50 C10-1*, *50 dap-1*, **22 dap-3**	Expl. MLP + fragments	345	2167
12	16 C4ol-1, **4 C4ol-8,** *50 C6ol-1*, *50 C8ol-1*, **1 C8ol-4**, **1 C10ol-3,** *50 C10ol-1*, *50 C6-1*, *50 C10-1*, **47 C10-4,** *50 dap-1*, **13 dap-3**	Expl. MLP + fragments	382	2549
13	**10 C4ol-8, 13 C6ol-6, 16 C8ol-4, 10 C10ol-3, 16 dap-3**	Expl. MLP	65	2614
14	**7 C4ol-8, 1 C8ol-4, 3 C10ol-3, 7 dap-3, 50 C4-2ol-6**	Expl. MLP + pcff+	68	2682
15	**3 C4ol-8, 12 C6ol-6, 11 C8ol-4, 9 C10ol-3, 1 C4-8, 5 C6-6, 9 C10-4, 16 dap-3**	Expl. MLP	66	2748
16	**3 C4ol-8, 7 dap-3**	Expl. MLP	10	2758
17	*127 dap-1, 24 C6-1, 33 C6ol-1*	fragments	183	2941
18		cleaning	-300	2641

Table S1 Structures added during the active learning and training set configurations at every iteration (step). Configurations created using PCFF+ driven molecular dynamics or random displacements are underlined, configurations originating from the "fragmentation" algorithm (stretching single molecules) are in italics, and configurations from the uncertainty-guided exploration *via* the MLP are given in bold.

S.2 Active Learning bond potential progression

In Figure S4 we show the energy of an hexanol molecule as a function of the C3–C4 separation while all other bond distances are kept constant. PCFF+ is shown as a black curve with points, as a representative of a classical bonded force field, displaying the quartic behavior of the bonding term. As expected, the DFT results (dashed red line) displayed a steep increase in energy as the bond distance was reduced below the equilibrium distance and a shallow, but steady, increase as the molecule was pulled apart towards a fragmentation energy of roughly 5 eV.

The MLP at different stages of the AL cycle is shown in the solid colored lines in Figure S4. At cyle 9, we observed that the MLP had the wrong fragmentation energy and a small local maximum at 2.5 Å, most likely due to the lack of training data in this region. After the introduction of fragments in cycle 11, the correct fragmentation energy was recovered, and the local maximum dissipated. The same local maximum was still present to a lesser extent in cycle 16, but the fragmentation energy was better reproduced, which we attribute to the inclusion of fragments in the training configurations. After refining the training set configurations, the final MLP displayed the expected behavior. The final curve in solid dark purple represents the ACE-MLP after all AL cycles, training set refinement, and hyperparameter optimizations. This curve closely results in good qualitative agreement with the DFT energy profile and no energy barrier, only deviating from the DFT energy curve at unphysically large bond lengths (*i.e.*, >3 Å), indicating an accurate prediction of the bonding interaction.

S.3 Definition of robustness in MLP exploration

The minimal dump (or sampling) period is one configuration every 40 fs, the maximum is one configuration every 500 fs.

A success ratio R_S is defined as the number of stable and non-extrapolated structures, divided by the number of target structures. A structure is stable if:

- No unphysical bond distances are present in the configuration, which are bond distances below 0.6 Å or above 2.6 Å.
- γ_{max} is below 100, as very high extrapolation grades are a sign of unstable configurations.

A structure is extrapolated if $\gamma_{max} > 1$.

Our goal was to keep this ratio at ≈ 0.5, in order to balance exploring new configurations without pushing the MLP into an extremely extrapolative regime. Therefore, we set a lower threshold at 0.25 and a higher threshold at 0.75. If the R_S dropped below 0.25, the algorithm reduced the sampling period by a factor of 2. If the R_s rose above 0.75, the algorithm doubled the sampling period.

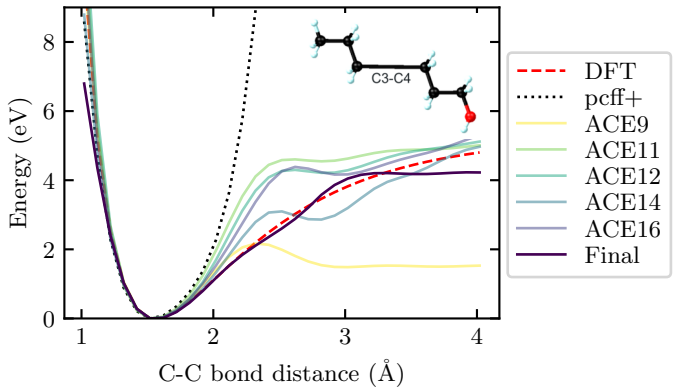

Fig. S4 The energy profile of C3-C4 carbon bond vector (see insert for illustration) of C6ol at various stages of training for the ACE-MLP. PCFF+ is shown as a dotted black curve, the DFT results are shown as a dashed red line, and the MLP results at various iterations of the AL cycle are shown as solid colored lines. The final curve represents the ACE-MLP after all AL cycles and hyperparameter optimization.

S.4 Active Learning

A typical input file to `pacemaker` used during the AL is:

```
cutoff: 6.0
seed: 1
metadata:
  origin: Automatically generated input
potential:
  deltaSplineBins: 0.001
  elements:
  - C
  - O
  - H
  embeddings:
    ALL:
      npot: FinnisSinclairShiftedScaled
      fs_parameters:
      - 1
      - 1
      - 1
      - 0.5
      ndensity: 2
  bonds:
    ALL:
      radbase: SBessel
      radparameters:
      - 5.25
      rcut: 3.0
      dcut: 0.1
      NameOfCutoffFunction: cos
      r_in: 0.4
      delta_in: 0.4
      core-repulsion:
      - 5.0
      - 5.0
    CC:
      radbase: SBessel
      radparameters:
      - 5.25
      rcut: 6.0
      dcut: 0.1
      NameOfCutoffFunction: cos
      r_in: 0.4
      delta_in: 0.4
      core-repulsion:
      - 5.0
```

```
        - 5.0
    functions:
      ALL:
        nradmax_by_orders:
        - 15
        - 6
        - 4
        - 2
        lmax_by_orders:
        - 0
        - 3
        - 2
        - 1
      number_of_functions_per_element: 500
  data:
    filename: step-12-label.pckl.gzip
    test_size: 0.1
  fit:
    loss:
      kappa: 0.001
      L1_coeffs: 0
      L2_coeffs: 1.0e-08
    optimizer: BFGS
    maxiter: 500
  backend:
    evaluator: tensorpot
    batch_size: 140
    display_step: 100
```

S.5 Hyperparameter optimization

We performed for a hyperparameter optimization *via* an extensive grid search. The following hyperparameter values were explored:

- κ: 0.1, 0.01, 0.001
- l_2: 1e-6, 1e-7
- n_m (l_m): 25,5,3 (0,3,2); 25,15,5 (0,3,2); 16,8,4,2 (0,3,2,1)
- r_c^s: 3, 4
- r_c^l: 6, 7, 8
- number of functions per element: 600, 800, 1000, 1250, 1500
- Long distance bond: "H-H", "C-C"
- nradmax / lradmax of 16,8,4,2/0,3,2,1 & 25,5,3/0,3,2

The optimal set of hyperparameters had a short range cutoff r_c^s at 3 Å, a long range cutoff r_c^l at 7 Å, a force weight κ of 0.001, 1500 independent functions per element, and the long distance interaction was on the C-C bond. We also saw that it is worth constraining the expansion of ACE to 4-body terms, and providing more functions to 2 body terms than the default settings. The input file employing the optimal set of hyperparameters is:

```
cutoff: 7.0
seed: 1
potential:
  deltaSplineBins: 0.001
  elements:
  - C
  - H
  - O
  embeddings:
    ALL:
      npot: FinnisSinclairShiftedScaled
      fs_parameters:
      - 1
      - 1
      - 1
      - 0.5
      ndensity: 2
  bonds:
    ALL:
      radbase: SBessel
      radparameters:
      - 5.25
      rcut: 3.0
      dcut: 0.01
    CC:
      radbase: SBessel
      radparameters:
      - 5.25
      rcut: 7.0
      dcut: 0.01
  functions:
    ALL:
      nradmax_by_orders:
      - 25
      - 15
      - 5
      lmax_by_orders:
      - 0
      - 3
      - 2
    number_of_functions_per_element: 1500
data:
  filename: labels_step_all-maxfor-10.pckl.gzip
  test_size: 0.02
fit:
  loss:
    kappa: 0.001
    L1_coeffs: 0
    L2_coeffs: 1.0e-06
  optimizer: BFGS
  maxiter: 1500
backend:
  evaluator: tensorpot
  batch_size: 512
  display_step: 100
```

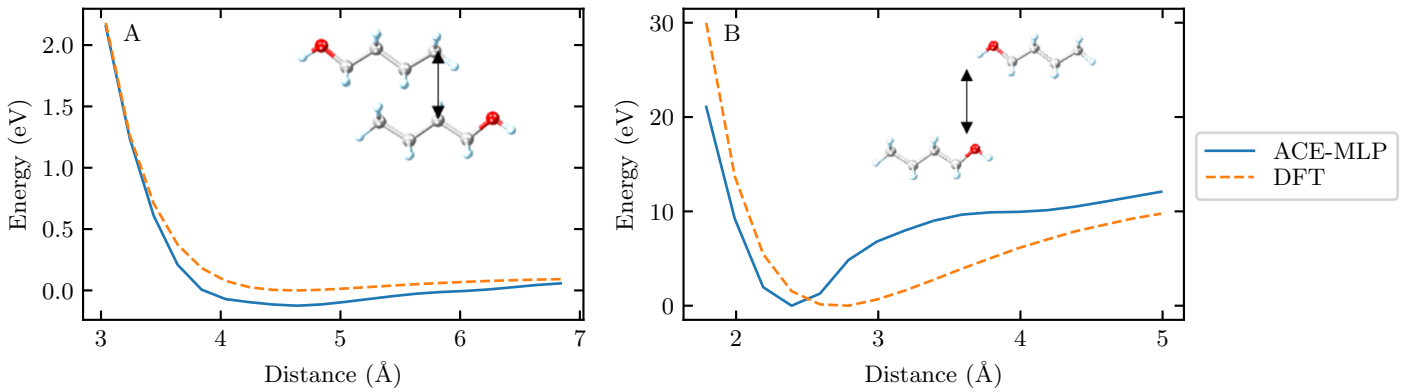

Fig. S5 Displacement energy profiles for two different configurations of interacting C4ol molecules. Insets illustrate the distance vector with a black line. The ACE-MLP results are shown as the solid blue lines, and the DFT results are shown as the dashed orange lines. (A) The energy of system was computed along the chain to chain (carbon to carbon) interaction. (B) The energy of system was computed along the OH-OH (hydrogen to oxygen) interaction.

S.6 Investigation of the OH-OH interaction

To test how accurately intermolecular interactions were reproduced by the MLP, we created two configurations of butanol dimers, as displayed in Figure S5 (insets). The chain-chain configuration contained two molecules placed in parallel along the chain, with the OH-groups as far apart as possible. We used this configuration to validate the vdW interaction. The second configuration placed the OH-OH groups close to each other to test hydrogen bonding. The energy profiles as a function of intermolecular distance are shown in Figure S5A for the chain-chain configuration and in Figure S5B for the OH-OH configuration. We observed good agreement between the MLP and DFT results for the chain-chain interactions, with the minimum at the correct position, albeit a bit deeper than the DFT energies. This could explain the small differences in the heat of vaporization. On the other hand, the MLP did not reproduce the hydrogen bonding interaction as well as the chain-chain interaction: the minimum was at a shorter distance, and the slope at larger distances was too steep. However, the depth of the minimum seemed to be captured well. Furthermore, the OH-OH interaction could not be described well by the MLP beyond 3 Å, as this was the short-range cutoff of the MLP. We believe that the steepness in slope comes from the fact that the MLP has to achieve the same depth of the minimum at a shorter distance, leading to a steeper slope. In addition, molecular bimers were not in the training set, and we believe that the MLP would perform better if trained on a larger set of configurations including dimers. Nevertheless, the MLP was able to capture the main features of the interaction, which is supported by the good performance computing the heat of vaporization.

S.7 Errors on energy and forces of MLP

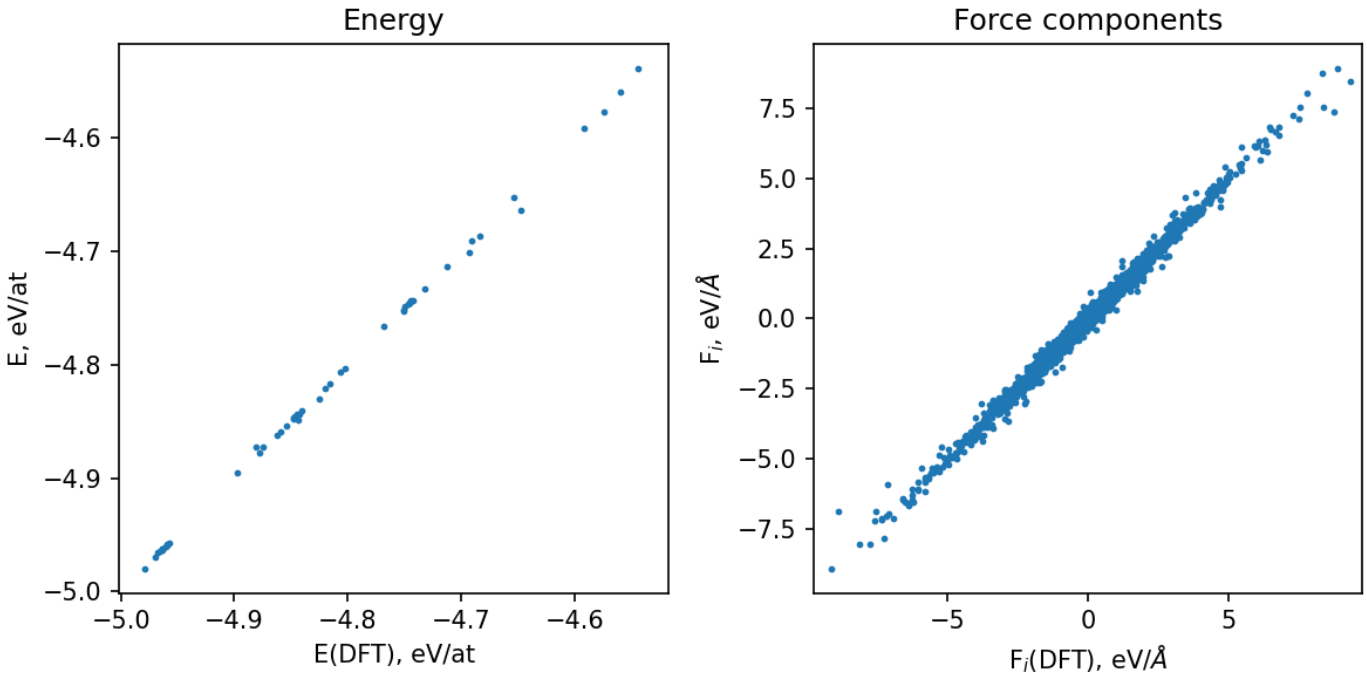

Fig. S6 (Left) The energies of test configurations as calculated in DFT against the energies for the same configurations by the best ACE-MLP. (Right) Forces in DFT against forces of the ACE-MLP.

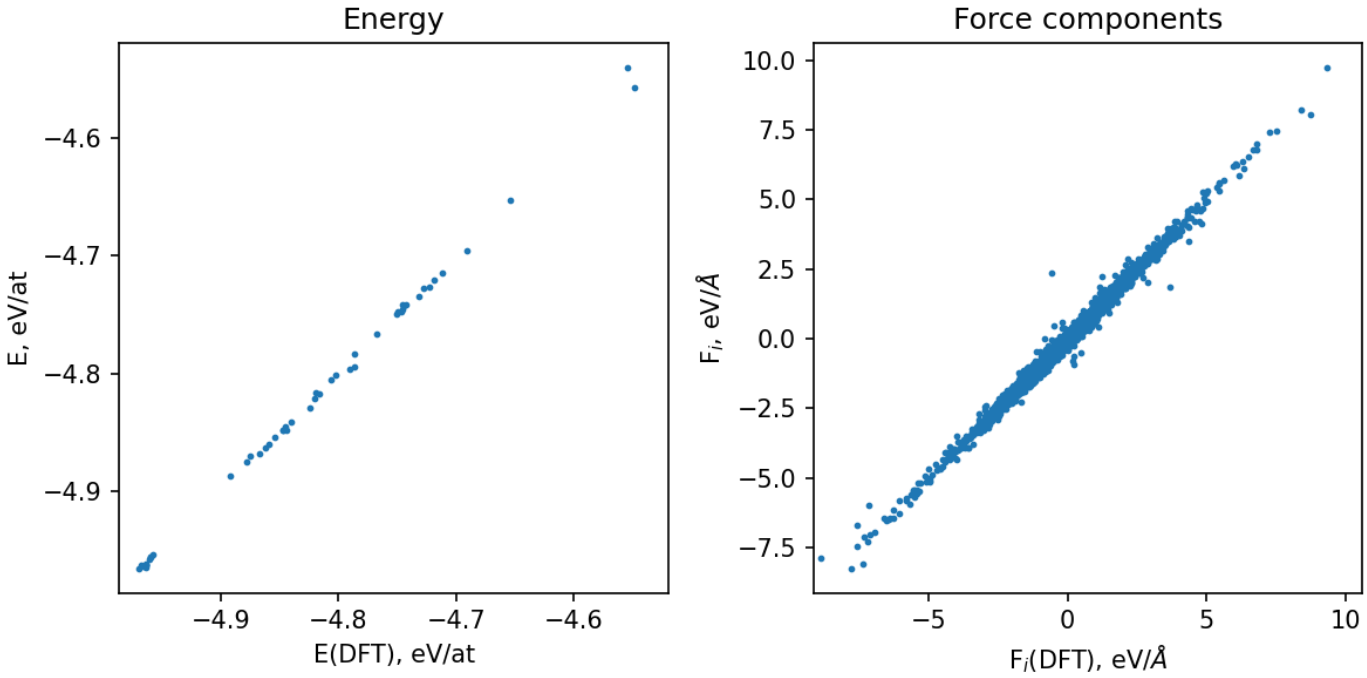

Fig. S7 The energies (left) and forces (right) of test configurations as calculated in DFT against the energies for the same configurations by the ACE-MLP with $r_c^s = r_c^l = 3$ Å.

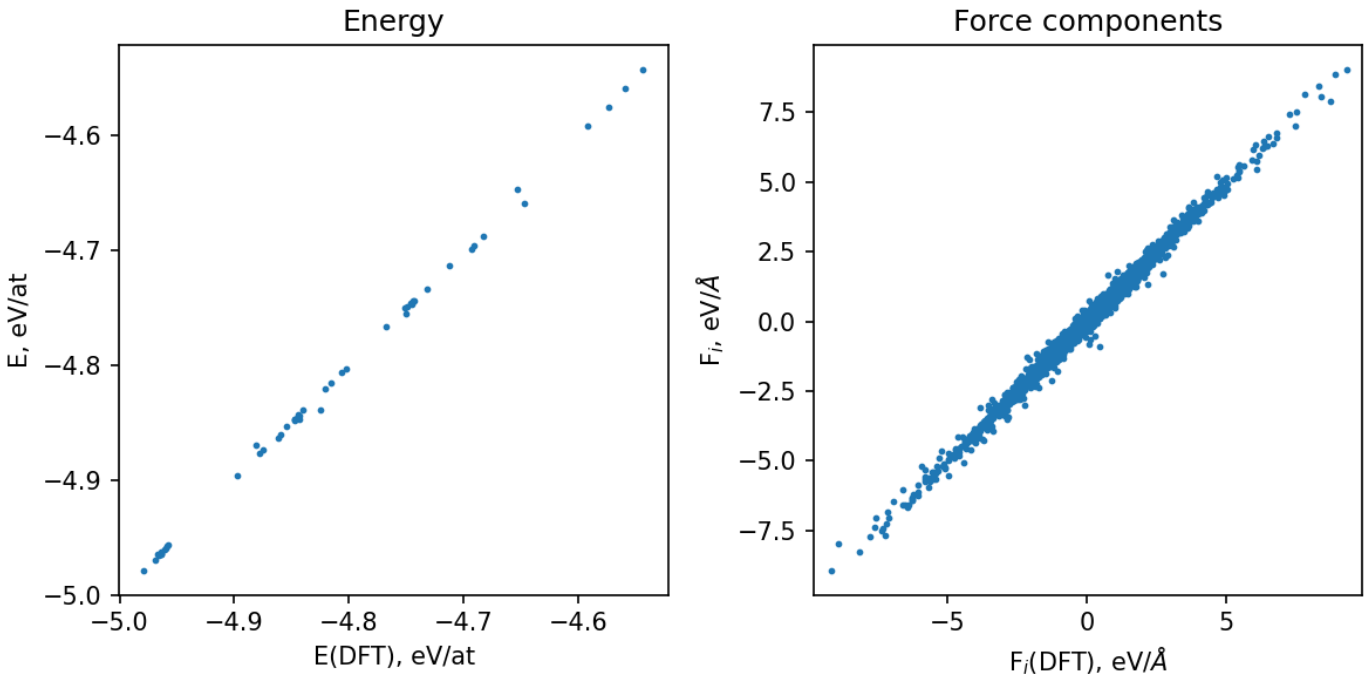

Fig. S8 The energies (left) and forces (right) of test configurations as calculated in DFT against the energies for the same configurations by the ACE-MLP with $r_c^s = r_c^l = 5$ Å. Note that this MLP was highly unstable in MD simulations, despite the excellent reproduction of energies and forces.

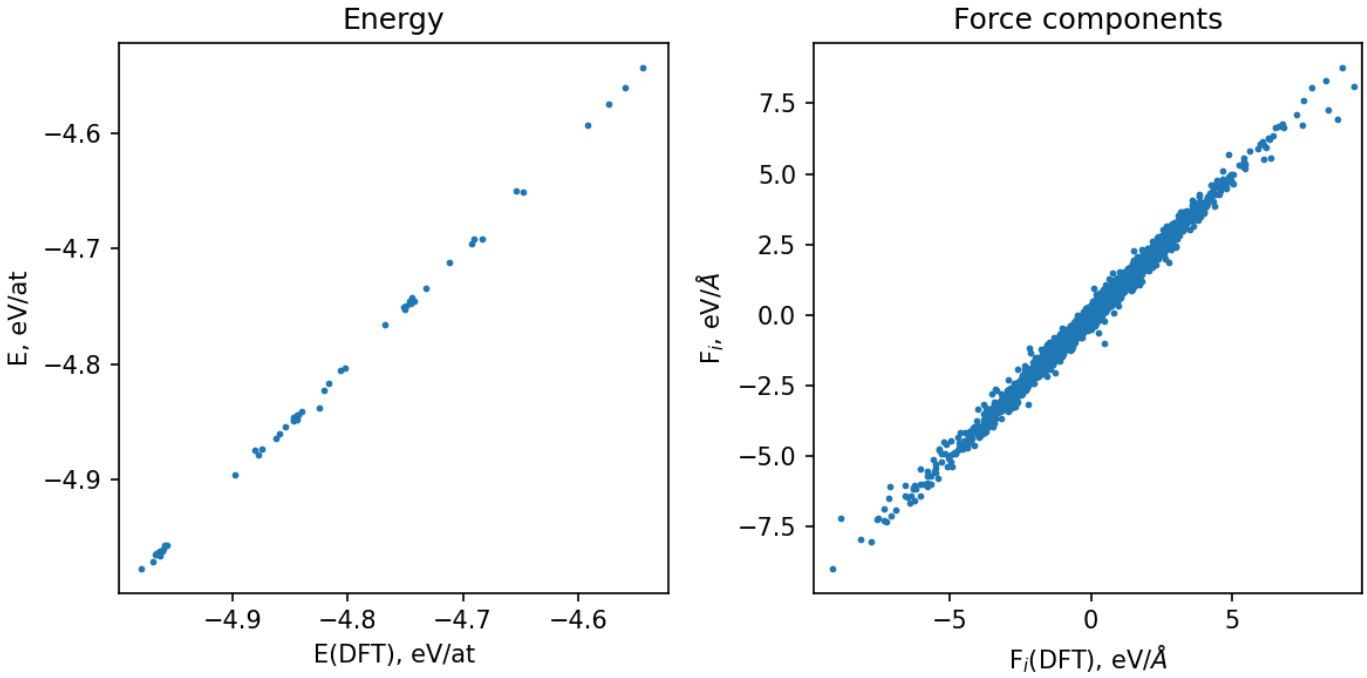

Fig. S9 The energies (left) and forces (right) of test configurations as calculated in DFT against the energies for the same configurations by the ACE-MLP with $r_c^s = r_c^l = 7$ Å. Note that this MLP was highly unstable in MD simulations, despite the excellent reproduction of energies and forces.