Supplementary Information for "Statistics makes a difference: Machine learning adsorption dynamics of functionalized cyclooctine on Si(001) at DFT accuracy"

Hendrik Weiske,[a] Rhyan Barrett,[a] Ralf Tonner-Zech,[a] Patrick Melix,[a] and Julia Westermayr[a,b,*]

[a] Wilhelm Ostwald Institute, Leipzig University, Leipzig, Germany.
[b] Center for Scalable Data Analytics and Artificial Intelligence Leipzig/Dresden, Humboldtstraße 25, 04105 Leipzig, Germany.

Contents

S1 Full Statistics of the Models

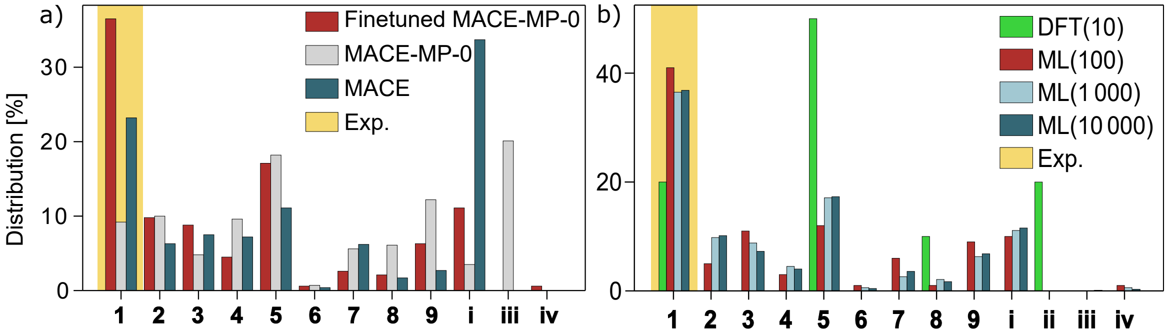

Figure S1: a) Distribution of the binding modes at the end of each MD simulation for the discussed ML models and the AIMD reference. The binding mode distribution of the fine-tuned model is given in red, the MACE-MP-0 model in grey, and the MACE model trained from scratch in dark blue. The experimentally observed binding mode is highlighted in yellow. b) Convergence of the distribution of adsorption sites for 10 (DFT) 100 (ML) 1 000 (ML) and 10 000 (ML) trajectories. The ML-driven trajectories are obtained using fine-tuned MACE-MP-0. Experimental observations are marked in yellow.

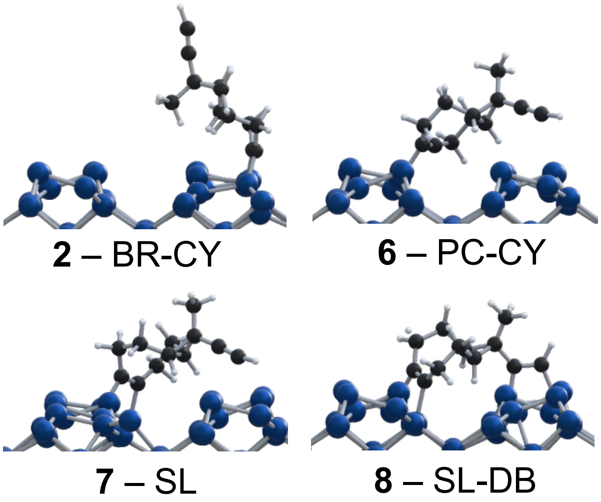

Figure S2: Binding modes detected at the end of MDs conducted using the fine-tuned MACE-MP-0 model, which are not observed in the reference AIMDs. Si atoms blue, C black and H white.

Several binding modes not observed as final states in AIMD trajectories are sampled in our machine learning-driven MDs (see Figure S2). Closer inspection of the AIMD trajectories, however, reveals these structures as transient intermediates. For instance, the prevalent doubly bound structure (**5**) is typically preceded by singly bound states. Binding modes BR-CY (**2**) and SL (**7**), as well as the PC (**6**) states, are found along AIMD trajectories as intermediates. The doubly bound sublayer structure SL-DB (**8**) and double-PC (**9**) states are the only modes detected in ML-MDs which do not appear in the DFT data.

S2 ML Model Comparison

Table S1: Mean absolute error of the energies and forces for different models. [a]Fine-Tuned model with the production/half/quarter of the production data. [b]Foundational MACE-MP-0 model without fine tuning. [c]From-scratch training without foundational model.

Calculator	MAE(E)	Max(E)	MAE(F)	Max(F)
FT(full)[a]	2.73×10^{-3}	6.44×10^{-3}	4.04×10^{-2}	1.87
FT(half)[a]	2.90×10^{-3}	7.20×10^{-3}	4.32×10^{-2}	1.90
FT(quarter)[a]	3.14×10^{-3}	7.91×10^{-3}	4.65×10^{-2}	1.91
MACE-MP0[b]	1.70×10^{-1}	1.79×10^{-1}	2.13×10^{-1}	2.71
regular[c]	1.75×10^{-3}	4.99×10^{-3}	2.97×10^{-2}	1.75

errors in eV and eV Å^{-1} for energies(E) and forces(F), respectively.

Table S2: Proportions test of of 1,000 vs 10,000 trajectories.

site	1k	10k	z-statistics	p-value
bridge(ac)	45	402	0.7330	0.4636
bridge(cy)	98	1015	-0.3499	0.7264
desorbed	111	1155	-0.4252	0.6707
double(BR)	11	114	-0.1138	0.9094
double(BR+OT)	125	1311	-0.5459	0.5851
double(OT+BR)	35	305	0.7840	0.4331
explode	6	30	1.5837	0.1133
ontop(ac)	88	726	1.7738	0.0761
ontop(cy)	365	3683	-0.2063	0.8365
other	63	681	-0.6123	0.5403
precursor	6	44	0.7172	0.4733
sublayer	26	358	-1.6098	0.1074
sublayer+other	21	170	0.9233	0.3558

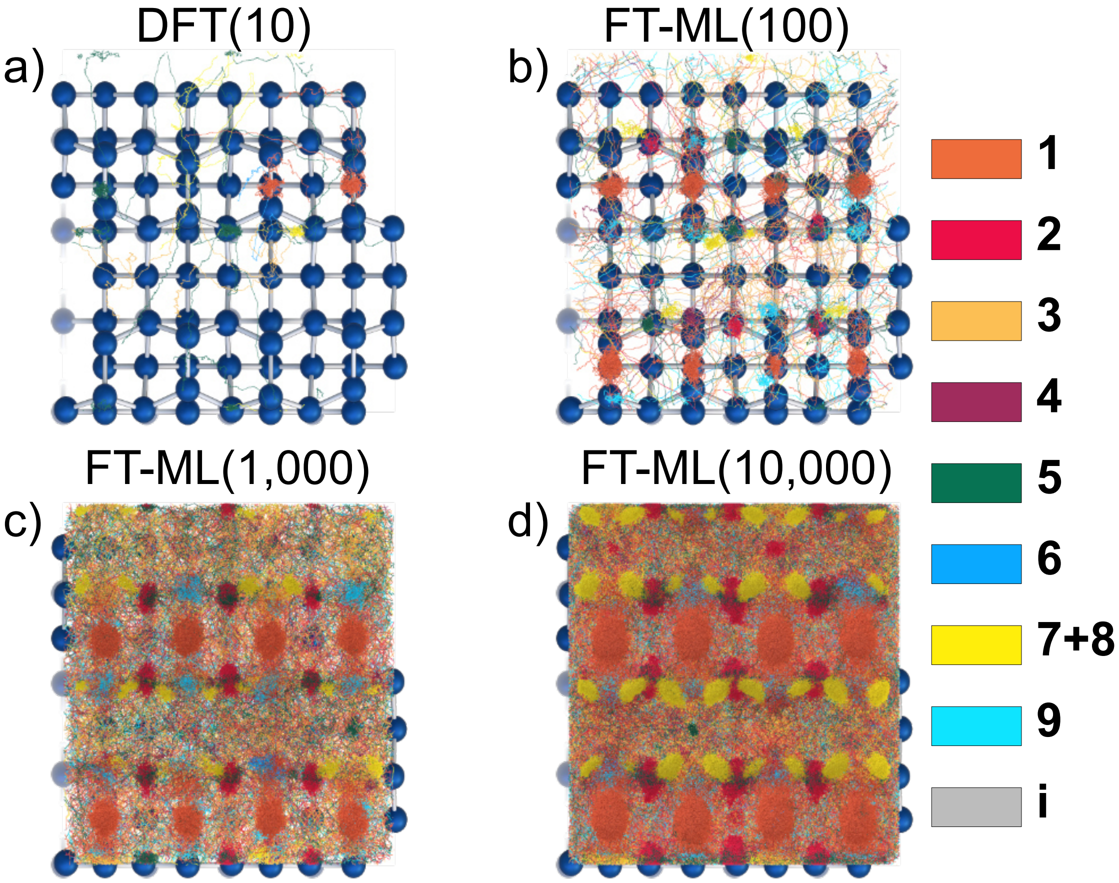

Figure S3: Median position of the cyclooctyne triple bond over all simulation runs for the 10 DFT (a), 100 FT (b) 1000 FT (c) and 10,000 FT (d) MDs. The bins are created with a size of 0.0015 Å^2. Coloring according to the final adsorption mode of trajectory.

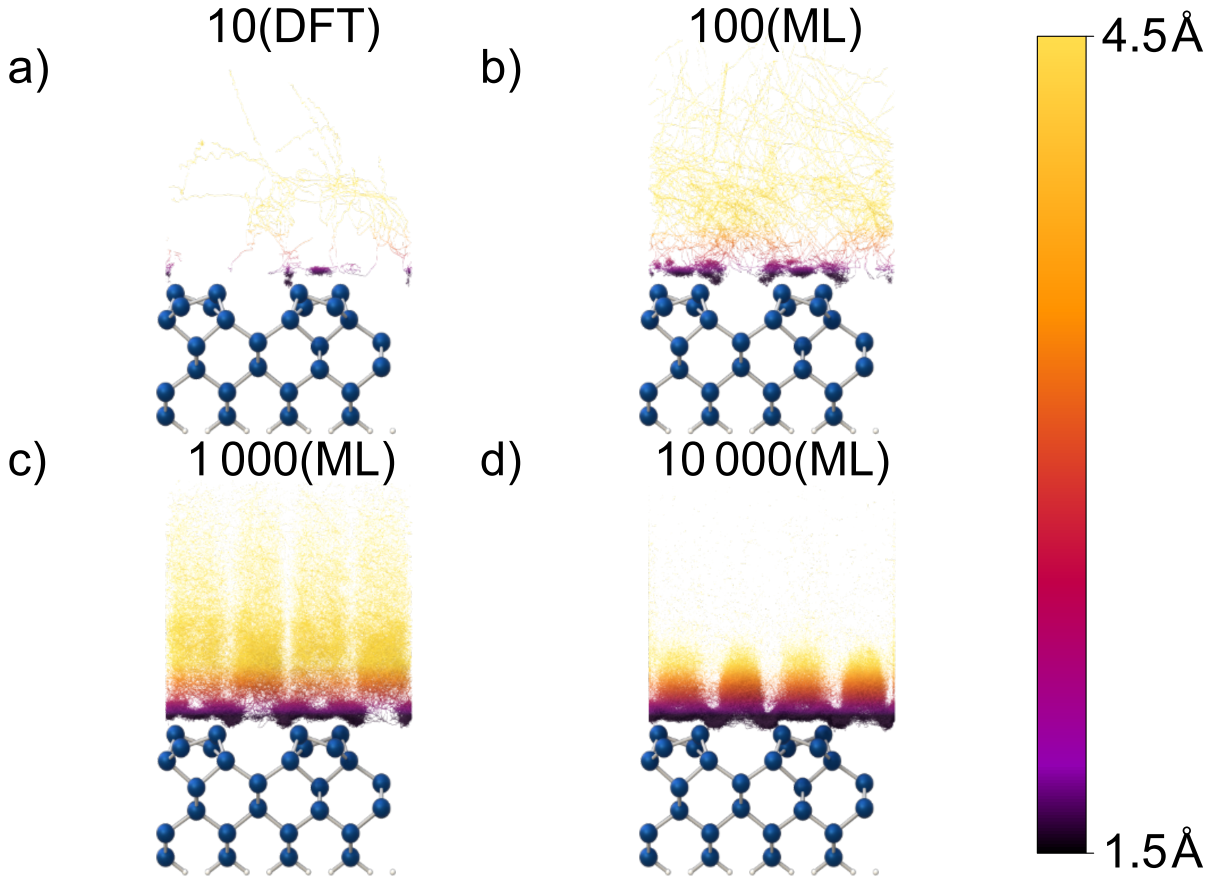

Figure S4: Representation of the sampling (side-view). The median positions of the cyclooctyne triple bond over all MD runs for the 10 DFT (a), 100 ML (b) 1000 ML (c) and 10,000 (d). The bins are created with a size of 0.0015 Å^2 and colored using their z-value. FT refers to fine-tuned models.

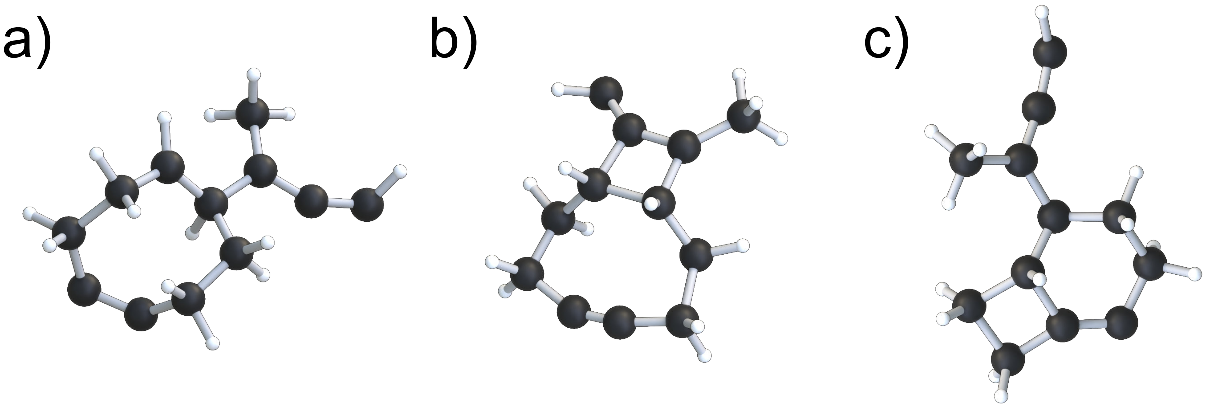

Figure S5: Examplary broken molecule structures, occurring using MACE-MP-0 without fine-tuning. Shown are a broken cyclopropanyl-structure (a), formation of a cyclobutanyl-ring at the acetylene group (b) and formation of a cyclobutanyl-ring at the cyclooctyne tripple bond (c)

S3 PCA analysis resolved by energy

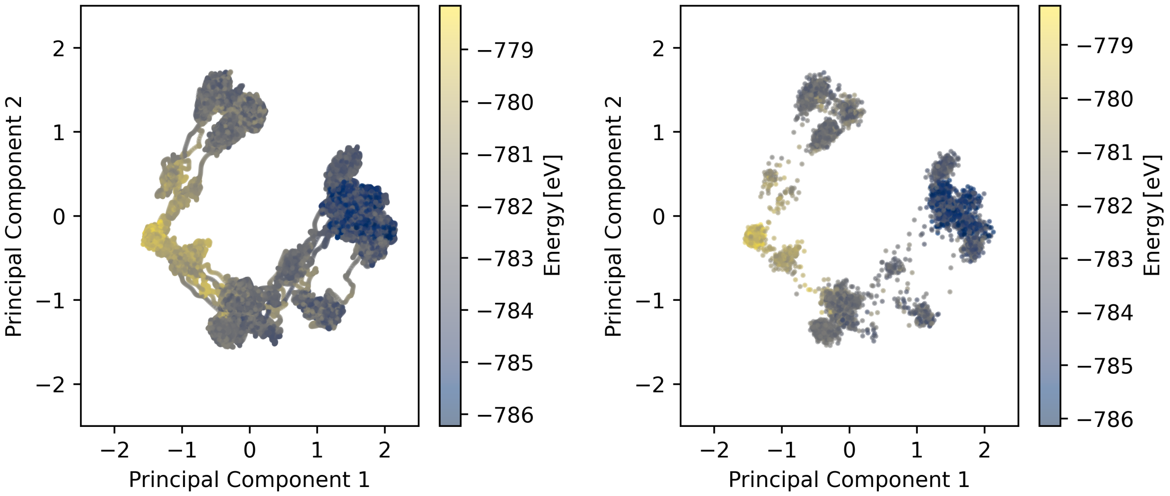

Figure S6: PCA analysis of full DFT (left) and production DFT (right) dataset using MACE descriptors.

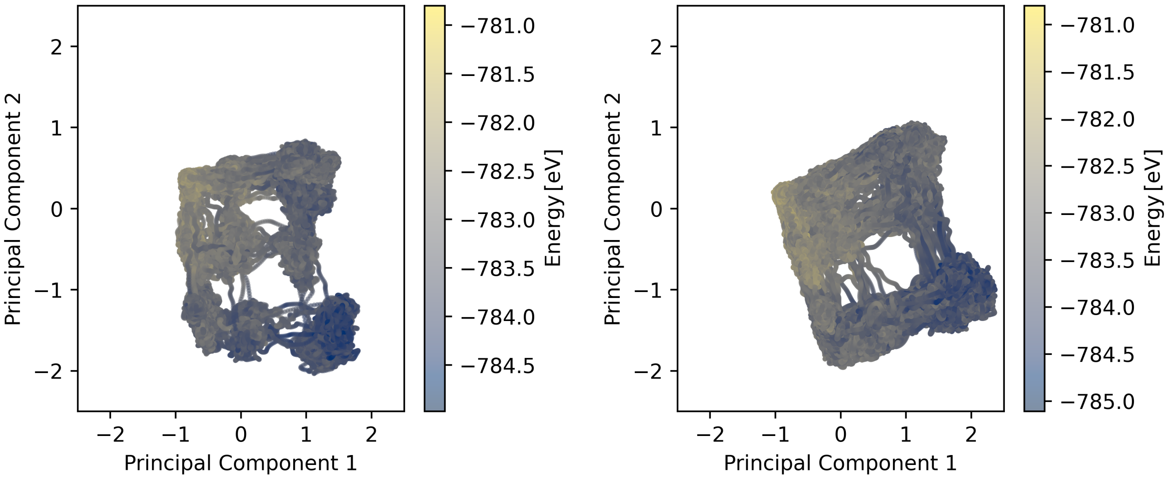

Figure S7: PCA analysis of 100 (left) and 1000 (right) ML trajectories (20.000 steps per trajectory).

S4 Unfunctionalized Cyclooctyne

To estimate the transferability of the fine-tuned and from-scratch models, we perform MD simulations using both models on a slightly different system: the **un**functionalized cyclooctyne molecule on the Si(001) surface. For this system, there is again previous available AIMD data from the group of RTZ.[1] The MD setup is the same as for the ECCO molecules, see the main text for details.

The fine-tuned model performs well, none of our MD simulations show any signs of breaking of the cyclooctyne molecules, that is often encountered during ML driven MDs (see also Figure S5). The resulting adsorption behaviour reproduces the trends from the AIMD study of Pecher *et al.*[1] well (see Figure S9. However, more bridge and sub-layer adsorption modes are found than

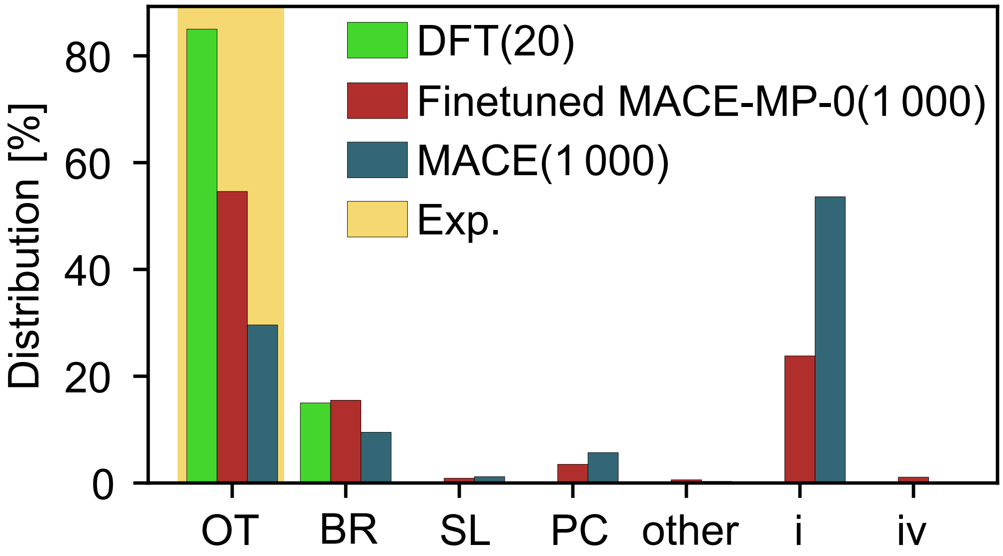

Figure S8: Distribution of adsorption sites of unfunctionalized cyclooctyne on Si(001) using the model fine-tuned on ECCO (FT) as well as the model trained from scratch (FS).

occur in the DFT reference. As shown in the main text, this might however only be due to the limited amount of AIMD trajectories.

In comparison, the from-scratch model performs, as expected, significantly worse. Mainly, this is due to an increase in desorbed molecules. The reduced amount of breaking ECCO molecules could be explained by the longer training time, that enables the from-scratch model to better represent strained variants of the ECCO molecule.

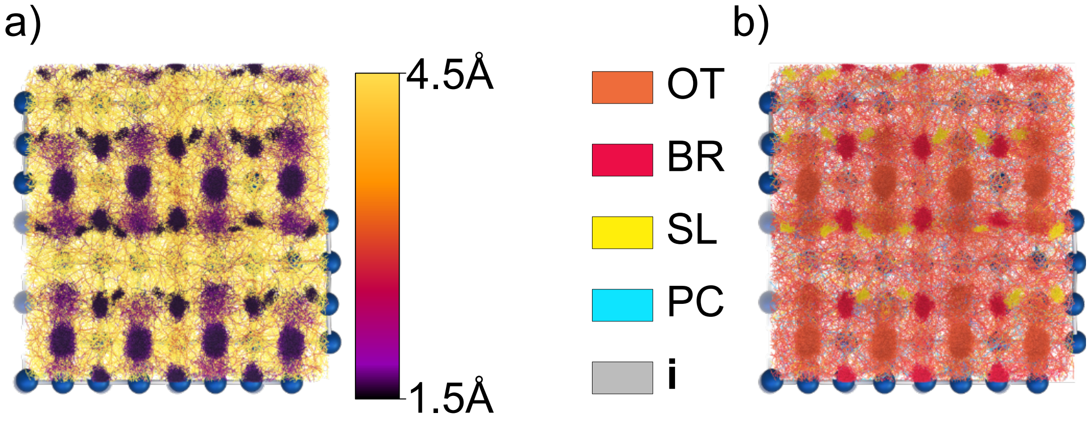

Figure S9: Binning of the measured lowest triple bond position at each xy-bin. a) color is representing the height of the measured points and b) the adsorption site of the respective trajectories.

S5 The Si(001) slab

Table S3: Silicon slab lattice parameters used in the DFT reference data[2].

	a	b	c
1	15.324	0	0
2	0	15.324	0
3	0	0	30.649

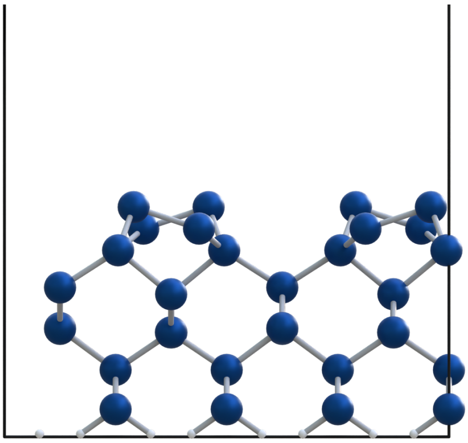

Figure S10: Silicon slab used in the DFT reference data[2]

.

S6 Side-by-side comparison of adsorption structures

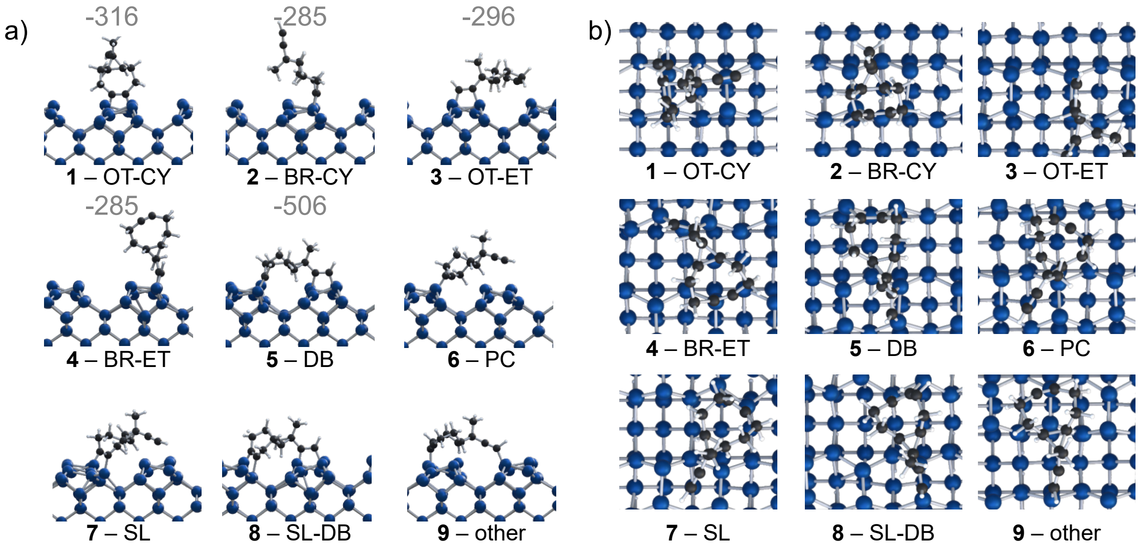

Figure S11: Side by side comparison in the a) frontview and b) topview with the available DFT energies from Pieck et al.[2] in gray.

S7 Statistics of 10 ML-trajectories

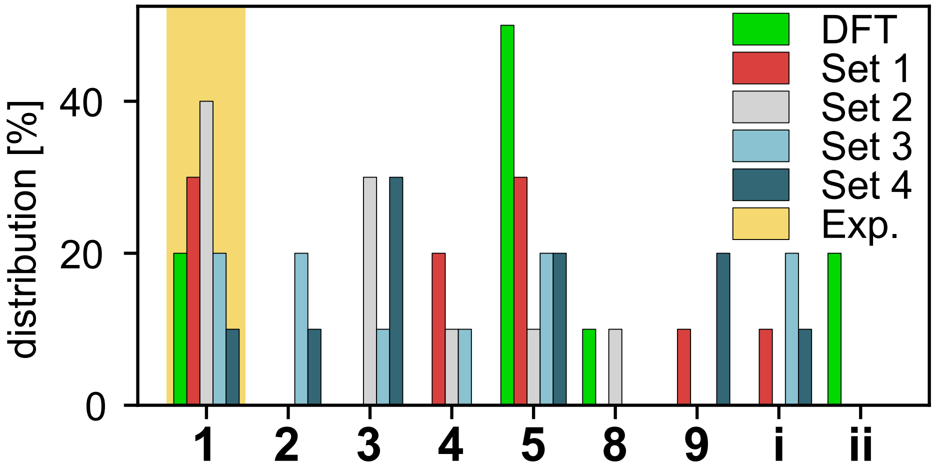

Figure S12: Comparison of 10 simulations using the ML model in order to compare the statistics of a low number of trajectories. The set 1-4 consists of 10 trajectories each, extracted from the 1000 trajectories used in the main results. The results show a high variance in distribution showing the unreliability of running a low number of trajectories.

References

[1] L. Pecher, S. Schmidt and R. Tonner, *Journal of Physical Chemistry C*, 2017, **121**, 26840–26850.

[2] F. Pieck and R. Tonner-Zech, *Molecules*, 2021, **26**, 6653.