npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials

Janice Lan[1,4], Aini Palizhati[2,4], Muhammed Shuaibi[1,4], Brandon M. Wood[1,4], Brook Wander[2], Abhishek Das[1], Matt Uyttendaele[1], C. Lawrence Zitnick[1]✉ and Zachary W. Ulissi[2,3]✉

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 ~2000× speedup in computation. To standardize benchmarking, we introduce the Open Catalyst Dense dataset containing nearly 1000 diverse surfaces and ~100,000 unique configurations.

npj Computational Materials (2023)9:172 ; https://doi.org/10.1038/s41524-023-01121-5

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., *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 Fig. 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].

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

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 (Fig. 1, bottom right). In addition, 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.

[1]Fundamental AI Research (FAIR), Meta AI, Meta, Menlo Park, CA, USA. [2]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. [3]Scott Institute for Energy Innovation, Carnegie Mellon University, Pittsburgh, PA, USA. [4]These authors contributed equally: Janice Lan, Aini Palizhati, Muhammed Shuaibi, Brandon M. Wood. ✉email: zitnick@meta.com; zulissi@andrew.cmu.edu

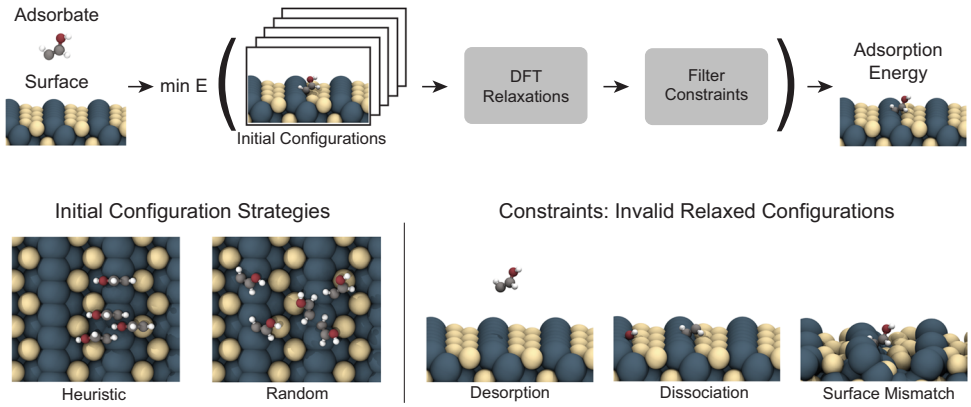

Fig. 1 An overview of the steps involved in identifying the adsorption energy for an adsorbate-surface combination. First, an adsorbate and surface combination are selected, then numerous configurations are enumerated heuristically and/or randomly. For each configuration, DFT relaxations are performed 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.

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 h 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 CO_2 Reduction Reaction (CO_2RR) 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 ~1000 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.

Considerable research effort has been dedicated to determining the lowest energy adsorbate-surface configuration through the 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 threefold:

- 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 ~1000 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 1. The dataset required ~4 million CPU-hrs of compute to complete. A more detailed discussion on OC20-Dense can be found in "Methods".

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/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.

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-Only (DFT-Heur). 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 "Methods".

Table 1. Size of OC20-Dense validation and test splits.

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

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.

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 "Methods".

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 2, 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 2, correlating well with success rates and OC20 *S2EF* metrics.

Table 2. Success rates evaluated using ML-predicted energies.

	OC20-Dense Test			
Model	Success rate* [%] ↑	Energy MAE [eV] ↓	OC20 S2EF MAE ↓	
			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.
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.

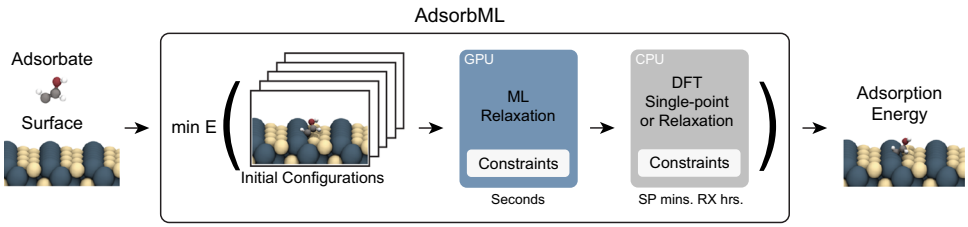

Fig. 2 The *AdsorbML* algorithm. Initial configurations are generated via heuristic and random strategies. ML relaxations are performed on GPUs and ranked in order of lowest to highest energy. The best k systems are passed on to DFT for either a single-point (SP) evaluation or a full relaxation (RX) 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.

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 ~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 (Fig. 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 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 "Methods". 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 1.81×.

ML+SP. The results of using single-point evaluations on ML-relaxed states are summarized in Fig. 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 1400× and 1500× 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 1384× and 1388×, respectively. If speed is of most importance, speedups as high as 6817× 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 2296× and 2299×, respectively. In Fig. 4, 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 1.21 eV/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 0.21 eV/Å and 0.15 eV/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 Fig. 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 relaxations, speedups unsurprisingly decrease. At $k = 5$, we see speedups of 215× and 172× for eSCN-MD-Large and GemNet-OC-MD-Large, respectively. Both SchNet and DimeNet++ see much smaller speedups at 42× and 55×, 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 models. Conversely, eSCN-MD-Large's much larger speedup can be attributed to the near relaxed states (average f_{max} ~0.15 eV/Å) it achieves in its predictions. With $k = 1$, speedups of 1064× 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 of

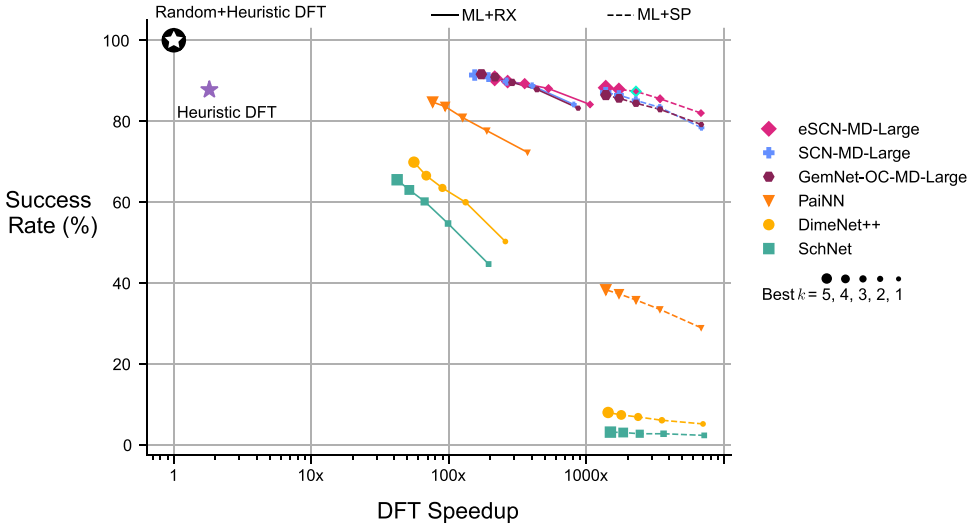

Fig. 3 Overview of the accuracy-efficiency trade-offs of the proposed *AdsorbML* methods across several baseline GNN models. For each model, DFT speedup and corresponding success rate are plotted for ML+RX and ML+SP across various best k. A system is considered successful if the predicted adsorption 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 represents 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.

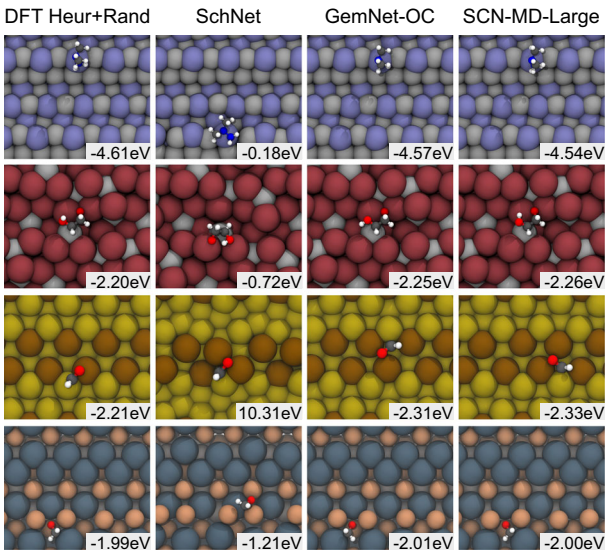

Fig. 4 Illustration of the lowest energy configurations as found by DFT-Heur+Rand, SchNet, GemNet-OC, and SCN-MD-Large on the OC20-Dense validation set. Corresponding adsorption energies are shown in the bottom right corner of each snapshot. ML-relaxed structures have energies calculated with a DFT single-point, 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).

356× and 288×, respectively.

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

In Table 3, 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. In addition, we evaluate success metrics across the different dataset subsplits. OC20-Dense uniformly samples from the four OC20 splits—ID, OOD-Adsorbate, OOD-Catalyst, and OOD-Both. Across our best models, we observe that performance remains consistent across the different distribution splits (Fig. 5). 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.

Table 3. Distribution of success rates for the proposed ML+SP and ML+RX strategies on the OC20-Dense test set.

Success rate						
DFT single-point on ML-relaxed structures (ML+SP)						
	$k=1$			$k=5$		
Model	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)						
	$k=1$			$k=5$		
Model	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%

"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.

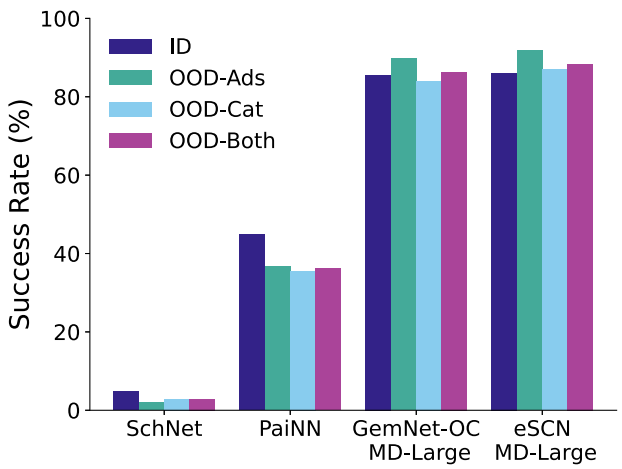

Fig. 5 ML+SP success rate at $k=5$ across the different subsplits of the OC20-Dense test set and several baseline models. Top performing models show marginal differences across the different distribution splits, suggesting good generalization performance to out-of-domain adsorbates and catalysts not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

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 ~3400× 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 and 15% of the time, depending on the efficiency trade-offs (Table 3). 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 2). 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 ~1000 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, ~1000 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 guarantee 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 $\sigma = 0.1$ 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 Equs. (2) and (3).

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

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

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.

Received: 27 February 2023; Accepted: 29 August 2023;
Published online: 22 September 2023

REFERENCES

1. Nørskov, J. K., Studt, F., Abild-Pedersen, F. & Bligaard, T. *Fundamental Concepts in Heterogeneous Catalysis* (John Wiley & Sons, 2014).
2. Chanussot, L. et al. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catal.* **11**, 6059–6072 (2021).
3. Dumesic, J. A., Huber, G. W. & Boudart, M. *Principles of Heterogeneous Catalysis* (Wiley Online Library, 2008).
4. Zitnick, C. L. et al. An introduction to electrocatalyst design using machine learning for renewable energy storage. Preprint at https://arxiv.org/abs/2010.09435 (2020).
5. Choudhary, K. et al. Recent advances and applications of deep learning methods in materials science. *NPJ Comput. Mater.* **8**, 59 (2022).
6. Wen, T., Zhang, L., Wang, H., Weinan, E. & Srolovitz, D. J. Deep potentials for materials science. *Mater. Futures* **1**, 022601 (2022).
7. Wei, J. et al. Machine learning in materials science. *InfoMat* **1**, 338–358 (2019).
8. Ulissi, Z. W., Medford, A. J., Bligaard, T. & Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and dft calculations. *Nat. Commun.* **8**, 1–7 (2017).
9. Tran, K. & Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for co2 reduction and h2 evolution. *Nat. Catal.* **1**, 696–703 (2018).
10. Zhong, M. et al. Accelerated discovery of co2 electrocatalysts using active machine learning. *Nature* **581**, 178–183 (2020).
11. Liu, X. et al. Understanding trends in electrochemical carbon dioxide reduction rates. *Nat. Commun* **8**, 1–7 (2017).
12. Nørskov, J. K. et al. Trends in the exchange current for hydrogen evolution. *J. Electrochem. Soc.* **152**, J23 (2005).
13. Wan, X. et al. Machine-learning-assisted discovery of highly efficient high-entropy alloy catalysts for the oxygen reduction reaction. *Patterns* **3**, 100553 (2022).
14. Seh, Z. W. et al. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science* **355**, eaad4998 (2017).
15. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864 (1964).
16. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical Review* **140**, A1133 (1965).
17. Sholl, D. S. & Steckel, J. A. *Density Functional Theory: A Practical Introduction* (John Wiley & Sons, 2022).
18. Teukolsky, S. A., Flannery, B. P., Press, W. & Vetterling, W. Numerical recipes in C. *SMR* **693**, 59–70 (1992).
19. Peterson, A. A. Global optimization of adsorbate-surface structures while preserving molecular identity. *Top. Catal.* **57**, 40–53 (2014).
20. Goedecker, S. Minima hopping: an efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **120**, 9911–9917 (2004).
21. Jung, H., Sauerland, L., Stocker, S., Reuter, K. & Margraf, J. T. Machine-learning driven global optimization of surface adsorbate geometries. *NPJ Comput. Mater.* **9**, 114 (2023).
22. Ong, S. P. et al. Python materials genomics (pymatgen): a robust, open-source Python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
23. Boes, J. R., Mamun, O., Winther, K. & Bligaard, T. Graph theory approach to high-throughput surface adsorption structure generation. *J. Phys. Chem. A* **123**, 2281–2285 (2019).
24. Andersson, M. P. et al. Toward computational screening in heterogeneous catalysis: Pareto-optimal methanation catalysts. *J. Catal.* **239**, 501–506 (2006).
25. Bligaard, T. et al. The Brønsted-Evans-Polanyi relation and the volcano curve in heterogeneous catalysis. *J. Catal.* **224**, 206–217 (2004).
26. Studt, F. et al. Identification of non-precious metal alloy catalysts for selective hydrogenation of acetylene. *Science* **320**, 1320–1322 (2008).
27. Nilekar, A. U., Sasaki, K., Farberow, C. A., Adzic, R. R. & Mavrikakis, M. Mixed-metal Pt monolayer electrocatalysts with improved CO tolerance. *J. Am. Chem. Soc.* **133**, 18574–18576 (2011).
28. Deshpande, S., Maxson, T. & Greeley, J. Graph theory approach to determine configurations of multidentate and high coverage adsorbates for heterogeneous catalysis. *NPJ Comput. Mater.* **6**, 1–6 (2020).
29. Tran, R. et al. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catal.* **13**, 3066–3084 (2023).
30. Jain, A. et al. Commentary: The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
31. Schütt, K. et al. Schnet: a continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing*

Systems (eds Guyon, I. et al.) 991–1001 (Neural Information Processing Systems Foundation, Inc. (NeurIPS) 2017).
32. Gasteiger, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. in *International Conference on Learning Representations (ICLR)* (ICLR, 2020).
33. Gasteiger, J., Giri, S., Margraf, J. T. & Günnemann, S. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. Preprint at https://arxiv.org/abs/2011.14115 (2020).
34. Gasteiger, J. et al. GemNet-OC: developing graph neural networks for large and diverse molecular simulation datasets. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=u8tvSxm4Bs (2022).
35. Zitnick, C. L. et al. Spherical channels for modeling atomic interactions. *Adv. Neural Inf. Process. Syst.* **35**, 8054–8067 (2022).
36. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
37. Kolluru, A. et al. Open challenges in developing generalizable large-scale machine-learning models for catalyst discovery. *ACS Catal.* **12**, 8572–8581 (2022).
38. Chang, C. & Medford, A. J. 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**, 18210–18216 (2021).
39. Chan, L., Hutchison, G. R. & Morris, G. M. Bayesian optimization for conformer generation. *J. Cheminform.* **11**, 32 (2019).
40. Fang, L., Makkonen, E., Todorović, M., Rinke, P. & Chen, X. Efficient amino acid conformer search with Bayesian optimization. *J. Chem. Theory Comput.* **17**, 1955–1966 (2021).
41. Xu, W., Reuter, K. & Andersen, M. Predicting binding motifs of complex adsorbates using machine learning with a physics-inspired graph representation. *Nat. Comput. Sci.* **2**, 443–450 (2022).
42. Ulissi, Z. W. et al. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for co2 reduction. *ACS Catal.* **7**, 6600–6608 (2017).
43. Ghanekar, P. G., Deshpande, S. & Greeley, J. Adsorbate chemical environment-based machine learning framework for heterogeneous catalysis. *Nat. Commun.* **13**, 1–12 (2022).
44. Schütt, K., Unke, O. & Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. in *International Conference on Machine Learning* 9377–9388 (PMLR, 2021).
45. S. Passaro and C. L. Zitnick, Reducing SO(3) convolu-tions to SO(2) for efficient equivariant GNNs. in *Proceedings of the 40th International Conference on MachineLearning, Proceedings of Machine Learning Research*, Vol. 202 (eds Krause, A. et al.) 27420–27438 (PMLR, 2023).
46. Schaarschmidt, M. et al. Learned force fields are ready for ground state catalyst discovery. Preprint at https://arxiv.org/abs/2209.12466 (2022).
47. Godwin, J. et al. Simple gnn regularisation for 3d molecular property prediction and beyond. in *International Conference on Learning Representations (ICLR)* (ICLR, 2021).
48. Ying, C. et al. Do transformers really perform badly for graph representation? *Adv. Neural Inf. Process. Syst.* **34**, 28877–28888 (2021).
49. Shuaibi, M. et al. Rotation invariant graph neural networks using spin convolutions. Preprint at https://arxiv.org/abs/2106.09575 (2021).
50. Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal-amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251–14269 (1994).
51. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
52. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758 (1999).
53. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comp. Mater. Sci.* **6**, 15–50 (1996).

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.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-023-01121-5.

Correspondence and requests for materials should be addressed to C. Lawrence Zitnick or Zachary W. Ulissi.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.