The Open Catalyst 2020 (OC20) Dataset and Community Challenges

E-mail:

Abstract

Catalyst discovery and optimization is key to solving many societal and energy challenges including solar fuels synthesis, long-term energy storage, and renewable fertilizer production. Despite considerable effort by the catalysis community to apply machine learning models to the computational catalyst discovery process, it remains an open challenge to build models that can generalize across both surface elemental composition and adsorbate identity/configuration. To address this we developed the OC20 dataset, consisting of 1,281,121 Density Functional Theory (DFT) relaxations (265,723,133 single point evaluations) across a wide swath of materials, surfaces, and adsorbates (nitrogen, carbon and oxygen chemistries). We augmented this dataset with randomly perturbed structures, short timescale molecular dynamics, and electronic structure analyses. The dataset is comprised of three central tasks indicative of day-to-day catalyst modeling tasks and comes with pre-defined train/test/validation splits to facilitate direct comparisons with future model development efforts. We applied three state-of-the-art graph neural network models (SchNet, Dimenet, CGCNN) to each of these tasks as baseline demonstrations for the community to build on. For example, models with 10M parameters trained on over 100M single-point calculations were only able to reproduce DFT energies/forces to within 0.02 eV/0.03 eV/Å respectively for 0.03% of validation configurations in the same distribution. Similarly, models to predict adsorption energies directly from initial unrelaxed states were able to achieve 0.65 eV MAE, significantly higher than the same models applied to far less diverse datasets. In almost every task, no upper limit on model size for accuracy was identified suggesting that even larger models are likely to improve on these results. The dataset and baseline models are both provided as open resources, as well as a public leader board to encourage community contributions to solve these important tasks.

Keywords

Catalysis, renewable energy, datasets, machine learning, graph convolutions, force field

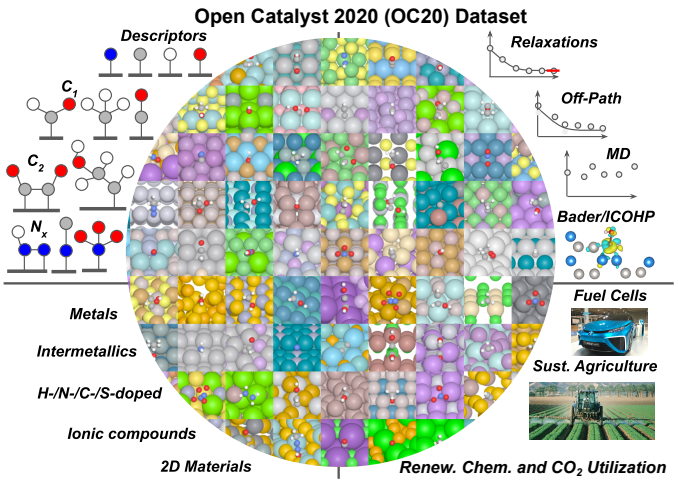

Figure 1: Adsorbates, materials, calculations, impact areas, and illustrative examples of the Open Catalyst (2020) (OC20) dataset. Images are a random sample of the dataset.

Introduction

Advancements to renewable energy processes are urgently needed to address climate change and energy scarcity around the world.[1,2] These include the generation of electricity through fuel cells, fuel generation from renewable resources, and the production of ammonia for fertilization. Catalysis plays a key role in each of these by enabling new reactions and improving process efficiency.[3,4] Unfortunately, discovering or optimizing catalysts remains a time-intensive process. The space of possible catalyst materials that can be synthesized or engineered is vast and modeling their full complexity under reaction conditions remains elusive. Simulation tools such as Density Functional Theory (DFT)[5] have greatly expanded our field's ability to develop reaction mechanisms for specific materials, rationalize experimental measurements, and suggest more active or selective structures for experimental testing. Despite steady growth in computing resources from Moore's law, the computational complexity of DFT remains a limiting factor in the large-scale exploration of new catalysts.[6,7] Given its societal importance, finding computationally efficient methods for molecular simulations is of utmost necessity. One potentially promising approach is the use of efficient ML models trained with data produced from computationally expensive models, such as DFT.

Indeed, the application of Artificial Intelligence and Machine Learning (AI/ML) to molecular simulations has increased in popularity recently, due to its ability to efficiently model complex functions in data-rich domains. There have been a number of demonstrations from domain scientists for specific challenges such as reaction network elucidation,[8–10] thermochemistry prediction,[11–19] structure optimization,[20–24] accelerating individual calculations,[25–28] and integration with characterization[29] (see recent reviews for a more thorough discussion[30–41]). Most of these tasks are variations on the same fundamental problem of predicting the energy and forces of various configurations of organic molecules at inorganic interfaces.

Modeling heterogeneous catalysts, which are used in the renewable energy applications described above, incorporates all the known difficulties of both organic and inorganic chemistry. In organic chemistry, there is an overwhelming space of molecules and reactions and many similar, low-energy conformers. In inorganic chemistry, there is a relatively large diversity in elements, coordination environments, lattice structures, and long-range interactions. The result is a complex space of compositions and chemistries for which computationally efficient modeling methods are needed for thorough exploration.

A critical factor in building ML models is the data used for training. Despite their importance and increased challenge, datasets in heterogeneous catalysis remain far smaller than in other related fields. Much of the progress in applying AI/ML in heterogeneous catalysis has been driven by increasingly large and diverse datasets of electronic structure calculations. In the past few years there has been a push towards larger datasets in catalysis, going from O(100s) to O(100,000)[42] relaxations. Most focus on relaxed adsorption energies of simple adsorbates with smaller datasets of transition state calculations. State-of-the-art ML methods are still improving as data is added to these datasets, so there is no indication that we have saturated the performance of these models. Further, models trained on these datasets have shown limited ability to generalize, which suggests that the models are not yet learning fundamental physical phenomena. As has been shown in other ML tasks,[43–45] we expect that significantly larger datasets will lead to improved accuracy and better generalization.

In this paper, we present the OC20, illustrated in Figure 1, comprised of over 1.2M DFT relaxations of molecular adsorptions onto surfaces (approx. 250 million single-point calculations) across a substantially larger structure and chemistry space than previously realized. While a dataset of this magnitude will lead to significant improvements in ML models, we note that this is still an extremely sparse sampling of all possibilities. We consider 82 different adsorbates (small adsorbates,

C_1/C_2 compounds, and N/O-containing intermediates) which are relevant for renewable energy and environmental applications. Relaxations were performed on randomly sampled low-Miller-index facets of stable materials from the Materials Project,[46] resulting in surfaces from 55 different elements and mixtures thereof. For each of the calculations, we include relaxation trajectories, Bader charges, and LOBSTER[47,48]-calculated orbital information. To aid in training more robust models, we additionally computed short, high-temperature ab-initio molecular dynamics trajectories on a randomly sampled subset of the relaxed states. We also randomly perturbed the atomic positions in a subset of the structures along the relaxation pathways and performed single point DFT calculations for these rattled structures. The dataset is publicly available at `http://anonymous.org` .

In addition to generating and sharing the dataset, we propose three related domain challenges as an open competition: (1) predict the energy and force for a given state, (2) predict a nearby relaxed state given an initial starting state, and (3) predict the relaxed adsorption energy given an initial state. (3) predict the relaxed state energy given an initial state. For the purposes of this manuscript, energy refers to adsorption energy unless otherwise noted. Adsorption energy is defined as the energy of the combined surface and adsorbate system (relaxed or not) minus the energy of the relaxed slab and the relaxed gas phase adsorbate molecule ($E_{ad} = E_{sys} – E_{slab} – E_{gas}$). For a further discussion of reference energies see the SI. The dataset is split into train/validation/test splits indicative of common situations in catalysis: predicting these properties for a previously unseen adsorbate, for a previously unseen crystal structure or composition, or both. To boot-strap research and the competition, we also provide an open software repository (`https://github.com/anonymous`) containing a set of baseline models, data loaders, and training scripts for each of these tasks. We believe that models capable of solving these tasks for the OC20 dataset will also be able to solve a large number of related catalysis problems.

Tasks

Our goal is to improve the efficiency with which inorganic and organic interfaces can be simulated for use in catalysis. Since the primary computational bottleneck are the DFT calculations used to compute a structure's forces and energy, we focus on the general challenge of efficient DFT approximation. We focus on structure relaxation—a fundamental calculation in catalysis used in determining a structure's activity and selectivity. We define three related tasks, in that success in one task may aid other tasks. These are not the only possibilities for this dataset, and future tasks may be added with additional data generation and input from the community.

In all our tasks, the structure contains a surface and adsorbate. The surface is defined by a unit cell that is periodic in the X and Y directions. Initial structures are heuristically determined. Ground truth data is computed for all tasks using DFT. Dataset details and evaluation metrics are provided in following sections.

Structure to Energy and Forces (*S2EF*) is the task where given the atomic positions of the atoms in a structure, the goal is to predict the energy and per-atom forces as calculated by Density Functional Theory (DFT). The force is defined as the negative gradient of the energy with respect to the atomic positions. This is our most general task and has the broadest applicability across catalysis and related fields. It is essentially identical to existing challenges in developing machine learning potentials.[49] However, the inclusion of both inorganic and organic materials and the dataset size, make this challenge unique.

Initial Structure to Relaxed Structure (*IS2RS*) takes as input an initial structure and predicts the atom positions in their final relaxed state. A relaxed structure can be confirmed with a single DFT calculation (all atom forces are close to zero), or used as input to a more detailed set of DFT calculations. Traditional relaxations are performed through an iterative process that estimates the atom forces using DFT, which are in turn used to update atom positions until convergence. This very compu-

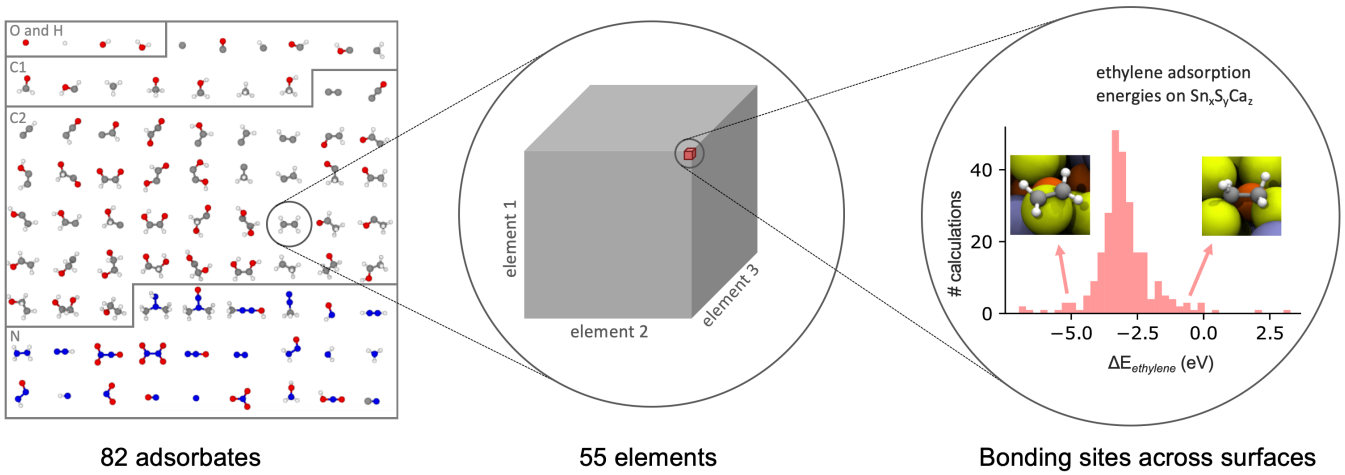

Figure 2: The adsorbates used to generate the Open Catalyst Dataset contain oxygen, hydrogen, C_1, C_2, and nitrogen molecules useful for renewable energy applications. Adsorbates that contain both carbon and nitrogen were counted both as C_X adsorbates and as nitrogen-containing adsorbates. For each adsorbate, up to 55^3 different catalyst compositions were considered, with up to dozens of adsorption energy calculations per adsorbate-composition pairing.

tationally expensive process typically requires hundreds of DFT calculations to converge.

If the *IS2RS* task is approached using ML approximations to DFT to estimate atom forces (*S2EF* task), evaluation on the *IS2RS* task may help determine whether models built for *S2EF* are sufficiently accurate for practical applications. Alternatively, it may be possible to predict the relaxed structure directly, without estimating a structure's energy or forces, as many of the changes during relaxation (say due to particular initial guess strategies) are systematic. These direct IS2RS approaches may lead to even further improvements in computational efficiency. However, such an approach might have trouble to generalizing situations with multiple nearby local minima.

Initial Structure to Relaxed Energy (*IS2RE*) task is given the initial structure as input, and predicts the structure's energy in the relaxed state. This is the most common task in catalysis, as the relaxed energies are often correlated with catalyst activity and selectivity, and the energies are important parameters for detailed microkinetic models. Similar to *IS2RS*, this task may be approached by estimating the relaxed structure and energy by iteratively applying *S2EF*, or by directly regressing the energy from the initial structure without estimating the intermediate or relaxed structures.

The OC20 Dataset

The OC20 dataset is constructed to provide both training and evaluation data for our three previously defined tasks involving DFT approximation and structure relaxation. Modern machine learning models, especially those employing deep learning, require sufficiently large datasets to learn accurate models. For training, we provide 640,118 relaxations across a wide variety of surfaces and adsorbates. The intermediate structures and their corresponding energy and forces are provided for each relaxation resulting in over 133 million structures. To potentially aid in training and to provide additional information for the catalysis community, we performed DFT calculations on rattled and *ab initio* Molecular Dynamics (AIMD) data. We also computed Bader charges and LOBSTER analyses (1.8M+ examples each).

Dataset Generation

The dataset is constructed in four stages: 1) adsorbate selection, 2) surface selection, 3) initial structure generation, and 4) structure relaxation. We describe each of these four stages

in turn, followed by a description of the additional data provided with the main dataset. All source code to generate the configurations are provided in the Open Catalyst Dataset repository (`https://github.com/anonymous`).

Adsorbate Selection

Adsorbates are randomly sampled from a set of 82 molecules that are chosen based on their utility to renewable energy applications. As shown in Figure 2, this includes adsorbates that contain only oxygen or hydrogen, C_1 molecules, C_2 molecules, and nitrogen-containing molecules. We enumerated the oxygen and hydrogen molecules for their ubiquitous presence in water-solvated electrochemical reactions. C_1 and C_2 molecules are important for solar fuel synthesis, while nitrogen-containing molecules have applicability in solar fuel and solar chemical synthesis. Note that some of the C_2 molecules have two binding sites; we refer to these as bidentate adsorbates. The list of all 82 adsorbates is provided in the Supplementary Information.

Surface Selection

Surfaces are sampled in three stages. First, the number of elements is selected with a 5% chance of choosing a unary material, 65% chance for a binary material, and a 30% chance for a ternary material. Greater emphasis is given to binary and ternary materials because these sets contain a wider variety of understudied materials. Next, a stable bulk material is randomly selected from the 11,010 materials in the Materials Project[46] with the number of elements chosen in the first step. Finally, all symmetrically distinct surfaces from the material with Miller indices less than or equal to 2 are enumerated, including possibilities for different absolute positions of surface plane. From this list of surfaces one is randomly selected. The surface atoms were replicated to a depth of at least 7 Å and a width of at least 8 Å.

Pymatgen[50] was used to search over all bulk materials in the Materials Project with non-positive formation energies and energies-above-lower-hulls of at most 0.1 eV/atom. The enumeration of symmetrically distinct surfaces was also performed using pymatgen.[50] Elements for the bulk materials were chosen from a set of 55 elements comprising reactive nonmetals, alkali metals, alkaline earth metals, metalloids, transition metals, and post-transition metals.

Note that DFT was used to re-relax the bulk structures prior to surface enumeration to ensure differences between the DFT settings used in the Materials Project and the Open Catalyst Dataset did not induce unintended stress or strain effects. Any bulks that we could not successfully relaxed were consequently omitted from this dataset.

Initial Structure Generation

The initial structures are generated by placing the selected adsorbates on the selected surfaces using CatKit.[51] Surface atoms are identified by their positions above the center-of-mass, their z-distance of 2 Angstroms below the uppermost atom, and by their under-coordination relative to the bulk atoms. Atomic coordination environments were calculated using pymatgen's Voronoi tesselation algorithm.[50] Next, we manually tagged the adsorbates' binding atoms for both mono- and bi-dentate adsorbates. Finally, we gave the surface structure, adsorbate, the identified surface atoms, and identified adsorbate binding sites to CatKit.[51] CatKit used this information to enumerate symmetrically distinct adsorption sites along with suggested per-site orientations for the adsorbates. Since one of our goals is to calculate adsorption energies, we generate two sets of VASP inputs for each system, (1) the adsorbate placed over the catalyst atoms, and (2) just the catalyst atoms without the adsorbate. This resulted in a total of 2,520,471 and 768,434 unique inputs for (1) and (2) respectively, which were later filtered and segregated into suitable train, test and validation splits.

Structure Relaxation

Structure relaxations were performed using VASP[52–54] until all per-atom forces are less

Table 1: Size of train/validation/test splits (number of structures for *S2EF* and initial structures for *IS2RS* and *IS2RE*). The structures for *S2EF* are sampled from 640,118 relaxations for train, and from 30k-70k relaxations for each validation and test split. Subsplits of validation and test are the same size, but are exclusive of each other. Subsplits include sampling from the same distribution as training (In Domain), unseen adsorbates (Out of Domain (OOD) Adsorbate), unseen element compositions for catalysts (OOD Catalyst), and unseen adsorbates and catalysts (OOD Both).

Task	Train	In Domain	OOD Adsorbate	OOD Catalyst	OOD Both
S2EF	133,953,162	1,000,000	1,000,000	1,000,000	1,000,000
IS2RS	460,364	24,946	24,966	24,963	24,988
IS2RE	460,364	24,946	24,966	24,963	24,988

than 0.03 eV/Å. A timeout was added to limit the total time it took VASP to perform a given relaxation. None of the relaxations in the dataset ran over 144 hours. Systems in which timed out without reaching the specified force threshold were set aside for the S2EF task. All intermediate structures, energies, and forces are stored for future training and evaluation. During the relaxations only the surface atoms (as defined during the generation above) were allowed to move; subsurface atoms were maintained at fixed positions. This was done to avoid unrealistic structure deformations and to simulate a semi-infinite condition with bulk material far below the catalyst surface. Relaxations generally followed previous high-throughput catalysis efforts with reasonable trade-offs between accuracy for surface chemistry and computational cost[15] (VASP, RPBE,[55] no spin polarization, etc). The energy of each adsorbate was referenced to gas-phase CO, H_2O, H_2, and N_2 and the bare slab relaxed energies. Full details of the relaxation procedures are provided in the SI. Resulting trajectories were further analyzed for per-atom force criterion, numerical issues or catastrophic reconstructions as described below in the Train, Validation, and Test Splits section.

AIMD and Rattled Calculations

The intermediate structures from the relaxations may result in a dataset biased towards structures with lower energies. To learn robust models, training samples with higher forces and greater configurational diversity may be needed. We adopted two strategies for generating additional training data (1) partial *ab initio* Molecular Dynamics (AIMD) in VASP[52] and (2) normally-distributed random position perturbation methods colloquially known in molecular simulations as "rattling."

AIMD calculations simulate the atomic interactions when heat is added to the system. Partial AIMD calculations were carried out on previously relaxed structures with random initial velocities generated from a Maxwell-Boltzmann distribution at a temperature of 900 K. We integrated the AIMD trajectories over 80 fs or 320 fs with integration steps of 2 fs in the NVE ensemble. Time-scales were selected to allow systems to explore local configurations while minding computational costs.

To diversify the distribution of single-point structures in the dataset, we "rattled" some of the structures by adding random displacements to the atomic positions. For each relaxation, 20% of the images in the trajectories were selected for rattling. The atomic displacements were sampled from a heuristically-generated normal distribution with a $\mu = 0$ and $\sigma = 0.05$. Single point DFT calculations were then performed on the rattled structures.

Similar to the relaxations, only the top surface atom layers were allowed to move in both the AIMD and rattled calculations with the rest of the atom positions held fixed. All calculations were performed at the same theoretical level and energy/forces convergence criteria as in the relaxation calculations. Approximately 950 thousand AIMD (*ca.* 64 million single-point energies/forces) and 30 million rat-

tled calculations were carried out.

Bader Charges and LOBSTER Analyses

We performed electronic structure calculations for general use by the catalysis research field. These calculations (i.e., Bader charges[47,56,57] and LOBSTER[58,59] analyses) were carried out on relaxed structures and also on randomly selected snapshots from both AIMD and rattled trajectories. Bader charges analyses provides charge density maxima at each atomic center and the Bader volume for each atom through the zero-flux partitioning method.[48] LOBSTER enables chemical-bonding analysis based on periodic DFT outputs.[58] LOBSTER calculates atom-projected densities of states (pDOS) or projected crystal orbital Hamilton population (pCOHP) curves, among others. Literature has demonstrated that such electronic structure information can provide valuable insights to the theoretical and the ML communities.[60–62]

Dataset profile

Approximately 873,000 adsorption energies were calculated successfully. Of these, 5.5% were calculations on unary catalysts; 59.5% were on binaries; and 35.0% were on ternaries. Among these calculations, 32.3% of them had reactive nonmetal elements in the catalyst; 12.8% of them had alkali metals; 11.5% had alkaline earth metals; 27.2% had metalloids; 71.1% had transition metals; and 40.9% had post-transition metals. Considering adsorbates: 5.0% of the calculations had adsorbates containing only oxygen or hydrogen; 24.9% of the calculations had C_1 adsorbates; 46.2% had C_2 adsorbates; and 30.6% had nitrogen-containing adsorbates.

Despite this dataset's large size compared to previous catalytic datasets, it still pales in comparison to the number of potential calculations. Of the $\binom{55}{3} + \binom{55}{2} + \binom{55}{1} = 27,775$ possible compositions, only 5,243 (18.9%) of them were successfully sampled here. Of the compositions sampled, there were an average of 249 successful adsorption calculations for each. Additionally: if we compare the number of sites we sampled here to rough estimates of the number of sites we could have sampled given our constraints on adsorbates, surfaces, and bulks, then we find that we performed *ca.* 0.07% of the possible calculations. This severe sparsity in the data compared to its large scale emphasizes the need for surrogate models.

Train, Validation and Test Splits

We split our dataset into training, validation, and testing sets. The training set is used to learn model parameters; the validation set is used to tune model hyperparameters and to perform ablation studies; and the test set is used to report model performance.

A careful choice of validation and test splits can help evaluate a model's performance on both interpolative and extrapolative tasks. Interpolative evaluation tests the ability to model variations of the training data, and is performed by sampling examples from the same distribution as the training dataset. Extrapolative evaluation tests a model's performance on unseen tasks, e.g., new materials or adsorbates. In the context of catalytic development, we strive to extrapolate beyond data we have already seen so that we can discover new materials and search spaces.[63,64]

We explore extrapolation along two dimensions; new adsorbates and new catalyst compositions. Adsorbate extrapolation is performed by holding out 14 adsorbates from the training dataset sampled from all types (O, H, C1, C2, and N) of adsorbates. Similarly for catalyst compositions, a subset of element combinations for catalysts is held out from the training dataset. These were sampled from the 1,485 binary and 26,235 ternary material combinations of the 55 elements used in the dataset. No surfaces with unary materials are in the catalyst compositions subsplits for training or validation. A full list of the adsorbates materials in train and validation splits are in the SI.

We used four subsplits for each of the validation and test sets by considering all combinations of potential extrapolations (Table 1). These include In-Domain (sampled from the

training distribution), Out-of-Domain Adsorbate (OOD Adsorbate), OOD Catalyst, and OOD Both (both unseen adsorbate and unseen catalyst compositions). As shown in Table 1, each subsplits in validation and testing contains *ca.* 25,000 relaxations. For the *S2EF* task we randomly select a one million structure subset from the relaxations in each subsplit. Note that the extrapolative subsplits of our validation set are completely exclusive to the extrapolative subsplits in the test set, e.g., the adsorbates in the validation adsorbate subsplit are unique from the adsorbates in the test adsorbate subsplit. This helps ensure overfitting to the test set does not occur during hyperparameter tuning on the validation set.

Baseline GNN Models

We evaluate our tasks using a set of baseline models that are representative of the current state-of-the-art. The set of models we evaluate is by no means comprehensive, but they demonstrate what is feasible with current models. Code and pretrained models for our baseline ML approaches implemented in PyTorch Geometric[65,66] are publicly available at the Open Catalyst Project (`http://anonymous.org`).

Our baseline ML approaches are all based on Graph Neural Networks (GNNs)[67] that operate over a graph structure containing nodes and edges. In our domain, the nodes represent atoms and edges represent the relationship between neighboring atoms. At each node, an atom embedding is iteratively updated based on messages passed along the edges. During this message-passing phase, GNNs employ neural networks to learn the atomic representations,[68,69] and unlike traditional descriptor-based models do not require hand-crafting. Node embeddings are initialized based on the atom's properties, such as their atomic number, group number, electronegativity, atomic volume, etc.[70] Outputs for the GNN may be computed from individual node (atom) embeddings for node-specific information (per-atom forces), or over the pooled node embeddings for system outputs (structure energy).

We benchmark three recent GNN methods – Crystal Graph Convolutional Neural Network (CGCNN),[70] SchNet[71] and DimeNet.[72] CGCNN is one of the first approaches to use GNNs on periodic crystal systems and uses a diverse set of features as input to the node embeddings. The original model encoded edge information using the discretized distances between atoms. SchNet proposed using continuous edge filters, which allows for the computation of per-atom forces through partial derivatives of the structure's energy with respect to the atom positions. To allow CGCNN to compute per-atom forces in the same manner, we updated the distance encoding to use gaussian basis functions but without the envelope distance function used in SchNet in our experiments. Finally, to not only encode distance information but also angular information between triplets of atoms, DimeNet introduced the use of directional message passing.

For all approaches, graph edges were determined by a nearest neighbor search limited by a cutoff radius of 6Å, retaining only the 50 nearest neighbors. When computing distances, periodic boundary conditions were taken into consideration. Atoms were tagged as three types, slab (fixed), surface (free), and adsorbate (free), to allow loss functions to emphasize free atoms over fixed atoms. The number of hidden channels is 128, 1024, 256 for CGCNN, SchNet and DimeNet respectively unless stated otherwise; resulting in 3.6 million (CGCNN), 7.4 million (SchNet) and 4.4 million (DimeNet) parameters. Note the size of the models was increased from their original implementations to account for OC20's larger size. Since the number of atom triplets per structure can be very large, directional messages in DimeNet were capped at 30,000. See the SI for additional details.

Since both the computed energies and forces are evaluated, the baseline loss function[26,72] uses the following form:

$$\mathcal{L} = \sum_i |E_i - E_i^{DFT}|$$

$$+\lambda \sum_{i,j} \frac{1}{N_i} |F_{i,j} - F_{i,j}^{DFT}|,$$

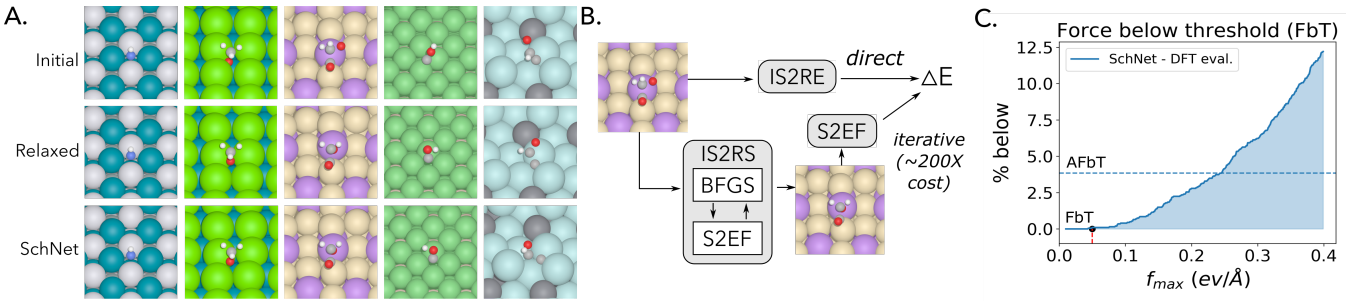

Figure 3: Demonstration of baselines SchNet model for solving the *IS2RE*, *S2EF*, and *IS2RS* tasks and the inter-relationships. (A) Snapshots of five representative initial adsorbate configurations before DFT relaxations, the same adsorbates after DFT relaxation, and the relaxed structures as relaxed by SchNet after fitting the *S2EF* task. (B) Illustration of three ways to predict the relaxed energy: directly through *IS2RE*, indirectly through *IS2RS*, and confirmation of the relaxed structure with a single DFT single-point. (C) SchNet performance as characterized by the percentage of structures within the desired max force threshold of 0.05 eV/Å(FbT) and average percentage of force below threshold (AFbT).

where λ is an empirical parameter, E_i is the energy of image i, and $F_{i,j}$ is the force of the jth free atom in image i, and N_i is the number of free atoms in image i. For the *IS2RE* task, in which only the energy is evaluated, only the first term of the loss function is used ($\lambda = 0$).

All of the models are ML-based as there are currently no physical models that operate over such a large composition space with reasonable accuracy and elemental parameterizations. In particular, the recently developed GFN0-xTB method[73] is parameterized for all of the elements in this dataset and is fast enough (approx 10,000X faster than DFT) to compete on these benchmarks and preliminary results are reported in the SI. However, since the method was not fit for inorganic surfaces and the xTB code[74] is still under active development for periodic boundary conditions, the results were excluded from the summaries here. We hope that the release of our dataset will inspire future efforts on parameterizating tight-binding DFT codes or reactive force field methods for these materials.

Experiments

We begin by describing the metrics used to evaluate our three tasks, followed by the results of our baseline models.

Evaluation metrics

For each task, we define evaluation metrics to track the progress in the field, as well as to measure the practical utility of the approaches. All ground truth values are computed using DFT. Our evaluation metrics are as follows:

***S2EF*:** The *S2EF* task has three metrics: the Mean Absolute Error (MAE) for energy, MAE for forces on free atoms and a combined metric. Our combined metric, Energy and Forces within Threshold (EFwT), is designed to measure the practical usefulness of a model for replacing DFT by evaluating whether both the computed energy and forces are close to the ground truth.

Energy MAE: Mean Absolute Error between the computed energy and the ground truth energy.

Force MAE: Mean Absolute Error between the computed per-atom forces and the ground truth forces. Errors are only computed for free catalyst and adsorbate atoms.

EFwT: The percentage of structures in which the computed energy is within $\epsilon = 0.02$ eV of the ground truth energy, and the maximum error in per-atom forces is below $\alpha = 0.03$ eV/Å. Both these criteria must be met for the structure to be labeled as "correct".

***IS2RS*:** Several methods exist for determining the accuracy of relaxed structures predicted by ML models. The simplest is to measure the distance between the predicted 3D positions of the atoms and those of the ground truth. However, small changes in position can lead to significant changes in the per-atom forces and a structure's energy. For this reason, a better measure of a proposed relaxed structure is the magnitude of its per-atom forces as measured by a single point DFT calculation. If the proposed relaxed structure represents a true local energy minimum, the forces should be close to zero.

ADwT: The Average DwT (Distance within Threshold) across thresholds ranging from $\beta = 0.01$Å to $\beta = 0.5$Å in increments of 0.001Å. DwT is computed as the percentage of structures with a atom position MAE below the threshold. MAE is only computed for free catalyst and adsorbate atom positions while taking into account periodic boundary conditions. We use ADwT as opposed to the MAE on 3D atom positions, since ADwT is robust to outliers and better indicates the percentage of relaxations that are likely to be successful.

FbT: The percentage of relaxed structures with maximum DFT calculated per-atom force magnitudes below a threshold of $\alpha = 0.05$ eV/Å. Force magnitudes of only free catalyst and adsorbate atoms are used. A value of $\alpha = 0.05$ eV/Å represents a practical threshold by which DFT relaxations are commonly assumed to have converged. To ensure that the ML relaxations find a relaxed structure that isn't significantly different from the ground truth relaxed structures, e.g., the adsorbate moves to a different binding site, an additional filtering step is applied. We filter on the atom position MAE (free catalyst and adsorbate atoms) with a threshold of $\beta = 0.5$Å. Thus, to be considered correct, a relaxed structure must meet both the FbT and the DwT criterion.

AFbT: The Average FbT (AFbT) over a range of thresholds ranging from $\alpha = 0.01$ eV/Å to $\alpha = 0.4$ eV/Å in increments of 0.001 eV/Å, Figure 3(C). This metric measures progress over a wider range of thresholds, which may be important for early algorithm development that may need thresholds more lenient than $\alpha = 0.05$ eV/Å to see improvement. Similar to FbT, the relaxed structures must also meet the same DwT criterion with $\beta = 0.5$Å.

Note that FbT and AFbT require the computation of single point DFT calculations, which are computationally expensive. For this reason, a random subset of 500 relaxed structures are chosen from the validation and test set splits (2000 total for each) for evaluating these metrics. If a DFT calculation fails to converge within 60 electronic steps or a wall time of 2 hrs, the system is assumed to be incorrect with forces beyond the thresholds for both FbT and AFbT.

***IS2RE*:** Similar to the *S2EF* task we propose two metrics for *IS2RE*. The first measures the MAE on the computed and ground truth energy. The second measures the ener-

Table 2: Predicting energy and forces from a structure (*S2EF*) as evaluated by Mean Absolute Error (MAE) of the energies, forces MAE, and the percentage of Energies and Forces within Threshold (EFwT). Results reported for models training on the entire training dataset.

	S2EF Validation			
Model	ID	OOD Ads	OOD Cat	OOD Both
	Energy MAE ↓			
Mean baseline	0.0000	0.0000	0.0000	0.0000
CGCNN[70]	0.5861	0.7088	0.6229	0.7680
SchNet[71]	0.5397	0.5845	0.6082	0.7942
DimeNet[72]	0.5880	0.7189	0.6629	0.8010
	Forces MAE ↓			
Mean baseline	0.0000	0.0000	0.0000	0.0000
CGCNN[70]	0.0656	0.0642	0.0629	0.0785
SchNet[71]	0.0438	0.0511	0.0458	0.0619
DimeNet[72]	0.0600	0.0579	0.0575	0.0702
	EFwT ↑			
Mean baseline	0.00%	0.00%	0.00%	0.00%
CGCNN[70]	0.01%	0.00%	0.00%	0.00%
SchNet[71]	0.03%	0.00%	0.06%	0.00%
DimeNet[72]	0.01%	0.00%	0.00%	0.00%

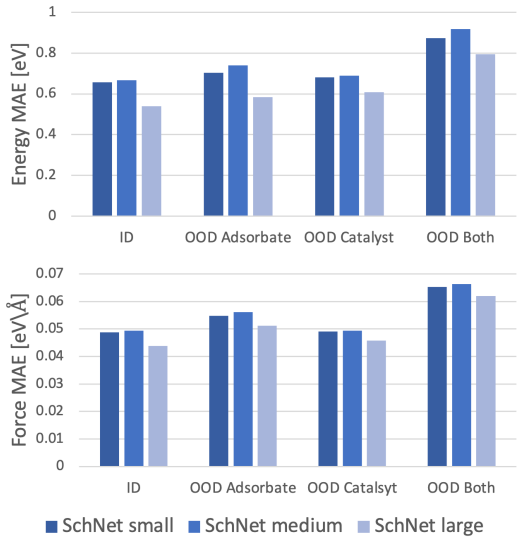

Figure 4: Predicting Structure to Energy and Forces (*S2EF*) as evaluated by Mean Absolute Error (MAE) of the energies and forces. The small, medium and large SchNet models have the following sizes: Small: 256 hidden, 4 message-passing layers, 1,316,097 params, Medium: 1024 hidden, 3 message-passing layers, 5,704,193 params, Large: 1024 hidden, 4 message-passing layers, 7,396,353 params. Results reported for models trained on the entire training dataset.

gies within a threshold (EwT) of the ground truth, which once again measures the percentage of estimated energies that are likely to be practically useful.

Energy MAE: Mean Absolute Error between the computed relaxed energy and the ground truth relaxed energy.

EwT: The percentage of computed relaxed energies within $\epsilon = 0.02$ eV of the ground truth relaxed energy.

While our evaluation metrics focus on accuracy, it is important to note that methods should also be significantly faster than conventional DFT. As a rough benchmark, we desire energy and force estimates at approximately 1 ms/atom which would significantly improve the applicability of DFT. Significantly faster than this (closer in speed to classical force fields) would open up even more interesting applications. We ask that users self-report timing results, but we are not going to make that a core part of the challenge as computation time can likely be further optimized for the best models and with hardware acceleration.

Table 3: Predicting relaxed structure from initial structure (*IS2RS*) as evaluated by Average Distance within Threshold (ADwT), Forces below Threshold (FbT), and Average Forces below Threshold (AFbT). All values in percentages, higher is better. Results reported for structure to force models trained on the All training dataset.

	IS2RS Validation			
Model	ID	OOD Ads	OOD Cat	OOD Both
		ADwT ↑		
SchNet[71]	35.68%	34.11%	29.42%	33.80%
DimeNet[72]	10.94%	9.80%	8.36%	10.21%
		FbT ↑		
SchNet[71]	0.402%	0.00%	0.00%	0.00%
DimeNet[72]	0.00%	0.00%	0.00%	0.00%
		AFbT ↑		
SchNet[71]	6.17%	3.38%	3.40%	2.44%
DimeNet[72]	0.00 %	0.00%	0.00%	0.00%

Leaderboard

To ensure consistent and fair evaluation, a public leaderboard is available on the Open Catalyst Project webpage (`http://anonymous.org`). Results on any of the tasks' test datasets may be uploaded for evaluation. Ground truth test data is not publicly released to reduce potential overfitting. Evaluation on the test set may only be done through the leaderboard. Ablation studies and hyper-parameter tuning may be done and reported on using the validation datasets.

Results

To provide baselines for the OC20 dataset, we report results using three state-of-the-art approaches: CGCNN,[70] SchNet,[71] and DimeNet.[72] Details of the models' implementations can be found in the Baselines Section.

***S2EF*:** Results on CGCNN,[70] SchNet,[71] and DimeNet[72] are evaluated. All approaches pre-

Table 4: Predicting relaxed state energy from initial structure (*IS2RE*) as evaluated by Mean Absolute Error (MAE) of the energies and the percentage of Energies within a Threshold (EwT) of the ground truth energy. Results reported for trained on the All training dataset.

		Energy MAE ↓				EwT ↑			
		Validation							
Model	Approach	ID	OOD Ads	OOD Cat	OOD Both	ID	OOD Ads	OOD Cat	OOD Both
Mean baseline	-	1.7668	1.7382	1.7469	1.5259	0.0000	0.0000	0.0000	0.0000
CGCNN[70]	Direct	0.6477	0.7012	0.6309	0.6688	2.95%	2.46%	3.42%	2.41%
SchNet[71]	Direct	0.6462	0.7063	0.6433	0.6630	3.00%	2.35%	3.08%	2.40%
DimeNet[72]	Direct	0.6469	0.7358	0.6402	0.6852	3.26%	2.24%	3.45%	2.32%
SchNet[71]	Relaxation	0.8796	0.9347	1.1595	0.9256	3.40%	2.82%	2.75%	2.36%
DimeNet[72]	Relaxation	0.8449	0.8909	1.0076	0.8281	2.60%	3.22%	1.96%	2.76%

dict structure energies in their forward pass and per-atom forces by the negative gradient of the predicted energy with respect to atomic positions.[75] Across all metrics SchNet performs the best, with DimeNet outperforming CGCNN on force MAE and CGCNN performing better on energy MAE. All approaches perform badly on the EFwT metric; indicating that the results are still far from being practically useful. Table 2 and Figure 4 show results across subsplits. As expected, the In Domain (ID) achieves the best results and the OOD Both performs the worst. However, results are not dramatically different, which shows some generalization to new adsorbates and catalysts. Increases in training data sizes results in significant improvements up until 20M, Figure 5(A). This may be due to the additional data beyond 20M being mostly redundant, i.e., it is adding nearby examples from the relaxation trajectories to those already in 20M. Finally, wider and deeper models are shown to improve accuracies in Figure 4. Increased depth (SchNet small) appears to give slight improvement over additional width (SchNet medium).

***IS2RS*:** For *IS2RS*, we use our *S2EF* baselines to drive ML relaxations from the given initial structures to estimate the relaxed structures using L-BGFS,[76] examples are shown in Figure 3(A). Table 3 shows that SchNet outperforms DimeNet in both the ADwT and AFbT metrics. However, the FbT metric indicates both methods do not produce relaxed structures with forces below thresholds used in practice. A plot of FbT across thresholds from 0.01 to 4.0 for SchNet is shown in Figure 3(C). Both methods show better generalization to new adsorbates vs new catalyst material compositions. Sin=milar to *S2EF* improved results are found with more training data up until 20M, Figure 5(B).

***IS2RE*:** For *IS2RE* we explore two pathways for computing the relaxed energy from the initial state, Figure 3(B). The first directly computes the relaxed energy given the initial state. The same model architectures are used as the *S2EF* task, but with new weights learned. The second approach uses models trained on the *S2EF* task to perform ML relaxations from which the resulting energy is returned. Note that the ML relaxation approach is about 200 times more expensive to compute, since energies needs to be computed at each relaxation step. For the relaxation-based approaches, if the energy prediction was above 10eV, which represents the cutoff above which relaxations are removed from the dataset, the energy prediction was set to the median of the training dataset to reduce outliers skewing MAE results. This occurred in ∼2% of the predicted structures.

As shown in Table 4, the direct approaches out performed those using relaxation across most metrics. The percentage of predicted energies within a tight threshold ranged from 2% to 3.5%; indicating that accuracies are still be-

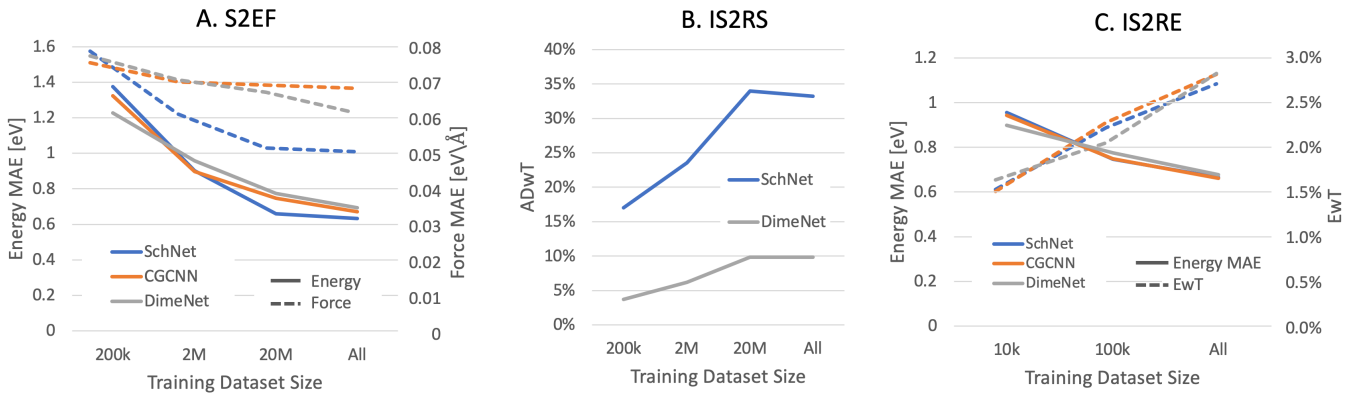

Figure 5: (A) Predicting energy and forces from a structure (*S2EF*) as evaluated by Mean Absolute Error (MAE) of the energies and forces. (B) Predicting relaxed structure from initial structure (*IS2RS*) as evaluated by Average Distance within Threshold (ADwT). (C) Predicting relaxed state energy from initial structure (*IS2RE*) as evaluated by Mean Absolute Error (MAE) of the energies and the percentage of Energies within a Threshold (EwT, $\epsilon = 0.02$ eV) of the ground truth energy. Results reported for *S2EF* and *IS2RS* trained on 200k, 2M, 20M and All dataset sizes. Results reported for *IS2RE* trained on 10k, 100k, All dataset sizes. All values averaged across validation subsplits.

low practical usefulness. Generalization to new adsorbates performed better than new catalyst materials. As shown in Figure 5(C), larger dataset sizes could significantly improve performance.

Outlook and future directions

The baseline models in this work give significant insights into the complexity of day-to-day challenges in catalysis and what it will take to achieve generalizable models. To illustrate the difference between catalysis and related efforts in small molecule property prediction and materials science and motivated by previous efforts,[78] we analyzed model performance for increasing dataset sizes. Figure 6(left) and Figure 2(middle) show the performance of GNN models similar to the baseline models in this work on datasets for small molecules (QM9) and materials (formation energies from the materials project). The scaling of model accuracy with respect to dataset size is related to the effective dimensionality of the task and the effective representation in the model. Comparing SchNet performance across all three tasks shows that the aggressive scaling for small molecules is reduced for inorganic materials, and further reduced for surfaces. Focusing on results from this study in Figure 6(right) shows that the scaling is similar for the same baseline models trained on the OC20 dataset and a related literature dataset of CO adsorption energies (see the SI). Importantly, the scaling suggests that achieving the desired accuracy using the current baseline models would require a dataset nearly 10 orders of magnitude larger than the current dataset. This implies that this problem will not be solved through brute-force methods alone, and that significantly improved ML representations are also necessary. This is an exciting opportunity for the broader community.

For the computer science and ML communities, we expect that this dataset will provide unique challenges and spur innovation in atomistic simulations. Many state-of-the-art methods for organic and inorganic materials are based on graph convolutional networks,[79] which have seen rapid progress. With the above perspective, we expect that additional creative solutions will be necessary to fully solve these tasks. While they have not been demonstrated for inorganic materials, physics-informed tensor representations for small molecules may be helpful.[80–83] Element embeddings and representations will be important to scale across ma-

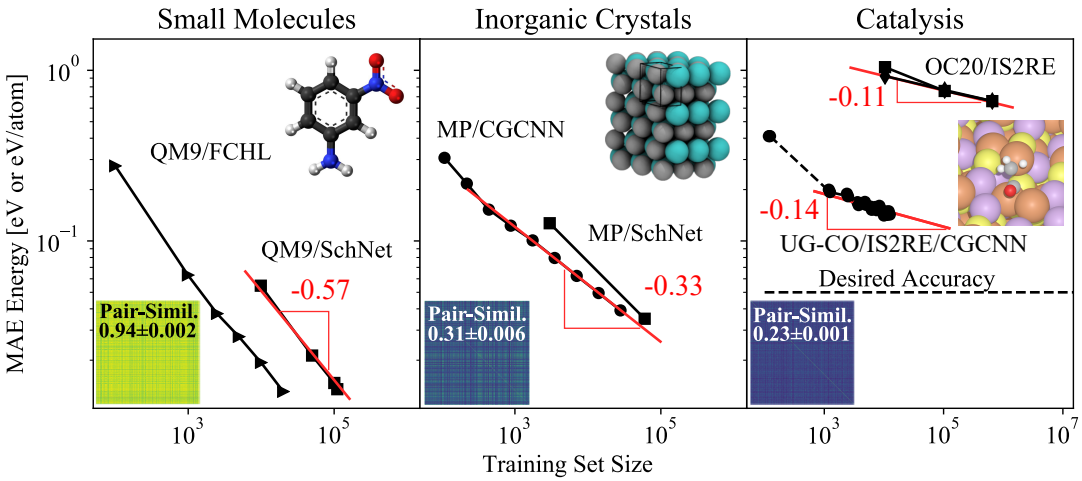

Figure 6: Model performance versus dataset size across three related atomistic domains. Insets are pairwise similarity for selected structures from the respective dataset using GraphDot (see the SI for details) (0/dark-blue/not-similar to 1/yellow/identical).[77] (left) Results[49] for FCHL/SchNet models trained on the QM9 small molecule dataset (slope -0.57). (middle) Models trained on Materials Project formation energies (slope -0.33, more difficult). (right) Results for catalysis including a literature dataset for CO adsorbates[15] and this work (slope -0.11 to -0.14, most difficult). Note that reaching the desired accuracy will require several orders of magnitude more data with current models.

terials. Incorporation of lower-level physics-based potentials is welcomed and encouraged. This includes the use of related datasets (organic molecules or inorganic materials) for pre-training or learning priors. Incorporating other electronic features in the training set, such as charge distribution to correctly localize effects is also an opportunity to effectively reduce the dimensionality of the problem.

Note that size of this dataset is larger by 2 orders of magnitude than previous catalyst DFT dataset efforts.[15,84] Along with the potential for more accurate ML models, it provides practical challenges to training atomistic machine learning models at scale, similar to software engineering challenges in image recognition and NLP.[85,86] The largest baseline models with 10M parameters were trained on upwards of 32 GPUs at a time, so we encourage the catalysis community to take advantage of these GPU-enabled resources. This is well-timed with the wave of large GPU-enabled supercomputers that are well-suited to these challenges, such as Perlmutter (DOE NERSC) or Summit (DOE OLCF), among many others.

The baseline models in this work represent the state-of-the-art for deep learning methods to predict thermochemistry for small molecules on inorganic surfaces. Solving this challenge with future model development efforts would enable a new generation of computational chemistry methods. In particular, on-the-fly thermochemistry for reaction intermediates would enable reaction mechanism prediction across materials or composition space. Accelerated methods would also enable the more routine use of more accurate computational methods (e.g. hybrid, exact-exchange, or RPA calculations) by focusing these efforts on the most promising and pre-relaxed structures. A solution to the S2EF task would also enable the more routine use of long timescale molecular dynamics for studying these systems. The potential applicability of the OC20 dataset is not just catalysis, but also has implications for areas where organic and inorganic materials interact, such as water quality remediation, geochemistry, advanced manufacturing, and durable energy materials.

Supporting Information Available

The supporting information contains details on the precise DFT calculation methods, the adsorbates and their assuming binding configurations, details on the graph similarity metrics, details on GFN0-XTB relaxations on a small number of sample systems, and comparison of baseline models on previously published literature dataset. The full open dataset is provided at `http://anonymous.org`, and the baseline models are provided as an open source repository at `https://github.com/anonymous`.

References

(1) Newell, R. G.; Raimi, D.; Villanueva, S.; Prest, B. *Global Energy Outlook 2020: Energy Transition or Energy Addition? With Commentary on Implications of the COVID-19 Pandemic*; 2020.

(2) *The Annual Energy Outlook explores long-term energy trends in the United States*.

(3) Nørskov, J. K.; Studt, F.; Abild-Pedersen, F.; Bligaard, T. Fundamental ConCepts in Heterogeneous Catalysis.

(4) She, Z. W.; Kibsgaard, J.; Dickens, C. F.; Chorkendorff, I.; Nørskov, J. K.; Jaramillo, T. F. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science* **2017**, *355*.

(5) Sholl, D. S.; Steckel, J. A. *Density Functional Theory*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2009.

(6) Matera, S.; Schneider, W. F.; Heyden, A.; Savara, A. Progress in Accurate Chemical Kinetic Modeling, Simulations, and Parameter Estimation for Heterogeneous Catalysis. *ACS Catalysis* **2019**, *9*, 6624–6647.

(7) Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting Knowledge from Data through Catalysis Informatics. *ACS Catalysis* **2018**, *8*, 7403–7429.

(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. *Nature Communications* **2017**, *8*, 1–7.

(9) Li, B.; Rangarajan, S. Designing compact training sets for data-driven molecular property prediction through optimal exploitation and exploration. *Molecular Systems Design & Engineering* **2019**, *4*, 1048–1057.

(10) Gu, G. H.; Plechac, P.; Vlachos, D. G. Thermochemistry of gas-phase and surface species via LASSO-assisted subgraph selection. *Reaction Chemistry & Engineering* **2018**, *3*, 454–466.

(11) Liu, Y.; Hong, W.; Cao, B. Machine learning for predicting thermodynamic properties of pure fluids and their mixtures. *Energy* **2019**, *188*, 116091.

(12) Jirasek, F.; Alves, R. A.; Damay, J.; Vandermeulen, R. A.; Bamler, R.; Bortz, M.; Mandt, S.; Kloft, M.; Hasse, H. Machine Learning in Thermodynamics: Prediction of Activity Coefficients by Matrix Completion. *Journal of Physical Chemistry Letters* **2020**, *11*, 981–985.

(13) Kauwe, S. K.; Graser, J.; Antonio Vazquez, ·.; Sparks, T. D. Integrating Materials and Manufacturing Innovation Machine Learning Prediction of Heat Capacity for Solid Inorganics. **2018**,

(14) Tran, K.; Neiswanger, W.; Yoon, J.; Zhang, Q.; Xing, E.; Ulissi, Z. W. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology* **2020**, *1*, 025006.

(15) Tran, K.; Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2

evolution. *Nature Catalysis* **2018**, *1*, 696–703.

(16) Gu, G. H.; Noh, J.; Kim, S.; Back, S.; Ulissi, Z.; Jung, Y. Practical Deep-Learning Representation for Fast Heterogeneous Catalyst Screening. *The Journal of Physical Chemistry Letters* **2020**, *11*, 3185–3191.

(17) Back, S.; Yoon, J.; Tian, N.; Zhong, W.; Tran, K.; Ulissi, Z. W. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *The Journal of Physical Chemistry Letters* **2019**, *10*, 4401–4408.

(18) Kim, M.; Yeo, B. C.; Park, Y.; Lee, H. M.; Han, S. S.; Kim, D. Artificial Intelligence to Accelerate the Discovery of N2 Electroreduction Catalysts. *Chemistry of Materials* **2019**, *32*, 709–720.

(19) Esterhuizen, J. A.; Goldsmith, B. R.; Linic, S. Theory-Guided Machine Learning Finds Geometric Structure-Property Relationships for Chemisorption on Subsurface Alloys. *Chem* **2020**,

(20) Sun, G.; Sautet, P. Metastable Structures in Cluster Catalysis from First-Principles: Structural Ensemble in Reaction Conditions and Metastability Triggered Reactivity. *Journal of the American Chemical Society* **2018**, *140*, 2812–2820.

(21) Li, B.; Huang, C.; Li, X.; Zheng, S.; Hong, J. Non-iterative structural topology optimization using deep learning. *CAD Computer Aided Design* **2019**, *115*, 172–180.

(22) Hayashi, K.; Ohsaki, M. Reinforcement Learning and Graph Embedding for Binary Truss Topology Optimization Under Stress and Displacement Constraints. *Frontiers in Built Environment* **2020**, *6*, 59.

(23) Schmidt, J.; Marques, M. R.; Botti, S.; Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. 2019; `https://doi.org/10.1038/s41524-019-0221-0`.

(24) Aksoz, Z.; Preisinger, C. *Impact: Design With All Senses*; Springer International Publishing, 2020; pp 18–31.

(25) Boes, J. R.; Kitchin, J. R. Neural network predictions of oxygen interactions on a dynamic Pd surface. *Molecular Simulation* **2017**, *43*, 346–354.

(26) Khorshidi, A.; Peterson, A. A. Amp: A modular approach to machine learning in atomistic simulations. *Computer Physics Communications* **2016**, *207*, 310–324.

(27) Peterson, A. A. Acceleration of saddle-point searches with machine learning. *The Journal of Chemical Physics* **2016**, *145*, 074106.

(28) Sun, G.; Sautet, P. Toward fast and reliable potential energy surfaces for metallic Pt clusters by hierarchical delta neural networks. *Journal of Chemical Theory and Computation* **2019**, *15*, 5614–5627.

(29) Timoshenko, J.; Frenkel, A. I. "Inverting" X-ray absorption spectra of catalysts by machine learning in search for activity descriptors. *ACS Catalysis* **2019**, *9*, 10192–10211.

(30) Kitchin, J. R. Machine learning in catalysis. *Nature Catalysis* **2018**, *1*, 230–232.

(31) Li, Z.; Wang, S.; Xin, H. Toward artificial intelligence in catalysis. *Nature Catalysis* **2018**, *1*, 641–642.

(32) Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting knowledge from data through catalysis informatics. *ACS Catalysis* **2018**, *8*, 7403–7429.

(33) Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C. Machine

learning for heterogeneous catalyst design and discovery. *AIChE Journal* **2018**, *64*, 2311–2323.

(34) Schlexer Lamoureux, P.; Winther, K. T.; Garrido Torres, J. A.; Streibel, V.; Zhao, M.; Bajdich, M.; Abild-Pedersen, F.; Bligaard, T. Machine learning for computational heterogeneous catalysis. *ChemCatChem* **2019**, *11*, 3581–3601.

(35) Li, H.; Zhang, Z.; Liu, Z. Application of artificial neural networks for catalysis: a review. *Catalysts* **2017**, *7*, 306.

(36) Tabor, D. P.; Roch, L. M.; Saikin, S. K.; Kreisbeck, C.; Sheberla, D.; Montoya, J. H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H., et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature Reviews Materials* **2018**, *3*, 5–20.

(37) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A Critical Review of Machine Learning of Energy Materials. *Advanced Energy Materials* **2020**, *10*, 1903242.

(38) Artrith, N. Machine learning for the modeling of interfaces in energy storage and conversion materials. *Journal of Physics: Energy* **2019**, *1*, 032002.

(39) Gu, G. H.; Noh, J.; Kim, I.; Jung, Y. Machine learning for renewable energy materials. *Journal of Materials Chemistry A* **2019**, *7*, 17096–17117.

(40) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K.-I. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catalysis* **2019**,

(41) Gu, G. H.; Choi, C.; Lee, Y.; Situmorang, A. B.; Noh, J.; Kim, Y.-H.; Jung, Y. Progress in Computational and Machine-Learning Methods for Heterogeneous Small-Molecule Activation. *Advanced Materials n/a*, 1907865.

(42) Mamun, O.; Winther, K. T.; Boes, J. R.; Bligaard, T. High-throughput calculations of catalytic properties of bimetallic alloy surfaces. *Scientific data* **2019**, *6*, 76.

(43) Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition. 2009; pp 248–255.

(44) Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: an asr corpus based on public domain audio books. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015; pp 5206–5210.

(45) Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; Parikh, D. Vqa: Visual question answering. Proceedings of the IEEE international conference on computer vision. 2015; pp 2425–2433.

(46) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. a. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013**, *1*, 011002.

(47) Henkelman, G.; Arnaldsson, A.; Jónsson, H. A fast and robust algorithm for Bader decomposition of charge density. *Computational Materials Science* **2006**, *36*, 354–360.

(48) Bader, R. Atoms in molecules: A quantum theory.

(49) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry* **2020**, 1–12.

(50) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.;

Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **2013**, *68*, 314–319.

(51) Boes, J. R.; Mamun, O.; Winther, K.; Bligaard, T. Graph theory approach to high-throughput surface adsorption structure generation. *The Journal of Physical Chemistry A* **2019**, *123*, 2281–2285.

(52) Kresse, G.; Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Physical Review B* **1994**, *49*, 14251–14269.

(53) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science* **1996**, *6*, 15–50.

(54) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **1996**, *54*, 11169–11186.

(55) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **1996**, *77*, 3865.

(56) Tang, W.; Sanville, E.; Henkelman, G. A grid-based Bader analysis algorithm without lattice bias. *Journal of Physics: Condensed Matter* **2009**, *21*, 084204.

(57) Sanville, E.; Kenny, S. D.; Smith, R.; Henkelman, G. Improved grid-based algorithm for Bader charge allocation. *Journal of Computational Chemistry* **2007**, *28*, 899–908.

(58) Nelson, R.; Ertural, C.; George, J.; Deringer, V. L.; Hautier, G.; Dronskowski, R. LOBSTER: Local orbital projections, atomic charges, and chemical-bonding analysis from projector-augmented-wave-based density-functional theory. *Journal of Computational Chemistry* **2020**,

(59) Deringer, V. L.; Tchougréeff, A. L.; Dronskowski, R. Crystal orbital Hamilton population (COHP) analysis as projected from plane-wave basis sets. *The Journal of Physical Chemistry A* **2011**, *115*, 5461–5466.

(60) Gong, S.; Xie, T.; Zhu, T.; Wang, S.; Fadel, E. R.; Li, Y.; Grossman, J. C. Predicting charge density distribution of materials using a local-environment-based graph convolutional network. *Physical Review B* **2019**, *100*, 184103.

(61) Nagai, R.; Akashi, R.; Sugino, O. Completing density functional theory by machine learning hidden messages from molecules. *npj Computational Materials* **2020**, *6*, 1–8.

(62) Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the electronic structure problem with machine learning. *npj Computational Materials* **2019**, *5*, 1–7.

(63) Kim, Y.; Kim, E.; Antono, E.; Meredig, B.; Ling, J. Machine-learned metrics for predicting thelikelihood of success in materials discovery. **2019**, 1–13.

(64) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Molecular Systems Design and Engineering* **2018**, *3*, 819–825.

(65) Fey, M.; Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428* **2019**, .

(66) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019; pp 8026–8037.

(67) Hamilton, W. L.; Ying, R.; Leskovec, J. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584* **2017**, .

(68) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *The Journal of Chemical Physics* **2016**, *145*, 170901.

(69) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters* **2010**, *104*, 136403.

(70) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters* **2018**, *120*, 145301.

(71) Schütt, K.; Kindermans, P.-J.; Felix, H. E. S.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. Advances in neural information processing systems. 2017; pp 991–1001.

(72) Klicpera, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. International Conference on Learning Representations (ICLR). 2020.

(73) Pracht, P.; Caldeweyher, E.; Ehlert, S.; Grimme, S. A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for large Molecules. **2019**, .

(74) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2020**, e01493.

(75) Pukrittayakamee, A.; Malshe, M.; Hagan, M.; Raff, L.; Narulkar, R.; Bukkapatnum, S.; Komanduri, R. Simultaneous fitting of a potential-energy surface and its corresponding force fields using feedforward neural networks. *The Journal of chemical physics* **2009**, *130*, 134101.

(76) Liu, D. C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical programming* **1989**, *45*, 503–528.

(77) Tang, Y.-H.; Selvitopi, R.; Popovici, D. A. T., et al. *GraphDot v0. 1*; 2019.

(78) Huang, B.; Symonds, N. O.; von Lilienfeld, O. A. The fundamentals of quantum machine learning. *arXiv preprint arXiv:1807.04259* **2018**, .

(79) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters* **2018**, *120*.

(80) Miller, B. K.; Geiger, M.; Smidt, T. E.; Noé, F. Relevance of Rotationally Equivariant Convolutions for Predicting Molecular Properties. *arXiv preprint arXiv:2008.08461* **2020**, .

(81) Bratholm, L. A.; Gerrard, W.; Anderson, B.; Bai, S.; Choi, S.; Dang, L.; Hanchar, P.; Howard, A.; Huard, G.; Kim, S., et al. A community-powered search of machine learning strategy space to find NMR property prediction models. *arXiv preprint arXiv:2008.05994* **2020**, .

(82) Anderson, B.; Hy, T. S.; Kondor, R. Cormorant: Covariant molecular neural networks. Advances in Neural Information Processing Systems. 2019; pp 14537–14546.

(83) Nigam, J.; Pozdnyakov, S.; Ceriotti, M. Recursive evaluation and iterative contraction of N-body equivariant features. *arXiv preprint arXiv:2007.03407* **2020**, .

(84) Hummelshøj, J. S.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nørskov, J. K. CatApp: A web application for surface chemistry and heterogeneous catalysis. *Angewandte Chemie - International Edition* **2012**, *51*, 272–274.

(85) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 770–778.

(86) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.

(87) Blöchl, P. E. Projector augmented-wave method. *Physical Review B* **1994**, *50*, 17953.

(88) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B* **1999**, *59*, 1758.

(89) Hammer, B.; Hansen, L. B.; Nørskov, J. K. Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals. *Physical Review B* **1999**, *59*, 7413.

(90) Monkhorst, H. J.; Pack, J. D. Special points for Brillouin-zone integrations. *Physical Review B* **1976**, *13*, 5188.

(91) Teter, M. P.; Payne, M. C.; Allan, D. C. Solution of Schrödinger's equation for large systems. *Physical Review B* **1989**, *40*, 12255.

(92) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. Numerical Recipes, Cambridge: Cambridge Univ. 1986.

(93) Feynman, R. P. Forces in molecules. *Physical Review* **1939**, *56*, 340.

(94) Tang, Y. H.; De Jong, W. A. Prediction of atomization energy using graph kernel and active learning. *Journal of Chemical Physics* **2019**, *150*.

Supplementary Material

Details of the DFT relaxations

DFT calculations were performed with the *Vienna Ab Initio Simulation Package* (VASP)[52–54] with periodic boundary conditions and the projector augmented wave (PAW) pseudopotentials.[87,88] The external electrons were expanded in plane waves with kinetic energy cut-offs of 350 eV. Exchange and correlation effects were taken into account via the generalized gradient approximation[55] and the revised Perdew-Burke-Ernzerhof (RPBE) functional, because of its improved description of the energetics of atomic and molecular bonding to surfaces.[89] Bulk and surface calculations were performed considering a K-point mesh for the Brillouin zone derived from the unit cell parameters as an on-the-spot method, employing the Monkhorst-Pack grid.[90] The ionic degrees of freedom were relaxed using a Conjugate Gradient minimization.[91,92] The relaxation was terminated when either the Hellmann-Feynman forces[93] were less than 0.03 eV/Å or the relaxation required more than 200 steps in a single uninterrupted VASP call. This limit was reset each time the calculation was checkpointed allowing some relaxations to exceed this 200 steps. The final distribution of residual forces is shown in Figure 7 in the SI. Relaxations still converging after approximately 5,000 core-hours were terminated and not included in the dataset. For the electronic degrees of freedom, the energy convergence criteria was fixed to 10^{-4} eV, where no spin magnetism or dispersion corrections were included.

Adsorbates included

The full list of adsorbates is indicated in 5. This list was constructed by considering the four monatomic species and adding common intermediates for renewable energy challenges. The number of possible organic molecules is combinatorially large, so this is not a comprehensive list. Larger molecules (e.g. C3) are also relevant but have an even larger number of possible configurations. Most adsorbates were monodentate (binding through a single adsorbate atom), but larger molecules known to bind in bi-dentate configurations were initialized that way. The atoms considered for either monodentate or bi-dentate adsorption location is indicated by *.

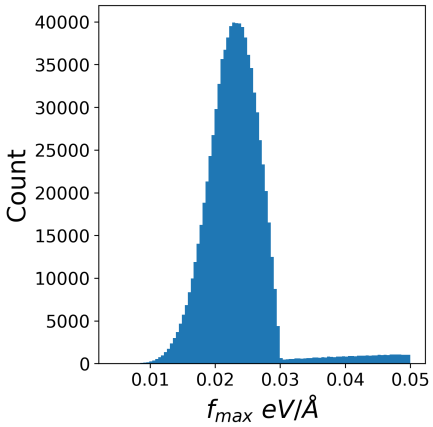

Figure 7: The distribution of max-absolute forces, f_{max}, for systems that converged and completed successfully. Systems in which $f_{max} > 0.05$ eV/Å were excluded from all tasks except S2EF.

Tight Binding Baseline

Obtaining reasonable energies, forces, and relaxed structures from tight binding codes is an enticing possibly because of the low computational cost compared to DFT; however, tight binding calculations on systems for catalysis remain a challenge, as demonstrated by SI Figure 8. We preformed tight binding calculations on 100 random systems from the validation set with extended tight binding (xTB) using the GFN0 parameters.[73] All calculations were done in accordance with our DFT calculations with two notable exceptions. For the combined systems, i.e. an adsorbate on a surface, all surface atoms were fixed during the relaxation. Additionally, the surface energies used for the computation of adsorption energies were approximated with single point energies. We did not allow surfaces to relax because of

Table 5: Adsorbates considered in OC20

Adsorbate class	# of adsorbates	Adsorbates
O/H Only	4	*H, *O, *OH, *OH_2
C_1	13	*C, *CO, *CH, *CHO, *COH, *CH_2, *CH_2*O, *CHOH, *CH_3, *OCH_3, *CH_2OH, *CH_4, *$OHCH_3$
C_2	41	*C*C, *CCO, *CCH, *CHCO, *CCHO, *COCHO, *CCHOH, *CCH_2, *CH*CH, CH_2*CO, *CHCHO, *CH*COH, *$COCH_2O$, *CHO*CHO, *COHCHO, *COHCOH, *CCH_3, *$CHCH_2$, *$COCH_3$, *$OCHCH_2$, *$COHCH_2$, *CHCHOH, *CCH_2OH, *CHOCHOH, *$COCH_2OH$, *COHCHOH, *CH_2*CH_2, *$OCHCH_3$, *$COHCH_3$, *$CHOHCH_2$, *$CHCH_2OH$, *OCH_2CHOH, *$CHOCH_2OH$, *$COHCH_2OH$, *CHOHCHOH, *CH_2CH_3, *OCH_2CH_3, *$CHOHCH_3$, *CH_2CH_2OH, *$CHOHCH_2OH$, *$OHCH_2CH_3$
Nitrogen-based	24	*$NH_2N(CH_3)_2$, *$ONN(CH_3)_2$, *$OHNNCH_3$, *$NNCH_3$, *ONH, *NHNH, *NHN_2, *N*NH, *$ONNO_2$, *NO_2NO_2, *N*NO, *N_2, *$ONNH_2$, *NH_2, *NH_3, *NONH, *NH, *NO_2, *NO, *N, *NO_3, *$OHNH_2$, *ONOH, *CN

unphysical behavior during optimization, which we likely attribute to periodic boundary conditions (PBCs). We are aware that the xTB code was designed for non-periodic systems and that incorporation of PBCs is an ongoing effort. Overall, the speed of the xTB was impressive and we look forward to future developments related to systems with PBCs.

Graph Generation Example

We generated graphs by converting atoms to nodes and detecting edges with a nearest neighbor search within a cutoff radius — node and edge features are model dependent. A simple 1D example is shown in SI Figure 9. PBCs complicate the process of determining edges and we take an atom centric view when determining if an edge is undirected or directed. Atoms within the unit cell all feel each other resulting in undirected edges, whereas unit cell atoms feel periodic image atoms but the reverse is not true, resulting in directed edges. The length of these interactions is of course limited by the radius of cutoff. In the example in SI Figure 9, atom 0 feels the periodic image atom 1, but the periodic image atom 1 does not feel atom 0. For this reason, we used mixed graphs that have both undirected and directed edges. In addition, a node can have multiple edges with another node (i.e. atom 0 and atom 1 have two different length connections because of PBCs) the result being a mixed multigraph.

Graph Pairwise Similarity

The mean pairwise similarity (mps) between a collection of graphs gives an indication of

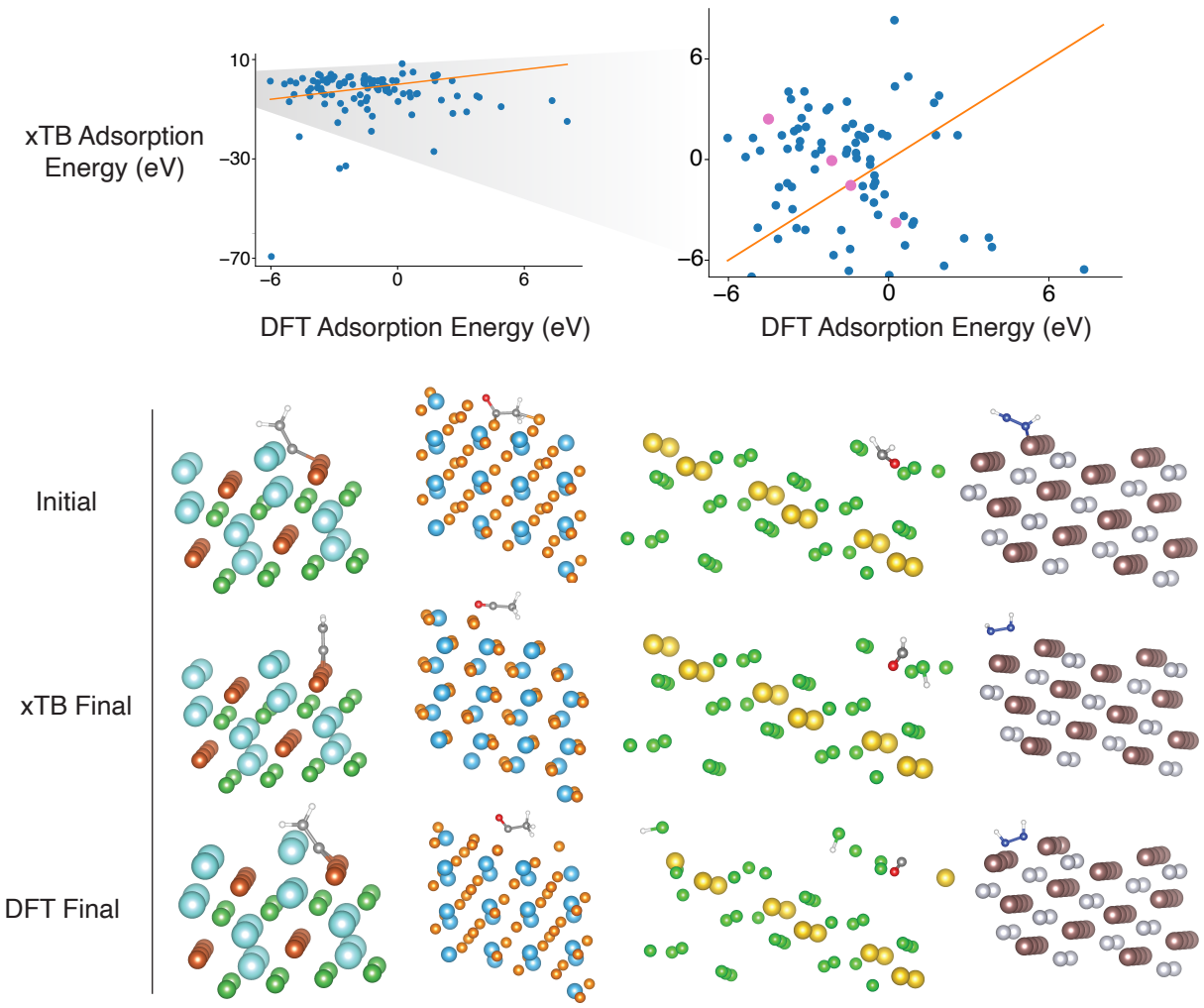

Figure 8: Top: A parity plot comparing xTB adsorption energies with DFT adsorption energies and an inset that limits xTB values to a range similar to that of DFT. Bottom: Initial and final structures corresponding to the pink markers in the plot above organized from left to right.

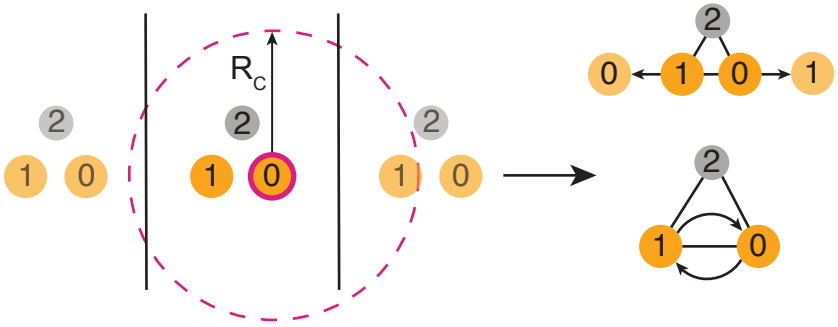

Figure 9: An example of graph generation for a 1-dimensional periodic system.

the diversity present in a given dataset and is comparable between different datasets. Pairwise similarity was computed as the mean of the elements in the upper triangle of the similarity matrix ($\mathbf{K}$) without the diagonal elements included (Equation below). The similarity matrix was calculated using graphs and the molecular kernel from the GraphDot package (https://graphdot.readthedocs.io/en/latest/), details of these methods are provided by Tang et al.[94] Mean pairwise similarity values range from 1, where all graphs are the same and decay to 0. The mean pairwise similarity can be compared between datasets if the graph and the kernel parameters are consistent. For the results in Figure of the main text, we randomly sampled 1000 systems (N) from a 10,000 subsample of each respective dataset and computed the mean pairwise similarity, this was repeated six times to collect statistics. Random subsampling was done to keep the similarity matrix the same size across datasets and to decrease the computational cost. For the similarity matrix calculation the adjacency length scale used to convert atomic structures to graphs was set to 6Åand the molecular kernel edge length scale was set to 18Å. All other parameters were set to default values.

$$\text{mps} = \frac{1}{N(N-1)/2} \sum_{i,j}^{N} \mathbf{K}_{ij}$$
$$\text{where } i < j$$

Train/test/validation splits

The following adsorbates were reserved for validation subsplits: *CH, *CHO, *$COCH_2OH$, *COH, *NH_2, *$NH_2N(CH_3)_2$, and *ONOH. Asterisks represent the binding atoms. The following adsorbates were reserved for the test subsilpts: *CH_2*CH_2, *CO, *$COHCH_2$, *NHN_2, *$NNCH_3$, *$OCHCH_2$, and *$ONNO_2$.

Baseline models implementation

All proposed baseline models were implemented using PyTorch Geometric. Several implementation changes, however, were necessary to make such models relevant to our dataset and tasks. We outline the modifications below:

SchNet

- Periodic boundary conditions (PBCs) were incorporated into the PyTorch Geometric implementation of SchNet.

DimeNet

- PBCs were incorporated into the PyTorch Geometric implementation of DimeNet.
- The number of triplets was capped to 30,000 during training to ensure tractability of our system sizes. All triplets were used during validation however. Previous work did not suffer the same computational challenges due to the significantly smaller system sizes of small organic molecules.

CGCNN

- Similar to SchNet, a Gaussian basis function was incorporated to the edge features. Although not contained within the original CGCNN implementation, a significant performance increase was observed.
- In order to make force predictions, a gradient call was included in the forward pass with respect to positions. The original CGCNN implementation was only concerned with energy predictions.