

AdsorbML: A Leap in Efficiency for Adsorption Energy Calculations using Generalizable Machine Learning Potentials

Janice Lan*,¹ Aini Palizhati*,² Muhammed Shuaibi*,¹ Brandon M. Wood*,¹ Brook Wander,² Abhishek Das,¹ Matt Uyttendaele,¹ C. Lawrence Zitnick^{†,1} and Zachary W. Ulissi^{†2,3}

¹*Fundamental AI Research (FAIR), Meta AI, Meta*

²*Department of Chemical Engineering, Carnegie Mellon University*

³*Scott Institute for Energy Innovation, Carnegie Mellon University*

Computational catalysis is playing an increasingly significant role in the design of catalysts across a wide range of applications. A common task for many computational methods is the need to accurately compute the adsorption energy for an adsorbate and a catalyst surface of interest. Traditionally, the identification of low energy adsorbate-surface configurations relies on heuristic methods and researcher intuition. As the desire to perform high-throughput screening increases, it becomes challenging to use heuristics and intuition alone. In this paper, we demonstrate machine learning potentials can be leveraged to identify low energy adsorbate-surface configurations more accurately and efficiently. Our algorithm provides a spectrum of trade-offs between accuracy and efficiency, with one balanced option finding the lowest energy configuration 87.36% of the time, while achieving a $\sim 2000\times$ speedup in computation. To standardize benchmarking, we introduce the Open Catalyst Dense dataset containing nearly 1,000 diverse surfaces and $\sim 100,000$ unique configurations.

INTRODUCTION

The design of novel heterogeneous catalysts plays an essential role in the synthesis of everyday fuels and chemicals. To accommodate the growing demand for energy while combating climate change, efficient, low-cost catalysts are critical to the utilization of renewable energy [1–4]. Given the enormity of the material design space, efficient screening methods are highly sought after [4–7]. Computational catalysis offers the potential to screen vast numbers of materials to complement more time- and cost-intensive experimental studies.

A critical task for many first-principles approaches to heterogeneous catalyst discovery is the calculation of adsorption energies. The adsorption energy is the energy associated with a molecule, or adsorbate, interacting with a catalyst surface. Adsorbates are often selected to capture the various steps, or intermediates, in a reaction pathway (e.g. $^*\text{CHO}$ in CO_2 reduction). Adsorption energy is calculated by finding the adsorbate-surface configuration that minimizes the structure’s overall energy. Thus, the adsorption energy is the global minimum energy across all potential adsorbate placements and configurations. These adsorption energies are the starting point for the calculation of the free energy diagrams to determine the most favorable reaction pathways on a catalyst surface [8]. It has been demonstrated that adsorption energies of reaction intermediates can be powerful

descriptors that correlate with experimental outcomes such as activity or selectivity [9–13]. This ability to predict trends in catalytic properties from first-principles is the basis for efficient catalyst screening approaches [1, 14].

Finding the adsorption energy presents a number of complexities. There are numerous potential binding sites for an adsorbate on a surface and for each binding site there are multiple ways to orient the adsorbate (see bottom-left in Figure 1). When an adsorbate is placed on a catalyst’s surface, the adsorbate and surface atoms will interact with each other. To determine the adsorption energy for a specific adsorbate-surface configuration, the atom positions need to be relaxed until a local energy minimum is reached. Density Functional Theory (DFT) [15–17] is the most common approach to performing this adsorbate-surface relaxation. DFT first computes a single-point calculation where the output is the system’s energy and the per-atoms forces. A relaxation then performs a local optimization where per-atom forces are iteratively calculated with DFT and used to update atom positions with an optimization algorithm (e.g. conjugate gradient [18]) until a local energy minimum is found. To find the global minimum, a strategy for sampling adsorbate-surface configurations and/or a technique such as minima hopping [19, 20] for overcoming energy barriers during optimization is required.

Adsorption energy (ΔE_{ads}) is calculated as the energy of the adsorbate-surface (E_{sys}) minus the energy of the clean surface (i.e. slab) (E_{slab}) and the energy of the gas phase adsorbate or reference species (E_{gas}), as defined by Chanussot, et al. and detailed in the Supporting Information (SI). [2, 4]

$$\Delta E_{\text{ads}} = E_{\text{sys}} - E_{\text{slab}} - E_{\text{gas}} \quad (1)$$

* **Equal Contribution**

† **Corresponding authors**

C.L.Z., email: zitnick@meta.com

Z.W.U., email: zulissi@andrew.cmu.edu

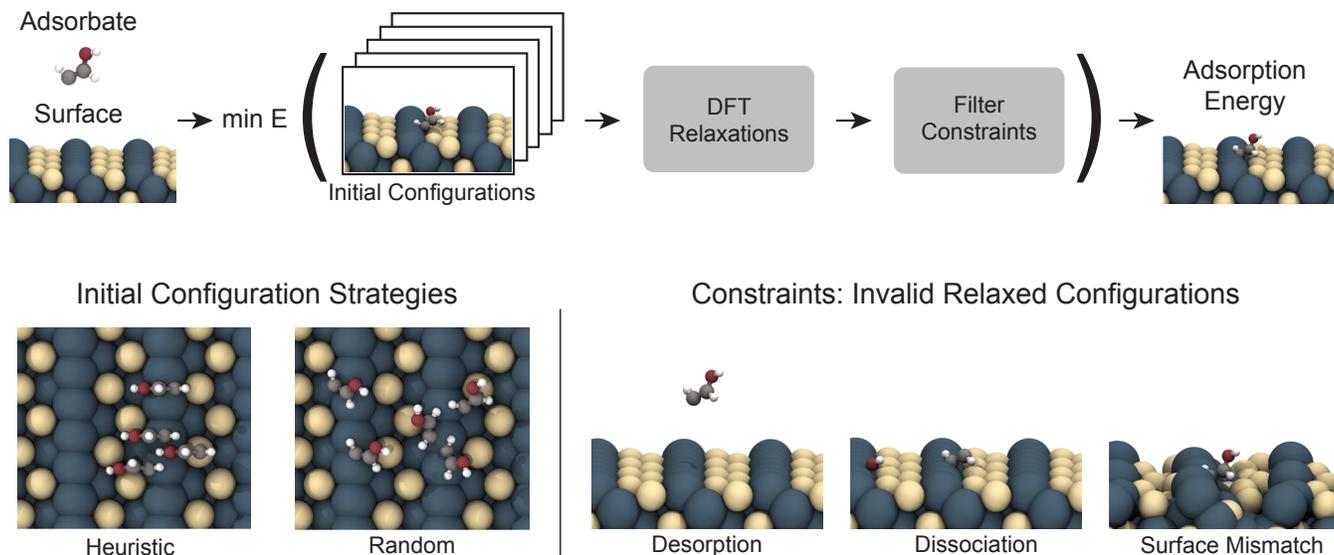


FIG. 1. An overview of the steps involved in identifying the adsorption energy for an adsorbate-surface combination. First, an adsorbate and surface combination are selected, then numerous configurations are enumerated heuristically and/or randomly. For each configuration, DFT relaxations are performed and systems are filtered based on physical constraints that ensure valid adsorption energies (i.e. desorption, dissociation, surface mismatch). The minimum energy across all configurations is identified as the adsorption energy.

Relaxed adsorbate-surface structures must respect certain desired properties in order for their adsorption energy to be both accurate and valid. One example of a constraint is the adsorbate should not be desorbed, i.e., float away, from the surface in the final relaxed structure (Figure 1 bottom-right). Additionally, if the adsorbate has multiple atoms it should not dissociate or break apart into multiple adsorbates because it would no longer be the adsorption energy of the molecule of interest [19, 21]. Similarly, if the adsorbate induces significant changes in the surface compared to the clean surface, the E_{slab} reference would create a surface mismatch. It is important to note that if a relaxed structure breaks one of these constraints it does not necessarily mean the relaxation was inaccurate; these outcomes do arise but they lead to invalid or inaccurate adsorption energies as it has been defined.

Identifying the globally optimal adsorbate-surface configuration has historically relied on expert intuition or more recently heuristic approaches. Intuition and trial and error can be used for one-off systems of interest but it does not scale to large numbers of systems. Commonly used heuristics are often based on surface symmetry [22, 23]. These methods have been used successfully in past descriptor-based studies [9, 10, 24–27]. More recently, a graph-based method has been used to identify unique adsorbate-surface configurations [28]. Nevertheless, as the complexity of the surfaces and adsorbates increase, the challenge of finding the lowest energy adsorbate-surface configuration grows substantially.

This is especially challenging when the adsorbate is flexible, having multiple configurations of its own, such that there are many effective degrees of freedom in the system.

While DFT offers the ability to accurately estimate atomic forces and energies, it is computationally expensive, scaling $O(N^3)$ with the number of electrons. Evaluating a single adsorbate-surface configuration with a full DFT relaxation can take ~ 24 hours to compute [2, 29]. Since numerous configurations are typically explored to find the adsorption energy, all the DFT calculations involved can take days or even weeks. Hypothetically, if one were to brute force screen 100,000 materials from the Materials Project database [30] for CO₂ Reduction Reaction (CO₂RR) using 5 adsorbate descriptors, ~ 90 surfaces/material, and ~ 100 sites/surface, one would need ~ 4.5 billion CPU-days of compute, an intractable problem for even the world’s largest supercomputers. To significantly reduce the required computation, a promising approach is to accelerate the search of lowest energy adsorbate-surface configurations with machine learned potentials.

Recently, machine learning (ML) potentials for estimating atomic forces and energies have shown significant progress on standard benchmarks while being orders of magnitude faster than DFT [2, 31–36]. While ML accuracies on the large and diverse Open Catalyst 2020 Dataset (OC20) dataset have improved to 0.3 eV for relaxed energy estimation, an accuracy of 0.1 eV is still desired for accurate screening [37]. This raises the question of whether a hybrid approach that uses both DFT

and ML potentials can achieve high accuracy while maintaining efficiency.

Assessing the performance of new methods for finding low energy adsorbate-surface configurations is challenging without standardized validation data. It is common for new methods to be tested on a relatively small number of systems, which makes generalization difficult to evaluate [19, 28, 38–40]. While OC20 contains $O(1M)$ “adsorption energies”, it did not sample multiple configurations per adsorbate-surface combination meaning the one configuration that was relaxed is unlikely to be the global minimum. This makes OC20 an inappropriate dataset for finding the minimum binding energy [2]. To address this issue, we introduce the Open Catalyst 2020 - Dense Dataset (OC20-Dense). OC20-Dense includes two splits - a validation and test set. The validation set is used for development; and the test set for reporting performance. Each split consists of approximately 1,000 unique adsorbate-surface combinations from the validation and test sets of the OC20 dataset. No data from OC20-Dense is used for training. To explore the generalizability of our approach, we take ~ 250 combinations from each of the four OC20 subsplits - In-Domain (ID), Out-of-Domain (OOD)-Adsorbate, OOD-Catalyst, and OOD-Both. For each combination, we perform a dense sampling of initial configurations and calculate relaxations using DFT to create a strong baseline for evaluating estimated adsorption energies.

We propose a hybrid approach to estimating adsorption energies that takes advantage of the strengths of both ML potentials and DFT. We sample a large number of potential adsorbate configurations using both heuristic and random strategies and perform relaxations using ML potentials. The best- k relaxed energies can then be refined using single-point DFT calculations or with full DFT relaxations. Using this approach, the appropriate trade-offs may be made between accuracy and efficiency.

Related Work

Considerable research effort has been dedicated to determining the lowest energy adsorbate-surface configuration through improvement of initial structure generation and global optimization strategies [19, 21, 28, 38–41]. Peterson [19] adopted the minima-hopping method and developed a global optimization approach that preserves adsorbate identity using constrained minima hopping. However, the method relies entirely on DFT to perform the search, still making it computationally expensive. More recently, Jung et al. [21] proposed an active learning workflow where a gaussian process is used to run constrained minima hopping simulations. Structures generated by their simulations are verified by DFT and iteratively added to the training set until model convergence is achieved. The trained model then runs parallel

constrained minima hopping simulations, a subset is refined with DFT, and the final adsorption energy identified. We note that prior attempts to use machine learning models to accelerate this process have typically relied on bespoke models for each adsorbate/catalyst combination, which limits broader applicability [42, 43]. One possibility to greatly expand the versatility of these methods while continuing to reduce the human and computational cost is using generalizable machine learning potentials to accelerate the search for low energy adsorbate-surface configurations.

The contributions of this work are three-fold:

- We propose the *AdsorbML* algorithm to identify the adsorption energy under a spectrum of accuracy-efficiency trade-offs.
- We develop the Open Catalyst 2020 - Dense Dataset (OC20-Dense) to benchmark the task of adsorption energy search.
- We benchmark literature Graph Neural Network (GNN) models on OC20-Dense using the proposed *AdsorbML* algorithm; identifying several promising models well-suited for practical screening applications.

RESULTS

OC20-Dense Evaluation

To evaluate methods for computing adsorption energies, we present the Open Catalyst 2020 - Dense Dataset (OC20-Dense) that closely approximates the ground truth adsorption energy by densely exploring numerous configurations for each unique adsorbate-surface system. Each OC20-Dense split comprises $\sim 1,000$ unique adsorbate-surface combinations spanning 74 adsorbates, 800+ inorganic bulk crystal structures, and a total of 80,000+ heuristically and randomly generated configurations. A summary of the two splits are provided in Table II. The dataset required ~ 4 million CPU-hrs of compute to complete. A more detailed discussion on OC20-Dense can be found in the Methods section.

We report results on a wide range of GNNs previously benchmarked on OC20 to evaluate the performance of existing models on OC20-Dense. These include SchNet [31], DimeNet++ [32, 33], PaiNN [44], GemNet-OC [34], GemNet-OC-MD [34], GemNet-OC-MD-Large [34], SCN-MD-Large [35], and eSCN-MD-Large [45] where MD corresponds to training on OC20 and its accompanying *ab initio* Molecular Dynamics (MD) dataset. Models were not trained as part of this work; trained models were taken directly from previously published work and can be found at <https://github.com/Open-Catalyst-Project/ocp/>

OC20-Dense Test

Model	Success Rate* [%] \uparrow	Energy MAE [eV] \downarrow	OC20 <i>S2EF</i> MAE \downarrow	
			Forces [eV/Å]	Energy [eV]
SchNet	1.01%	0.5150	0.0496	0.4445
DimeNet++	1.72%	0.4329	0.0446	0.4753
PaiNN	10.92%	0.2994	0.0294	0.2459
GemNet-OC	46.51%	0.1849	0.0179	0.1668
GemNet-OC-MD	50.05%	0.1966	0.0173	0.1694
GemNet-OC-MD-Large	48.03%	0.1935	0.0164	0.1665
SCN-MD-Large	51.87%	0.1758	0.0160	0.1730
eSCN-MD-Large	56.52%	0.1739	0.0139	0.1709

*ML predictions that lead to valid configurations and are within 0.1 eV of their DFT evaluation

TABLE I. Success rates evaluated using ML predicted energies. ML predictions are only considered if their predicted energies are within 0.1 eV of its DFT evaluation. Energy MAE is also computed between predicted ML and DFT energy minima. We also show OC20 *S2EF* Val-ID results, with metrics correlating well with success rates and energy MAE.

blob/main/MODELS.md. Of the models, (e)SCN-MD-Large and GemNet-OC-MD-Large are currently the top performers on both OC20 and Open Catalyst 2022 Dataset (OC22). Exploring the extent these trends hold for OC20-Dense will be important to informing how well progress on OC20 translates to more important downstream tasks like the one presented here.

Split	Unique Systems	Unique Configurations	Adsorbates	Bulks
Validation	973	85, 658	74	833
Test	989	105,714	74	837

TABLE II. Size of OC20-Dense validation and test splits. Unique adsorbate-surface systems are selected from the respective OC20 validation and test splits. Each split samples ~ 250 systems from each of its respective distribution subsplits - ID, OOD-Ads, OOD-Catalyst, OOD-Both.

Ideally, the ground truth for OC20-Dense would be the minimum relaxed energy over all possible configurations for each adsorbate-surface system. Since the number of possible configurations is combinatorial, the community has developed heuristic approaches to adsorbate placement on a catalyst surface [22, 23]. When evaluating only heuristic configurations, we refer to this as DFT-Heuristic-Only (DFT-Heur). To add to the configuration space, we also uniformly sample sites on the surface at random with the adsorbate placed on each of those sites with a random rotation along the z-axis and slight wobble around the x and y axis. When evaluating against both heuristic and random configurations, we refer to this as DFT-Heuristic+Random (DFT-Heur+Rand). Although computationally more expensive, this benchmark provides a more thorough search of configurations and a more accurate estimate of the adsorption energies than using only heuristic configurations, a common baseline used by the community. More details on the two benchmarks can be found in the Methods section.

ML Relaxations

We explore to what extent ML predictions can find the adsorption energy within a threshold of the DFT minimum energy, or lower. While a perfect ML surrogate to DFT will only be able to match DFT, small errors in the forces and optimizer differences have the potential to add noise to relaxations and result in configurations previously unexplored [46]. For each model, relaxations are performed on an identical set of adsorbate configurations. Initial configurations are created based off heuristic strategies commonly used in the literature [22, 23] and randomly generated configurations on the surface. ML-driven relaxations are run on all initial configurations; systems not suitable for adsorption energy calculations due to physical constraints are removed, including dissociation, desorption, and surface mismatch. An in-depth discussion on relaxation constraints can be found in the Methods section.

When evaluating performance, we define success as finding an adsorption energy within an acceptable tolerance (0.1 eV in this work [2, 37, 46]) or lower of the DFT adsorption energy in OC20-Dense. Note that the ground truth adsorption energies in OC20-Dense are an upper bound, since it is possible that a lower adsorption energy may exist. When evaluating ML predicted adsorption energies, the results must be verified using a single-point DFT calculation, since an evaluation metric without a lower bound could be easily gamed by predicting low energies (see SI). To reliably evaluate ML we consider an ML adsorption energy successful if its within 0.1 eV of the DFT adsorption energy or lower, **and** a corresponding DFT single-point evaluation of the predicted ML structure is within 0.1 eV of the predicted ML energy. This ensures that a ML prediction not only found a low adsorption energy but is accurate and not artificially inflated. Results are reported in Table I, where top OC20 models including eSCN-MD-Large and GemNet-OC-MD-Large achieve success rates of 56.52% and 48.03%, respectively.

Energy MAE between ML and DFT adsorption energies are also reported in Table I, correlating well with success rates and OC20 *S2EF* metrics.

While the current state of models have made incredible progress [37], higher success rates are needed for everyday practitioners. In a high-throughput setting where successful candidates go on to more expensive analyses or even experimental synthesis, a success rate of $\sim 50\%$ could result in a substantial waste of time and resources studying false positives. As model development will continue to help improve metrics, this work explores hybrid ML+DFT strategies to improve success rates at the cost of additional compute.

AdsorbML Algorithm

We introduce the *AdsorbML* algorithm to use ML to accelerate the adsorbate placement process (Figure 2). For each model, we explore two strategies that incorporate ML followed by DFT calculations to determine the adsorption energy. We note that this strategy is general and can be used with any initial configuration algorithm.

In both approaches the first step is to generate ML relaxations. However, rather than taking the minimum across ML relaxed energies, we rank the systems in order of lowest to highest energy. The best k systems with lowest energies are selected and (1) DFT single-point calculations are done on the corresponding structures (ML+SP) or (2) DFT relaxations are performed from ML relaxed structures (ML+RX). The first strategy aims to get a more reliable energy measurement of the ML predicted relaxed structure, while the second treats ML as a pre-optimizer with DFT completing the relaxation. By taking the k lowest energy systems, we provide the model with k opportunities to arrive at an acceptably accurate adsorption energy. As we increase k , more DFT compute is involved, but compared to a full DFT approach, we still anticipate significant savings. The adsorption energy for a particular system is obtained by taking the minimum of the best k DFT follow-up calculations.

In both strategies, ML energies are used solely to rank configurations, with the final energy prediction coming from a DFT calculation. While computationally it would be ideal to fully rely on ML, the use of DFT both improves accuracy and provides a verification step to bring us more confidence in our adsorption energy predictions.

Experiments

Our goal is to find comparable or better adsorption energies to those found using DFT alone in OC20-Dense. The metric we use to quantify this task is success rate, which is the percentage of OC20-Dense systems where our ML+DFT adsorption energy is within 0.1 eV or lower

than the DFT adsorption energy. A validation of the ML energy is not included in these experiments since all final adsorption energies will come from at least a single DFT call, ensuring all values are valid. Another metric we track is the speedup compared to the DFT-Heur+Rand baseline. Speedup is evaluated as the ratio of DFT electronic steps used by DFT-Heur+Rand to the proposed hybrid ML+DFT strategy. A more detailed discussion on the metrics can be found in the Methods section. Unless otherwise noted, all results are reported on the test set, with results on the validation set found in the SI. When evaluating the common baseline of DFT-Heur that uses only DFT calculations, a success rate of 87.76% is achieved at a speedup of 1.81x.

ML+SP The results of using single-point evaluations on ML relaxed states are summarized in Figure 3. eSCN-MD-Large and GemNet-OC-MD-Large achieve a success rate of 86+% at $k = 5$ with eSCN-MD-Large outperforming all models with a success rate of 88.27%, slightly better than the DFT-Heur baseline. Other models including SchNet and DimeNet++ do significantly worse with success metrics as low as 3.13% and 7.99%, respectively; suggesting the predicted relaxed structures are highly unfavorable. The speedups are fairly comparable across all models, ranging between 1400x and 1500x for $k=5$, orders of magnitude faster than the DFT-Heur baseline. Specifically, eSCN-MD-Large and GemNet-OC-MD-Large give rise to speedups of 1384x and 1388x, respectively. If speed is of most importance, speedups as high as 6817x are achievable with $k = 1$ while still maintaining success rates of 82% for eSCN-MD-Large. At a more balanced trade-off, $k = 3$, success rates of 87.36% and 84.43% are attainable for eSCN-MD-Large and GemNet-OC-MD-Large while maintaining speedups of 2296x and 2299x, respectively. In Figure 5 the minimum energy binding sites of several systems are compared as identified with ML+SP across different models.

ML+RX While single-point evaluations offer a fast evaluation of ML structures, performance is heavily reliant on the accuracy of the predicted relaxed structure. This is particularly apparent when evaluating the max per-atom force norm of ML relaxed structures with DFT. SchNet and DimeNet++ have on average a max force, f_{max} , of 2.00 eV/Å and 1.21eV/Å, respectively, further supporting the challenge these models face in obtaining valid relaxed structures. On the other hand, models like GemNet-OC-MD-Large and eSCN-MD-Large have an average f_{max} of 0.21eV/Å and 0.15eV/Å, respectively. While these models are a lot closer to valid relaxed structures (i.e. $f_{max} \leq 0.05$ eV/Å), these results suggest that there is still room for further optimization. Results on DFT relaxations from ML relaxed states are plotted in Figure 3. eSCN-MD-Large and GemNet-OC-MD-Large outperform all models at all k values, with a 90.60% and 91.61% success rate at $k = 5$, respectively. Given the additional DFT costs associated with refining

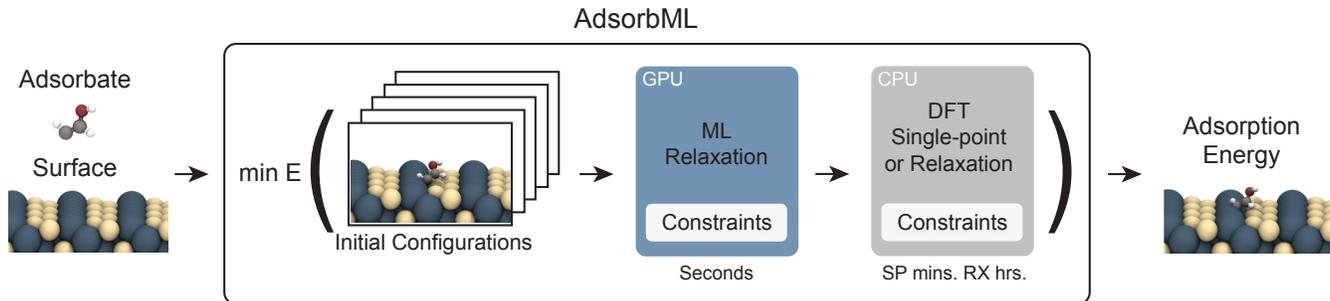


FIG. 2. The *AdsorbML* algorithm. Initial configurations are generated via heuristic and random strategies. ML relaxations are performed on GPUs and ranked in order of lowest to highest energy. The best k systems are passed on to DFT for either a single-point (SP) evaluation or a full relaxation (RX) from the ML relaxed structure. Systems not satisfying constraints are filtered at each stage a relaxation is performed. The minimum is taken across all DFT outputs for the final adsorption energy.

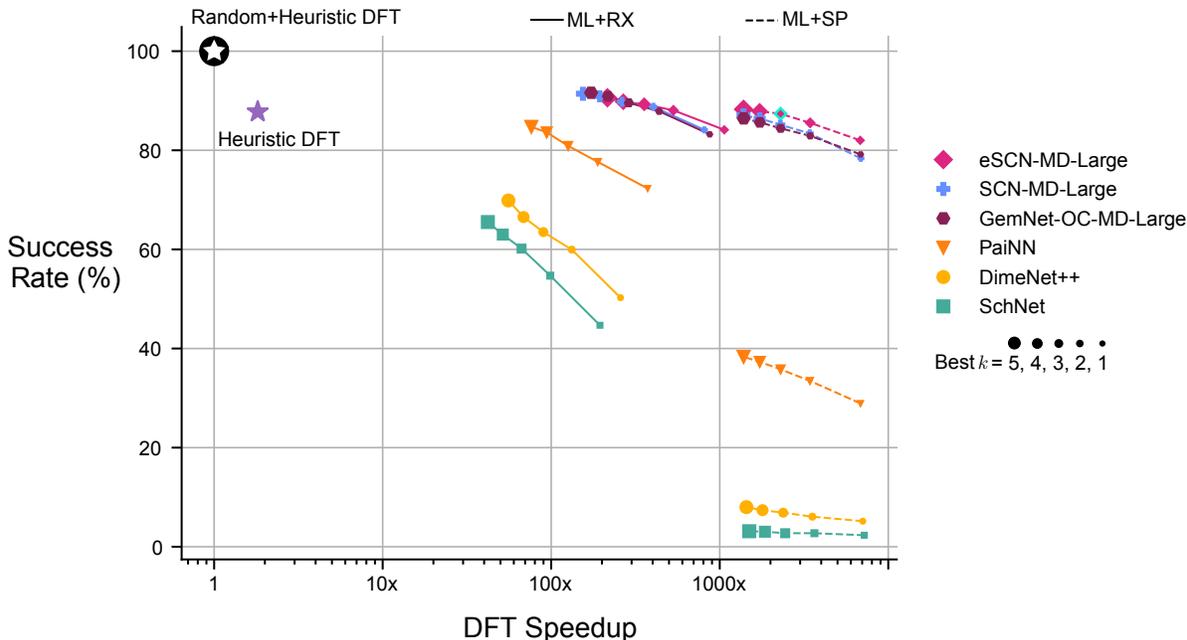


FIG. 3. Overview of the accuracy-efficiency trade-offs of the proposed *AdsorbML* methods across several baseline GNN models. For each model, DFT speedup and corresponding success rate are plotted for ML+RX and ML+SP across various best- k . A system is considered successful if the predicted adsorption energy is within 0.1 eV of the DFT minimum, or lower. All success rates and speedups are relative to Random+Heuristic DFT. Heuristic DFT is shown as a common community baseline. The upper right-hand corner represent the optimal region - maximizing speedup and success rate. The point highlighted in teal corresponds to the balanced option reported in the abstract - a 87.36% success rate and 2290x speedup. A similar figure for the OC20-Dense validation set can be found in the SI.

relaxations, speedups unsurprisingly decrease. At $k = 5$, we see speedups of 215x and 172x for eSCN-MD-Large and GemNet-OC-MD-Large, respectively. Both SchNet and DimeNet++ see much smaller speedups at 42x and 55x, respectively. The much smaller speedups associated with SchNet and DimeNet++ suggest that a larger number of DFT steps is necessary to relax the previously unfavorable configurations generated by the mod-

els. Conversely, eSCN-MD-Large’s much larger speedup can be attributed to the near relaxed states (average $f_{max} \sim 0.15\text{eV}/\text{\AA}$) it achieves in its predictions. With $k = 1$, speedups of 1064x are achievable while still maintaining a success rate of 84.13% for eSCN-MD-Large. At a more balanced trade-off, $k = 3$, success rates of 89.28% and 89.59% are attainable for eSCN-MD-Large and GemNet-OC-MD-Large while maintaining speedups

Success Rate

DFT single-point on ML relaxed structures (ML+SP)						
Model	$k=1$			$k=5$		
	Much better	Parity	Much worse	Much better	Parity	Much worse
SchNet	0.40%	1.92%	97.67%	0.71%	2.43%	96.87%
DimeNet++	0.91%	4.25%	94.84%	1.31%	6.67%	92.01%
PaiNN	2.12%	26.79%	71.08%	3.34%	34.98%	61.68%
GemNet-OC	6.47%	66.13%	27.40%	6.88%	74.12%	19.01%
GemNet-OC-MD	6.27%	70.17%	23.56%	7.58%	76.24%	16.18%
GemNet-OC-MD-Large	5.86%	73.31%	20.83%	7.18%	79.27%	13.55%
SCN-MD-Large	6.67%	71.69%	21.64%	7.58%	79.47%	12.94%
eSCN-MD-Large	5.06%	76.95%	18.00%	6.27%	82.00%	11.73%

DFT relaxations on ML relaxed structures (ML+RX)						
Model	$k=1$			$k=5$		
	Much better	Parity	Much worse	Much better	Parity	Much worse
SchNet	10.82%	33.87%	55.31%	18.71%	46.81%	34.48%
DimeNet++	9.40%	40.85%	49.75%	15.57%	54.30%	30.13%
PaiNN	9.81%	62.49%	27.70%	14.26%	70.48%	15.27%
GemNet-OC	9.81%	72.30%	17.90%	12.23%	75.73%	12.03%
GemNet-OC-MD	8.29%	74.12%	17.59%	11.63%	78.26%	10.11%
GemNet-OC-MD-Large	7.48%	75.73%	16.78%	10.11%	81.50%	8.39%
SCN-MD-Large	8.90%	75.23%	15.87%	12.94%	78.46%	8.59%
eSCN-MD-Large	6.47%	77.65%	15.87%	9.10%	81.50%	9.40%

TABLE III. Distribution of success rates for the proposed ML+SP and ML+RX strategies on the OC20-Dense test set. “Parity” corresponds to being within 0.1 eV of the DFT adsorption energy; “Much better” corresponds to being less than 0.1 eV than DFT; and “Much worse” being higher than 0.1 eV of DFT.

of 356x and 288x, respectively.

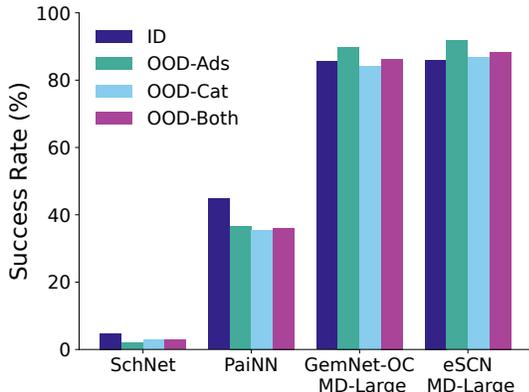


FIG. 4. ML+SP success rate at $k=5$ across the different subsplits of the OC20-Dense test set and several baseline models. Top performing models show marginal differences across the different distribution splits, suggesting good generalization performance to out-of-domain adsorbates and catalysts not contained in the OC20 training dataset.

The results suggest a spectrum of accuracy and efficiency trade-offs that one should consider when selecting a strategy. For our best models, ML+SP results are al-

most 8x faster than ML+RX with only a marginal performance decrease in success rates (3-4%), suggesting a worthwhile compromise. This difference is much more significant for worse models.

In Table III we measure the distribution of predictions that are much better, in parity, or much worse than the ground truth, where much better/worse corresponds to being lower/higher than 0.1 eV of the DFT adsorption energy. Across both strategies, we observe that the most accurate models do not necessarily find much better minima. For instance, at $k=5$ ML+RX, eSCN-MD-Large finds 9.10% of systems with much lower minima, compared to DimeNet++ finding 15.57%. Similarly, while eSCN-MD-Large outperformed models in ML+SP, it observes less of an improvement with ML+RX; a consequence of the model arriving at a considerable local minima that a subsequent DFT relaxation has minimal benefit. This further suggests that some form of noise in models can aid in finding better minima. The full set of tabulated results for ML+SP and ML+RX experiments can be found in the SI for the OC20-Dense test and validation sets.

Distribution splits Additionally, we evaluate success metrics across the different dataset subsplits. OC20-Dense uniformly samples from the four OC20 splits -

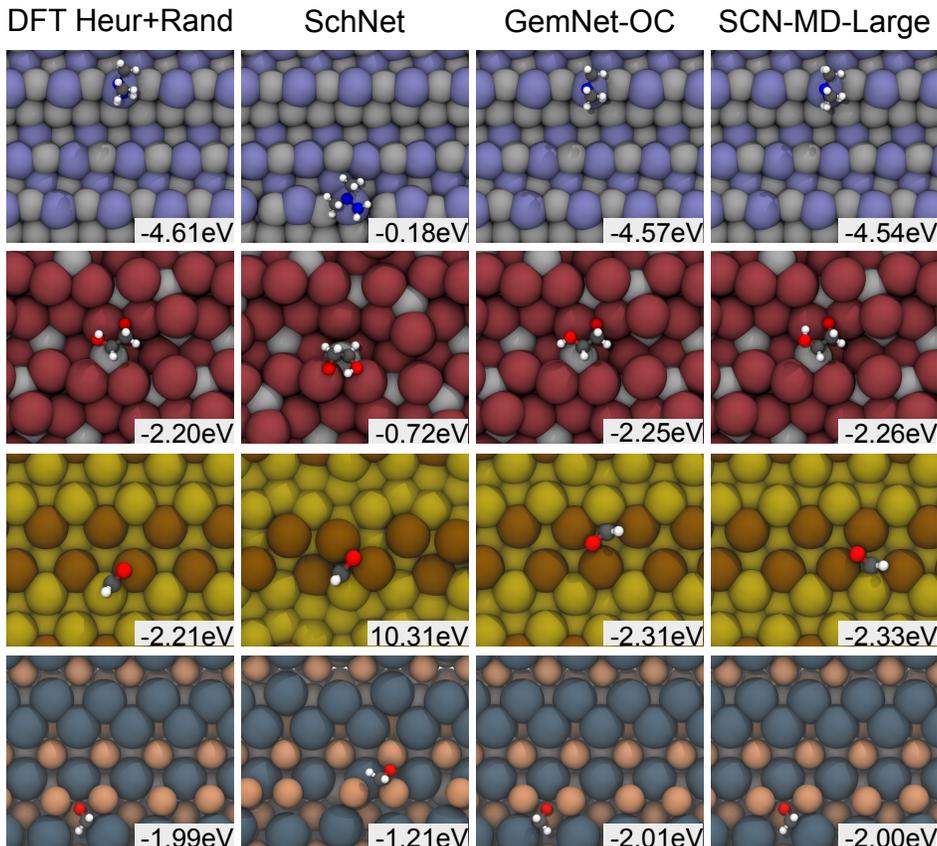


FIG. 5. Illustration of the lowest energy configurations as found by DFT-Heur+Rand, SchNet, GemNet-OC, and SCN-MD-Large on the OC20-Dense validation set. Corresponding adsorption energies are shown in the bottom right corner of each snapshot. ML relaxed structures have energies calculated with a DFT single-point, ML+SP. A variety of systems are shown including ones where ML finds lower, higher, and comparable adsorption energies to DFT. Notice that several of the configurations in the third and fourth systems are symmetrically equivalent, and that SchNet induces a large surface reconstruction in the third system resulting in the extremely large DFT energy (10.31 eV).

ID, OOD-Adsorbate, OOD-Catalyst, and OOD-Both. Across our best models, we observe that performance remains consistent across the different distribution splits (Figure 4). This suggests that for applications including adsorbates or surfaces that are not contained in OC20, AdsorbML still provides accurate and meaningful results. While we expect results to be consistent with OC20 where ID outperforms OOD, that is not necessarily the case here. eSCN-MD-Large, ML+SP at $k = 5$, achieves 86.00% on ID while a 88.35% success rate on OOD-Both, with similar trends on ML+RX. We attribute this discrepancy to the fairly small sample size per split (250). The full set of results can be found in the SI.

Configuration analysis Alongside the main results, we explore the performance of using only heuristic or only random ML configurations on the OC20-Dense validation set. Results are reported on SCN-MD-Large, for the ML+SP strategy. At $k = 5$, when only random configurations are used, success drops from 87.77% to 82.94%. More drastically, when only considering heuristic configurations, success drops significantly to 62.18%. This

suggests that random configurations can have a larger impact. Additional results can be found in the SI.

DISCUSSION

We envision this work as an important but initial step towards reducing the computational cost of DFT for not just catalysis applications, but computational chemistry more broadly. *AdsorbML* provides a spectrum of accuracy and efficiency trade-offs one can choose depending on the application and computational resources available. For example, if we are interested in screening the largest number of CO_2 reduction reaction catalysts possible, given a fixed compute budget, we could choose ML+SP at $k = 2$ for a 85% success rate while screening $\sim 3400\times$ more materials than would have been possible with DFT alone. On the other hand, if depth of study is more important, ML+RX is a good alternative as the structures are fully optimized with DFT and the computational speedup comes from reducing the total number

of relaxation steps required. In this scenario, the ML potential serves as an efficient pre-optimization step. Even though ML models comprise a small portion of the overall compute (see SI for details), we expect these requirements to be reduced even further as more effort is placed on inference efficiency in the future.

One observation that merits additional studies is that ML models found much better minima between 5%-15% of the time, depending on the efficiency trade-offs (Table III). If our ML models were perfect there would be no instances with lower adsorption energies; however, implicit noise in the form of inaccurate force predictions allows the ML models to traverse unexplored regions of the potential energy surface. Exploring to what extent implicit and explicit noise [46, 47] impact ML relaxations and downstream tasks such as success rate is an important area of future research.

Another natural extension to this work is focusing on alternative methods of global optimization and initial configuration generation. Here, we focused on accelerating brute force approaches to finding the global minimum by enumerating initial adsorbate-surface configurations. However, there are likely to be much more efficient approaches to global optimization such as minima hopping [20], constrained optimization [19, 21], Bayesian optimization, or a directly learned approach. It is worth noting that while our enumeration spanned a much larger space than traditional heuristic methods, it was not exhaustive and all-encompassing. We found that increasing the number of random configurations beyond what was sampled had diminishing returns, as the change in success rate from heuristic + 80% random DFT to heuristic + 100% random DFT was only 1.6% (see the SI for more details). If screening more ML configurations continues to be advantageous, thinking about how we handle duplicate structures could further help accuracy and efficiency. We explore this briefly in the SI, where removing systems with nearly the same ML energies resulted in marginal benefit.

While current models like GemNet-OC and eSCN-MD-Large demonstrate impressive success rates on OC20-Dense, ML relaxations without any subsequent DFT are still not accurate enough for practical applications (Table I). In order for future modeling work to address this challenge there are a number of observations worth highlighting. First, there is a positive correlation between success rate on OC20-Dense and both the *S2EF* and relaxation based Initial Structure to Relaxed Energy (*IS2RE*) OC20 tasks. Thus, relaxation based *IS2RE* and *S2EF* metrics can be used as proxies when training models on OC20. Another important note on model development is that OC20-Dense’s validation set is a subset of the OC20 validation set; as a result, the OC20 validation data should not be used for training when evaluating on OC20-Dense. Lastly, it is strongly encouraged that results reported on the OC20-Dense validation set be evaluated using a DFT

single-point calculation because the success rate metric can be manipulated by predicting only low energies. This could be done with as few as $\sim 1,000$ single-point calculations. Alongside the release of the OC20-Dense test set, we will explore releasing a public evaluation server to ensure consistent evaluation and accessibility for DFT evaluation, if there’s interest.

Tremendous progress in datasets and machine learning for chemistry has enabled models to reach the point where they can substantially enhance and augment DFT calculations. Our results demonstrate that current state-of-the-art ML models not only accelerate DFT calculations for catalysis but enable more accurate estimates of properties that require global optimization such as adsorption energies. While the models used in this work are best suited for idealized adsorbate-surface catalysts, fine-tuning strategies can help enable applications to other chemistries including metal-organic frameworks and zeolites [29]. Similarly, the models used in this work were trained on a consistent level of DFT theory (revised Perdew-Burke-Ernzerhof, no spin-polarization), generalizing to other functionals and levels of theory could also be enabled with fine-tuning or other training strategies. Given the timeline of ML model development these results would not have been possible even a couple of years ago. We anticipate this work will accelerate the large-scale exploration of complex adsorbate-surface configurations for a broad range of chemistries and applications. Generalizing these results to more diverse materials and molecules without reliance on DFT is a significant community challenge moving forward.

METHODS

Open Catalyst 2020 - Dense Dataset (OC20-Dense)

The evaluation of adsorption energy estimations requires a ground truth dataset that thoroughly explores the set of potential adsorption configurations. While OC20 computed adsorption energies for $O(1M)$ systems, the energies may not correspond to the minimum of that particular adsorbate-surface combination. More specifically, for a given catalyst surface, OC20 considers all possible adsorption sites but only places the desired adsorbate on a randomly selected site in one particular configuration. The tasks presented by OC20 enabled the development of more accurate machine learned potentials for catalysis [34, 35, 47–49], but tasks like *IS2RE*, although correlate well, are not always sufficient when evaluating performance as models are penalized when finding a different, lower energy minima - a more desirable outcome. As a natural extension to OC20’s tasks, we introduce OC20-Dense to investigate the performance of models to finding the adsorption energy.

OC20-Dense is constructed to closely approximate the

adsorption energy for a particular adsorbate-surface combination. To accomplish this, a dense sampling of initial adsorption configurations is necessary. OC20-Dense consists of two splits - a validation and test set. For each split, $\sim 1,000$ unique adsorbate-surface combinations from the respective OC20 validation/test set are sampled. A uniform sample is then taken from each of the subsplits (ID, OOD-Adsorbate, OOD-Catalyst, OOD-Both) to explore the generalizability of models on this task. For each adsorbate-surface combination, two strategies were used to generate initial adsorbate configurations: heuristic and random. The heuristic strategy serves to represent the average catalysis researcher, where popular tools like CatKit [23] and Pymatgen [22] are used to make initial configurations. Given an adsorbate and surface, Pymatgen enumerates all symmetrically identical sites, an adsorbate is placed on the site, and a random rotation along the z axis followed by slight wobbles in the x and y axis is applied to the adsorbate. While heuristic strategies seek to capture best practices, they do limit the possible search space with no guarantees that the true minimum energy is selected. To address this, we also randomly enumerate M sites on the surface and then place the adsorbate on top of the selected site. In this work, M=100 is used and a random rotation is applied to the adsorbate in a similar manner. In both strategies we remove unreasonable configurations - adsorbates not placed on the slab and/or placed too deep into the surface. DFT relaxations were then run on all configurations with the results filtered to remove those that desorb, dissociate or create surface mismatches. The minimum energy across those remaining is considered the adsorption energy. While random is meant to be a more exhaustive enumeration, it is not perfect and could likely miss some adsorbate configurations. The OC20-Dense validation set was created in a similar manner but contained notable differences, details are outlined in the SI.

The OC20-Dense test set comprises 989 unique adsorbate+surface combinations spanning 74 adsorbates and 837 bulks. Following the dense sampling, a total of 56,282 heuristic and 49,432 random configurations were calculated with DFT. On average, there were 56 heuristic and 50 random configurations per system (note - while M=100 random sites were generated, less sites were available upon filtering.) In total, ~ 4 million hours of compute were used to create the dataset. All DFT calculations were performed using *Vienna Ab initio Simulation Package* (VASP) [50–53]. A discussion on DFT settings and details can be found in the SI.

Evaluation Metrics

To sufficiently track progress, we propose two primary metrics - success rate and DFT speedup. **Success rate** is the proportion of systems in which a strategy returns

energy that is within σ , or lower of the DFT adsorption energy. A margin of $\sigma = 0.1\text{eV}$ is selected as the community is often willing to tolerate a small amount of error for practical relevance [2, 37]. Tightening this threshold for improved accuracy is a foreseeable step once models+strategies saturate. While high success rates are achievable with increased DFT compute, we use **DFT speedup** as a means to evaluate efficiency. Speedup is measured as the ratio of DFT electronic, or self-consistency (SC), steps used by DFT-Heur+Rand and the proposed strategy. Electronic steps are used as we have seen them correlate better with DFT compute time than the number of ionic, or relaxation, steps. DFT calculations that failed or resulted in invalid structures were included in speedup evaluation as they still represent realized costs in screening. We chose not to include compute time in this metric as results are often hardware dependent and can make comparing results unreliable. ML relaxation costs are excluded from this metric as hardware variance along with CPU+GPU timings make it nontrivial to normalize. While ML timings are typically negligible compared to the DFT calculations, a more detailed analysis of ML timings can be found in the SI. Metrics are reported against the rigorous ground truth - DFT-Heur+Rand, and compared to a community heuristic practice - DFT-Heur. Formally, metrics are defined in Equations 2 and 3.

$$\text{Success Rate} = \frac{\sum_i^N \mathbb{1}[\min(\hat{E}_i) - \min(E_i) \leq \sigma]}{N} \quad (2)$$

$$\text{DFT Speedup} = \frac{\sum_N N_{SCsteps}}{\sum_N \hat{N}_{SCsteps}} \quad (3)$$

where i is an adsorbate-surface system, N the total number of unique systems, $\mathbb{1}(x)$ is the indicator function, $\hat{\square}$ is the proposed strategy, $N_{SCsteps}$ is the number of self-consistency, or electronic steps, and $\min(E)$ is the minimum energy across all configurations of that particular system. For both metrics, higher is better.

Relaxation Constraints

It is possible that some of the adsorbate-surface configurations we consider may relax to a state that are necessary to discard in our analysis. For this work we considered three such scenarios: (1) desorption, (2) dissociation, and (3) significant adsorbate induced surface changes. Desorption, the adsorbate molecule not binding to the surface, is far less detrimental because desorbed systems are generally high energy. Still, it is useful to understand when none of the configurations considered have actually adsorbed to the surface. Dissociation, the breaking of an adsorbate molecule into different atoms

or molecules, is problematic because the resulting adsorption energy is no longer consistent with what is of interest, i.e., the adsorption energy of a single molecule, not two or more smaller molecules. Including these systems can appear to correspond to lower adsorption energies, but due to the energy not representing the desired system it can result in false positives. Lastly, we also discard systems with significant adsorbate induced surface changes because, just as with dissociation, we are no longer calculating the energy of interest. In calculating adsorption energy, a term is included for the energy of the clean, relaxed surface. An underlying assumption in this calculation is that the corresponding adsorbate-surface system’s resulting surface must be comparable to the corresponding clean surface, otherwise this referencing scheme fails and the resulting adsorption energy is inaccurate. For each of these instances we developed detection methods as a function of neighborhood connectivity, distance information, and atomic covalent radii. Depending on the user’s application, one may decide to tighten the thresholds defined within. Details on each of the detection methods and further discussion can be found in the SI.

DATA AVAILABILITY

The full open dataset is provided at <https://github.com/Open-Catalyst-Project/AdsorbML>.

CODE AVAILABILITY

All accompanying code is provided at <https://github.com/Open-Catalyst-Project/AdsorbML>.

AUTHOR CONTRIBUTIONS

A.P., L.Z., and Z.U. conceptualized the project and performed preliminary experiments. J.L., M.S., and B.M.W. substantially expanded the scope of the project, developed the final methodology, conducted all experiments, analyzed the results, and prepared the codebase and dataset for release under the guidance of Z.U. and L.Z. B.W. contributed to the methodology for detecting invalid configurations. A.D. contributed to the AdsorbML methodology and provided guidance on models and experiments. L.Z. and M.U. supervised the project. All authors contributed to the writing and editing of the paper. J.L., A.P., M.S., and B.M.W. contributed equally as co-first authors.

COMPETING INTERESTS

The authors declare no competing interests.

SUPPLEMENTARY INFORMATION

The supplementary information contains all tabulated results, results figure for the OC20-Dense validation set, ML model and compute details, OC20-Dense placement details, DFT calculation details, details on the relaxation constraints, model constraint counts, unvalidated ML success rates, and additional results on configuration analysis and random baselines.

-
- [1] J. K. Nørskov, F. Studt, F. Abild-Pedersen, and T. Bligaard, *Fundamental concepts in heterogeneous catalysis* (John Wiley & Sons, 2014).
- [2] L. Chanussot *et al.*, Open catalyst 2020 (oc20) dataset and community challenges, *ACS Catal.* **11**, 6059 (2021).
- [3] J. A. Dumesic, G. W. Huber, and M. Boudart, *Principles of heterogeneous catalysis* (Wiley Online Library, 2008).
- [4] C. L. Zitnick *et al.*, An introduction to electrocatalyst design using machine learning for renewable energy storage, Preprint at <https://arxiv.org/abs/2010.09435>. (2020).
- [5] K. Choudhary *et al.*, Recent advances and applications of deep learning methods in materials science, *NPJ Comput. Mater.* **8**, 59 (2022).
- [6] T. Wen, L. Zhang, H. Wang, E. Weinan, and D. J. Srolovitz, Deep potentials for materials science, *Mater. Futures* **1**, 022601 (2022).
- [7] J. Wei *et al.*, Machine learning in materials science, *InfoMat* **1**, 338 (2019).
- [8] Z. W. Ulissi, A. J. Medford, T. Bligaard, and J. K. Nørskov, To address surface reaction network complexity using scaling relations machine learning and dft calculations, *Nat. Commun.* **8**, 1 (2017).
- [9] K. Tran and Z. W. Ulissi, Active learning across intermetallics to guide discovery of electrocatalysts for co2 reduction and h2 evolution, *Nat. Catal.* **1**, 696 (2018).
- [10] M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.-T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, *et al.*, Accelerated discovery of co2 electrocatalysts using active machine learning, *Nature* **581**, 178 (2020).
- [11] X. Liu, J. Xiao, H. Peng, X. Hong, K. Chan, and J. K. Nørskov, Understanding trends in electrochemical carbon dioxide reduction rates, *Nat. Commun* **8**, 1 (2017).
- [12] J. K. Nørskov *et al.*, Trends in the exchange current for hydrogen evolution, *J. Electrochem. Soc.* **152**, J23 (2005).
- [13] X. Wan, Z. Zhang, W. Yu, H. Niu, X. Wang, and Y. Guo, Machine-learning-assisted discovery of highly efficient high-entropy alloy catalysts for the oxygen reduction reaction, *Patterns* **3**, 100553 (2022).
- [14] Z. W. Seh, J. Kibsgaard, C. F. Dickens, I. Chorkendorff, J. K. Nørskov, and T. F. Jaramillo, Combining theory and experiment in electrocatalysis: Insights into materials design, *Science* **355**, eaad4998 (2017).
- [15] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, *Phys. Rev.* **136**, B864 (1964).
- [16] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, *Physical Review* **140**, A1133 (1965).
- [17] D. S. Sholl and J. A. Steckel, *Density functional theory: a practical introduction* (John Wiley & Sons, 2022).
- [18] S. A. Teukolsky, B. P. Flannery, W. Press, and W. Vetterling, Numerical recipes in c, *SMR* **693**, 59 (1992).
- [19] A. A. Peterson, Global optimization of adsorbate-surface structures while preserving molecular identity, *Top. Catal.* **57**, 40 (2014).
- [20] S. Goedecker, Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems, *J. Chem. Phys.* **120**, 9911 (2004).
- [21] H. Jung, L. Sauerland, S. Stocker, K. Reuter, and J. T. Margraf, Machine-learning driven global optimization of surface adsorbate geometries, *NPJ Comput. Mater.* **9**, 114 (2023).
- [22] S. P. Ong *et al.*, Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, *Comput. Mater. Sci.* **68**, 314 (2013).
- [23] J. R. Boes, O. Mamun, K. Winther, and T. Bligaard, Graph theory approach to high-throughput surface adsorption structure generation, *J. Phys. Chem. A* **123**, 2281 (2019).
- [24] M. P. Andersson *et al.*, Toward computational screening in heterogeneous catalysis: Pareto-optimal methanation catalysts, *J. Catal.* **239**, 501 (2006).
- [25] T. Bligaard, J. K. Nørskov, S. Dahl, J. Matthiesen, C. H. Christensen, and J. Sehested, The Brønsted–Evans–Polanyi relation and the volcano curve in heterogeneous catalysis, *J. Catal.* **224**, 206 (2004).
- [26] F. Studt, F. Abild-Pedersen, T. Bligaard, R. Z. Sørensen, C. H. Christensen, and J. K. Nørskov, Identification of non-precious metal alloy catalysts for selective hydrogenation of acetylene, *Science (New York, N.Y.)* **320**, 1320 (2008).
- [27] A. U. Nilekar, K. Sasaki, C. A. Farberow, R. R. Adzic, and M. Mavrikakis, Mixed-metal Pt monolayer electrocatalysts with improved CO tolerance, *J. Am. Chem. Soc.* **133**, 18574 (2011).
- [28] S. Deshpande, T. Maxson, and J. Greeley, Graph theory approach to determine configurations of multidentate and high coverage adsorbates for heterogeneous catalysis, *NPJ Comput. Mater.* **6**, 1 (2020).
- [29] R. Tran *et al.*, The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts, *ACS Catal.* **13**, 3066 (2023).
- [30] A. Jain *et al.*, Commentary: The materials project: A materials genome approach to accelerating materials innovation, *APL Materials* **1**, 011002 (2013).
- [31] K. Schütt *et al.*, Schnet: A continuous-filter convolutional neural network for modeling quantum interactions, in *Adv. Neural Inf. Process. Syst.* (2017) pp. 991–1001.
- [32] J. Gastegger, J. Groß, and S. Günnemann, Directional message passing for molecular graphs, in *International Conference on Learning Representations (ICLR)* (2020).
- [33] J. Gastegger, S. Giri, J. T. Margraf, and S. Günnemann, Fast and uncertainty-aware directional message passing for non-equilibrium molecules, Preprint at <https://arxiv.org/abs/2011.14115>, p. N/A (2020).
- [34] J. Gastegger *et al.*, GemNet-OC: Developing Graph Neural Networks for Large and Diverse Molecular Simulation Datasets, *Trans. Mach. Learn. Res. (TMLR)* (2022).
- [35] C. L. Zitnick *et al.*, Spherical Channels for Modeling Atomic Interactions, in *Adv. Neural Inf. Process. Syst. (NeurIPS)* (2022).
- [36] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Machine learning of accurate energy-conserving molecular force fields, *Science Advances* **3**, e1603015 (2017).
- [37] A. Kolluru *et al.*, Open challenges in developing generalizable large-scale machine-learning models for catalyst discovery, *ACS Catal.* **12**, 8572 (2022).
- [38] C. Chang and A. J. Medford, Application of Density Functional Tight Binding and Machine Learning to Evaluate the Stability of Biomass Intermediates on the Rh(111) Surface, *J. Phys. Chem. C* **125**, 18210 (2021).

- [39] L. Chan, G. R. Hutchison, and G. M. Morris, Bayesian optimization for conformer generation, *J. Cheminform.* **11**, 32 (2019).
- [40] L. Fang, E. Makkonen, M. Todorović, P. Rinke, and X. Chen, Efficient Amino Acid Conformer Search with Bayesian Optimization, *J. Chem. Theory Comput.* **17**, 1955 (2021).
- [41] W. Xu, K. Reuter, and M. Andersen, Predicting binding motifs of complex adsorbates using machine learning with a physics-inspired graph representation, *Nat. Comput. Sci.* **2**, 443 (2022).
- [42] Z. W. Ulissi *et al.*, Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for co2 reduction, *ACS Catal.* **7**, 6600 (2017).
- [43] P. G. Ghanekar, S. Deshpande, and J. Greeley, Adsorbate chemical environment-based machine learning framework for heterogeneous catalysis, *Nat. Commun.* **13**, 1 (2022).
- [44] K. Schütt, O. Unke, and M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra, in *ICML (2021)* pp. 9377–9388.
- [45] S. Passaro and C. L. Zitnick, Reducing SO(3) convolutions to SO(2) for efficient equivariant GNNs, in *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 202, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett (PMLR, 2023) pp. 27420–27438.
- [46] M. Schaarschmidt *et al.*, Learned force fields are ready for ground state catalyst discovery, Preprint at <https://arxiv.org/abs/2209.12466>. (2022).
- [47] J. Godwin, M. Schaarschmidt, A. L. Gaunt, A. Sanchez-Gonzalez, Y. Rubanova, P. Veličković, J. Kirkpatrick, and P. Battaglia, Simple gnn regularisation for 3d molecular property prediction and beyond, in *International Conference on Learning Representations (ICLR) (2021)*.
- [48] C. Ying *et al.*, Do transformers really perform badly for graph representation?, *Adv. Neural Inf. Process. Syst.* **34**, 28877 (2021).
- [49] M. Shuaibi *et al.*, Rotation invariant graph neural networks using spin convolutions, Preprint at <https://arxiv.org/abs/2106.09575>. (2021).
- [50] G. Kresse and J. Hafner, Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium, *Phys. Rev. B* **49**, 14251 (1994).
- [51] G. Kresse and J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, *Phys. Rev. B* **54**, 11169 (1996).
- [52] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, *Phys. Rev. B* **59**, 1758 (1999).
- [53] G. Kresse and J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, *Comp. Mater. Sci.* **6**, 15 (1996).
- [54] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, DbSCAN revisited, revisited: why and how you should (still) use dbSCAN, *ACM Transactions on Database Systems (TODS)* **42**, 1 (2017).
- [55] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, The atomic simulation environment—a python library for working with atoms, *Journal of Physics: Condensed Matter* **29**, 273002 (2017).
- [56] R. García-Muelas and N. López, Statistical learning goes beyond the d-band model providing the thermochemistry of adsorbates on transition metals, *Nat. Commun.* **10**, 1 (2019).
- [57] W. Gao, Y. Chen, B. Li, S.-P. Liu, X. Liu, and Q. Jiang, Determining the adsorption energies of small molecules with the intrinsic properties of adsorbates and substrates, *Nat. Commun.* **11**, 1 (2020).

SUPPLEMENTARY TABLES

Main Paper Results

Tabulated results are provided for both the OC20-Dense validation and test set. Supplementary Table I and Supplementary Table II evaluate against DFT-Heuristic+Random. Additionally, Supplementary Table III and Supplementary Table IV evaluates against the less exhaustive, but more common DFT-Heuristic-Only baseline.

OC20-Dense Test										
ML+DFT Singlepoints (ML+SP)										
Model	<i>k=1</i>		<i>k=2</i>		<i>k=3</i>		<i>k=4</i>		<i>k=5</i>	
	Success	Speedup								
SchNet	2.33%	7202.16	2.73%	3645.62	2.73%	2446.02	3.03%	1854.64	3.13%	1496.95
DimeNet++	5.16%	7057.96	6.07%	3544.06	6.88%	2380.24	7.38%	1794.72	7.99%	1439.84
PaiNN	28.92%	6841.82	33.47%	3428.48	35.79%	2297.60	37.31%	1727.91	38.32%	1385.31
GemNet-OC	72.60%	6828.17	76.85%	3428.86	79.27%	2302.96	80.18%	1732.48	80.99%	1389.94
GemNet-OC-MD	76.44%	6803.10	80.79%	3425.00	82.61%	2291.70	83.32%	1724.68	83.82%	1383.52
GemNet-OC-MD-Large	79.17%	6856.28	82.91%	3440.35	84.43%	2299.03	85.64%	1729.76	86.45%	1388.03
SCN-MD-Large	78.36%	6878.87	83.42%	3420.15	85.14%	2290.25	86.45%	1725.47	87.06%	1383.26
eSCN-MD-Large	82.00%	6817.21	85.54%	3437.52	87.36%	2296.16	87.87%	1724.63	88.27%	1384.10
ML+DFT Relaxations (ML+RX)										
Model	<i>k=1</i>		<i>k=2</i>		<i>k=3</i>		<i>k=4</i>		<i>k=5</i>	
	Success	Speedup								
SchNet	44.69%	194.66	54.70%	98.72	60.16%	66.70	62.99%	51.61	65.52%	42.08
DimeNet++	50.25%	257.78	59.96%	132.39	63.50%	89.82	66.53%	68.52	69.87%	55.73
PaiNN	72.30%	373.54	77.65%	189.02	80.89%	126.04	83.62%	94.07	84.73%	76.26
GemNet-OC	82.10%	727.68	85.64%	372.27	87.06%	252.36	87.06%	190.00	87.97%	151.89
GemNet-OC-MD	82.41%	759.83	86.55%	392.89	88.27%	260.95	89.18%	193.01	89.89%	154.99
GemNet-OC-MD-Large	83.22%	872.63	87.87%	437.01	89.59%	288.80	90.90%	216.86	91.61%	172.46
SCN-MD-Large	84.13%	811.03	88.78%	403.70	89.79%	262.88	90.90%	194.10	91.41%	154.31
eSCN-MD-Large	84.13%	1064.42	88.07%	530.80	89.28%	356.25	89.79%	267.78	90.60%	215.58

Supplementary Table I. Model success and speedup results as evaluated against DFT-Heuristic+Random across varying k for the OC20-Dense test set. This evaluation corresponds to a more exhaustive, but expensive approach - reflected by the increased speedups.

OC20-Dense Validation										
ML+DFT Single-points (ML+SP)										
Model	<i>k=1</i>		<i>k=2</i>		<i>k=3</i>		<i>k=4</i>		<i>k=5</i>	
	Success	Speedup								
SchNet	2.77%	4266.13	3.91%	2155.36	4.32%	1458.77	4.73%	1104.88	5.04%	892.79
DimeNet++	5.34%	4271.23	7.61%	2149.78	8.84%	1435.21	10.07%	1081.96	10.79%	865.20
PaiNN	27.44%	4089.77	33.61%	2077.65	36.69%	1395.55	38.64%	1048.63	39.57%	840.44
GemNet-OC	68.76%	4185.18	77.29%	2087.11	80.78%	1392.51	81.50%	1046.85	82.94%	840.25
GemNet-OC-MD	68.76%	4182.04	78.21%	2092.27	81.81%	1404.11	83.25%	1053.36	84.38%	841.64
GemNet-OC-MD-Large	73.18%	4078.76	79.65%	2065.15	83.25%	1381.39	85.41%	1041.50	86.02%	834.46
SCN-MD-Large	77.80%	3974.21	84.28%	1989.32	86.33%	1331.43	87.36%	1004.40	87.77%	807.00
ML+DFT Relaxations (ML+RX)										
Model	<i>k=1</i>		<i>k=2</i>		<i>k=3</i>		<i>k=4</i>		<i>k=5</i>	
	Success	Speedup								
SchNet	35.25%	131.26	46.04%	68.64	51.08%	47.24	55.50%	36.19	58.07%	29.58
DimeNet++	43.47%	175.54	54.57%	90.04	60.84%	61.01	64.65%	46.06	67.93%	37.21
PaiNN	61.66%	262.38	71.12%	131.30	75.75%	86.64	79.03%	64.27	81.19%	51.88
GemNet-OC	73.59%	448.50	83.14%	231.76	86.84%	152.57	88.18%	117.40	89.41%	95.24
GemNet-OC-MD	72.25%	503.62	81.40%	251.69	85.10%	167.71	87.05%	124.21	88.49%	100.21
GemNet-OC-MD-Large	76.16%	543.48	83.04%	272.66	86.13%	183.47	88.18%	139.20	88.90%	112.29
SCN-MD-Large	80.58%	489.29	86.95%	252.77	89.31%	170.15	90.13%	126.72	90.65%	100.92

Supplementary Table II. Model success and speedup results as evaluated against DFT-Heuristic+Random across varying k for the OC20-Dense validation set. This evaluation corresponds to a more exhaustive, but expensive approach - reflected by the increased speedups.

OC20-Dense Test										
ML+DFT Singlepoints (ML+SP)										
Model	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	Success	Speedup								
SchNet	2.88%	4006.55	3.40%	2027.43	3.50%	1359.96	3.81%	1030.36	4.01%	831.24
DimeNet++	6.48%	3917.46	7.61%	1964.28	8.64%	1318.71	9.05%	993.91	9.47%	797.17
PaiNN	33.64%	3817.50	38.07%	1909.48	40.43%	1278.98	41.87%	961.09	42.90%	770.42
GemNet-OC	76.85%	3799.78	81.28%	1906.74	83.54%	1279.51	84.36%	961.73	84.88%	771.19
GemNet-OC-MD	79.84%	3792.77	83.74%	1909.18	85.70%	1275.16	86.73%	958.61	87.04%	768.24
GemNet-OC-MD-Large	82.82%	3816.00	86.11%	1914.08	87.65%	1278.30	88.48%	961.07	89.40%	770.73
SCN-MD-Large	82.61%	3838.32	87.24%	1905.32	88.79%	1275.08	89.92%	959.76	90.33%	768.69
eSCN-MD-Large	85.60%	3795.10	88.79%	1913.05	90.53%	1277.51	90.95%	958.69	91.46%	768.80
ML+DFT Relaxations (ML+RX)										
Model	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	Success	Speedup								
SchNet	47.43%	108.43	57.92%	55.02	63.17%	37.14	66.15%	28.72	68.42%	23.43
DimeNet++	54.22%	143.15	63.99%	73.61	67.80%	49.86	70.78%	38.00	74.07%	30.90
PaiNN	76.54%	208.76	81.89%	105.98	84.88%	70.64	87.35%	52.68	88.68%	42.66
GemNet-OC	85.91%	414.29	89.40%	209.98	90.84%	141.75	90.95%	106.41	91.67%	85.01
GemNet-OC-MD	86.01%	429.26	89.40%	222.61	91.05%	148.26	91.67%	109.18	92.08%	87.38
GemNet-OC-MD-Large	86.21%	502.76	90.53%	248.12	92.08%	163.33	93.11%	122.23	93.52%	96.86
SCN-MD-Large	87.45%	454.12	91.98%	226.07	92.80%	147.38	93.62%	108.85	94.03%	86.29
eSCN-MD-Large	87.55%	594.32	91.15%	296.49	92.28%	198.63	92.80%	149.40	93.62%	120.13

Supplementary Table III. Model success and speedup results as evaluated against DFT-Heuristic across varying k for the OC20-Dense test set. This evaluation corresponds to a more common community approach.

OC20-Dense Validation										
ML+DFT Single-points (ML+SP)										
Model	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	Success	Speedup								
SchNet	4.82%	1520.54	6.75%	768.60	7.60%	519.72	8.14%	393.38	8.78%	317.71
DimeNet++	10.39%	1518.59	14.03%	766.37	15.85%	511.41	17.34%	385.57	18.20%	308.20
PaiNN	39.08%	1464.56	45.40%	742.08	48.82%	497.86	50.64%	373.62	52.03%	299.01
GemNet-OC	75.70%	1502.03	84.15%	746.69	87.37%	497.62	88.01%	373.47	89.08%	299.44
GemNet-OC-MD	76.77%	1494.44	85.44%	747.21	88.22%	501.47	89.83%	375.80	90.79%	299.95
GemNet-OC-MD-Large	80.73%	1455.42	85.97%	736.25	89.29%	492.34	91.11%	370.49	91.86%	296.54
SCN-MD-Large	85.12%	1430.68	91.01%	714.02	92.29%	477.07	92.93%	359.51	93.36%	288.46
ML+DFT Relaxations (ML+RX)										
Model	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	Success	Speedup								
SchNet	44.54%	47.68	55.46%	24.75	59.74%	17.07	64.45%	13.06	67.45%	10.68
DimeNet++	52.03%	63.92	64.56%	32.90	70.13%	22.22	73.98%	16.74	77.52%	13.48
PaiNN	70.24%	97.51	79.44%	48.50	84.05%	32.01	86.83%	23.69	89.08%	19.04
GemNet-OC	79.55%	166.33	88.44%	86.65	90.90%	56.63	92.29%	43.62	93.04%	35.30
GemNet-OC-MD	78.80%	185.49	87.58%	93.00	91.11%	62.08	92.18%	45.84	93.58%	36.93
GemNet-OC-MD-Large	81.80%	201.28	88.44%	100.31	91.54%	67.66	93.15%	51.11	93.68%	41.21
SCN-MD-Large	86.72%	185.47	92.83%	96.46	94.43%	64.01	95.07%	47.43	95.50%	37.72

Supplementary Table IV. Model success and speedup results as evaluated against DFT-Heuristic across varying k for the OC20-Dense validation set. This evaluation corresponds to a more common community approach.

Subsplit Results

Results evaluated across different subsplits are shown in Supplementary Table V and Supplementary Table VI.

		OC20-Dense Test									
Split	Model	ML+DFT Singlepoints (ML+SP)					ML+DFT Relaxations (ML+RX)				
		$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
ID	SchNet	2.40%	4.00%	4.00%	4.40%	4.80%	45.20%	55.20%	60.80%	62.40%	65.60%
	DimeNet++	5.60%	6.80%	8.00%	9.20%	9.60%	49.60%	60.00%	64.00%	68.00%	71.20%
	PaiNN	36.00%	40.40%	43.60%	44.40%	44.80%	72.80%	78.00%	78.80%	80.80%	81.20%
	GemNet-OC	74.40%	79.20%	81.60%	81.60%	81.60%	82.40%	86.40%	87.60%	87.60%	88.00%
	GemNet-OC-MD	78.00%	80.00%	80.80%	81.60%	82.40%	81.60%	84.40%	85.60%	86.80%	87.60%
	GemNet-OC-MD-Large	78.80%	83.20%	84.80%	85.20%	85.60%	83.60%	86.80%	87.20%	88.40%	88.80%
	SCN-MD-Large	77.20%	80.40%	82.40%	84.00%	84.80%	82.40%	86.40%	88.40%	88.40%	89.20%
	eSCN-MD-Large	80.40%	83.20%	84.40%	84.80%	86.00%	82.80%	84.80%	86.40%	86.80%	88.40%
OOD-Ads	SchNet	1.63%	1.63%	1.63%	2.04%	2.04%	46.53%	53.88%	58.78%	62.45%	64.08%
	DimeNet++	3.27%	4.08%	4.08%	4.49%	5.71%	46.94%	57.55%	60.82%	63.27%	65.71%
	PaiNN	28.16%	33.47%	34.69%	36.33%	36.73%	71.84%	80.00%	84.08%	85.71%	86.53%
	GemNet-OC	77.14%	79.59%	81.22%	82.04%	83.67%	84.49%	86.53%	87.76%	87.76%	88.57%
	GemNet-OC-MD	82.04%	86.53%	88.57%	88.57%	88.57%	84.08%	89.39%	91.43%	92.24%	93.06%
	GemNet-OC-MD-Large	83.27%	86.94%	88.16%	89.39%	89.80%	86.53%	90.20%	92.65%	93.47%	93.88%
	SCN-MD-Large	83.67%	88.57%	89.80%	90.20%	91.02%	88.16%	92.65%	93.06%	94.69%	94.69%
	eSCN-MD-Large	86.12%	89.80%	91.43%	91.84%	91.84%	87.76%	91.84%	93.06%	93.47%	93.47%
OOD-Cat	SchNet	2.86%	2.86%	2.86%	2.86%	2.86%	45.71%	55.92%	61.63%	63.67%	66.53%
	DimeNet++	4.49%	5.71%	6.53%	6.94%	7.35%	53.06%	63.27%	65.71%	68.98%	73.06%
	PaiNN	24.90%	28.98%	31.84%	34.29%	35.51%	75.10%	77.55%	81.22%	84.90%	86.12%
	GemNet-OC	64.90%	71.02%	73.06%	75.10%	76.73%	78.78%	82.86%	84.49%	84.49%	86.53%
	GemNet-OC-MD	70.61%	76.73%	79.59%	80.00%	80.41%	81.63%	86.53%	88.16%	88.57%	88.98%
	GemNet-OC-MD-Large	75.51%	78.78%	81.22%	82.86%	84.08%	82.04%	87.35%	88.98%	90.20%	90.61%
	SCN-MD-Large	75.92%	82.04%	83.67%	85.31%	86.12%	82.86%	87.35%	88.57%	90.20%	91.02%
	eSCN-MD-Large	79.18%	84.49%	86.53%	86.53%	86.94%	82.86%	87.76%	88.57%	88.57%	90.20%
OOD-Both	SchNet	2.41%	2.41%	2.41%	2.81%	2.81%	41.37%	53.82%	59.44%	63.45%	65.86%
	DimeNet++	7.23%	7.63%	8.84%	8.84%	9.24%	51.41%	59.04%	63.45%	65.86%	69.48%
	PaiNN	26.51%	30.92%	32.93%	34.14%	36.14%	69.48%	75.10%	79.52%	83.13%	85.14%
	GemNet-OC	73.90%	77.51%	81.12%	81.93%	81.93%	82.73%	86.75%	88.35%	88.35%	88.76%
	GemNet-OC-MD	75.10%	79.92%	81.53%	83.13%	83.94%	82.33%	85.94%	87.95%	89.16%	89.96%
	GemNet-OC-MD-Large	79.12%	82.73%	83.53%	85.14%	86.35%	80.72%	87.15%	89.56%	91.57%	93.17%
	SCN-MD-Large	76.71%	82.73%	84.74%	86.35%	86.35%	83.13%	88.76%	89.16%	90.36%	90.76%
	eSCN-MD-Large	82.33%	84.74%	87.15%	88.35%	88.35%	83.13%	87.95%	89.16%	90.36%	90.36%

Supplementary Table V. Success rates evaluated on DFT-Heur+Rand across the different in-domain and out-of-domain subsplits for the OC20-Dense test set. Results reported for both ML+SP and ML+RX strategies across different k values.

		OC20-Dense Validation									
Split	Model	ML+DFT Single-points (ML+SP)					ML+DFT Relaxations (ML+RX)				
		$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
ID	SchNet	1.64%	2.87%	3.69%	3.69%	3.69%	37.70%	50.41%	54.92%	60.25%	62.30%
	DimeNet++	3.69%	4.51%	4.92%	6.56%	7.38%	46.31%	55.33%	60.66%	65.57%	67.62%
	PaiNN	31.97%	39.34%	43.03%	45.08%	46.31%	60.66%	72.54%	75.82%	78.28%	79.92%
	GemNet-OC	76.23%	81.15%	84.84%	85.66%	86.48%	74.18%	84.84%	90.57%	92.62%	92.62%
	GemNet-OC-MD	71.31%	81.15%	84.02%	86.48%	87.30%	73.77%	84.84%	86.89%	88.11%	88.52%
	GemNet-OC-MD-Large	77.87%	82.79%	85.25%	87.30%	87.30%	77.05%	83.61%	86.07%	88.52%	88.52%
	SCN-MD-Large	81.15%	85.66%	87.30%	88.52%	88.52%	82.38%	86.48%	87.70%	88.93%	89.34%
OOD-Ads	SchNet	4.51%	6.56%	6.56%	6.97%	7.38%	37.70%	48.77%	54.10%	59.43%	60.66%
	DimeNet++	5.74%	7.79%	9.84%	11.48%	12.70%	45.49%	56.15%	63.93%	67.62%	71.31%
	PaiNN	29.51%	36.07%	38.93%	40.57%	41.80%	63.52%	72.95%	77.05%	79.51%	81.56%
	GemNet-OC	68.44%	79.51%	84.02%	84.43%	86.07%	74.18%	81.97%	85.66%	86.48%	88.93%
	GemNet-OC-MD	70.90%	81.15%	84.43%	86.07%	86.89%	71.72%	81.15%	86.07%	88.11%	88.93%
	GemNet-OC-MD-Large	74.18%	83.20%	85.66%	88.52%	89.75%	78.28%	84.02%	88.11%	90.98%	91.39%
	SCN-MD-Large	77.87%	84.02%	86.89%	88.11%	88.93%	79.10%	86.89%	88.93%	90.16%	90.98%
OOD-Cat	SchNet	1.68%	2.10%	2.94%	4.20%	4.62%	34.87%	44.12%	50.00%	54.62%	57.56%
	DimeNet++	7.14%	10.08%	10.50%	11.34%	11.76%	40.34%	54.20%	57.56%	61.76%	65.97%
	PaiNN	25.63%	31.93%	35.29%	37.39%	38.24%	67.23%	76.05%	79.83%	84.03%	86.13%
	GemNet-OC	70.59%	78.99%	82.35%	83.19%	85.29%	77.73%	86.97%	89.08%	89.50%	90.76%
	GemNet-OC-MD	71.43%	80.67%	84.03%	84.45%	85.71%	74.79%	82.35%	86.97%	89.50%	92.02%
	GemNet-OC-MD-Large	70.17%	76.05%	82.35%	84.45%	84.87%	76.47%	82.77%	86.55%	88.66%	89.50%
	SCN-MD-Large	80.67%	89.08%	89.92%	90.76%	91.18%	83.61%	92.02%	94.54%	95.38%	95.38%
OOD-Both	SchNet	3.24%	4.05%	4.05%	4.05%	4.45%	30.77%	40.89%	45.34%	47.77%	51.82%
	DimeNet++	4.86%	8.10%	10.12%	10.93%	11.34%	41.70%	52.63%	61.13%	63.56%	66.80%
	PaiNN	22.67%	27.13%	29.55%	31.58%	31.98%	55.47%	63.16%	70.45%	74.49%	77.33%
	GemNet-OC	59.92%	69.64%	72.06%	72.87%	74.09%	68.42%	78.95%	82.19%	84.21%	85.43%
	GemNet-OC-MD	61.54%	70.04%	74.90%	76.11%	77.73%	68.83%	77.33%	80.57%	82.59%	84.62%
	GemNet-OC-MD-Large	70.45%	76.52%	79.76%	81.38%	82.19%	72.87%	81.78%	83.81%	84.62%	86.23%
	SCN-MD-Large	71.66%	78.54%	81.38%	82.19%	82.59%	77.33%	82.59%	86.23%	86.23%	87.04%

Supplementary Table VI. Success rates evaluated on DFT-Heur+Rand across the different in-domain and out-of-domain subsplits for the OC20-Dense validation set. Results reported for both ML+SP and ML+RX strategies across different k values.

Model Implementation and Compute Details

Models used for this work included SchNet [31], DimeNet++ [32, 33], PaiNN [44], GemNet-OC [34], GemNet-OC-MD [34], GemNet-OC-MD-Large [34], and SCN-MD-Large [35]. Note, while Gasteiger, et al.[34] used two trained GemNet-OC-MD-Large models optimized for energy and forces to run relaxations and make *IS2RE* predictions, we use only a single model, the force variant. No models were trained as part of this work, pretrained checkpoints were obtained directly from <https://github.com/Open-Catalyst-Project/ocp/blob/main/MODELS.md> or by contacting the authors directly (SCN-MD-Large). All models used identical optimization parameters and ran for 300 relaxation steps or until that max per-atom force norm was less than or equal to 0.02 eV/Å, whichever comes first. All model configuration files can be found at <https://github.com/Open-Catalyst-Project/AdsorbML/tree/main/configs>.

While speedup metrics are defined solely based off DFT electronic steps, the compute associated with ML relaxations are reported in Supplementary Table VII alongside the DFT compute necessary for the example of evaluating the top $k = 5$ systems. All model relaxations were done on 32GB NVIDIA V100 cards.

Model	ML RX (GPU-hrs)	DFT SP (CPU-hrs)	DFT RX (CPU-hrs)
SchNet	24.2	2,199.16	51,989.30
DimeNet++	249.2	2,576.00	54,785.46
PaiNN	60.4	2,225.63	38,409.31
GemNet-OC	133.0	2,824.25	25,073.62
GemNet-OC-MD	133.0	2,441.12	26,411.08
GemNet-OC-MD-Large	638.3	2,448.27	19,265.97
SCN-MD-Large	1129.2	2,645.39	17,313.90
DFT Heuristic	-	-	806,351.19
DFT Random	-	-	1,096,396.77

Supplementary Table VII. Total compute time associated with ML relaxations, DFT singlepoints (SP) and relaxations (RX) on best $k = 5$ ML predictions for the OC20-Dense validation set. Baseline DFT Heuristic and Random ground truths are also shown for reference.

To consider both GPU and CPU timing we can compute an alternative speedup metric based off their total compute time:

$$\text{Alternative DFT Speedup} = \frac{\text{Total DFT Time}}{\text{Total ML+DFT Time}}$$

To compare the impact of ML compute time we consider the alternative speedup metric with and without factoring in ML compute in the total time for the OC20-Dense validation set. Results are reported in Supplementary Table VIII.

For larger model variants like SCN-MD-Large and GemNet-OC-MD-Large we see that ML compute time is non-negligible, with speedups dropping from 3596x and 3885x to 1147x and 1686x, respectively when evaluating ML+SP at $k = 1$. Smaller models like GemNet-OC, GemNet-OC-MD, and PaiNN see marginal drops in speedups. When considering ML+RX, the overall DFT time involved in refining relaxations makes ML compute a lot less significant, with the largest models like SCN-MD-Large and GemNet-OC-MD-Large seeing only a 24.6% and 14.2% slowdown. Also shown in Supplementary Table VIII, as k is increased to 5, the compute associated with ML becomes more insignificant. While ML is often treated as negligible in workflows, it is important to be aware of the real cost, particularly when working at scale. These results suggest that strategies that leverage minimal DFT (ML+SP) can often be bottlenecked by ML compute if large, complex models are used like SCN-MD-Large. While leveraging the state-of-the-art model is often favorable, these results suggest that sacrificing a few percentage points on success rate could be a meaningful trade-off if we can increase throughput at inference (e.g. GemNet-OC vs SCN-MD-Large). We note that the models used in this work were used off the shelf, without optimizing for inference. There is significant potential to improve ML throughput with adequate optimizations.

Alternative DFT Speedup								
Model	$k=1$				$k=5$			
	ML+SP		ML+RX		ML+SP		ML+RX	
	without ML	with ML						
SchNet	4326.08	4100.65	182.99	182.57	865.22	855.81	36.60	36.58
DimeNet++	3693.23	2489.08	173.65	169.79	738.65	673.48	34.73	34.57
PaiNN	4274.63	3763.54	247.69	245.76	854.93	832.32	49.54	49.46
GemNet-OC	3368.59	2726.64	379.43	369.63	673.72	643.42	75.89	75.49
GemNet-OC-MD	3897.29	3062.98	360.22	351.37	779.46	739.19	72.04	71.68
GemNet-OC-MD-Large	3885.90	1686.86	493.81	423.63	777.18	616.45	98.76	95.59
SCN-MD-Large	3596.35	1147.45	549.49	414.37	719.27	504.10	109.90	103.17

Supplementary Table VIII. Alternative speedup metric as computed by total runtime across all models on the OC20-Dense validation set. Speedup is computed with and without factoring in ML runtime to compare results. Results are evaluated for both ML+SP and ML+RX strategies at $k = 1$ and $k = 5$.

Deduplication

It is possible that different initial configurations relax to identical, or symmetrically identical sites with nearly identical ML energies. As a result, this means that redundant DFT calculations may be performed if such systems appear in the best k ranking. Another way to look at this is that it is beneficial to have diverse candidates in the best k . This becomes more important if we increase the number of random placements.

One way to address this is through a deduplication step before selecting the best k in the proposed algorithm. This would enable us to increase the number of random placements without the concern of redundant calculations. To explore this, we incorporate a deduplication step via density-based spatial clustering of applications with noise (DBSCAN) [54] to cluster configurations based off ML relaxed energies. The best k systems are then selected by looping through each cluster, taking the lowest energy of the group, and then removing it from the cluster until k placements have been selected. Clusters are controlled by a hyperparameter ΔE , specifying the maximum energy difference between points in a cluster. Too small of a ΔE can result in little deduplication while too large can result in unique systems being clustered together. Results on SCN-MD-Large ML+SP are reported in Supplementary Table IX for various ΔE , with $\Delta E = 0$ corresponding to no deduplication.

ΔE	Success				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
0	77.80%	84.28%	86.33%	87.36%	87.77%
1.00E-09	77.80%	84.28%	86.33%	87.36%	87.77%
0.005	77.80%	84.28%	86.43%	87.36%	87.98%
0.01	77.80%	84.69%	86.43%	87.46%	87.98%
0.02	77.80%	83.25%	84.79%	85.61%	86.13%

Supplementary Table IX. Deduplication results with SCN-MD-Large ML+SP under different ΔE cluster thresholds. Success rates computed against the DFT-Heur+Rand ground truth on the OC20-Dense validation set.

While we observe some improvements with deduplication, overall we see marginal benefit across all k . A ΔE of 0.01eV provides a minor improvement compared to no deduplication. More substantial improvements could come from exploring other strategies (e.g. structure-based) or increasing the number of placements. We leave these questions as potential future directions.

Varying Heuristic+Random ratios

While a fixed set of random configurations was generated for each system ($M = 100$), an obvious question arises if more random configurations will aid in finding better minima. To explore whether a saturation point exists, we

report results on DFT-Heur + varying proportion of random configurations in Supplementary Table X. While success rates unsurprisingly increase, we see diminishing returns with only a 1.6% difference between 80% and 100% random configurations as compared to the 8% improvement between 0% and 10% additional random configurations.

+% Random	Success Rate	Speedup
0%	71.12%	2.87
10%	79.45%	2.45
20%	86.33%	2.11
30%	88.90%	1.86
40%	91.06%	1.66
50%	93.42%	1.49
60%	94.86%	1.36
70%	96.20%	1.25
80%	98.36%	1.16
90%	99.08%	1.08
100%	100.00%	1.00

Supplementary Table X. Success rate and speedup of varying proportions of random configurations added to the heuristics. DFT-Heur corresponds to the 0.0% data point, and DFT-Heur+Rand corresponds to the ground truth used throughout the paper. Results reported for the OC20-Dense validation set.

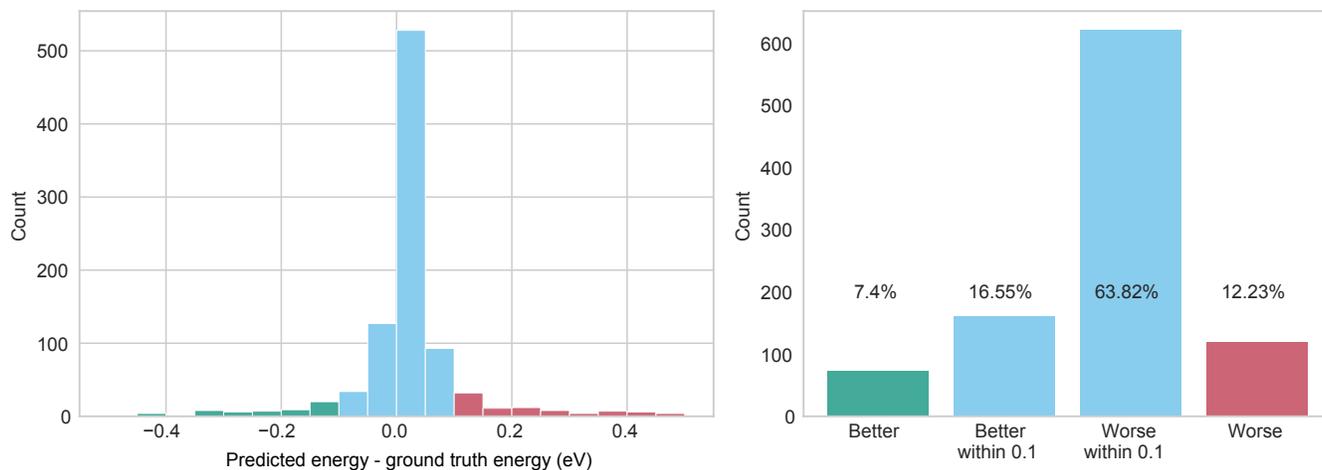
Additional Results

To better visualize the distribution of success rates, Supplementary Figure 1 shows the breakdown for SCN-MD-Large. Even though the success rates of single-points and relaxations are similar, the more nuanced histogram shows how the predicted energies are lower with relaxations.

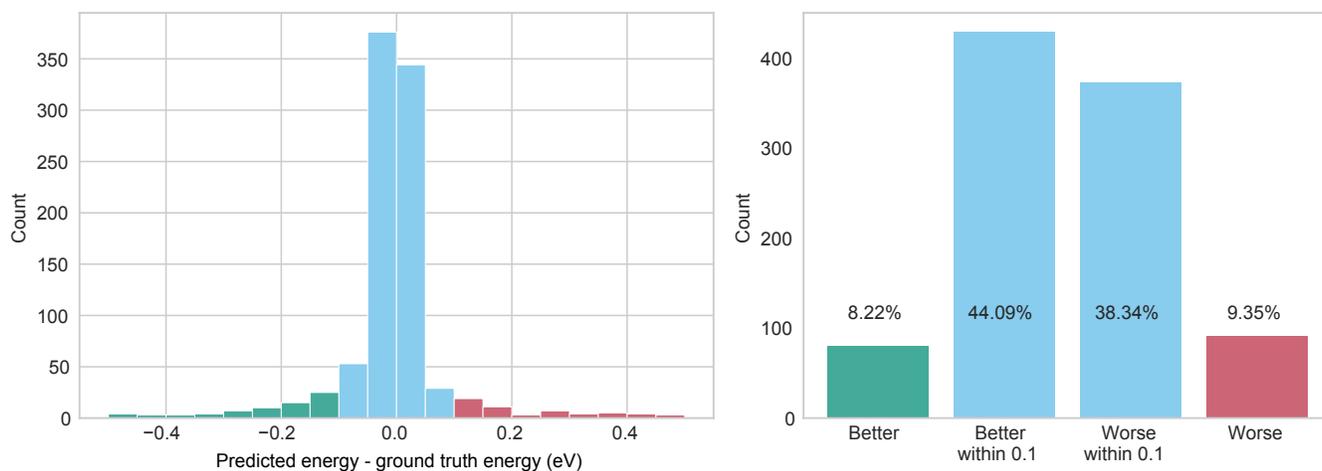
Configuration analysis

Supplementary Table XI compares the use of random and heuristic configurations independently. Random alone does slightly worse and heuristic alone does significantly worse when compared to the same ground truth. However, when limiting ground truth to the same set of initial configurations, success rates return to higher values.

SCN Single-points, k=5



SCN Relaxations, k=5



Supplementary Figure 1. Results for SCN-MD-Large, single-points (top) and relaxations (bottom) at $k = 5$. Left: distribution of differences between predicted and ground truth adsorption energies. Lower is better, meaning that *AdsorbML* found a better binding site. Differences within 0.1 eV are also considered comparable and a success, represented in teal. Red bars are failure cases. Right: an aggregation of the major categories of energy differences. Results reported on the OC20-Dense validation set.

Configuration type	Success				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Heuristic ML, GT-both	57.04%	60.64%	61.36%	61.87%	62.18%
Random ML, GT-both	73.48%	79.03%	81.91%	82.73%	82.94%
Heuristic ML, GT-heuristic	77.94%	83.30%	84.05%	84.69%	85.12%
Random ML, GT-random	78.11%	83.61%	86.41%	87.24%	87.45%

Supplementary Table XI. Comparing random and heuristic configurations on the OC20-Dense validation set.. Heuristic ML represents using the *AdsorbML* algorithm but only on heuristic initial configurations, and Random ML uses only random configurations. GT-both considers both heuristic and random for ground truth (which was done for the main results), and GT-heuristic and GT-random mean that ground truth only uses heuristic or random configurations, respectively. Results show that removing random configurations decreases the success rates more. When switching to GT-heuristic, Heuristic ML becomes competitive again, indicating that random configurations helps both AdsorbML and ground truth.

Random baselines

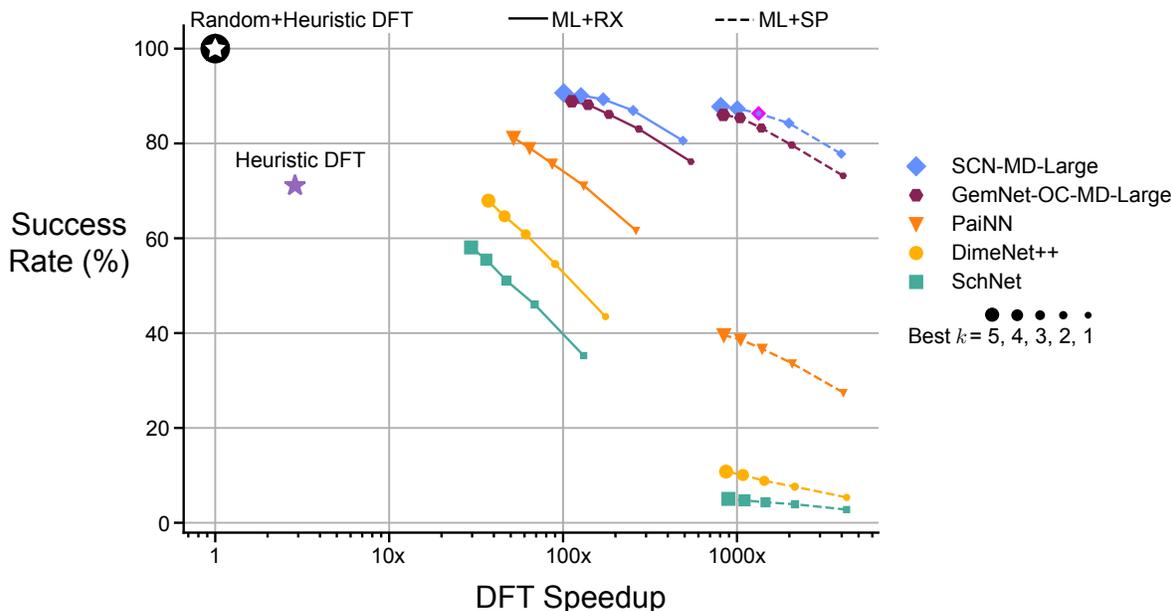
Supplementary Table XII shows success rates if we use ML to choose a different set of k configurations, namely a random set and the worst set. These sanity checks confirm that the ML ranking of the best k are indeed crucial, and that random and worst k perform badly as expected.

Binding site selection	Success				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Best k	77.80%	84.28%	86.33%	87.36%	87.77%
Random k	20.90%	31.86%	40.08%	45.53%	50.15%
Worst k	1.85%	3.39%	4.32%	5.14%	6.27%

Supplementary Table XII. Comparison against random and worst ranking baselines for SCN-MD-Large single-points on the OC20-Dense validation set. “Best k” refers to the regular algorithm, with the same results as in Table I. For “Random k”, we choose a random set of k placements and averaged success rates across three seeds. For “Worst k”, we choose the placements with highest ML predicted energies rather than lowest. As expected, random performs badly and choosing high energy placements performs the worst.

SUPPLEMENTARY FIGURES

OC20-Dense Validation Success v. Speedup



Supplementary Figure 2. Overview of the accuracy-efficiency trade-offs of the proposed *AdsorbML* methods across several baseline GNN models on the OC20-Dense validation set. For each model, DFT speedup and corresponding success rate are plotted for ML+RX and ML+SP across various best- k . A system is considered successful if the predicted adsorption energy is within 0.1 eV of the DFT minimum, or lower. All success rates and speedups are relative to Random+Heuristic DFT. Heuristic DFT is shown as a common community baseline. The upper right-hand corner represent the optimal region - maximizing speedup and success rate. The point outlined in pink corresponds to the balanced option - a 86.33% success rate and 1331x speedup.

SUPPLEMENTARY NOTES

Relaxation Constraints

To ensure proposed algorithms are accurately computing adsorption energies of the desired molecule, we filter problematic, or anomalous structures. These include dissociation, desorption, and adsorbate-induced surface changes. To accomplish this, we rely on neighborhood detection methods implemented in the Atomic Simulation Environment (ASE) [55] detailed below.

To detect dissociation (1), a connectivity matrix is constructed for the adsorbate prior to relaxation and another is constructed for the adsorbate after relaxation. Two atoms are considered connected if the covalent radii have any overlap. The two matrices are compared and must be identical, otherwise it is classified as dissociated. To detect desorption, a connectivity matrix is constructed for the relaxed adsorbate-surface configuration. In this case, atoms are considered connected if there is any overlap of the atomic radii with a small cushion. This cushion is a 1.5 multiplier to the covalent radii. We did this so that we would only discard systems where the adsorbate has no interaction with the surface to avoid discarding physisorbed systems. To detect significant adsorbate induced surface changes (3), a connectivity matrix is constructed for the relaxed surface and another is constructed for the relaxed surface-adsorbate configuration. For the surface-adsorbate configuration, the subset of atoms belonging to the surface are considered. The process of constructing the connectivity matrices is repeated twice. First, with a cushion applied to the relaxed surface but no cushion applied to the relaxed adsorbate-surface configuration. Second, with a cushion applied to the relaxed adsorbate-surface configuration but no cushion applied to the relaxed surface. This cushion is a

1.5 multiplier to the covalent radii. For each of these cases, we check that the connected atoms for the system without the cushion is a subset of those found with the cushion. Considering both cases ensures that we are considering both bond breaking and bond forming events and are not ignoring cases where bonds are only broken as would occur if a surface atom moved up into the vacuum layer.

Constraint Counts

For all models, ML relaxations were removed that violated certain physical constraints (dissociation, desorption, surface mismatch). Supplementary Table XIII shares a breakdown of the filtered counts for different models on the OC20-Dense validation set. Unsurprisingly, top performing models like SCN-MD-Large and GemNet-OC have a lot fewer removed and more comparable to DFT than models like SchNet and DimeNet++.

Model	Dissociation	Desorption	Surface mismatch	Total
SchNet	10,287	4,183	40,639	48,546
DimeNet++	9,706	6,389	17,491	29,913
PaiNN	8,273	5,211	8,656	20,539
GemNet-OC	8,944	4,815	10,105	22,143
GemNet-OC-MD	8,781	5,019	9,526	21,676
GemNet-OC-MD-Large	8,871	4,693	10,000	21,829
SCN-MD-Large	8,524	4,972	9,048	20,860
DFT-Heur+Rand	9,075	3,491	8,407	19,432

Supplementary Table XIII. Breakdown of relaxations removed for violating proposed constraints for ML and DFT ground truth relaxations. Note, a system may have more than one violation type, hence the total may not correspond to the sum across all types. Counts reported on the validation set.

DFT and Calculation Details

DFT relaxations were performed consistent with OC20’s methodology. *Vienna Ab initio Simulation Package* (VASP) with projector augmented wave (PAW) pseudopotentials and the revised Perdew-Burke-Ernzerhof (RPBE) functional were used for all calculations [50–53]. All relaxations were performed with a maximum number of electronic steps of 60. All single-point evaluations were allowed a maximum of 300 electronic steps. This was done to ensure that the initialized wavefunction had sufficient steps to converge. Single-point calculations in which electronic steps were unconverged were discarded. The same was done for unconverged electronic steps at relaxed structures for relaxation calculations. All other settings and details are consistent with the OC20 manuscript [2].

Similarly, adsorption energy calculations are also done consistent with OC20. We note that there is some ambiguity in the catalysis literature for the choice of the gas phase reference, E_{gas} . If the adsorbate is itself a stable gas phase molecule then the adsorption energy might be calculated referenced to itself in the gas phase. However, this quantity is less helpful when calculating thermodynamically consistent free energy diagrams. As used in this work, E_{gas} is often chosen as a linear combination of reference gas phase species [2, 56, 57].

OC20-Dense Placement Details

For a unique adsorbate-surface combination, multiple adsorbate configurations were enumerated as part of the proposed *AdsorbML* pipeline. The OC20-Dense validation and test set were created at different stages of the manuscript, with some notable placement code improvements happening between the two. We highlight those changes here.

The OC20-Dense validation set was created using the code provided at <https://github.com/Open-Catalyst-Project/Open-Catalyst-Dataset/tree/86b5254fe5>. There, the heuristic strategy used CatKit[23] to enumerate all symmetrically distinct sites and provide a suggested adsorbate orientation. The random strategy randomly enumerated M=100 configurations on the surface and placed the adsorbate 2 Å above (in the z direction) the selected site. A random rotation is then applied to the adsorbate along the (0,0,1) adsorption

site vector. The OC20-Dense test set was created slightly different following improvements to code provided at <https://github.com/Open-Catalyst-Project/Open-Catalyst-Dataset/tree/628c5136d0>. For random, sites are defined by first constructing a Delaunay meshgrid with surface atoms as nodes. The positions of the sites are uniformly randomly sampled along the Delaunay triangles. For heuristic, we use the functionality built in Pymatgen[22], which similarly makes a Delaunay meshgrid. We consider sites on the nodes (atop), between 2 nodes (bridge) and in the centers of the triangles (hollow). For both approaches, the adsorbate is uniformly randomly rotated around the z direction, and provided a slight wobble around x and y, which amounts to randomized tilt within a certain cone around the north pole. The adsorbate database includes information about which atoms are expected to bind. The binding atom of the adsorbate is placed at the site. After being placed at the site, the adsorbate is translated along the surface normal until it is no longer overlapping with the surface and the minimum distance between any adsorbate and surface atom is 0.1 Å. Despite the differences, results across all models between the two splits retain the same trends, supporting the use of the validation set for development. The improved heuristic strategy is reflected in the difference in the DFT-Heur baseline - 87.76% and 1.81x speedup vs. 71.12% and 2.87x for the OC20-Dense test and validation set, respectively.

CHANGELOG

This section tracks the changes to this document since the original release.

v1. Initial version.

v2.

- Updated DFT-Heuristic and DFT-Random total compute times, ignoring systems in which were run but ignored from evaluation.
- Updated all speedup numbers as a result of the updated DFT compute times.
- Updated the OC20-Dense dataset statistics, ignoring systems that were removed from evaluation due to problematic inputs.

v3. Published in *npj Comput. Mater.*

- Introduced the OC20-Dense Test set, curated in a similar manner to the previous validation set.
- Evaluated all models on the new OC20-Dense Test set, updating all evaluation metrics across the manuscript's tables and figures.
- Included eSCN-MD-Large, a more recent state-of-the-art OC20 model, to the set of models for OC20-Dense evaluation.