Generalizable Machine Learning Models for Electrocatalyst Discovery

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in the
Chemical Engineering Department

Muhammed Shuaibi

M.A.S. ChE, Illinois Institute of Technology
B.S. ChE, Illinois Institute of Technology

Carnegie Mellon University
Pittsburgh, PA

July, 2022

Generalizable Machine Learning Models for Electrocatalyst Discovery

by

Muhammed Shuaibi

Submitted to the Chemical Engineering Department
on 14 July 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

With the global population on the rise, increasing energy demands, and resulting climate change, the future of our energy infrastructure has become one of society's most pressing problems. Decreasing prices of renewable energy offers a promising path towards a sustainable future. However, the sun does not always shine nor the wind always blows, and addressing how we store intermittent energy sources will play a key role in our transition. One approach is to store energy in chemical forms. Unfortunately, these processes often rely on expensive rare metal catalysts, making them ill suited for commercial scale. The discovery of catalysts that can efficiently, selectively, and economically take part in these processes will be critical for society.

This thesis is centered around building generalizable machine learning (ML) models that span chemical and material space for catalyst discovery. A vital component in achieving this includes the curation of large-scale catalyst datasets. We first present how we can leverage active learning methods and physical biases to build ML models in the low data regime to accelerate density functional theory (DFT). We then present the largest catalyst dataset of its kind, *Open Catalyst 2020 (OC20)*, accompanied by baseline models and challenges to stimulate research in the catalysis and ML communities. With this dataset we explore the extent to which building a general purpose machine learning model is feasible. We then develop *SpinConv*, a graph neural network (GNN) that uniquely captures 3D atomic information to improve predictions on OC20. Next, we expand OC20 to present the *Open Catalyst 2022 (OC22)* dataset, consisting of oxide materials and more general purpose tasks. We also explore the extent existing datasets complement one another through alternative training strategies. Lastly, we discuss some of the challenges, trends, and general findings the community and ourselves have faced in building generalizable machine learning models.

Thesis Supervisor: Zachary W. Ulissi
Title: Associate Professor

Acknowledgments

First and foremost I would like to recognize the various institutions that have provided me the financial support to make this work possible, including: the Department of Chemical Engineering at Carnegie Mellon University; the Department of Energy, Basic Energy Sciences (DE-SC0019441); and Meta Platforms, Inc. (previously Facebook, Inc.). This research used resources of the National Energy Research Scientific Computing Center (NERSC); and the Summit supercomputer at Oak Ridge National Laboratory. I also acknowledge my doctoral committee, including Professor John Kitchin, Associate Professor Chrysanthos Gounaris, Trustee Professor Alan McGaughey, and Research Professor Jeff Schneider for their time and thoughtful feedback.

Throughout my Ph.D. I am grateful to have met so many incredible people that helped shaped the scientist I am today. At the top of that list is Zachary (Zack) Ulissi, my research advisor. There isn't a single day that went by in which I regret taking my chances and joining his, at the time, small, young new group. Zack has been nothing but supportive and patient in every aspect of my Ph.D., both professionally and personally. Whether it was asking him to switch research projects or take some time off to get married the summer of my Qualifying exams, Zack was always compassionate, understanding, and respectful. Zack was always a fountain of ideas I could turn to for an inspiring discussion. Thank you for being the best advisor I could have asked for. I want to also thank Larry Zitnick for playing such an important role in my Ph.D. I am grateful for our collaboration, the internship opportunities, the countless modeling discussions, and the honor to call you a mentor in my life.

I am grateful for all my collaborators, colleagues, and friends that I couldn't have done this work without. From my research group, Kevin Tran, Pari Palizhati, and Jun Yoon for being pioneers of the group and exceptional friends; Javier Heras-Domingo for teaching me about DFT and catalysis; and Adeesh Kolluru for all the insightful discussions. I am grateful for the opportunity to work alongside Weihua Hu and Johannes Gasteiger who I learned so much from about graph neural networks. I

owe a great deal of gratitude to my *Meta AI* colleagues, for their countless insightful discussions, their engineering standards, the seamless coordination, and constant willingness to assist in training models I could have never dreamed of training in an academic setting. Specifically, I want to highlight Abhishek Das, Siddharth Goyal, Janice Lan, and Brandon Wood. Brandon and Abhishek, specifically, have been great mentors and role models to me throughout my journey. From my cohort, I want to also thank Aliakbar Izadkhah who has been an incredible friend I could always rely on.

No amount of words can express how thankful I am for my family's love, support, and motivation to aim high. I want to thank my parents Abdallah Shuaibi and Amal Asfour for the sacrifices they made and instilling a passion for learning in me at such an early age. They have been pillars of inspiration and support for every aspect of my life. I want to also thank my siblings, Ahmed, Asma, Ayman, and Yousof for who I could always count on to make me laugh and have a good time. I am grateful for all the love and support my in-laws Raied and Lina Abdullah have provided me. Between their generous accommodations and spontaneous vacations, they made sure I always had a means to focus and have fun.

Most importantly, I cannot express how grateful I am for my wife, Dana Abdullah. Between the constant checking of slack messages, evening meetings, and mental distractions of research, I will forever be thankful for her ability to put up with it during my Ph.D. Her love and support made it easy to give my research my all. I will cherish all the experiences we had throughout my Ph.D. and look forward to the new ones we'll make on our next adventure.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation: Renewable energy storage

The push towards a sustainable energy future has become a global focus as the ecological and economical consequences of climate change continue to rise [175]. The transition away from a fossil-fuel reliant society will require alternatives to not just electricity generation, but production of society's most common fuels and chemicals [236]. As the cost of renewable energy sources (solar, wind, hydro, etc.) decreases, investments in these technologies are becoming more common [186, 98]. By 2050, the nation's electricity generation from renewable energy sources is expected to double, comprising 42% of the total electricity generation - far greater than today's 21% [62]. However, the sun does not always shine nor the wind always blows. Naively scaling renewable energy sources may result in excess, wasted, generation at times of low demand (Figure 1-1). While storage solutions like batteries and pumped hydro are available, they are often limited by their costs, scalability, and geographical constraints [298]. To ensure the successful transition to a sustainable future, the development of a variety of energy storage methods will be critical.

Electrochemical processes offer a promising route by storing energy chemically (Figure 1-2) [236, 207, 121, 281, 91, 163, 249]. Energy storage in the form of hydrogen or small hydrocarbons are advantageous in their ability to scale, be transported, and make use of existing infrastructure. In coupling such processes with renewable

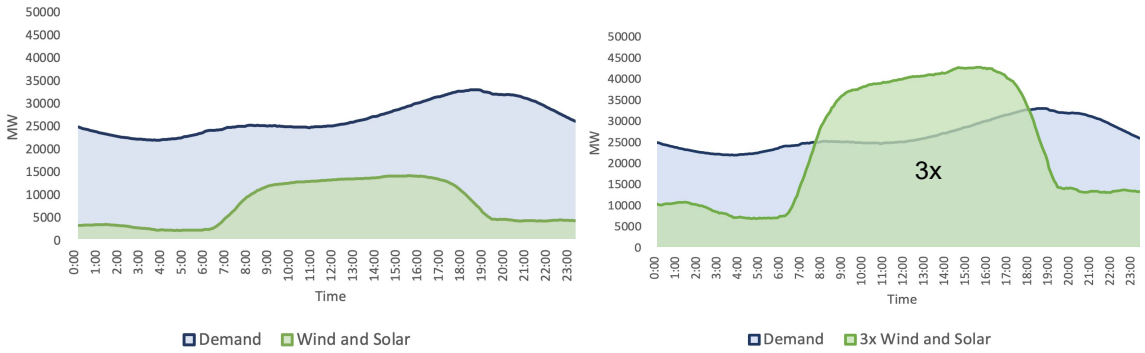


Figure 1-1: The challenge with naively scaling renewable energy sources. In times of peak generation, excess energy is going to waste as high-demand often occurs at a different time. Solutions are necessary to store energy at times of peak generation to be used at times of peak demand. Data obtained from California ISO[37] for an ordinary summer day (August 6, 2020). Adapted from Zitnick, L., Shuaibi, M. et al. [298].

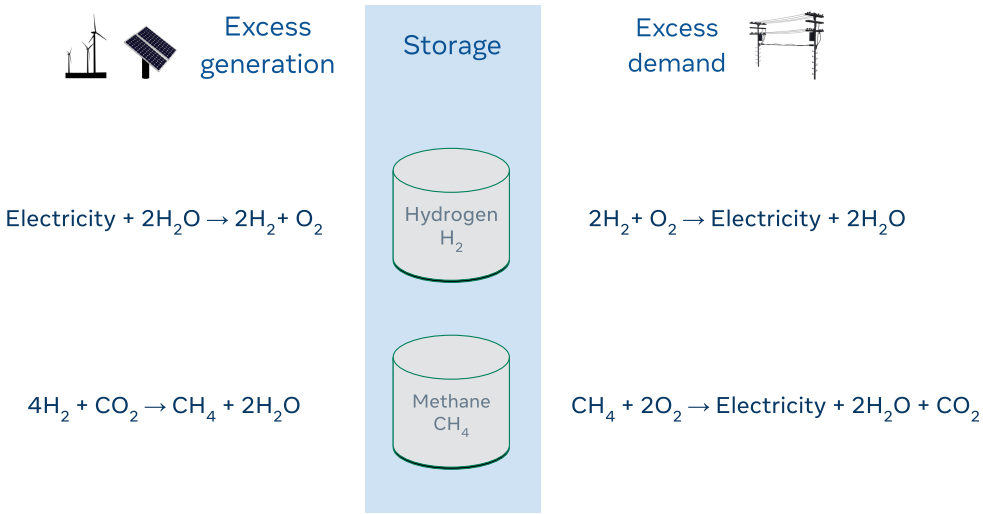


Figure 1-2: Electrochemical cells offer a potential alternative to energy storage. During times of excess generation, electricity can be used to drive chemical reactions to produce meaningful fuels and products. For example, electricity can split water into hydrogen and oxygen, storing hydrogen for fuels at a later time. Stored hydrogen can also be used to react harmful CO_2 into hydrocarbon fuels.

energy, essential products are produced from fossil-fuel free sources while also offering a means to store the intermittency of renewable energy sources. However, renewable energy technologies, such as electrochemical and fuel cells, are currently limited by the availability of catalysts that can efficiently, selectively, and economically perform the necessary reactions [236].

1.2 Computational tools for catalyst screening

The discovery of new catalysts has traditionally been done by experiments. Through experimentation, important catalyst properties including reactivity, stability, and selectivity can be directly measured. However, given the number of unique material combinations, the search space for new catalysts can quickly exceed billions or even trillions of possibilities. Although effective, experimentation is often limited to a handful of systems per month, making them infeasible for the large scale screening desired.

Alongside advancements in high performance computing, quantum mechanical tools such as Density Functional Theory (DFT) have aided in accelerating the catalyst discovery process [192]. With DFT, molecular systems can be simulated and studied on catalyst surfaces. While experimental properties are not directly available, intermediate properties like adsorption energy can serve as meaningful descriptors for chemical reaction rates [40, 298, 183, 185]. Despite the success and progress quantum mechanics has allowed researches to make [120, 210, 288, 142, 215, 196], tools like DFT scale very poorly - $O(N^3)$ in the number of electrons, or worse for more accurate theories. With simulations taking anywhere on the order of hours to days, more efficient methods are still necessary.

"Machine learning potentials" (MLPs) have emerged in the past decade to bridge the gap between DFT-level accuracy and traditional force-field-level efficiency [11, 222, 174, 198, 27, 125, 21]. Trained on DFT or other quantum mechanical data, MLPs take in an atomic configuration and return a system total energy and per-atom forces. MLP outputs allow one to study the energy landscape, or Potential

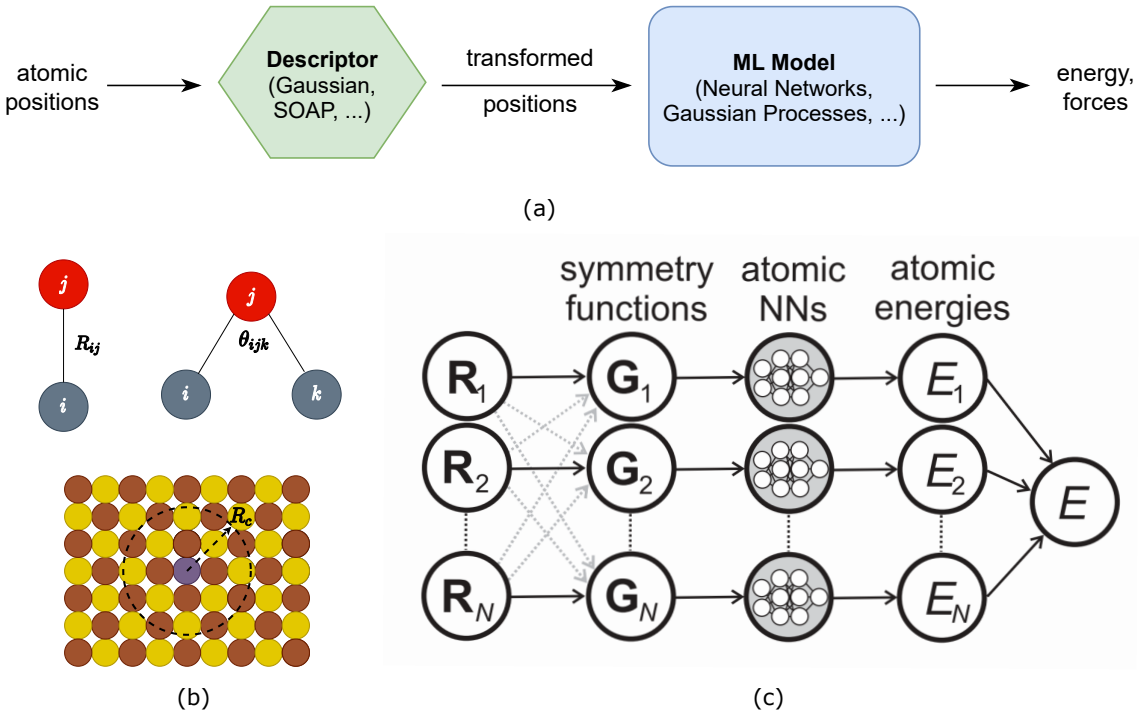


Figure 1-3: (a) A typical machine learning potential (MLP) workflow involves a transformation of coordinates to capture symmetry invariances before being used in a ML model. (b) Features commonly contain two and three-body interactions in their representations. Cutoff radii are employed to describe the local environment of any given atom. (c) An example MLP, Behler-Parinello Neural Network[29], contains element specific neural networks contributing to the total energy of the system; obtained from [26].

Energy Surface (PES), of an atomic system orders of magnitude faster than first-principles methods. A key step in the development and success of MLPs is representing the atomic system into meaningful features [125, 29]. Representations must be constructed while maintaining physical symmetries of rotation and translation invariance. A common representation may include explicit two-body and three-body terms (Figure 1-3b) [232, 133]. Hand-crafting representations is often a tricky and non-trivial task that may not always generalize to new atomic environments; a representation for one system may not be sufficient for another. As MLPs are entirely data-driven, their ability to generalize is limited by the diversity and quality of their training dataset. Achieving a generalizable machine learning potential, that accurately predicts catalyst properties across chemical and material space, will require answering two important questions - how do we curate diverse datasets to model any

arbitrary catalyst chemistry and how do we best represent and model atomic systems?

1.3 Graph Neural Networks

Figure 1-4: (left) A graph neural network first constructs a graph from the raw atomic structure, embedding atoms as nodes and bonds as edges, each with corresponding features. The graph goes through several message passing layers to arrive at (right) a final representation that is then fed through a simple neural network to arrive at the desired property.

A new class of models, graph neural networks (GNNs), have grown in popularity for molecular applications. Models of this sort have been used to predict crystal properties [286] and study small organic molecules [232, 133, 130, 25]. Here, MLPs no longer need to rely on hand-crafted features to accurately represent the atomic environment. Instead, graphs, almost naturally, are constructed with atoms as nodes and interactions between atoms as edges. Node representations are updated based off "messages" exchanged by neighboring nodes, referred to as "message passing" [79] (Figure 1-4). This process is repeated for several iterations, or interaction layers. At each interaction layer, messages are sent between nodes, aggregated, and used to update the node's representation. Nodes are updated in parallel, with messages corresponding to some non-linear function of the nodes' representations. More formally, for an atom v, the inbound message at interaction layer t is expressed as:

$$m_v^{t+1} = \sum_w^{N_{ngh}^v} M_t(h_v^t, h_w^t, e_{vw}) \tag{1.1}$$

Where h is the atomic representation, w a neighboring atom, e_{vw} the edge between nodes v and w, N_{ngh}^v the atoms in the neighborhood of v, and M_t the parameterized

message function to be learned. The atomic representation h_v is then updated based off a learnable update function, U:

$$h_v^{t+1} = U(h_v^t, m_v^{t+1}) \tag{1.2}$$

After T interaction layers, N atomic representations are fed into a readout, or output function, R, for the final property prediction, P:

$$P = \sum_v^N R(h_v^T) \tag{1.3}$$

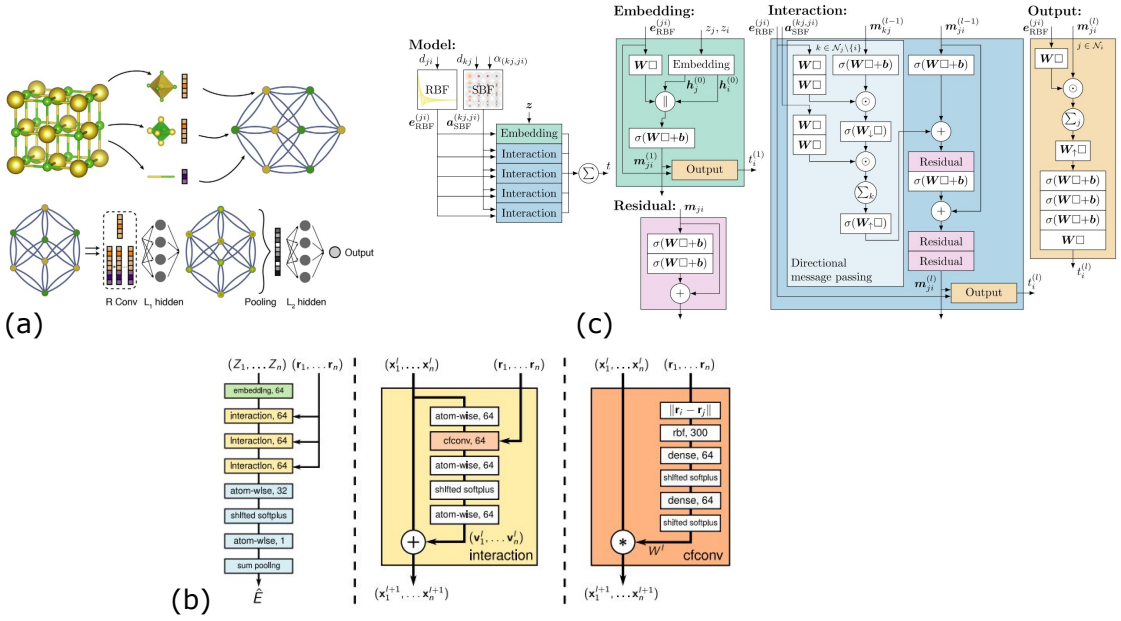


Figure 1-5: GNNs continue to grow in complexity to better model atomic systems. Models primarily differ in the ways they update and exchange messages between nodes. Both (a) CGCNN [286] and (b) SchNet [232] present early architectures only capturing two-body interactions. (c) DimeNet[133] builds on this by also capturing angular interactions in its messages. Figures obtained directly from Xie et al.[286], Schütt et al.[232], Gasteiger et al.[133].

When building MLPs, the property desired is the total energy E. Per-atom forces can then be obtained through the gradient with respect to atomic positions:

$$\hat{F}_i = -\frac{d\hat{E}}{d\mathbf{x}_i} \tag{1.4}$$

GNN architectures often distinguish themselves in their node/edge representations and message, update, and readout functions. Models often involve highly parameterized, complex network architectures and require substantial amounts of data to be successful (Figure 1-5). As a result, GNNs are able to achieve state of the art performances on small molecule datasets and are well suited for catalyst applications.

The development of GNNs for atomistic applications, however, has been primarily limited to small organic molecules [48, 212, 231, 130, 132, 25, 206]. There lacks a thorough exploration of GNNs for catalyst applications. Developed for crystalline materials, Crystal Graph Convolution Neural Network (CGCNN) [286], continues to be a popular baseline for material applications despite the development of stronger baselines for small molecules [148, 190, 217]. While the development of GNNs for small molecules has led to notable improvements, catalyst systems are often much larger and more complex in nature. While models like SchNet[232] and DimeNet++[133] explicitly capture bonded and angular interactions, they rely on message passing to implicitly capture higher order interactions (e.g. dihedral or long range). This can pose a challenge for large catalyst systems to sufficiently capture the full 3D environment. Additionally, the prediction of gradient-derived forces often corresponds to a 2-4x overhead in compute and may be constraining the model's predictive ability rather than assisting it. While the trends and design choices made for small molecules have been well documented, the catalyst community lacks well established baselines and benchmarks for building accurate models. Only then are we able to study model trends and develop more efficient and expressive architectures for catalysis.

1.4 Datasets

At the core of any ML approach is the quality and magnitude of reliable training data. In the context of atomistic modeling, training data often refers to semi-empirical or quantum mechanical data (e.g. DFTB[200], DFT[192], CCSD(T)[209], etc.). However, models developed on finite datasets are limited in their applications to similar systems. For example, a model trained on only Copper-based materials would not

Dataset	Description	Elements	Avg. size	Train set size
MD17[48]	Eight separate molecules	H, C, N, O	12.5 (9-21)	$8 \times 1{,}000$
ISO17[234, 232, 212]	$C_7O_2H_{10}$ isomers	H, C, O	19	404,000
S_N2[270]	Methyl halides, halide ions	H, C, F, Cl, Br, I	5.4 (2-6)	400,000
ANI-1x[245]	Selected MD samples	H, C, N, O	15.3 (2-63)	4,956,005
QM7-X[107]	Small molecules	H, C, N, O, S, Cl	16.7 (4-23)	4,175,037
COLL[131]	Molecule collisions	H, C, O	10.2 (2-26)	120,000

Table 1.1: A comparison of small organic molecular datasets. Sizeable datasets aided in the development and application of GNNs for molecular simulations. However, datasets are limited in their chemical diversity and atomic size, prohibiting the applications to out of domain applications like material and catalyst discovery. Adapted from Gasteiger, J., Shuaibi, M., et al.[76]

be expected to perform well on small hydrocarbons. Additionally, as GNNs grow in complexity, datasets need to be sufficiently large enough for highly parameterized models to learn meaningful atomic representations. The advancements in GNNs for small organic molecules can be partly attributed to the availability of such datasets (Table 1.1). Despite the sizes of the mentioned datasets, their chemical diversity is often limited to only a handful of elements - H, C, N, O; inhibiting their use for catalyst applications, which require supervision on a much larger set of elements.

Datasets created for catalysis suffer in both their chemical diversity and size. With datasets spanning as little as $\sim$100 [38, 2, 160, 179, 5] to at most $\sim$100,000 [161, 265, 283], GNNs have often been overlooked in lieu of smaller, more simple architectures [125, 21]. More troubling, datasets being curated for a range of chemistries using different computational tools (i.e. DFT settings) makes it challenging to centralize a diverse dataset for model development. As a result, the research community has relied on self-curated datasets to develop models that are only relevant to their specific applications. The primary challenge in building a general purpose machine learning potential is the availability of large, diverse catalyst datasets spanning chemical and material space.

1.5 Active Learning

One of the challenges in building reliable MLPs is the availability of sufficient training data. This issue is particularly emphasized when studying new systems, where no data may be available. Active learning (AL), a branch of machine learning, can aid in such regimes as it explores how to systematically select data points to improve the learning algorithm (Figure 1-6). Rather than arbitrarily curating datasets, AL can be used to select training data points from a pool of candidates [238]. On-the-fly learning, or online active learning (OAL), has shown particular potential for molecular applications [272, 116, 30]. In such a framework, models are updated on-the-fly as the simulation evolves. MLP predictions are used when the model is confident in its prediction, otherwise a quantum mechanical call is made. Calls made to the oracle (e.g. DFT) are then added to the training data, the model is retrained, and the simulation proceeds - all with no terminations to the simulation. Since the cost of DFT is so high, AL is well-suited for accelerating catalyst discovery.

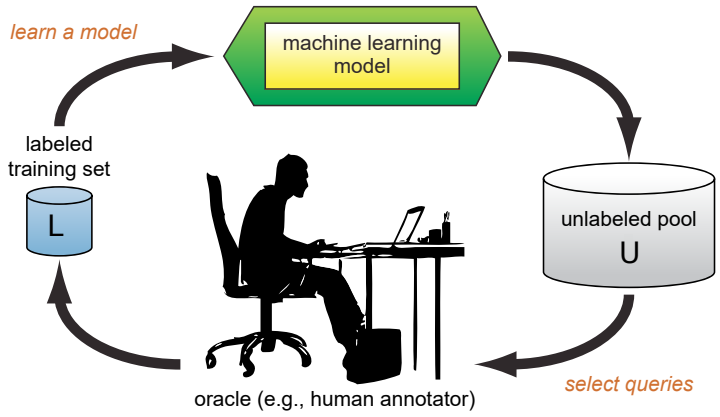


Figure 1-6: Overview of an active learning framework. An initial machine learning model samples data from an unlabeled pool based off a variety of querying strategies. Sampled candidates are evaluated by an oracle and added to the training dataset. The ML model is then retrained and the process repeated until a desired accuracy is achieved. Figure obtained directly from Settles [238].

The development of generalizable machine learning models, by definition, would alleviate the challenges of studying unseen systems. However, the curation of datasets that span *all* of chemical and material space is impossible. In such instances, AL

methods can help accelerate the screening process while still minimizing computational costs. AL frameworks developed for molecular applications have traditionally relied on Gaussian Processes (GP) [272, 21]. While effective in the small data regime, GPs can scale rather poorly [22]. On the other hand, neural networks scale more efficiently and have seen significant architectural developments in the last decade. Bridging the gap between more complex architectures and AL frameworks has the potential to improve catalyst discovery workflows by filling in the gaps of generalizable modeling efforts.

1.6 Research objective

The goal of my thesis research is to accelerate the catalyst discovery process through generalizable machine learning models and methods. I accomplish this through the creation of large-scale datasets that span chemical and material space. Accompanying these datasets, I formulate challenges relevant to every-day catalyst tasks and present baseline models to bootstrap community engagement. Through the curation of these datasets, I develop new model architectures, methods, and strategies to improve property predictions across catalyst chemistry.

Chapter 2

Enabling robust offline active learning for machine learning potentials using simple physics-based priors

This work originally appeared as: Shuaibi, M., Sivakumar, S., Chen, R.Q. and Ulissi, Z.W., 2020. Enabling robust offline active learning for machine learning potentials using simple physics-based priors. Machine Learning: Science and Technology, 2(2), p.025007. It has been edited to include the supplementary information in Appendix A.

2.1 Abstract

Machine learning surrogate models for quantum mechanical simulations has enabled the field to efficiently and accurately study material and molecular systems. Developed models typically rely on a substantial amount of data to make reliable predictions of the potential energy landscape or careful active learning and uncertainty estimates. When starting with small datasets, convergence of active learning approaches is a major outstanding challenge which limited most demonstrations to online active learning. In this work we demonstrate a Δ-machine learning approach that enables stable convergence in offline active learning strategies by avoiding unphysical configurations with initial datasets as little as a single data point. We demonstrate our

framework's capabilities on a structural relaxation, transition state calculation, and molecular dynamics simulation, with the number of first principle calculations being cut down anywhere from 70-90%. The approach is incorporated and developed alongside AMP*torch*, an open-source machine learning potential package, along with interactive Google Colab notebook examples.

2.2 Introduction

The last decade has seen a surge in machine learning applications to material science, physics, and chemistry [11, 222, 174, 198, 27, 125, 21]. Characterizing a molecular system's potential energy surface (PES) has been a crucial step to the development of new catalysts and materials. Structure relaxation, molecular dynamics, and transition state calculations rely almost entirely on an accurate PES to serve their functions. Machine Learning Potential (MLP)s have demonstrated chemical accuracy at orders of magnitude faster computation times than traditional *ab-initio* methods including density functional theory (DFT) and coupled cluster single double triple (CCSDT) [300]. However, these demonstrations have generally required large datasets and careful uncertainty estimates. More importantly, the models developed have struggled to generalize to new systems and faced convergence issues when adding data, making the practicality of their day-to-day applications challenging [44, 170, 27, 227]. The potential of active learning in molecular simulations has not been fully realized due to convergence and implementation challenges.

The careful curation of training datasets for accurate molecular simulations has recently given way to active learning [272, 116, 73, 72]. Active learning (AL) is the branch of machine learning concerned with systematically querying data points to be be part of the training set [238]. The iterative process queries new data, trains a model, and repeats until a model performance is achieved. AL methods are particularly useful when the cost of querying data is substantial - as in the case of computing DFT. There are two main classes of strategies with relevance to molecular simulations. In Online-AL, configurations are generated sequentially using a MLP and for

each a decision is made whether to accept the estimate, perhaps using an uncertainty estimate. In Offline-AL, a pool of candidates is generated and a decision is made which of the pool to add to the training set.

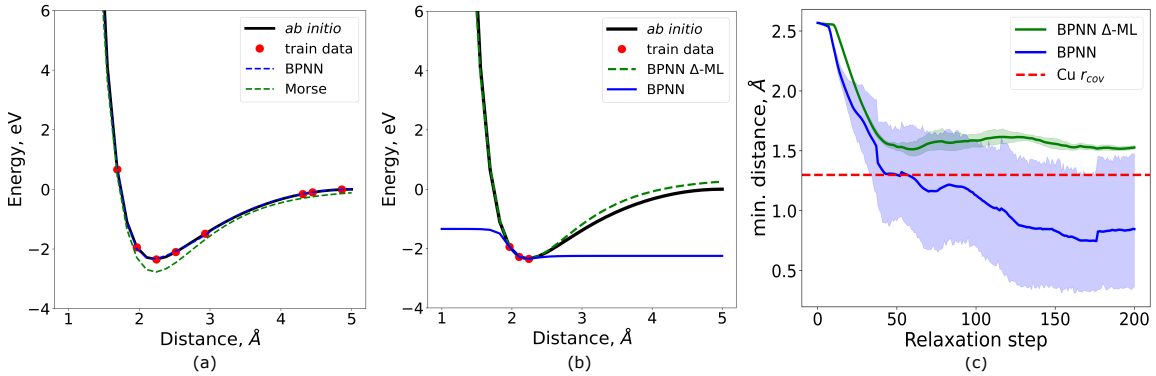


Figure 2-1: A traditional Behler-Parinello neural network (BPNN) trained to replicate the potential energy surface (PES) of a Cu-Cu bond with **(a)** a dataset spanning the PES and **(b)** a limited dataset trained with and without a Morse prior. **(c)** The minimum pair-wise distance of a structure relaxation carried out with a BPNN model, with and without the Morse prior. Relative to the covalent radius of Cu, our model consistently predicts more physically-consistent configurations as compared to the more unstable BPNN. Error bars represent the 95% confidence interval.

Although there are many strategies available for both Online-AL and Offline-AL, both commonly assume that all generated candidates are feasible to be queried and that adding data will not reduce accuracy on previous training data. Both of these assumptions are difficult with MLP: DFT often fails to converge on far-from-equilibrium structures, and many MLP suffer if even a small number of configurations with large energies/forces are added to the training dataset [73]. These concerns are especially problematic when dealing with little to no initial data. The most common approach to address these challenges is to carefully monitor uncertainty in the active learning process and prevent extrapolation to unphysical regions. This strategy is relatively straightforward to implement in Online-AL: if the uncertainty estimate is below a threshold, accept the prediction, otherwise run the DFT calculations. If the step size is small enough, the new configuration should be not so different from configurations in the training set. However, in Offline-AL it is difficult to ensure the queried configurations will converge with DFT and won't contaminate the dataset

once added. Instead of solving this problem, we show that it is possible to mostly fix the underlying issues leading to unrealistic configurations.

In this work, we demonstrate that stable convergence in Offline-AL with MLP is possible by adding simple repulsive potentials and robust training procedures. This approach is implemented for the common combination of Behler-Parinello MLP fingerprints with neural network atomic energy models [29]. We show that a Δ-ML approach with a base pairwise Morse potential and linear mixing rules is capable of sufficiently capturing the repulsive interactions between atoms that lead to DFT errors. Since this Morse potential is not responsible for capturing the full potential, the parameterization only needs to be done once for each element. We demonstrate this approach for several types of calculations common in catalysis: structure relaxations, molecular dynamics, and transition state calculations. In each case, convergence with the addition of training data is essentially impossible with the base potential and well behaved with the Δ-ML approach. In most cases this process allows for a reduction of 70-90% in the number of DFT single-point evaluations necessary. This process is further improved using standard neural network training approaches in the ML community to reduce the impact of random initial weights on small datasets. All of these are demonstrated in open-source and accessible AMP*torch* GitHub repository with Google Colab ASE examples [240, 243].

2.3 Methods

The ML community continues to make advancements in the optimization and implementation of neural network based models [159, 65, 193]. To leverage some of these approaches, we employ a Behler-Parinello neural network (BPNN). BPNNs construct element specific neural networks with the energy of the system the sum of atomic energy contributions. Per-atom forces are directly obtained from the negative gradient of the energy with respect to the atomic positions. We refer the readers to several reviews for a more detailed discussion on the BPNN model [29, 27, 125]. Additionally, neural network based models don't suffer the same kernel selection and scalability

challenges that can come with Gaussian processes (GP) and other bayesian models [22]. Training neural networks, however, can be an extremely challenging task we hope to address in this work.

In the presence of an abundance of data, BPNN-like models have shown great success in replicating the PES of various systems [125, 198, 230]. In the small data limit, however, neural network based models are unable to successfully characterize the energy surface, Figure 2-1b. More notably, model predictions are entirely "physics-free", such that simple repulsive interactions are only ever learned by the model once enough data has been provided. As a result, a considerable amount of time may be wasted learning simple, widely understood, characteristics of the PES. Hybrid physics-based machine learning models can provide an important path forward to making reliable, physically-consistent discoveries in the sciences [282, 123]. To address this, we incorporate a Δ-ML approach [212, 297] to learn the correction, E_{NN}, between a simple Morse potential, ΔE_{morse}, and *ab-initio* level theory - namely, DFT, $\Delta E_{DFT}(\mathbf{x})$:

$$\Delta E_{DFT}(\mathbf{x}) = E_{DFT}(\mathbf{x}) - E_{DFT}(x_{ref}) \tag{2.1}$$

$$\Delta E_{morse}(\mathbf{x}) = E_{morse}(\mathbf{x}) - E_{morse}(x_{ref}) \tag{2.2}$$

$$E_{NN}(\mathbf{x}) = \Delta E_{DFT}(\mathbf{x}) - \Delta E_{morse}(\mathbf{x}) \tag{2.3}$$

$$E_{morse+NN}(\mathbf{x}) = \Delta E_{morse}(\mathbf{x}) + E_{DFT}(x_{ref}) + E_{NN}(\mathbf{x}) \tag{2.4}$$

Where $E_{DFT}(x_{ref})$ and $E_{morse}(x_{ref})$ correspond to reference energies necessary to correct for differences in their absolute energies. Reference energies are computed from a same arbitrary structure, x_{ref}; the dataset's first structure was used in our applications. Per-element parameters of the Morse potential, D_e, r_e, and a, are fitted to DFT data *a priori*. A more detailed description of the fitting procedure is included in Appendix A. By leveraging the Morse potential as the backbone to the model, the ML component is allowed to learn the remaining functional form while still capturing physics-based repulsive interactions previously missed. Additionally, learning a correction can allow the neural network to learn a much smoother function

than the underlying PES, improving training stability and convergence.

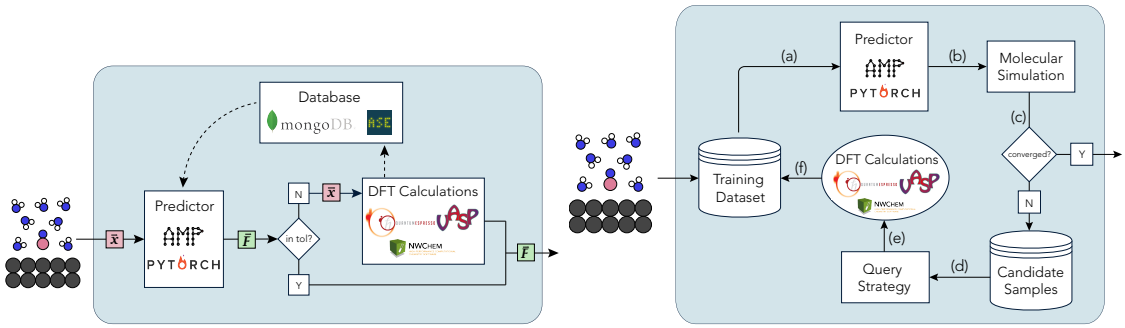


Figure 2-2: Online and Offline active learning frameworks to accelerate molecular simulations. **Left**: Online-AL. At each time step, our ML model makes a prediction of the energy and forces and assesses the uncertainty of its estimate. If confident, the ML results are used to take a step in the molecular simulation. Otherwise, a DFT call is made, added to a database, and the model retrained. **Right**: Proposed Offline-AL. **(a)** An initial training dataset is used to train the ML model; **(b)** the trained ML model runs the atomistic simulation of interest; **(c)** termination if converged, otherwise, the generated data is stored as a pool of potential candidates; **(d)** a query strategy identifies what points to be added to the training set; **(e)** *ab-initio* calculations are performed on selected candidates; **(f)** queried points are added to the training set. Repeat until convergence is reached.

We illustrate the benefits of this simple Morse potential by running a structure relaxation of carbon on copper (C/Cu) with our model trained on a single image (Figure 2-1c). The minimum pair-wise distance of the resulting trajectory are compared to that not employing a morse potential. Our model consistently predicts configurations above the covalent radius of copper, a good indication repulsive forces are being captured. On the other hand, a traditional BPNN shows wide variations while on average predicting configurations well below the more stable covalent radius.

The fitting of MLPs is an important process in our AL framework, as they are responsible for generating candidates for training data. A poorly fit MLP may generate unfeasible candidates that DFT can not converge on. This is especially true when working without a physics-based potential. Working within the small data regime allows us to leverage quasi-newton optimizers, namely LBFGS. LBFGS and other second order optimizers provides us with improved convergence of our model training over standard first order methods such as SGD and Adam. This advantage, however, is only really feasible in the small data limit where the computational cost of such

methods can be afforded. Additionally, we incorporate a cosine annealing learning rate scheduler with warm restarts [159] to aid in the convergence of the Offline-AL framework. A more detailed comparison can be found in Appendix A.

Similar to previous works [272, 116, 116], our Online-AL framework begins with little to no data and must identify the right points to query and improve the model over the course of a molecular simulation (Figure 2-2). Rather than relying on kernel-based models, our Online-AL framework utilizes the proposed physics-coupled BPNN. We incorporate bootstrap-ensembling, or bagging, in order to quantify our model's uncertainty. Bagging involves training multiple, randomly initialized, independent models with training sets randomly sampled, *with* replacement, from an original dataset [199]. Predictions and uncertainty estimations are then calculated from the ensemble statistics.

An offline active learning can offer model and computational advantages over Online-AL frameworks. Rather than making query decisions in a dynamic process, we present a method to select from a pool of generated candidates. Prior works have incorporated offline active learning to various extents. Sivaraman, et al. [244] used active learning to downselect from an existing hafnium dioxide AIMD simulation to train a GAP model. Novikov, et al. [184] used active learning and Moment Tensor Potentials (MTP) to run atomistic simulations. We show, however, that a standard neural network is unable to follow a similar framework without careful modifications. Rossi, et al. [220] use an ensemble of neural networks to estimate uncertainty along an atomistic simulation. Having begun from an extensively sampled training dataset, their need for retraining was avoided, a problem we address for neural networks in the small data regime. While the use of active learning has shown incredible success in training models with fractions of the dataset, it assumes such datasets exist to begin with. We propose a framework to enable accurate atomistic simulations beginning with as little as a single data point. We accomplish this by iteratively running an ML-driven molecular simulation. After each iteration, a querying strategy samples from the generated trajectory. Queried points are then evaluated with DFT, added to an original dataset, and the ML model retrained (Figure 2-2). The process is

repeated until a defined convergence criteria is met. Despite the ML model resulting in inaccurate simulations early on, diverse, informative configurations are generated to train the ML model. In dealing with a pool of query candidates, the framework allows us to explore alternative querying strategies other than uncertainty estimates of Online-AL [238]. The reliance on uncertainty estimates can pose more fundamental questions surrounding energy conservation from a retrained potential [184] and how trustworthy a model's estimates really are [265].

We demonstrate the proposed framework on several common catalysis applications: structure relaxations, transition-state calculations, and molecular dynamics with system sizes between ∼12-30 atoms. A random sample query strategy is introduced in the Offline-AL schemes to demonstrate the effectiveness of even the simplest of query strategies over Online-AL. More problem-specific query strategies are proposed for structural relaxations and transition-state calculations, further improving the convergence. To show the generality of this approach in small-data applications, we also use two common DFT packages - Vienna Ab initio Simulation Package (VASP) and Quantum Espresso (QE) [137, 138, 78]. The use of QE allows for interactive and open demonstrations of this approach. Several Google Colab notebooks have been included in Appendix A allowing users to easily experiment and explore new systems with AMP*torch* and QE without needing to locally install and manage dependencies.

2.4 Results and discussion

A structural relaxation is performed for C/Cu(100) with cell size $2 \times 2 \times 3$. An initial guess of 3Å from the surface is made for the adsorbate. Periodic boundary conditions are applied in the x and y directions and the last slab layer is fixed from relaxations. The training dataset begins with a single initial structure.

Performance is measured by the final structure and energy mean-absolute-errors (MAE). A random sample query strategy selects configurations from the generated relaxations to be queried. We run the Offline-AL framework under a variety of batching scenarios, terminating after N iterations, sampling M configurations per iteration,

Batching Scenario		Energy MAE (eV)	Structure MAE (Å)
Iterations	Samples per iteration		
20	1	0.0063	0.0037
10	2	0.0069	0.0063
5	4	0.0080	0.0067

Table 2.1: Comparison of different offline active learning batching scenarios on the structural relaxation of C/Cu(100). At each iteration, a varying number of queries are randomly made from the generated relaxation. A tradeoff in performance and the number of samples per iteration is observed for a fixed total number of DFT calls = 20. All models trained here incorporated the proposed Morse prior.

for an arbitrary total of $NM = 20$ DFT calls. Results are summarized in Table 2.1.

Under the above random query strategy, systematic termination of the Offline-AL loop is quite heuristic. To address this, we incorporate alternative query and termination strategies - quasi-random and uncertainty sampling. In quasi-random, at each iteration, in addition to a random configuration, the predicted relaxed structure is also queried. Similarly, uncertainty sampling samples the k-most uncertain points in addition to the relaxed structure. In both strategies, if the predicted relaxed structure's max per-atom force, as evaluated by DFT, is below the optimizer's convergence criteria, the AL loop is terminated. Otherwise, the configurations are added to the original dataset, and the framework cycles. In querying the model's predicted relaxed structure we are assured in our framework's ability to accurately reach a local minima.

We compare the performance of this Offline-AL scheme and Online-AL with and without the Δ-ML in Table 2.2. Offline-AL and Online-AL tolerances correspond to the max per-atom force termination criteria and max force variance tolerated by the ensemble, respectively. Force termination criteria of 0.03 and 0.05 eV/Å are compared to explore the tradeoff between accuracy and number of DFT calls. Online-AL was empirically set to query a DFT call when the ensemble based force uncertainty reached above a threshold of 0.05 eV/Å. The energy and structure MAE associated with the system's initial structure is 2.82 eV and 0.15 Å, respectively. Our best performing framework - Offline-AL with Δ-ML (0.03 eV/Å), reported average energy and structure MAEs of 0.0039 eV and 0.0032 Å with 17 total DFT calls - a 66.7%

Framework (tolerance)	MLP	Energy MAE (eV)	Structure MAE (Å)	DFT calls
DFT	-	-	-	51
Offline-AL (0.03 eV/Å)	BPNN Δ-ML	0.0039	0.0032	17
Offline-AL (0.05 eV/Å)	BPNN Δ-ML	0.0049	0.0059	15
Offline-AL (0.05 eV/Å)	BPNN only	does not converge		
Online-AL (0.05 eV/Å)	BPNN Δ-ML	0.0073	0.0107	30
Online-AL (0.05 eV/Å)	BPNN only	0.2884	0.0263	22

Table 2.2: Summary of the various strategies' performance on the structural relaxation of C/Cu(100). The effects of the Morse prior on the convergence of both the offline and online active learning are also shown. The querying strategy employed by the Offline-AL framework relies on a quasi-random strategy, additionally sampling and assessing convergence on the framework's generated relaxed structure.

Query Strategy	Energy MAE (eV)	Structure MAE (Å)	DFT calls
Random	0.0063	0.0037	20
Quasi-Random	0.0049	0.0059	15
Uncertainty Sampling	0.0061	0.0050	19

Table 2.3: Comparison of different offline active learning query strategies on the structural relaxation of C/Cu(100). All models trained here incorporated the proposed Morse prior.

reduction. Without the inclusion of the Morse prior, a standard BPNN was unable to converge, generating configurations that DFT was unable to evaluate in almost all our experiments. We compare several query strategies in Table 2.3, demonstrating similar success in all scenarios.

Next, we demonstrate an application to transition state calculations, specifically, nudge-elastic-band (NEB) methods [103, 102]. NEB calculations require defining the initial and final structures for the transition state to be calculated. Machine-learning accelerated NEB calculations have typically relied on *ab-initio* relaxed initial and final structures, a costly step of a NEB calculation [8]. In fixing the initial and final structures, the machine learning objective is simplified to an interpolation problem. We demonstrate our framework's ability to accelerate the complete NEB calculation, including initial and final structure relaxations, to find the surface diffusion energy barrier of oxygen on Cu(100). To illustrate our framework, we use five images to build the NEB including the initial and final states which have not been relaxed previously. The initial training dataset includes only the unrelaxed initial and final structures.

The convergence evolution of our Offline-AL framework is illustrated in Figure 2-3d, approaching the true energy barrier after a few iterations. Similarly, convergence was not achieved, with often failing DFT evaluations, without the inclusion of the Morse prior. A simple random strategy is first employed. Here the images are randomly sampled from generated NEBs and evaluated using DFT before being added to the training data. Termination is achieved after a fixed number of iterations. Additionally, we compare the efficiency of two more systematic querying and early stopping methodologies. An uncertainty sampling strategy queries images with the highest uncertainty, which are then evaluated with DFT, and added to the training data. Termination is reached when the difference between the predicted energies from ML and DFT at the saddle point are less than a tolerance. An additional strategy is also tailor-made for the NEBs, where the highest energy point, along with the initial and final points are sampled at each iteration. The loop is terminated once the difference between the ML predicted energies and DFT evaluated energies of the three points is less than a specified threshold. All three cases demonstrate a significant

reduction in the number of DFT calls required to construct the NEB as shown in (Figure 2-3f).

Machine learning surrogates to DFT are considerably favorable in the context of long time-scale simulations, namely, molecular dynamics (MD). Unlike structural relaxations, MD simulations are typically carried out on orders of magnitudes more steps. Several works have addressed these challenges through GP-based Online-AL frameworks [272, 116]. We demonstrate that our proposed Offline-AL framework is capable of converging to an accurate MD simulation. A 2ps MD simulation of CO on Cu(100) in a 300K NVT ensemble is used for our demonstration.

Beginning with a dataset containing only the initial structure, our framework cycles for several iterations, randomly querying 50 configurations for a total of 500 DFT calls by the end of our experiment. Unlike structural relaxations with a well defined target, MD simulations are more stochastic in nature and are unlikely to follow an identical trajectory over multiple iterations. To demonstrate the effectiveness of our framework, we verify the performance, at each iteration, by randomly sampling 400 configurations from our ML predicted trajectory and validate their corresponding energy and force predictions with DFT. We illustrate the iterative convergence of our framework in Figure 2-4. Despite the upper limit of 10 iterations, we observe good agreement with DFT by iteration 6 - a reduction of 85% in DFT calls. Additionally, we demonstrate consistency in the radial distribution function of our framework's generated simulation to that of the original DFT simulation (Figure 2-5). Although our demonstration takes place at a moderate 300K, the extremely limited data of our ML model results in highly perturbed configurations within the first few iterations of the simulation. Without the presence of our proposed Morse prior these configurations are far off equilibrium and often fail to converge by DFT. A similar demonstration under a more perturbed, higher temperature system is included in the SI with comparable success as early as the 3rd iteration - a 92% reduction in DFT calls.

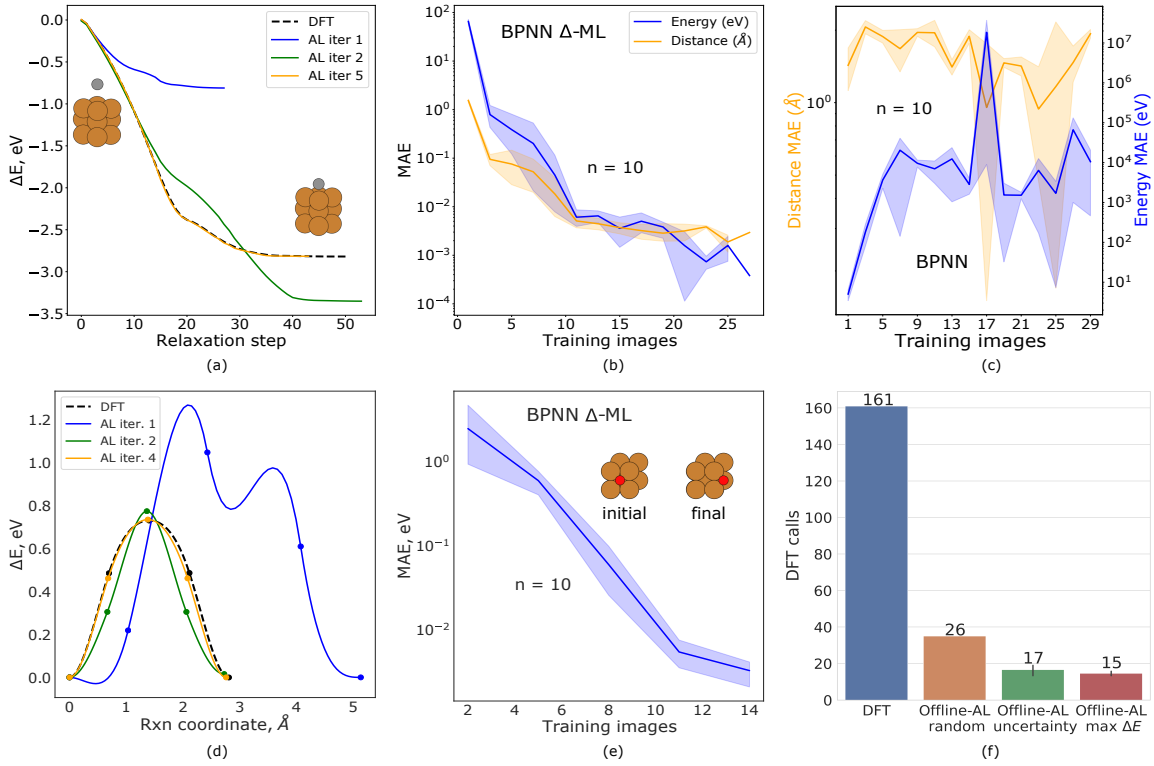


Figure 2-3: Offline-AL applications to structural relaxations and transition state calculations. **(a)** Evolution of the structural relaxation of C on Cu(100) over a few cycles of the Offline-AL **(b)** Relaxed structure and energy learning curves of the Offline-AL framework, using the BPNN Δ-ML model. **(c)** Convergence instability associated with not incorporating the Morse potential prior in an Offline-AL context. **(d)** Evolution of the transition state calculation for the surface diffusion of O on Cu(100). Despite the poor performance of the first iteration, the framework is able to recover and converge to an accurate prediction. **(e)** Learning curve associated with the energy barrier of the oxygen diffusion example of (d). **(f)** Total number of DFT calls queried by the Offline-AL under different querying strategies for the energy barrier associated with the diffusion of oxygen on copper. Error bars represent the 95% confidence interval.

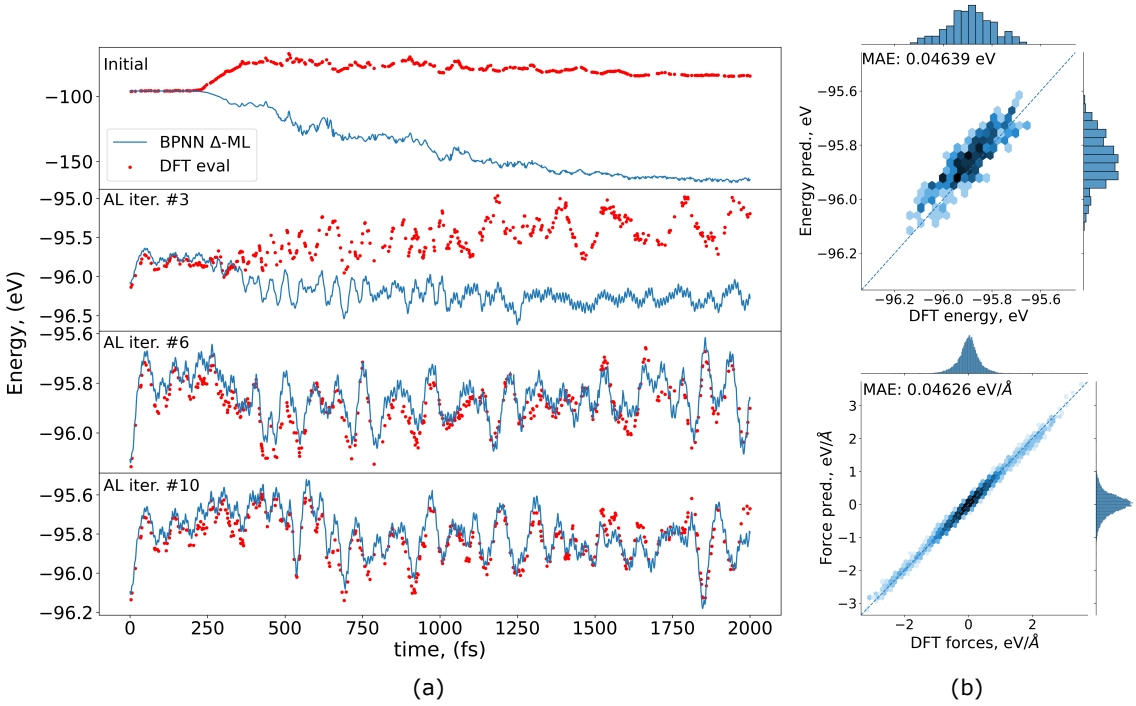


Figure 2-4: Offline-AL demonstration to a 2ps MD simulation of CO on Cu(100) **(a)** Evolution of the MD trajectory over several iterations of the active learning framework. We verify the effectiveness of our framework by randomly sampling configurations and comparing DFT evaluated energy and forces with that of our model's predictions. **(b)** Parity plots associated with the DFT evaluated configurations and our model's predictions of the 6th iteration, demonstrating good agreement. Shading was scaled logarithmically with darker shading corresponding to a higher density of points.

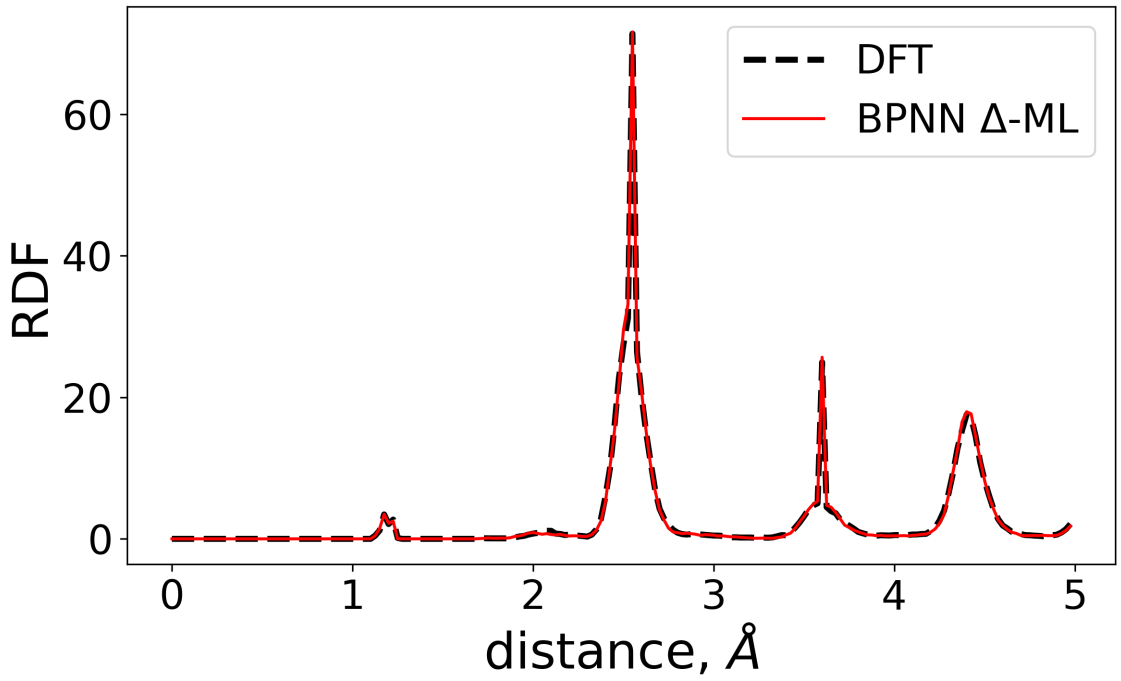


Figure 2-5: Radial distribution function (RDF) of the ground truth DFT and our framework's 6th iteration for the MD simulation of CO/Cu(100). Demonstrating good consistency even before the allotted number of iterations.

2.5 Conclusion

The development of accurate and reliable MLP has been a challenging task for the community. The careful curation of datasets is especially difficult in trying to generalize to new systems. Active learning has provided promising results in accelerating molecular simulations while minimizing risks of extrapolation. Neural-network based models, however, have struggled with such demonstrations for their reliance on large amounts of data. As deep learning research continues to make significant strides, understanding how to better incorporate neural network based MLPs into active learning pipelines can help provide more accurate and robust models.

This manuscript presented a neural-network based offline active-learning framework to accelerate a variety of molecular simulations beginning with extremely limited data. We introduced a physics-based prior, Morse potential, into our model in a Δ-ML manner, to capture basic repulsive interactions crucial in the convergence of our framework. We demonstrate the framework's ability to accurately converge simulations including structural relaxations, molecular dynamics simulations, and transition-state calculations. In each of these, the proportion of DFT calls reduced were 71%, 75%, and 91%, respectively. The framework presented is extremely flexible, allowing users to define their own querying strategies, termination criteria, and incorporate their own, more complex molecular simulations they wish to accelerate with AMP*torch*. Similar to other works, the nature of our active learning framework introduces assumptions and limitations surrounding the feasibility of DFT queries. While our framework helps in accelerating atomistic simulations, it's applicability is limited by the time it takes to query DFT points. Systems in which DFT calls may be infeasible (10,000+ atoms or far-from-equilibrium) will fail under this and other active learning strategies, leaving opportunities for the development of robust models trained on large datasets [40]. At this time we make no guarantees that the performance of the ML model will always improve when a queried data point is added to the dataset. Our experiments recognize this as a particular issue in the small data regime but was often mitigated in our work by the presence of the Morse potential

and more sophisticated learning rate schedulers, where otherwise would have failed. Future directions will explore more systematic querying strategies and termination criteria to further accelerate the framework while being robust to larger, more complex systems still compute feasible under DFT. Additionally, exploring alternative model priors and adversarial training techniques can help improve the performance, consistency, and generalizability of active learning frameworks [64, 251].

2.6 Calculation Settings

Single-point DFT calculations were performed *Quantum Espresso (QE) [78]* implemented in *ASE* [143]; using the *PBE* exchange-correlation functional [197]; a plane wave basis set with an energy-cutoff of 500 eV; k-points of $4 \times 4 \times 1$; and the pseudopotentials provided by Garrity, et al. [74]. The same settings were also used for DFT calculations in fitting the Morse potential parameters. The following tools and settings were used for our DFT calculations: *VASP 5.4.4.18* [137, 138]; using the *PBE* exchange-correlation functional; a plane wave basis set with an energy-cutoff of 400eV; and k-points of $4 \times 4 \times 1$. *VASP* was used for all structure relaxation and MD examples and *QE* for the NEB examples. AMP*torch* [240] was used for all machine learning and active learning components of the framework.

Chapter 3

The Open Catalyst 2020 (OC20) Dataset and Community Challenges

This work originally appeared as: Chanussot, L., Das, A.*, Goyal, S.*, Lavril, T.*, Shuaibi, M.*, Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., Palizhati, A., Sriram, A., Wood, B., Yoon, J., Parikh, D., Zitnick, C.L., and Ulissi, Z., 2021. Open catalyst 2020 (OC20) dataset and community challenges. ACS Catalysis, 11(10), pp.6059-6072. It has been edited to include the supplementary information in Appendix B. *These authors contributed equally.*

My contribution in this work included baseline model and repository implementation, data preprocessing, task and metric development, model training and evaluation, and writing corresponding sections in the manuscript. I was also the primary editor of the entire manuscript, handling all reviewer correspondence and revisions.

3.1 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 elemental compo-

sitions of surfaces and adsorbate identity/configurations, perhaps because datasets have been smaller in catalysis than related fields. To address this we developed the OC20 dataset, consisting of 1,281,040 Density Functional Theory (DFT) relaxations (∼264,890,000 single point evaluations) across a wide swath of materials, surfaces, and adsorbates (nitrogen, carbon, and oxygen chemistries). We supplemented this dataset with randomly perturbed structures, short timescale molecular dynamics, and electronic structure analyses. The dataset comprises three central tasks indicative of day-to-day catalyst modeling and comes with pre-defined train/validation/test splits to facilitate direct comparisons with future model development efforts. We applied three state-of-the-art graph neural network models (CGCNN, SchNet, DimeNet++) to each of these tasks as baseline demonstrations for the community to build on. In almost every task, no upper limit on model size was identified, suggesting that even larger models are likely to improve on initial 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.

3.2 Introduction

Advancements to renewable energy processes are needed urgently to address climate change and energy scarcity around the world [177, 61]. 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 efficiencies.[183, 236, 181] 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 DFT [239] 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.[165, 166] 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[269, 146, 90], thermochemistry prediction [156, 117, 124, 265, 266, 89, 15, 126, 63], structure optimization [252, 145, 99, 229, 4], accelerating individual calculations[33, 125, 198, 253], and integration with characterization[261] (see recent reviews for a more thorough discussion [129, 151, 166, 81, 228, 147, 254, 44, 10, 88, 264, 87, 255]). Most of these tasks are variations on the same fundamental problem: modeling heterogeneous catalysis. The dataset developed seeks to target a specific subclass of this problem, periodic slab models. Such modeling involves predicting the energy and forces of various configurations of adsorbate molecules at inorganic interfaces.

Unfortunately, modeling of heterogeneous catalysts entails all the known difficulties of modeling both organic and inorganic chemistry. In organic chemistry modeling involves an overwhelming space of molecules and reactions and many similar, low-energy conformers. Inorganic chemistry involves a 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 the importance of heterogeneous catalysis, datasets for it remain smaller than those in other related fields[51, 128] due to additional complexity and higher computational cost. Much of the progress in applying AI/ML in heterogeneous catalysis has been driven by increasingly large and diverse datasets of electronic structure calculations.

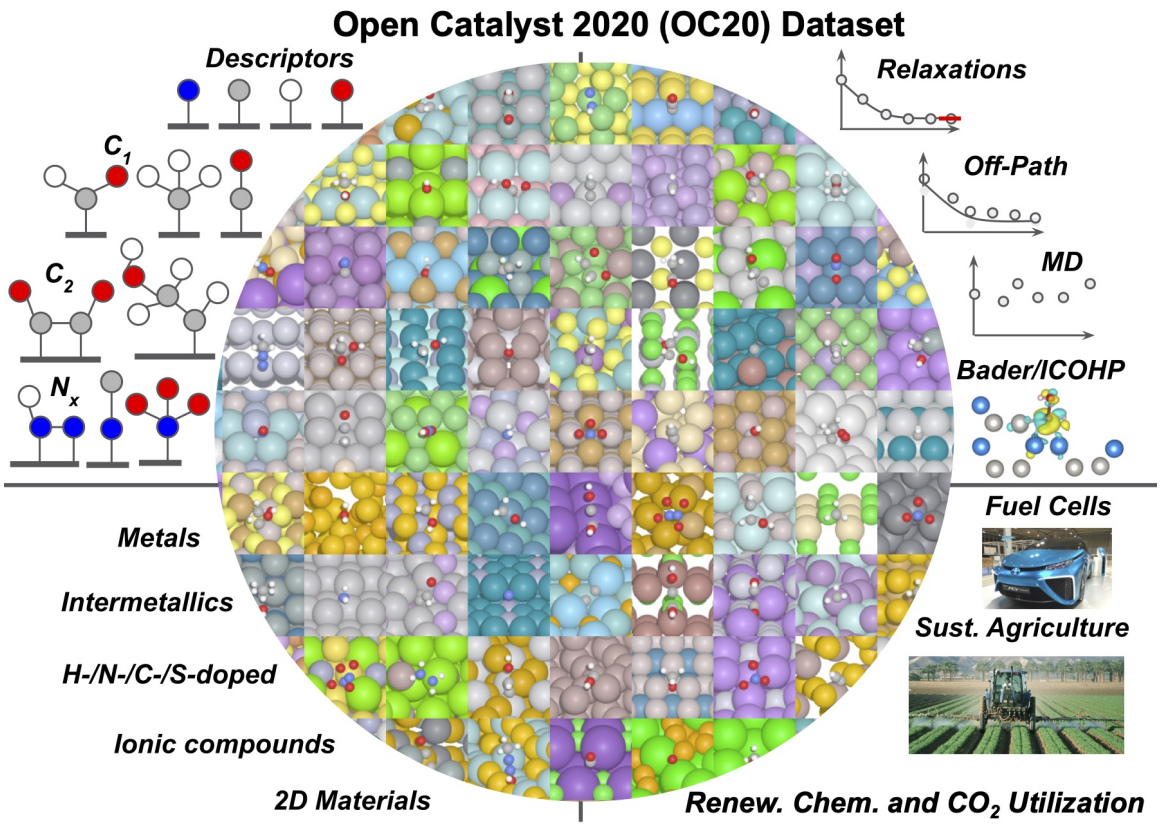


Figure 3-1: Adsorbates, materials, calculations, and impact areas of the OC20 dataset. Images are a random sample of the dataset.

In the past few years there has been a push towards larger datasets in catalysis, going from O(100) [38, 2, 160, 179, 5] to O(1,000) [58, 150, 23] then O(100,000) [161, 265, 283] 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 representations. As has been shown in other ML tasks [55, 191, 9], we expect that significantly larger datasets will lead to improved accuracy and better generalization.

In this paper, we present the OC20 dataset, (Figure 3-1) which comprises over 1.2 million DFT relaxations of molecular adsorptions onto surfaces (*ca.* 250 million single-point calculations) across a substantially larger structure and chemistry space

than previously realized. We envision OC20 to serve as a crucial stepping stone in the development of ML models for practical catalysis applications.

While a dataset of this magnitude will lead to significant improvements in ML models, 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) that are relevant for renewable energy and environmental applications. Relaxations are performed on randomly sampled low-Miller-index facets of stable materials from the Materials Project [114], resulting in surfaces from 55 different elements and mixtures thereof. For each of the calculations, we include relaxation trajectories, Bader charges, and LOBSTER [101, 16]-calculated orbital information. To aid in training more robust models, we additionally compute short, high-temperature *ab initio* Molecular Dynamics (MD) trajectories on a randomly sampled subset of the relaxed states. We also randomly perturb the atomic positions in a subset of the structures along the relaxation pathways and perform single point DFT calculations for these perturbed/rattled structures. We recognize that OC20 addresses a simplified version of heterogeneous catalysis - single adsorbates on idealized structures. Although useful as a first step to informing reaction pathways, the reality involves a number of additional complexities that impact catalyst performance, including reaction conditions, solvation effects, kinetics, etc. While we believe OC20's approximations to be a reliable step forward, it is important to understand the limits of models developed from this dataset. Future work that incorporates more of the complexities mentioned will undoubtedly benefit from the developments related to OC20. The dataset is publicly available at `http://opencatalystproject.org`. We also plan to upload the dataset to other open systems (e.g. NOMAD or Zenodo) for long-term availability.

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. 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/Open-Catalyst-Project/ocp`) containing a set of baseline models, data loaders, and training scripts for each of these tasks. While we focus on a subset of tasks, we believe that models capable of solving these tasks on the OC20 dataset will also be able to address a large number of related catalysis problems.

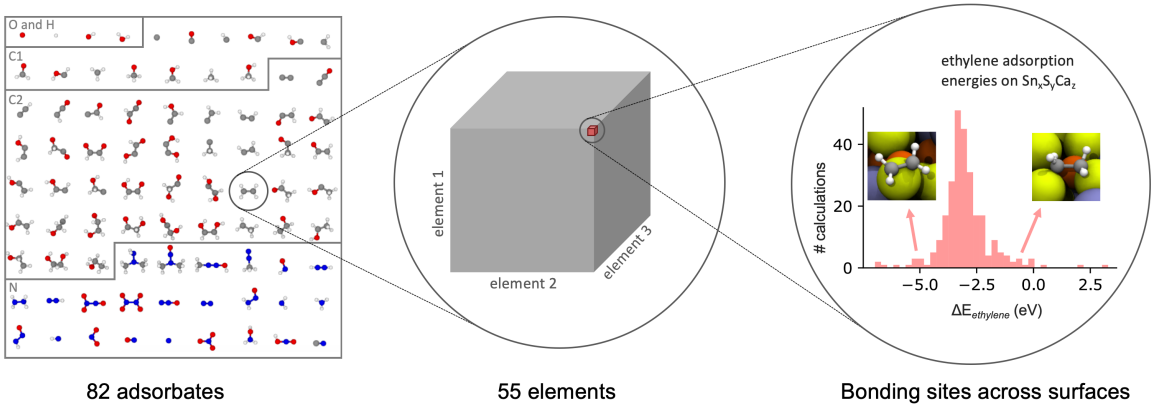


Figure 3-2: The adsorbates used to generate the Open Catalyst Dataset contain oxygen, hydrogen, C_1, C_2, and nitrogen molecules useful for renewable energy applications. Adsorbates that contain both carbon and nitrogen were counted both as C_X adsorbates and as nitrogen-containing adsorbates. For each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

3.3 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 bottlenecks 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 all directions with a vacuum layer of at least 20Å applied in the **z** direction. 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.

S2EF is to take the positions of the atoms as input and predict the energy and per-atom forces as calculated by DFT. For the purposes of this manuscript, energy refers to adsorption energy unless otherwise noted. The 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. 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 [276]. However, the inclusion of both inorganic and organic materials and the dataset size make this challenge unique.

IS2RS takes as input an initial structure and predicts the atomic positions in their final, relaxed state. Traditional relaxations are performed through an iterative process that estimates the atomic forces using DFT, which are in turn used to update atom positions until convergence. This very computationally expensive process typically requires hundreds of DFT calculations to converge.

If the *IS2RS* task is approached using ML approximations of DFT to estimate atomic 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 (Figure 3-3(B)), 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.

IS2RE task is to take the initial structure as input and predict 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.

3.4 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 640,081 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 training 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 (MD) data. We also computed Bader charges and LOBSTER analyses (over 1.8 million examples each) as these computed properties may be useful for models by explaining why the energies are what they are.

3.4.1 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/Open-Catalyst-Project/Open-Catalyst-Dataset`).

Adsorbate Selection

Adsorbates are sampled randomly from a set of 82 molecules that are chosen for their utility to renewable energy applications. As shown in Figure 3-2, 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 11,451 materials in the Materials Project[114] 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[187] 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[187]. 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-optimize the bulk structures prior to surface enu-

meration to ensure differences between the DFT settings used in the Materials Project and OC20 did not induce unintended stress or strain effects. Any bulks that we could not successfully relax were omitted from this dataset.

Task	Train	In Domain	OOD Adsorbate	OOD Catalyst	OOD Both
S2EF	133,934,018	987,036	999,838	987,343	997,922
IS2RS	460,328	24,733	24,961	24,738	24,971
IS2RE	460,328	24,733	24,961	24,738	24,971

Table 3.1: Size of train/validation splits (number of structures for *S2EF* and initial structures for *IS2RS* and *IS2RE*). The structures for *S2EF* are sampled from 640,081 relaxations for train, and from 30k-70k relaxations for each validation and test split. Subsplits of validation and test are roughly the same size, but are exclusive of each other. Subsplits include sampling from the same distribution as training (In Domain), unseen adsorbates (OOD Adsorbate), unseen element compositions for catalysts (OOD Catalyst), and unseen adsorbates and catalysts (OOD Both). Test sizes are similar.

Initial Structure Generation

The initial structures are generated by placing the selected adsorbates on the selected surfaces using CatKit [34] and the atomic simulation environment (ASE) [143]. Surface atoms are identified by their positions above the center-of-mass, their z-distance within 2 Å of the upper-most atom, and by their under-coordination relative to the bulk atoms. Atomic coordination environments were calculated using pymatgen's Voronoi tesselation algorithm [187]. 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.[34] CatKit used this information to enumerate a list of symmetrically distinct adsorption sites along with suggested per-site orientations for the adsorbates. From this list, an adsorption configuration is randomly selected. The sites selected are not necessarily the most stable adsorption site on each surface. Since one of our goals is to calculate adsorption energies, we generate two sets of 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 1,919,165 and 616,124 unique inputs for

(1) and (2), respectively, which were later filtered and segregated into suitable train, validation, and test validation splits as described later in this section.

Structure Relaxation

All structure relaxations were performed using the Vienna Ab Initio simulation Package (VASP) [140, 138, 139, 273, 141] until all per-atom forces are less than 0.03 eV/Å. Calculations were allowed up to 144 hours (12 cores) for the relaxation. Systems that timed out before 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 adsorbate and 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. Given the intended scale of OC20, the careful consideration of DFT settings was a non-trivial challenge. Relaxations generally followed previous high-throughput catalysis efforts with reasonable trade-offs between accuracy for surface chemistry and computational cost[266] (VASP [140, 138, 139, 273, 141], RPBE[197], no spin polarization, etc). The choices made for DFT were a result of several important considerations: ensuring calculations were representative, concerns associated with inconsistent cutoffs/settings, and representative of typical numerical/convergence issues the computational chemistry field faces. The assumptions made were necessary to achieve the dataset's scale. Detecting small numerical or convergence errors is a non-trivial problem that could be improved with this dataset. Most importantly, we anticipate models and methods that solve the S2EF, IS2RE, or IS2RS tasks for this dataset are very likely to solve future challenges for future surface science datasets with different DFT modeling choices.

System DFT energies were referenced to represent adsorption energies. Adsorption energies were calculated according to the Equation below, where E_{sys} is the DFT energy of the combined surface (i.e. slab) and adsorbate — this energy can be from both relaxed and intermediate structures. The reference energies for each system,

E_{slab} and E_{gas} are the DFT energy of the relaxed surface and adsorbate molecule respectively. The value of E_{gas} for each adsorbate was computed as a linear combination of N_2, H_2O, CO, and H_2 resulting in the atomic energies found in the supplementary.

$$E_{ad} = E_{sys} - E_{slab} - E_{gas}$$

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.

MD 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 MD in VASP [140, 138, 139, 273, 141] and (2) normally-distributed random position perturbation methods colloquially known in molecular simulations as "rattling."

MD calculations simulate the atomic interactions when heat is added to the system. Partial MD 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 MD 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 with ASE [143]. 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 MD 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 MD (*ca.* 64 million single-point energies/forces) and 30 million rattled 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 [256, 224, 101] and LOBSTER [176, 56] analyses) were carried out on relaxed structures and also on randomly selected snapshots from both MD and rattled trajectories. Bader charge analyses provides charge density maxima at each atomic center and the Bader volume for each atom through the zero-flux partitioning method [16]. LOBSTER enables chemical-bonding analysis based on periodic DFT outputs [176]. 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 [82, 173, 39].

Dataset profile

Approximately 872,000 adsorption energies were calculated successfully. Of these, 3.7% were calculations on unary catalysts; 61.4% were on binaries; and 34.9% were on ternaries. Among these calculations, 28.9% of them had reactive nonmetal elements in the catalyst; 8.1% of them had alkali metals; 10.2% had alkaline earth metals; 26.4% had metalloids; 81.3% had transition metals; and 37.2% had post-transition metals. Considering adsorbates: 6.6% of the calculations had adsorbates containing only oxygen or hydrogen; 25.2% of the calculations had C_1 adsorbates; 44.4% had C_2 adsorbates; and 29.0% had nitrogen-containing adsorbates.

Despite this dataset's large size compared to previous catalytic datasets, it still very small 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.

3.4.2 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 [127, 167].

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 extrapolative subsplits for validation and testing. 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 3.1). 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 5.2, 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.

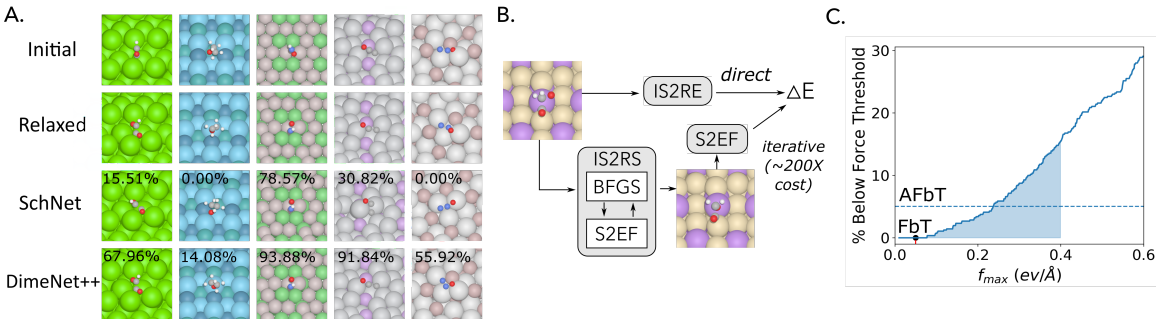


Figure 3-3: Demonstration of baselines SchNet and DimeNet++ models for solving the *IS2RE*, *S2EF*, and *IS2RS* tasks and the inter-relationships. (A) Snapshots of five representative initial adsorbate configurations before DFT relaxations, the same adsorbates after DFT relaxation, and the relaxed structures as relaxed by SchNet and DimeNet++ after fitting the *S2EF* task. ADwT metrics are overlaid on the model snapshots. (B) 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 force-only 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) of 0.4 eV/Å(shaded area).

3.5 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, 194] are publicly available at the Open Catalyst Project (`http://opencatalystproject.org`).

Our baseline ML approaches are all based on Graph Neural Networks (GNNs) [95] 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 [27, 21], 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. [286] 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) [286], SchNet [232] and DimeNet++ [131, 133]. 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. DimeNet++, an extension to DimeNet, replaces the Bilinear layer with a Hadamard product and additional multilayer perceptrons; providing reported speed improvements of 8x and a 10% accuracy boost on QM9 [212].

For all approaches, graph edges were determined by a nearest neighbor search limited by a cutoff radius of 6Å, retaining up to 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, 192 for CGCNN, SchNet and DimeNet++ respectively unless stated otherwise; resulting in 3.6 million (CGCNN), 7.4 million (SchNet) and 1.8 million (DimeNet++) parameters. Model sizes were chosen so that runtimes were roughly equivalent. Note the size of the models was increased from their original implementations to account for OC20's larger size. Model hyperparameters and additional modifications can be found in the supplementary.

Since both the computed energies and forces are evaluated, the baseline loss function [133, 125] uses the following form:

$$\mathcal{L} = \lambda_E \sum_i |E_i - E_i^{DFT}|$$
$$+\lambda_F \sum_{i,j} \frac{1}{N_i} |F_{i,j} - F_{i,j}^{DFT}|,$$

where λ_E and λ_F are empirical parameters, 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 ($\lambda_F = 0$).

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 [201] is parameterized for all of the elements in this dataset and is fast enough (approx 1,000X 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 [17] 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.

3.6 Experiments

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

3.6.1 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.

Force cosine: Mean cosine of the angle between the computed per-atom forces and the ground-truth forces. Similar to MAE, these are only computed for free 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 an 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 (Forces below Threshold) 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 3-3(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.

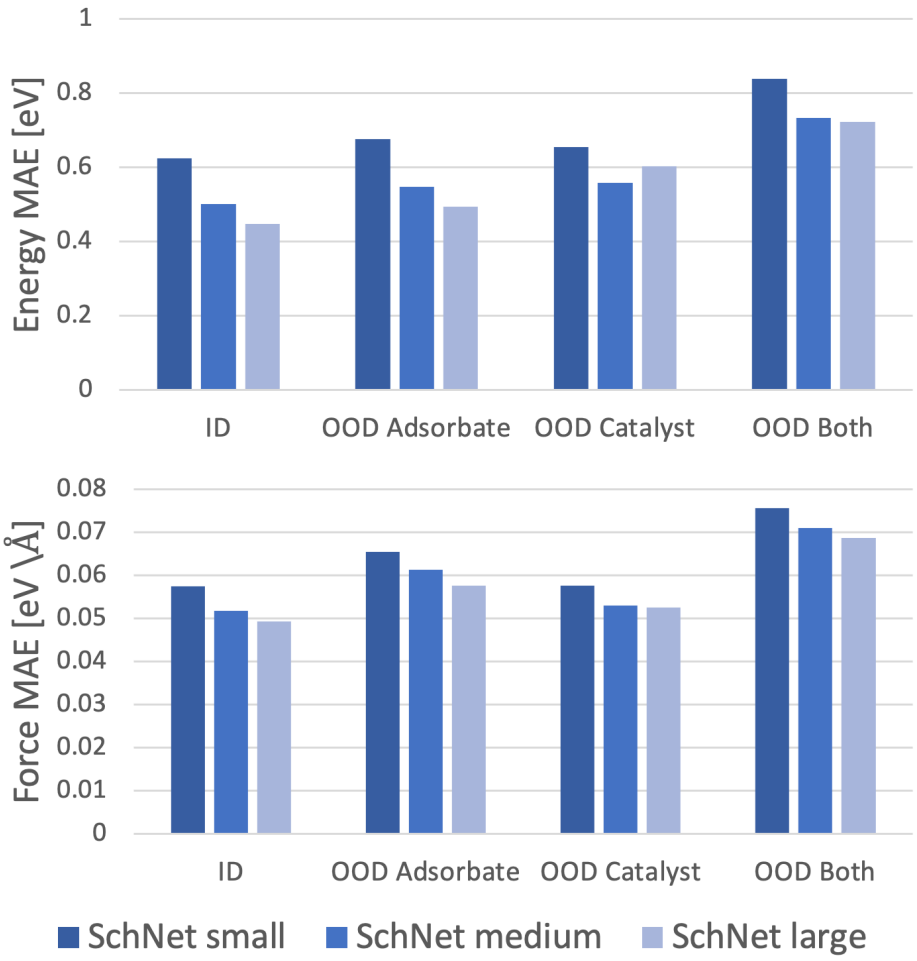


Figure 3-4: Predicting Structure to Energy and Forces (*S2EF*) as evaluated by Mean Absolute Error (MAE) of the energies and forces. The small, medium and large SchNet models have the following sizes: Small: 256 hidden, 4 message-passing layers, 1,316,097 params, Medium: 1024 hidden, 3 message-passing 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.

***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

	S2EF Test			
Model	ID	OOD Ads	OOD Cat	OOD Both
	Energy MAE [eV] ↓			
Median baseline	2.0596	2.4188	2.0110	2.5460
CGCNN [286]	0.5105	0.6321	0.5202	0.7681
SchNet [232]	0.4421	0.4858	0.5279	0.7057
SchNet [232] – force-only	34.0689	33.7670	35.2701	38.4607
SchNet [232] – energy-only	0.3975	0.4533	0.5626	0.7241
DimeNet++ [133, 131]	0.4579	0.4701	0.5056	0.6489
DimeNet++ [133, 131] – force-only	28.2214	28.9404	28.8636	34.9118
DimeNet++ [133, 131] – energy-only	0.3585	0.4022	0.5041	0.6549
DimeNet++ [133, 131]-Large – force-only	29.3504	30.0338	30.0074	36.7665
	Force MAE [eV/Å] ↓			
Median baseline	0.0808	0.0801	0.0787	0.0978
CGCNN [286]	0.0683	0.0728	0.0670	0.0851
SchNet [232]	0.0493	0.0529	0.0509	0.0655
SchNet [232] – force-only	0.0442	0.0469	0.0459	0.0591
SchNet [232] – energy-only	0.5794	0.5974	0.5852	0.6463
DimeNet++ [133, 131]	0.0442	0.0458	0.0444	0.0559
DimeNet++ [133, 131] – force-only	0.0331	0.0341	0.0340	0.0417
DimeNet++ [133, 131] – energy-only	0.3399	0.3395	0.3395	0.3643
DimeNet++ [133, 131]-Large – force-only	0.0280	0.0289	0.0312	0.0371
	Force cosine ↑			
Median baseline	0.0000	0.0000	0.0000	0.0000
CGCNN [286]	0.1541	0.1369	0.1492	0.1444
SchNet [232]	0.3184	0.2954	0.2956	0.2987
SchNet [232] – force-only	0.3595	0.3391	0.3279	0.3403
SchNet [232] – energy-only	0.0845	0.0798	0.0804	0.0830
DimeNet++ [133, 131]	0.3628	0.3476	0.3465	0.3684
DimeNet++ [133, 131] – force-only	0.4870	0.4717	0.4607	0.4954
DimeNet++ [133, 131] – energy-only	0.1066	0.0959	0.1048	0.1015
DimeNet++ [133, 131]-Large – force-only	0.5638	0.5502	0.5115	0.5516
	EFwT ↑			
Median baseline	0.00%	0.00%	0.00%	0.00%
CGCNN [286]	0.01%	0.00%	0.01%	0.00%
SchNet [232]	0.11%	0.04%	0.06%	0.01%
SchNet [232] – force-only	0.00%	0.00%	0.00%	0.00%
SchNet [232] – energy-only	0.00%	0.00%	0.00%	0.00%
DimeNet++ [133, 131]	0.10%	0.03%	0.05%	0.01%
DimeNet++ [133, 131] – force-only	0.00%	0.00%	0.00%	0.00%
DimeNet++ [133, 131] – energy-only	0.00%	0.00%	0.00%	0.00%
DimeNet++ [133, 131]-Large – force-only	0.00%	0.00%	0.00%	0.00%

Table 3.2: 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.

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

	IS2RS Test			
Model	ID	OOD Ads	OOD Cat	OOD Both
		ADwT ↑		
IS baseline	21.37%	19.09%	21.42%	26.28%
SchNet [232]	15.92%	12.83%	14.63%	14.78%
SchNet [232] – force-only	32.47%	28.59%	30.94%	35.09%
DimeNet++ [133, 131]	30.62%	26.66%	30.01%	32.29%
DimeNet++ [133, 131] – force-only	48.73%	45.19%	48.54%	53.17%
DimeNet++ [133, 131]-Large – force-only	52.43%	48.47%	50.91%	54.85%
		FbT ↑		
IS baseline	0.00%	0.00%	0.00%	0.00%
SchNet [232]	-	-	-	-
SchNet [232] – force-only	0.00%	0.00%	0.00%	0.00%
DimeNet++ [133, 131]	0.00%	0.20%	0.00%	0.00%
DimeNet++ [133, 131] – force-only	0.61%	0.20%	0.00%	0.20%
DimeNet++ [133, 131]-Large – force-only	1.02%	0.40%	0.00%	0.20%
		AFbT ↑		
IS baseline	0.06%	0.34%	0.21%	0.00%
SchNet [232]	-	-	-	-
SchNet [232] – force-only	5.31%	2.82%	2.66%	2.73%
DimeNet++ [133, 131]	3.60%	3.01%	2.61%	2.33%
DimeNet++ [133, 131] – force-only	17.42%	14.67%	14.12%	14.46%
DimeNet++ [133, 131]-Large – force-only	25.58%	20.73%	20.05%	20.62%

Table 3.3: 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. The initial structure was used as a naive baseline (IS baseline). FbT and AFbT metrics are only computed when ADwT metrics are greater than 20.26%.

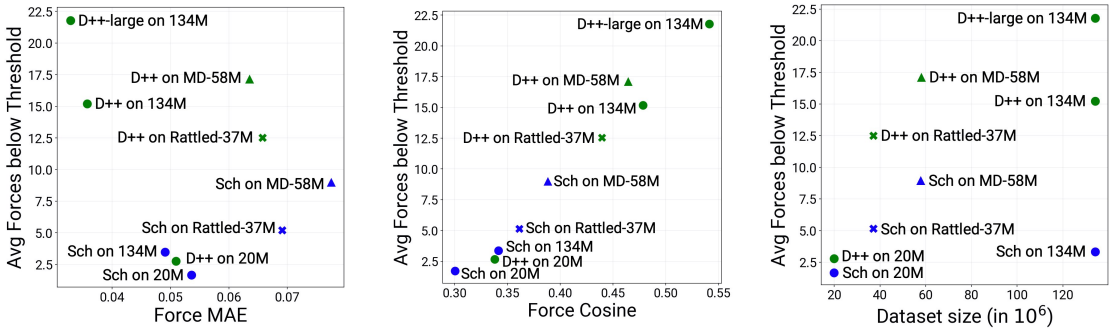
Energy MAE: Mean Absolute Error between the computed relaxed energy and the ground truth relaxed energy.

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 10 ms which would significantly improve the applicability of DFT. Significantly faster than this (closer in speed to

		IS2RE Test							
		Energy MAE [eV] ↓				EwT ↑			
Model	Approach	ID	OOD Ads	OOD Cat	OOD Both	ID	OOD Ads	OOD Cat	OOD Both
Median baseline	-	1.7489	1.8911	1.7107	1.6807	0.75%	0.69%	0.83%	0.78%
CGCNN [286]	Direct	0.6135	0.9155	0.6211	0.8506	3.41%	1.93%	3.11%	1.99%
SchNet [232]	Direct	0.6372	0.7342	0.6611	0.7035	2.96%	2.33%	2.95%	2.22%
DimeNet++ [133, 131]	Direct	0.5605	0.7252	0.5750	0.6613	4.26%	2.06%	4.10%	2.42%
SchNet [232]	Relaxation	0.7088	0.7741	0.7665	0.8055	4.23%	2.63%	3.52%	2.52%
SchNet [232] – force-only + energy-only	Relaxation	0.7066	0.7420	0.7966	0.7493	4.18%	2.98%	3.39%	2.70%
DimeNet++ [133, 131]	Relaxation	0.6687	0.6864	0.6858	0.6835	4.29%	3.36%	3.79%	3.51%
DimeNet++ [133, 131] – force-only + energy-only	Relaxation	0.5112	0.5744	0.5922	0.6130	6.14%	4.29%	5.10%	3.84%
DimeNet++ [133, 131] – large force-only + energy-only	Relaxation	0.5022	0.5430	0.5780	0.6117	6.58%	4.34%	5.09%	3.93%

Table 3.4: 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 models trained on the All training dataset.



(a) (b) (c)

Figure 3-5: Results of force-only SchNet (denoted by 'Sch') and DimeNet++ ('D++') *S2EF* models trained on S2EF-20M, S2EF-100M, S2EF-20M + Rattled ('Rattled-37M') and S2EF-20M + MD ('MD-58M') dataset splits used to drive relaxations from given initial structures (*IS2RS*). We plot *IS2RS* AFbT performance against *S2EF* force cosine, *S2EF* force MAE and number of training samples for the different variants. 3-5a,3-5b: *IS2RS* AFbT seems to correlate better with *S2EF* force cosine than *S2EF* force MAE, especially when analyzing models trained on Rattled-37M or MD-58M data. 3-5c: Further, both DimeNet++ and SchNet achieve higher AFbT when trained on MD-58M than S2EF-134M. Additional MD data seems to offer a stronger learning signal than additional S2EF data.

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.

3.6.2 Leaderboard

To ensure consistent and fair evaluation, a public leaderboard is available on the Open Catalyst Project webpage (`http://opencatalystproject.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.

3.6.3 Results

To provide baselines for the OC20 dataset, we report results using three state-of-the-art approaches: CGCNN [286], SchNet [232], and DimeNet++ [131, 133]. Details of the models' implementations can be found in the Baselines Section.

***S2EF*:** Results on CGCNN [286], SchNet [232], and DimeNet++ [133, 131] are evaluated. All approaches predict structure energies in their forward pass and per-atom forces by the negative gradient of the predicted energy with respect to atomic positions [203]. Across most metrics DimeNet++ performs the best, with SchNet marginally outperforming DimeNet++ and CGCNN on EFwT. SchNet outperforms CGCNN across all metrics. Since tradeoffs exist in the prediction of energy and forces, we trained three variants of SchNet and DimeNet++ with $\{\lambda_E, \lambda_F\} = \{1, 30\}, \{0, 100\}, \{100, 1\}$ for SchNet/DimeNet++, SchNet/DimeNet++ force-only and SchNet/DimeNet++ energy-only respectively. As expected, the energy-only model performs best on energy MAE, while the force-only performs best on force MAE. DimeNet++ and SchNet both provide a balance between the two and the best results on EFwT. All approaches perform badly on the EFwT metric; indicating that the results are still far from being practically useful. Table 3.2 and Figure 3-4 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 between In Domain, OOD Adsorbate and OOD Catalyst, which shows some generalization to new adsorbates and catalysts. Increases in training data sizes results in significant improvements,

Figure 3-6(A). The rate and amount of improvement varies based on the model. Finally, wider and deeper models are shown to improve accuracies in Figure 3-4. Both increased depth (Medium to Large) and width (Small to Medium) show improvements.

Figure 3-6: (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) using force-only models. (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, and All dataset sizes. *S2EF* and *IS2RE* values averaged across validation subsplits. *IS2RS* values evaluated on the test in-domain (ID) subsplit.

***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 [154], examples are shown in Figure 3-3(A). Table 3.3 shows that DimeNet++ outperforms SchNet in the ADwT and AFbT metrics. However, the FbT metrics indicate both methods do not produce relaxed structures with forces below thresholds used in practice. Since only the computed forces are used for the IS2RS task and not the energies, it is not surprising that the DimeNet++ force-only model performs the best. It was trained using only force losses and performs significantly better on AFbT and ADwT, but still is near zero when measured by FbT. A plot of FbT across thresholds from 0.01 to 0.6 for SchNet is shown in Figure 3-3(C). Both methods show better generalization to new adsorbates vs new catalyst material compositions. Similar to *S2EF*

improved results are found with more training data, especially for DimeNet++ and SchNet, Figure 3-6(B). Experiments using the additional rattled and MD data are shown in Figure 3-5. Interestingly, the force cosine metric appears to better correlate with AFbT scores than force MAE. A discussion on these results may be found in the supplementary.

***IS2RE*:** For *IS2RE* we explore two pathways for computing the relaxed energy from the initial state, Figure 3-3(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.

As shown in Table 3.4, the hybrid relaxation approaches outperformed the direct across all metrics. The percentage of predicted energies within a tight threshold (EwT) ranged from 2% to 6%; indicating that accuracies are still below practical usefulness. Generalization to new catalyst compositions performed better than new adsorbates. As shown in Figure 3-6(C), larger dataset sizes could significantly improve performance. The best direct-based approach, DimeNet++, was evaluated via the relaxation-based approach. The use of DimeNet++ force-only to perform the relaxation, followed by DimeNet++ energy-only to compute the relaxed energy significantly outperformed the use of a single model (optimized for EFwT) to compute both. Best metrics were achieved using the large DimeNet++ force-only model, followed by DimeNet++ energy-only.

3.7 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. Motivated by previous efforts[109], we analyzed model performance for increasing dataset sizes to illustrate the differences between catalysis and related efforts—e.g.,

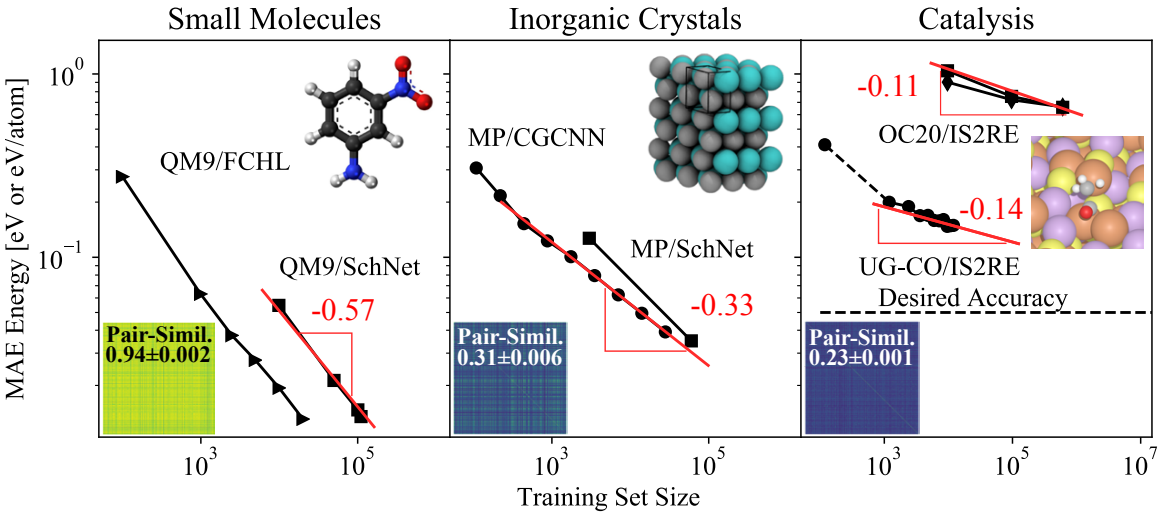


Figure 3-7: Model performance versus dataset size across three related atomistic domains. Insets are pairwise similarity for selected structures from the respective dataset using GraphDot (see the SI for details) (0/dark-blue/not-similar to 1/yellow/identical)[257, 258]. (left) Results [276] for FCHL/SchNet models trained on the QM9 small molecule dataset (slope -0.57). (middle) Models[286, 232] trained on Materials Project formation energies (slope -0.33, more difficult). (right) Results for catalysis including a literature dataset for CO adsorbates [266] 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.

materials sciences or small molecule property prediction. Figure 3-7(left) and Figure 3-7(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 DimeNet++ 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 3-7(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, this 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 [286], 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 [168, 35, 7, 178]. Element embeddings and representations will be important to scale across materials. 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 the size of this dataset is larger by 2 orders of magnitude than previous catalyst DFT dataset efforts [266, 112]. 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 [100, 208]. The largest baseline models with *ca.* 10 million 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 enable transition state calculations, kinetic approximations, vibrational frequency calculations, and the more routine use of long timescale molecular dynamics for studying these systems. Sensitivity analyses will be necessary to understand the level of accuracy needed for models to be practically relevant for varying applications. Given the sparsity and breadth of OC20, the availability of relevant experimental data will also be a crucial challenge in the next stage of validating model results with experiments. 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.

3.8 Supporting Information Available

The supporting information contains details on the precise DFT calculation methods, the adsorption energy reference energies, the adsorbates and their assuming binding configurations, details on graph construction, description of the graph similarity metrics, a few sample GFN0-XTB relaxations, the precise train/test/validation splits, details on the modified CGCNN/SchNet/DimeNet++ implementations, results on the Rattled/MD experiments, hyperparameters for baseline models, a list of adsorbates in OC20, and full results on the validation splits. The full open dataset is provided at `http://opencatalystproject.org` in accessible extxyz format, and the baseline models are provided as an open source repository at `https://github.com/Open-Catalyst-Project/ocp`.

Chapter 4

Rotation Invariant Graph Neural Networks using Spin Convolutions

This work originally appeared as: Shuaibi, M., Kolluru, A., Das, A., Grover, A., Sriram, A., Ulissi, Z. and Zitnick, C.L., 2021. Rotation invariant graph neural networks using spin convolutions. arXiv preprint arXiv:2106.09575.

4.1 Abstract

Progress towards the energy breakthroughs needed to combat climate change can be significantly accelerated through the efficient simulation of atomic systems. Simulation techniques based on first principles, such as Density Functional Theory (DFT), are limited in their practical use due to their high computational expense. Machine learning approaches have the potential to approximate DFT in a computationally efficient manner, which could dramatically increase the impact of computational simulations on real-world problems.

Approximating DFT poses several challenges. These include accurately modeling the subtle changes in the relative positions and angles between atoms, and enforcing constraints such as rotation invariance or energy conservation. We introduce a novel approach to modeling angular information between sets of neighboring atoms in a graph neural network. Rotation invariance is achieved for the network's edge messages

through the use of a per-edge local coordinate frame and a novel spin convolution over the remaining degree of freedom. Two model variants are proposed for the applications of structure relaxation and molecular dynamics. State-of-the-art results are demonstrated on the large-scale Open Catalyst 2020 dataset. Comparisons are also performed on the MD17 and QM9 datasets.

4.2 Introduction

Many of the world's challenges such as finding energy solutions to address climate change [298, 40] and drug discovery [212, 237] are fundamentally problems of atomic-scale design. A notable example is the discovery of new catalyst materials to drive chemical reactions that are essential for addressing energy scarcity, renewable energy storage, and more broadly climate change [298, 216]. Potential catalyst materials are typically modeled using Density Functional Theory (DFT) that estimates the forces that are exerted on each atom and the energy of a system or structure of atoms. Unfortunately, the computational complexity of DFT limits the scale at which it can be applied. Efficient machine learning approximations to DFT calculations hold the potential to significantly increase the discovery rate of new materials for these important global problems.

Graph Neural Networks (GNNs) [84, 296] are a common approach to modeling atomic structures, where each node represents an atom and the edges represent the atom's neighbors [234, 79, 119, 232, 235, 286, 206, 133]. A significant challenge in designing models is utilizing relative angular information between atoms, while maintaining a model's invariance to system rotations. Numerous approaches have been proposed, such as only using the distance between atoms [232, 235, 286], or limiting equivariant angular representations to linear transformations to maintain equivariance [280, 25, 7, 260]. One promising approach is the use of triplets of neighboring atoms to define local coordinate frames that are invariant to system rotations [133, 132]. The relative angles between the three atoms may be used to update the GNN's messages while maintaining the network's invariance to rotations. It has been shown that this

additional angular information results in significantly improved accuracies on several tasks [133, 132, 40].

We propose encoding angular information using a local reference frame defined by only two atoms; the source and target atoms for each edge in a GNN. Using this reference frame, a spherical representation of the incoming messages to the source atom is created, Figure 4-1. The representation has the benefit of encoding all neighboring atom information, and not just information between atom triplets, which may result in higher-order information being captured. The complication is a reference frame defined by two atoms (or two 3D points) still has one remaining degree of freedom - the roll rotation about the axis defined by the two 3D points. If this final degree of freedom is not accounted for, the model will not be invariant to system rotations. Our solution is to perform a convolution on the spherical representation across this final rotation, called a "spin convolution". By globally pooling the convolution's features, the resulting SpinConv model maintains rotation invariance while enabling the capture of rich angular information.

We describe two model variations that are used depending on the importance of energy conservation in the final application. We propose an energy-centric model that enforces energy conservation by calculating the forces using the negative partial derivative of the energy with respect to the atoms' positions [48]. Our second approach is a force-centric model that directly estimates the atom forces that is not energy conserving. While the force-centric model's energy estimation is rotation invariant, the model's final force estimation layer is not strictly rotation equivariant, but through its architectural design it is encouraged to learn rotation equivariance during training.

Results are demonstrated on the Open Catalyst 2020 (OC20) dataset [40] aimed at simulating catalyst materials that are useful for climate change related applications. The OC20 dataset contains over 130M training examples for approximating the DFT-estimated forces and energies. Our SpinConv model achieves state-of-the-art performance for both energy and force estimation. Notably, the force-centric variant, which is not energy conserving, outperforms the energy-centric models. Significant gains in accuracy are achieved for predicting relaxed energies from initial structures,

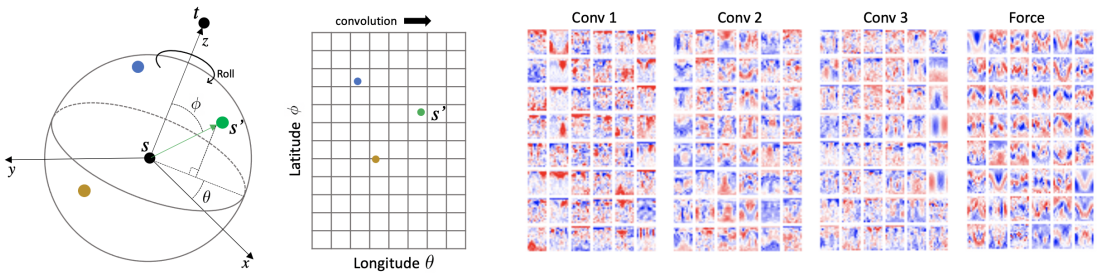


Figure 4-1: Illustration of projecting an atom $\acute{s}$ in the neighborhood of s onto a sphere in a local coordinate frame defined by atom s and t (left). For each projected atom, a corresponding latitude ϕ (inclination) and longitude θ (azimuth) is computed for its projection onto a 2D reference frame (middle). The spin convolution is done in the longitudinal direction, corresponding to a roll is 3D space. (right) Example channel filters that are learned using the grid-based approach for the first through third message blocks and the force block.

by using the force-centric approach to predict the relaxed structure followed by its energy. Ablation studies are performed on numerous architectural choices, such as the choice of spherical representation and the size of the model. For completeness, we also evaluate our model on the MD17 [48, 47] and QM9 [212] datasets that measure accuracy for molecular dynamics and property prediction tasks respectively for small molecules. Results compare favorably with respect to state-of-the-art methods.

4.3 Approach

We model a system or structure of atoms using a Graph Neural Network (GNN) [84, 149, 296], where the nodes represent atoms and the edges represent the atoms' neighbors. In this section, we describe both an energy-centric and force-centric model to estimating atomic forces, which vary in how they estimate forces and whether they are energy conserving. We begin by describing the components shared by each approach, followed by how these components are used. Code will be released upon acceptance under a permissive open-source license.

4.3.1 Inputs and Outputs

The inputs to the network are the 3D positions $\boldsymbol{x}_i$ and the atomic numbers a_i for all $i \in n$ atoms. The outputs are the per atom forces $\boldsymbol{f}_i \in \mathbb{R}^3$ and the overall structure's energy E. The 3D distance offset between a pair of source and target atoms s and t respectively is $\boldsymbol{x}_{st} = \boldsymbol{x}_s - \boldsymbol{x}_t$ with a distance of $d_{st} = \|\boldsymbol{x}_{st}\|_2$. Directional information is encoded using the normalized unit vector $\hat{\boldsymbol{x}}_{st} = \boldsymbol{x}_{st}/d_{st}$.

The graph neural network is constructed with each atom t as a node and the edges representing the atom's neighbors $s \in N_t$, where N_t contains all atoms s with $d_{st} < \delta$. Each edge has a corresponding message m_{st} that passes information from atom s to t. The output forces and energy are computed as a function of edge messages m_{st} that we describe next.

4.3.2 Energy and force estimation

The energy-centric and force-centric models compute the structure's energy E as an output. Our GNN model updates for each edge an M-dimensional hidden message $\boldsymbol{h}_{st}^{(k)} \in \mathbb{R}^M$ for K iterations. The structure's energy $E \in \mathbb{R}$ is computed as a function of the final layer of the edge messages in the GNN:

$$E(\boldsymbol{x}, a) = \sum_t \boldsymbol{F}_e(a_t, \sum_s \boldsymbol{h}_{st}^{(K)}), \tag{4.1}$$

where $\boldsymbol{F}_e$ is a single embedding block described later. As we also discuss later, the edge messages $\boldsymbol{h}_{st}$ are invariant to system rotations, so the estimated energy E is also invariant.

The estimation of the forces varies for the energy-centric and force-centric models. The energy-centric model estimates the forces using the negative partial derivative of the energy with respect to the atom positions. This approach to force estimation has the benefit of enforcing energy conservation [48], i.e., the forces along any closed path sum to zero. The calculation of the partial derivative [48, 232, 235] requires an additional step similar to performing backpropagation when updating the network's

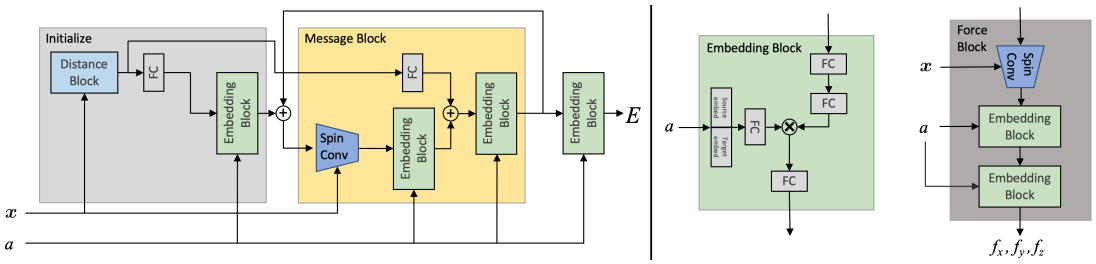


Figure 4-2: (left) Overall model diagram for energy-centric model taking atom positions $\boldsymbol{x}$ and atomic numbers a as input and estimating the energy E. (right) Diagram of the embedding and force blocks. The force block is only used in the force-centric model to estimate the per-atom forces after the message blocks.

weights:

$$\boldsymbol{f} = -\frac{\partial}{\partial \boldsymbol{x}} E(\boldsymbol{x}, a) \tag{4.2}$$

The force-centric model estimates forces directly for an atom t using:

$$\boldsymbol{f}_t = \boldsymbol{F}_f(a_t, \hat{\boldsymbol{x}}_t, \boldsymbol{h}_t^{(K)}), \tag{4.3}$$

where $\boldsymbol{F}_f$ is the force block we describe later, $\hat{\boldsymbol{x}}_t$ are all the normalized unit vectors for the neighbors of t and $\boldsymbol{h}_t^{(K)}$ are all incoming messages to atom t. This has the benefit of improved efficiency since it does not require an extra backward pass to estimate the forces. The tradeoff is that it does not enforce energy conservation, i.e., the sum of the forces along a closed path may not equal zero. Depending on the application, an energy-centric or force-centric approach may be most suitable. In either model, losses may be applied to both the energy and force estimates with weights determined by the needs of the application.

4.3.3 Messages

The edge messages are iteratively updated to allow information from increasingly distant atoms to be captured. Each message is represented by a tuple, $m_{st} = \{\hat{\boldsymbol{x}}_{st}, d_{st}, \boldsymbol{h}_{st}^k\}$, where $\boldsymbol{h}_{st}^k$ is the message's hidden state at iteration k. Both $\hat{\boldsymbol{x}}_{st}$ and d_{st} are used to update the message's hidden state $\boldsymbol{h}_{st}$, which is itself rotation invariant due to the spin convolution that we describe later. The hidden state $\boldsymbol{h}_{st} \in \mathbb{R}^M$ is

updated using:

$$\boldsymbol{h}_{st}^{(k+1)} = \boldsymbol{h}_{st}^{(k)} + \boldsymbol{F}_h \left(a_s, a_t, m_{st}^{(k)}, m_s^{(k)} \right), \tag{4.4}$$

where $m_s^{(k)}$ is the set of messages coming into node s, i.e., all $m_{\acute{s}s}$ with $\acute{s} \in N_s$. The form of $\boldsymbol{F}_h$ is illustrated in Figure 4-2. It contains three parts; the spin convolution that transforms a spherical projection of the messages into a rotation invariant representation, the distance block that encodes the distance d_{st} between atoms, and the embedding block that incorporates information about the atoms' atomic numbers. The output of the spin convolution is passed through an embedding block, added to the output of the distance block and finally passed through another embedding block. We describe each of these parts in turn. The hidden messages are initialized using just a distance block followed by and embedding block, Figure 4-2.

Spin Convolution

The spin convolution captures information about the neighbors $\acute{s} \in N_s$ of atom s when updating the message hidden state $\boldsymbol{h}_{st}$. The spin convolution has three stages that we describe in turn; projection, convolution and pooling. The convolution captures the relative angular information between the neighboring atoms, and the pooling ensures the output D-dimensional feature representation is invariant to system rotations.

An important feature is the angular information of the neighboring atoms in N_s relative to s and t. This information is encoded by creating a local reference frame in which atom s is the center $(0, 0, 0)$ and the z-axis points from atom s to atom t. As shown in Figure 4-1(left), this fixes all degrees of freedom except the roll rotation about the vector from s to t. The spin convolution is performed across a discretized set of rotations about the roll rotation axis. At each rotation, the atoms $\acute{s}$ are projected onto a sphere centered on s and used to create a spherical representation of the hidden states $\boldsymbol{h}_{\acute{s}s}$. Each atom $\acute{s} \in N_s$ is projected using a polar coordinate frame (ϕ, θ) where ϕ may be viewed as the latitude (inclination) and θ as the longitude (azimuth). The polar coordinates are computed in the local edge coordinate frame using $\bar{\boldsymbol{x}}_{\acute{s}s} = \mathbf{R}_{st}\hat{\boldsymbol{x}}_{\acute{s}s}$ where $\mathbf{R}_{st}$ is a 3D rotation matrix that satisfies $\mathbf{R}_{st}\hat{\boldsymbol{x}}_{st} = (0, 0, 1)$. To capture the

rich information encoded in the relative angular information between atoms, a set of filters is applied to the spherical representation (Figure 4-1(right)), similar to how a filter is applied to an image patch with traditional CNNs.

We explore two potential spherical representations: spherical harmonics and a grid-based approach. Spherical harmonics represent a spherical function using a set of basis functions that are equivariant to rotations. The degree ℓ indicates the number of basis functions $L = (\ell + 1)^2$ used. The spherical representation of the incoming messages for each atom is $\mathbb{R}^L \times \mathbb{R}^M$, where M is the size of the message hidden states in $\boldsymbol{h}$. The second approach uses the computed polar coordinates (ϕ, θ) for all $\acute{s} \in N_s$ to create a grid-based representation, Figure 4-1(middle). The polar coordinates are discretized creating a $\mathbb{R}^\phi \times \mathbb{R}^\theta \times \mathbb{R}^M$ feature representation. Each message hidden state $\boldsymbol{h}_{s\acute{s}}^{(k)} \in \mathbb{R}^M$ is added to the 3D feature representation using bilinear interpolation with its corresponding (ϕ, θ).

A 1D convolution is performed with either spherical representation in the longitudinal direction. Filters have the same size as the feature representation, $\mathbb{R}^L \times \mathbb{R}^M$ or $\mathbb{R}^\phi \times \mathbb{R}^\theta \times \mathbb{R}^M$ for spherical harmonics and the grid-based approach respectively. Full coverage filters are used since the angular relationship between atoms at distant angles is important, e.g., the forces of atoms at exactly 180° from each other may cancel out. Large filters also enable the network to learn the complex relationships between numerous neighboring atoms. Rotations are performed using Wigner D-matrices for the spherical harmonic representation, while a simple translation is used for the grid-based representation. The result of the convolution is a $\mathbb{R}^\theta \times \mathbb{R}^D$ feature vector corresponding to D filters applied to each longitudinal orientation. To make the representation invariant to rotations, average pooling is performed in the longitudinal direction resulting in a final $\mathbb{R}^D$ feature vector.

Distance Block

The distance block encodes the distance between two atoms. The distance is encoded using a set of evenly distributed Gaussian basis functions $\mathcal{G}$ with means μ_i and standard deviation σ. The means of the basis functions are evenly distributed from 0 to δ

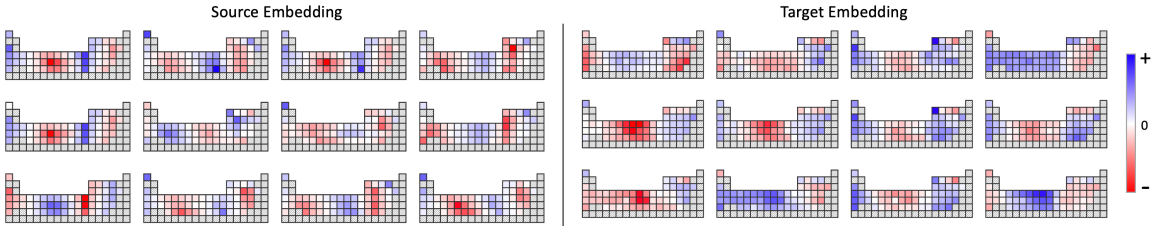


Figure 4-3: Illustration of learned embeddings (weights on the one-hot embeddings) for the source a_s and target a_t atomic numbers plotted on a periodic table. A random sample of 12 values from each embedding are shown. Embeddings are from the first embedding block in the first message update. Note that neighboring atoms in the periodic table with similar properties have similar weights. Elements not in the OC20 dataset are marked with a light grey checkerboard pattern.

angstroms. Since the atomic radii of each element varies, the relative position of two atoms s and t is highly dependent on their atomic numbers a_s and a_t. To account for this, gain $v_{a_s a_t}$ and offset $u_{a_s a_t}$ scalars for the distance d_{st} are learned for each potential pair of atomic numbers:

$$\boldsymbol{b}_i = \mathcal{G}_i(v_{a_s a_t} d_{st} + u_{a_s a_t} - \mu_i, \sigma) \tag{4.5}$$

The resulting feature $\boldsymbol{b}$ is passed through a linear transformation to create a D-dimensional feature vector that is passed to the next block.

Embedding Block

The embedding block incorporates the atomic number information a_s and a_t into the update of the message's hidden state. The embedding operation may be interpreted as a mixture of experts [164] approach that computes B different variations of the input, which are weighted by an embedding computed from the atoms' atomic numbers. The block's inputs are used to compute B sets of hidden values $\boldsymbol{V}_{st} \in \mathbb{R}^D \times \mathbb{R}^B$. A one-hot embedding for the atomic numbers a_s and a_t are concatenated and used to compute an B dimensional vector, $\boldsymbol{v}_{st} \in \mathbb{R}^B$, for weighting the B different sets of hidden values. An illustration of the learned embeddings are shown in Figure 4-3. $\boldsymbol{v}_{st}$ is computed using a two layer network and softmax. The matrix $\boldsymbol{V}_{st}$ is multiplied by vector $\boldsymbol{v}_{st}$ resulting in a vector of length D. As shown in Figure 4-2, the result is passed through

an additional fully connected layer before being passed to the next block. The output of the block is either D if it is used in the message update. If the embedding block is used to compute the final energy, only the atomic number a_t embedding is used, the input dimension is M instead of D, and the output is size 1.

4.3.4 Force Block

The force block computes the per-atom 3D forces f from a_t, $\hat{\boldsymbol{x}}_t$, and $\boldsymbol{h}_t^{(K)}$ using Equation (4.3). The force block uses a similar spin convolution as the message block, except the sphere is centered on the target atom t and is orientated along the x, y and z axes to compute f_x, f_y and f_z respectively. That is, the force block is used three times to compute the force magnitude in each orthogonal direction for each atom. The force block uses the same embedding blocks as message passing, Figure 4-2.

The same weights are used to compute forces in each of the three directions, only the orientation of the sphere used to create the convolutional features changes. To add more robustness to the force estimation and encourage rotational equivariance, the overall structure may be randomly rotated several times and the forces estimated. The multiple estimates may then be rotated back to the original reference frame and averaged. For both training and testing, five random rotations are used. Empirically, this approach encourages the networks to learn an approximate rotation equivariant representation even though rotation equivariance is not strictly enforced.

4.4 Experiments

In this section, we begin by presenting our primary results on the Open Catalyst 2020 (OC20) dataset [40] and compare against state-of-the-art models. This is followed by results on the smaller datasets of MD17 [48, 47] and QM9 [212] for additional model comparison.

Implementation details. For all models, the edge messages have size $M = 32$ with $K = 3$ layers, the hidden dimension $D = 256$ and embedding dimension $B = 8$.

Model	Hidden dim	#Msg layers	#Params	Train time	Inference time	OC20 Test Energy MAE [eV] ↓	Force MAE [eV/Å] ↓	Force Cos ↑	EFwT [%] ↑
Median	–	–	–			2.258	0.08438	0.0156	0.005
SchNet[235, 40]	1024	5	9.1M	194d	0.8h	–	0.04903	0.3413	0
DimeNet++[132, 40]	192	3	1.8M	587d	8.5h	0.5343	0.04758	0.3560	0.05
DimeNet++ energy-only[132, 40]	192	3	1.8M	587d	8.5h	0.4802	0.3459	0.1021	0.0
DimeNet++ force-only[132, 40]	192	3	1.8M	587d	8.5h	–	0.03573	0.4785	–
DimeNet++-large[132, 40]	512	3	10.7M	1600d	27.0h	–	0.03275	**0.5408**	–
ForceNet[108]	512	5	11.3M	31d	1.3h	–	0.03432	0.4770	–
ForceNet-large[108]	768	7	34.8M	194d	3.5h	-	0.03113	0.5195	-
SpinConv (energy-centric)	256	3	6.1M	275d	22.7h	0.4114	0.03888	0.4299	0.16
SpinConv (energy-centric) force-only	256	3	6.1M	380d	22.7h	–	0.03258	0.4976	–
SpinConv (force-centric)	256	3	8.5M	275d	9.1h	**0.3363**	**0.02966**	0.5391	**0.45**

Table 4.1: Comparison of SpinConv to existing GNN models on the S2EF task. Average results across all four test splits are reported. We mark as bold the best performance and close ones, *i.e.*, within 0.0005 MAE, which according to our preliminary experiments, is a good threshold to meaningfully distinguish model performance. Training time is in GPU days, and inference time is in GPU hours. Median represents the trivial baseline of always predicting the median training force across all the validation atoms.

Unless otherwise stated, the convolutional filters are of size 16x12 and 12x8 for the force-centric and energy-centric models respectively. A smaller filter size was used for the energy-centric model due to memory constraints. GroupNorm [285] is applied after the spin convolution with group size 4. An L1 loss is used for all experiments. The force loss was weighed by 100 with respect to the energy loss, except for the force-only model where the energy loss is set to 0. All models were trained with Adam (amsgrad) to convergence with the learning rate multiplied by 0.8 when the validation error plateaus. Training was performed using batch sizes ranging from 64 to 96 samples across 32 Volta 32GB GPUs. The Swish [211] function is used for all non-linear activation functions. The neighbors $s \in N_t$ of each atom t are found using a distance threshold of $\delta = 6$Å. If more than 30 atoms are within the distance threshold, only the closest 30 are used. The distance block uses 256 to 512 Gaussian basis functions with σ's equal to three times the distance between Gaussian means.

4.4.1 OC20

The OC20 dataset [40] contains over 130 million structures used to train models for predicting forces and energies during structure relaxations that is released under a CC Attribution 4.0 License. Since the goal of a structure relaxation is to find a local energy minimum, energy conservation in optional for this task. We report results for

Model	Energy MAE (eV) ↓				Force MAE (eV/Å) ↓			
	ID	OOD Ads.	OOD Cat.	OOD Both	ID	OOD Ads.	OOD Cat.	OOD Both
Median	2.043	2.420	1.992	2.577	0.0809	0.0801	0.0787	0.0978
Energy Loss Only								
SchNet	0.395	0.446	0.551	0.703	-	-	-	-
DimeNet++	0.359	0.402	0.506	0.654	-	-	-	-
Force Loss Only								
SchNet	-	-	-	-	0.0443	0.0469	0.0459	0.0590
DimeNet++	-	-	-	-	0.0331	0.0341	0.0340	0.0417
DimeNet++-large	-	-	-	-	0.0281	0.0289	0.0312	0.0371
ForceNet	-	-	-	-	0.0313	0.0320	0.0331	0.0409
ForceNet-large	-	-	-	-	0.0278	0.0283	0.0309	0.0375
SpinConv (energy-centric)	-	-	-	-	0.0309	0.0321	0.0315	0.0393
Energy and Force Loss								
SchNet	0.443	0.491	0.529	0.716	0.0493	0.0527	0.0508	0.0652
DimeNet++	0.486	0.470	0.533	0.648	0.0443	0.0458	0.0444	0.0558
SpinConv (energy-centric)	0.351	0.367	0.411	0.517	0.0358	0.0374	0.0364	0.0460
SpinConv (force-centric)	**0.261**	**0.275**	**0.350**	**0.459**	**0.0269**	**0.0277**	**0.0285**	**0.0356**

Table 4.2: Comparison of SpinConv to existing GNN models on different test splits. We mark as bold the best performance and close ones, *i.e.*, within 0.0005 MAE, which according to our preliminary experiments, is a good threshold to meaningfully distinguish model performance. Training time is in GPU days, and inference time is in GPU hours. Median represents the trivial baseline of always predicting the median training force across all the validation atoms.

the Structure to Energy and Forces (S2EF), the Initial Structure to Relaxed Energy (IS2RE) and the Initial Structure to Relaxed Structure (IS2RS) tasks.

Structure to Energy and Forces (S2EF)

There are four metrics for the S2EF task, the energy and force Mean Absolute Error (MAE), the Force Cosine similarity, and the Energy and Forces within a Threshold (EFwT). The EFwT metric is meant to indicate the percentage of energy and force predictions that would be useful in practice. Results for three model variants are shown in Table 4.1 on the test set. The SpinConv force-centric approach has the lowest energy MAE and force MAE of all models. While still low in absolute terms, the SpinConv models are improving over other models on the EFwT metric. DimeNet++-large slightly out performs SpinConv on the force cosine metric. The training time for the SpinConv is significantly faster than DimeNet++, while being a little slower than ForceNet [108] or SchNet [235].

In Table 4.2 we examine the performance of SpinConv across different test splits. Note that the energy prediction of SpinConv is signficantly better than SchNet or

Model	Hidden dim	#Msg layers	#Params	Train time	OC20 Val ID 30k Energy MAE [eV] ↓	Force MAE [eV/Å] ↓	Force Cos ↑	EFwT [%] ↑
Median								
Energy-Centric								
SpinConv (grid 12x8)	128	2	1.3M	54d	–	0.0417	0.401	–
SpinConv (spherical harmonics, $\ell = 5$)	256	3	6.4M	119d	–	0.0405	0.411	–
SpinConv (grid 12x8)	256	3	6.1M	87d	–	0.0406	0.426	–
Force-Centric								
SpinConv (grid 12x8)	128	2	1.8M	54d	0.376	0.0370	0.436	0.15%
SpinConv (grid no conv 16x12)	256	3	8.5M	56d	0.341	0.0348	0.462	**0.20%**
SpinConv (spherical harmonics, $\ell = 5$)	256	3	8.1M	113d	0.321	**0.0328**	**0.484**	**0.22%**
SpinConv (grid 16x12)	256	3	8.5M	76d	**0.317**	**0.0326**	**0.484**	**0.20%**

Table 4.3: Ablation studies for SpinConv model variations trained for 560k steps (32-48 batch size, 0.2 epochs) with 16 Volta 32 GB GPUs. Training time is in GPU days and the validation set is a 30k random sample of the OC20 ID Validation set.

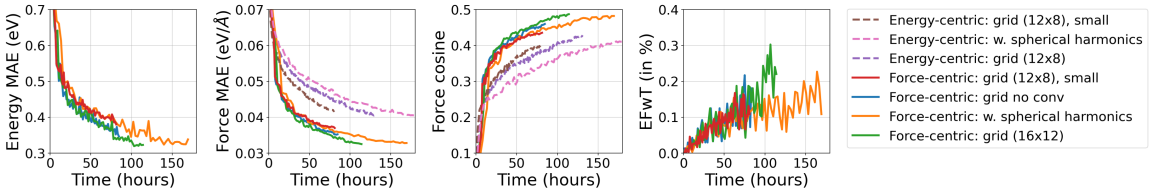


Figure 4-4: Performance of SpinConv ablations on OC20 Val ID 30k (Table 4.3). All models trained for 560k steps and plotted against wall-clock training time. Note force-centric models and grid-based approaches converge more quickly than energy-centric models and those using spherical harmonics.

DimeNet++. Across all models the accuracy for the in domain split are highest and decline for the three Out of Domain (OOD Adsorbate, OOD Catalyst, OOD Both) splits. SpinConv outperforms all models on each of the different domain splits. When comparing energy-centric approaches trained with both force and energy losses (bottom rows), the SpinConv model does significantly better at predicting both. In fact, the energy-centric approach trained on forces and energy outperforms the DimeNet++ [132] model when trained on only energy, or energy and forces.

We examine variations of the SpinConv model in Table 4.3 and Figure 4-4 through ablation studies. We trained three variants of the energy-centric model and four variants of the force-centric model. The grid-based and spherical harmonic approaches produced similar accuracies. However, the grid-based approach was significantly faster to train, so it was used in the remaining experiments. Smaller models lead to reduced performance on the OC20 dataset, but we found for smaller datasets such as MD17 or QM9 smaller model sizes can be beneficial to avoid overfitting. Finally, we test the impact of not performing the convolution (no conv) and only applying

Model	Approach	Energy MAE [eV] ↓				EwT ↑			
		ID	OOD Ads	OOD Cat	OOD Both	ID	OOD Ads	OOD Cat	OOD Both
Median baseline	-	1.7499	1.8793	1.7090	1.6636	0.71%	0.72%	0.89%	0.74%
CGCNN [286]	Direct	0.6149	0.9155	0.6219	0.8511	3.40%	1.93%	3.10%	2.00%
SchNet [232]	Direct	0.6387	0.7342	0.6616	0.7037	2.96%	2.33%	2.94%	2.21%
DimeNet++ [133]	Direct	0.5620	0.7252	0.5756	0.6613	4.25%	2.07%	4.10%	2.41%
SpinConv	Direct	0.5583	0.7230	0.5687	0.6738	4.08%	2.26%	3.82%	2.33%
DimeNet++	Relaxation	0.6908	0.6842	0.7027	0.6834	4.25%	3.36%	3.76%	3.52%
DimeNet++ – force-only + energy-only	Relaxation	0.5124	0.5744	0.5935	0.6126	6.12%	4.29%	5.07%	3.85%
DimeNet++ – large force-only + energy-only	Relaxation	0.5034	0.5430	0.5789	0.6113	6.57%	4.34%	5.09%	3.93%
SpinConv (force-centric)	Relaxation	**0.4235**	**0.4415**	**0.4572**	**0.4245**	**9.37%**	**6.75%**	**8.49%**	**6.76%**

Table 4.4: Initial Structure to Relaxed Energy (IS2RE) results on the OC20 test split as evaluated by the Energy MAE (eV) and Energy within Threshold (EwT) [40] (see OC20 discussion board). Comparisons made for the direct and relaxation approaches using various models.

Model	**Inference time** ↓	**AFbT (%)** ↑					**ADwT (%)** ↑				
		ID	OOD Ads.	OOD Cat.	OOD Both	**Average**	ID	OOD Ads.	OOD Cat.	OOD Both	**Average**
SchNet [232]	54.1h	5.28	2.82	2.62	2.73	3.36	32.49	28.59	30.99	35.08	31.79
DimeNet++ [132]	407.6h	17.52	14.67	14.32	14.43	15.23	48.76	45.19	48.59	53.14	48.92
DimeNet++-large [132]	814.6h	**25.65**	**20.73**	**20.24**	**20.67**	**21.82**	52.45	48.47	50.99	54.82	51.68
ForceNet [108]	75.1h	10.75	7.74	7.54	7.78	8.45	46.83	41.26	46.45	49.60	46.04
ForceNet-large [108]	186.9h	14.77	12.23	12.16	11.46	12.66	50.59	45.16	49.80	52.94	49.62
SpinConv (force-centric)	263.2h	21.10	15.70	15.86	14.01	16.67	**53.68**	**48.87**	**53.92**	**58.03**	**53.62**

Table 4.5: Relaxed structure from initial structure (IS2RS) results on the OC20 test split, as evaluated by Average Distance within Threshold (ADwT) and Average Forces below Threshold (AFbT). All values in percentages, higher is better. Results computed via the OCP evaluation server. Inference times are total across the 4 splits.

the filter at a single rotation. Rotation invariance was maintained by orienting the filter based on the mean angle of the neighboring atoms weighted by distance. The result of not performing the convolution is significantly reduced accuracy. However, its faster training time may make it suitable for some applications.

Finally, for the force-centric SpinConv model we explore results when varying the number of random rotations used in the force block. The force MAE when using a single random rotation is 0.0276 and improves slightly to 0.0270 when using 5 random rotations. Increasing the number of rotations beyond 5 leads to negligible gains. The standard deviation of the force estimates at different random rotations is 0.004 eV/Å. This is equal to 15% of the force MAE, which indicates the amount of error due to the model not being strictly rotation equivariant is small relative to the overall error of the model.

Initial Structure to Relaxed Energy (IS2RE)

The Initial Structure to Relaxed Energy (IS2RE) task takes an initial atomic structure and attempts to predict the energy of the structure after it has been relaxed. Two approaches may be taken to address this problem, the direct and relaxation approaches [40]. The direct treats the task as a standard regression problem and directly estimates the relaxed energy from the initial structure. The relaxation approach computes the relaxed structure using the ML predicted forces to update the atom positions. Next, given the ML relaxed structure the energy is estimated. We show results for both approaches in the OC20 dataset using SpinConv in Table 4.4.

The results of the SpinConv model significantly outperform all previous approaches using the relaxation approach for both energy MAE and Energy within Threshold (EwT) metrics. DimeNet++ also shows improved results for the relaxation approach with the best approach using two models; DimeNet++-large for force estimation and DimeNet++ (energy-only) for the energy estimation. Note in contrast to other approaches, SpinConv shows good results across all test splits, including those with out of domain adsorbates and catalysts. Using the direct approach, SpinConv is comparable to DimeNet++'s direct approach.

Initial Structure to Relaxed Structure (IS2RS)

Our final results on the OC20 dataset are on the IS2RS task where predicted forces are used to relax an atom structure to a local energy minimum. The is performed by iteratively estimating the forces that are in turn used to update the atoms positions. This process is repeated until convergence or 200 iterations. Results are shown in Table 4.5. The suggested metrics are Average Distance within Threshold (ADwT) metric, which measures whether the atom positions are close to those found using DFT and Average Forces below Threshold (AFbT), which measures whether a true energy minimum was found (i.e., forces are close to zero). On the ADwT metric, SpinConv outperforms other approaches (53.62% averaged across splits). On the AFbT metric, DimeNet++-large outperforms SpinConv (21.82% *vs*. 16.67%), but is

Molecule	GDML	PhysNet	PhysNet-ens5	SchNet	DimeNet*	SpinConv
Aspirin	0.02	0.06	0.04	0.33	0.09	**0.07**
Benzene	0.24	0.15	0.14	0.17	**0.15**	0.17
Ethanol	0.09	0.03	0.02	0.05	0.03	**0.02**
Malonaldehyde	0.09	0.04	0.03	0.08	**0.04**	**0.04**
Naphthalene	0.03	0.04	0.03	0.11	0.06	**0.04**
Salicylic	0.03	0.04	0.03	0.19	0.09	**0.05**
Toluene	0.05	0.03	0.03	0.09	0.05	**0.03**
Uracil	0.03	0.03	0.03	0.11	0.04	**0.03**
Mean	0.073	0.053	0.044	0.141	0.069	**0.058**

Table 4.6: Forces MAE (kcal/molÅ) on MD17 for models trained using 50k samples. Best results for models not using domain specific information are in bold. *The DimeNet results were trained in-house as the original authors did not use the 50k dataset. DimeNet was found to outperform DimeNet++ on this task.

Task Units	α bohr^3	$\Delta\epsilon$ meV	ϵ_{HOMO} meV	ϵ_{LUMO} meV	μ D	C_ν cal/mol K	G meV	H meV	R^2 bohr^3	U meV	U_0 meV	ZPVE meV
NMP [79]	.092	69	43	38	.030	.040	19	17	.180	20	20	1.50
Schnet [232]	.235	63	41	34	.033	.033	14	14	**.073**	19	14	1.70
Cormorant [7]	.085	61	34	38	.038	.026	20	21	.961	21	22	2.03
L1Net [168]	.088	68	46	35	.043	.031	14	14	.354	14	13	1.56
LieConv [67]	.084	49	30	25	.032	.038	22	24	.800	19	19	2.28
TFN [260]	.223	58	40	38	.064	.101	-	-	-	-	-	-
SE(3)-Tr. [70]	.142	53	35	33	.051	.054	-	-	-	-	-	-
EGNN [225]	.071	48	29	25	.029	.031	12	12	.106	12	11	1.55
DimeNet++ [132]	**.044**	33	25	20	.030	.023	**8**	7	.331	**6**	**6**	1.21
SphereNet [155]	.047	**32**	**24**	**19**	**.027**	**.022**	**8**	**6**	.292	7	**6**	**1.12**
SpinConv	.058	47	26	22	**.027**	.028	12	12	.156	12	12	1.50

Table 4.7: Mean absolute error results for QM9 dataset [212] on 12 properties for small molecules.

more than ∼3x slower (814.6h *vs.* 263.2h) during inference. SpinConv outperforms all other models.

4.4.2 MD17

The MD17 dataset [48, 47] contains molecular dynamic simulations for eight small molecules. Two training datasets are commonly used, one containing 1k examples and another containing 50k examples. We found the 1k training dataset to be too small for the SpinConv model, and may be more appropriate for approaches that incorporate

prior chemistry knowledge, such as hand-coded features or force fields [48, 270]. The 50k dataset provides significantly more training data, but the remaining validation and test data are highly similar to those found in training, and may not guarantee independent samples in the test set[49]. Nevertheless, we report results on MD17 for comparison to prior work on the molecular dynamics task. Research in this domain would greatly benefit from the generation of a larger dataset.

Results are shown in Table 4.6. SpinConv is on par or better for 7 of the 8 molecules when compared to DimeNet [133]. Both SpinConv and DimeNet perform well with respect to the GDML [48] and PhysNet [270] models that take advantage of domain-specific information. Given the smaller dataset size, the SpinConv model uses a reduced 8x8 grid-based spherical representation. Other model parameters are the same as previously described.

4.4.3 QM9

Our final set of results are on the popular QM9 dataset [212] that tests the prediction of numerous properties for small molecules. While the SpinConv model was designed to estimate energies and per-atom forces, we may use the same model to predict other proprieties. Results are shown in Table 4.7 on a random test split for an energy-centric 8x8 grid-based SpinConv model. The results of DimeNet++ and the recent SphereNet[155] outperform those of others. However, DimeNet++, SphereNet and SpinConv perform well with respect to other approaches across many properties.

4.5 Related work

A common approach to estimating molecular and atomic properties is the use of GNNs [234, 79, 119, 232, 235, 286, 206, 133] where nodes represent atoms and edges connect neighboring atoms. One of the first approaches for force estimation was SchNet [232], which computed forces using only the distance between atoms without the use of angular information. Unlike previous approaches that used discrete distance filters [286], SchNet proposed the used of differentiable edge filters. This

enabled the construction of an energy-conserving model for molecular dynamics that estimates forces by taking the negative gradient of the energy with respect to the atom positions [48]. DimeNet extended this approach to also represent the angular information between triplets of atoms [133, 132]. The more recent SphereNet further extends this by capturing dihedral angles [155]. SpinConv is able to model relative angular relationships between all neighboring atoms, and not just triplets of atoms, due to the use of the spin convolutional filter. In parallel to invariant models, rotational equivariant networks are explored in depth by [280, 25, 7, 260, 225]. This was accomplished by decoupling the network-fed invariant information (distance), from the equivariant information (distance vector), followed by the careful combination via tensor products. The energy-centric SpinConv model is invariant to rotations due to the use of global pooling after the spin convolution. The final force block of the force-centric model is not strictly rotation equivariant, but is encouraged to learn rotation equivariance during training.

Another approach to force estimation is to directly regress the forces as an output of the network. This doesn't enforce energy conservation or rotational equivariance, but as shown by ForceNet [108], such models can still produce accurate force estimates.

Numerous approaches incorporate more domain specific information into machine learning models. These include GDML [48] and PhysNet [270] that use handcrafted features and force-fields respectively. OrbNet [206] is a hybrid approach that utilizes proprietary orbital features that improves accuracy while achieving significant efficiency gains over DFT. While these approaches can lead to improved accuracy, they typically result in increased computational expense over ML models.

4.6 Discussion

A common approach to estimating molecular and atomic properties is the use of GNNs [234, 79, 119, 232, 235, 286, 206, 133] where nodes represent atoms and edges connect neighboring atoms. One of the first approaches for force estimation was

SchNet [232], which computed forces using only the distance between atoms without the use of angular information. Unlike previous approaches that used discrete distance filters [286], SchNet proposed the used of differentiable edge filters. This enabled the construction of an energy-conserving model for molecular dynamics that estimates forces by taking the negative gradient of the energy with respect to the atom positions [48]. DimeNet extended this approach to also represent the angular information between triplets of atoms [133, 132]. The more recent SphereNet further extends this by capturing dihedral angles [155]. SpinConv is able to model relative angular relationships between all neighboring atoms, and not just triplets of atoms, due to the use of the spin convolutional filter. In parallel to invariant models, rotational equivariant networks are explored in depth by [280, 25, 7, 260, 225]. This was accomplished by decoupling the network-fed invariant information (distance), from the equivariant information (distance vector), followed by the careful combination via tensor products. The energy-centric SpinConv model is invariant to rotations due to the use of global pooling after the spin convolution. The final force block of the force-centric model is not strictly rotation equivariant, but is encouraged to learn rotation equivariance during training.

Another approach to force estimation is to directly regress the forces as an output of the network. This doesn't enforce energy conservation or rotational equivariance, but as shown by ForceNet [108], such models can still produce accurate force estimates.

Numerous approaches incorporate more domain specific information into machine learning models. These include GDML [48] and PhysNet [270] that use handcrafted features and force-fields respectively. OrbNet [206] is a hybrid approach that utilizes proprietary orbital features that improves accuracy while achieving significant efficiency gains over DFT. While these approaches can lead to improved accuracy, they typically result in increased computational expense over ML models.

4.7 Societal Impact

This work is motivated by the problems we face due to climate change [298], many of which require innovative solutions to reduce energy usage and replace traditional chemical feedstocks with renewable alternatives. For example, one of the most energy intensive chemical processes is the development of new electrochemical catalysts for ammonia fertilizer production that helped to feed the world's growing population during the 20th century [94]. This is also an illustrative example of possible unintended consequences as advancements in chemistry and materials may be used for numerous purposes. As ammonia fertilization increased in use, its overuse in today's farming has led to ocean "dead zones" and its production is very carbon intensive. Knowledge and techniques used to create ammonia were also transferred to the creation of explosives during wartime. We hope to steer the use of ML for atomic simulations to societally-beneficial uses by training and testing our approaches on datasets, such as OC20, that were specifically designed to address chemical reactions useful for addressing climate change.

Chapter 5

The Open Catalyst 2022 (OC22) Dataset and Challenges for Oxide Electrocatalysis

This work originally appeared as: Tran, R., Lan, J.*, Shuaibi, M.*, Wood, B.M.*, Goyal, S.*, Das, A., Heras-Domingo, J., Kolluru, A., Rizvi, A., Shoghi, N., Sriram, A., Ulissi, Z., Zitnick, C.L, 2022. The Open Catalyst 2022 (OC22) Dataset and Challenges for Oxide Electrocatalysis. arXiv preprint arXiv:2206.08917. ACS Catalysis, under review. It has been edited to include the supplementary information in Appendix C. *These authors contributed equally.*
My contribution in this work included task formulation, planning and coordinating all modeling and training experiments, data preprocessing and split creation, model evaluations, and the primary writer and editor of the manuscript.

5.1 Abstract

Computational catalysis and machine learning communities have made considerable progress in developing machine learning models for catalyst discovery and design. Yet, a general machine learning potential that spans the chemical space of catalysis is still out of reach. A significant hurdle is obtaining access to training data across a

wide range of materials. One important class of materials where data is lacking are oxides, which inhibits models from studying the OER and oxide electrocatalysis more generally. To address this we developed the OC22 dataset, consisting of 62,331 DFT relaxations (∼9,854,504 single point calculations) across a range of oxide materials, coverages, and adsorbates (*H, *O, *N, *C, *OOH, *OH, *OH_2, *O_2, *CO). We define generalized tasks to predict the total system energy that are applicable across catalysis, develop baseline performance of several graph neural networks (SchNet, DimeNet++, ForceNet, SpinConv, PaiNN, GemNet-dT, GemNet-OC), and provide pre-defined dataset splits to establish clear benchmarks for future efforts. For all tasks, we study whether combining datasets leads to better results, even if they contain different materials or adsorbates. Specifically, we jointly train models on OC20 and OC22, or fine-tune pretrained OC20 models on OC22. In the most general task, GemNet-OC sees a ∼32% improvement in energy predictions through fine-tuning and a ∼9% improvement in force predictions via joint training. Surprisingly, joint training on both the OC20 and much smaller OC22 datasets also improves total energy predictions on OC20 by ∼19%. The dataset and baseline models are open sourced, and a public leaderboard will follow to encourage continued community developments on the total energy tasks and data.

5.2 Introduction

One of the most challenging scientific problems facing humanity in the 21st century is the development of suitable technologies to produce, store, and use clean energy. Renewable energy is often produced by intermittent sources (e.g. sunlight, wind, or tides) so efficient grid-scale storage is required to transfer power from times of excess generation to times of excess demand. There are a number of promising storage techniques including the conversion of renewable energy to a chemical form, e.g. water splitting to H_2, or CO_2 conversion to liquid fuels. These applications rely on the availability of efficient electrocatalysts. In many cases the most stable and active catalysts for these reactions are inorganic oxides which present a number of challenges to cat-

alyst design compared to simpler metal surfaces. Developing generalizable machine learning methods to quickly and accurately predict the activity and stability of oxide catalysts would have a major impact on renewable energy storage and utilization.

As a motivating example of the need and challenges for oxide electrocatalysts, consider water splitting for the generation of clean H_2; an energy-dense fuel that is used in fuel cells or ammonia synthesis. Electrochemical water splitting consists of two coupled half-reactions,

$$\begin{aligned} &\text{OER:} && 2H_2O \longrightarrow O_2 + 4\,(H^+ + e^-) \\ &\text{HER:} && 4\,(H^+ + e^-) \longrightarrow 2\,H_2 \\ \hline & && 2\,H_2O \longrightarrow O_2 + H_2, \end{aligned}$$

which split two water molecules to evolve H_2 and O_2 gas. This process is extremely energy intensive. The OER is largely responsible for the total inefficiency of this reaction and is quite complicated due to bond rearrangements and the formation of an O−O bond. Water splitting typically uses very harsh acidic conditions to reduce gas solubility and improve proton conductivity, and for which high performance proton exchange membranes are widely available. Unfortunately, for these conditions there are very few known materials that are stable and active, except extremely expensive metal oxides, such as those using Ir or Ru [267]. Currently, there are significant efforts to design complex multi-component oxide OER catalysts to reduce the cost and improve their activity and stability [115, 293]. Computational chemistry can play a critical role in helping screen, discover, and understand such materials.

Computational methods can be used to predict the activity and stability of a proposed oxide catalyst, but these techniques are significantly more complicated than for metal catalysts and present many additional challenges. First, there are many oxide polymorphs (crystal structures) for any given chemical composition that must be considered to identify the most stable catalyst structure[68]. Second, the surface of an oxide catalyst is often prone to reconstruction, leaching, doping, and defects [42]. Third, the environment can lead to a number of possible surface terminations. Fourth,

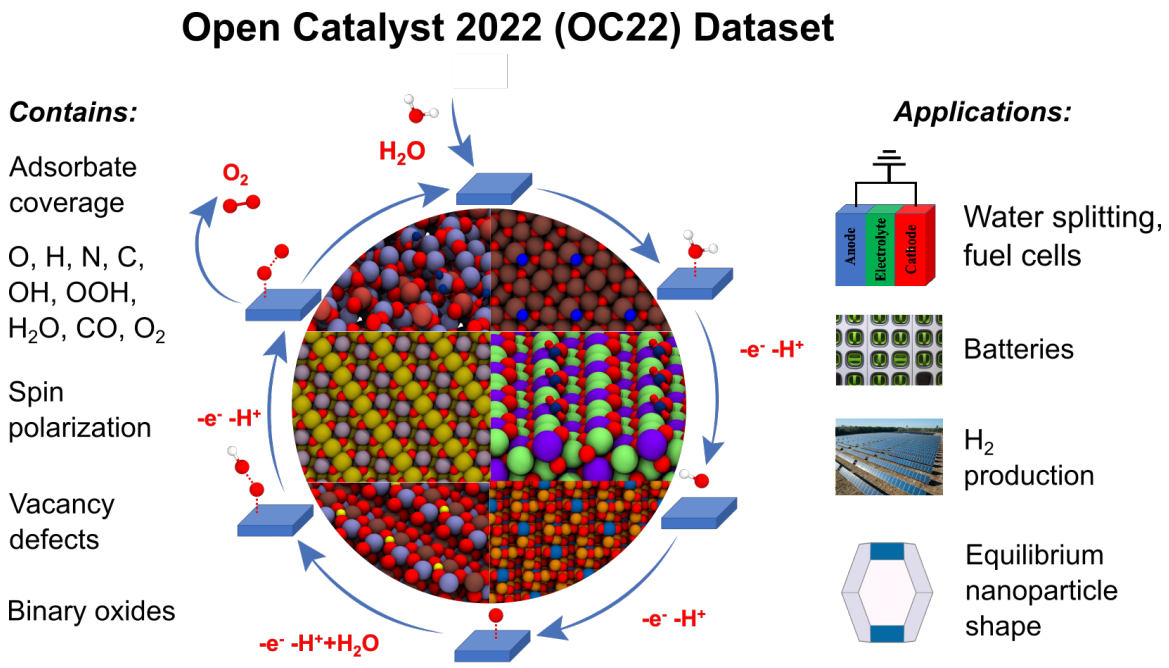


Figure 5-1: Overview of the contents and impact areas of the OC22 dataset. The water nucleophilic attack mechanism is highlighted for the OER reaction, with H_2O and O_2 as reactants and products, respectively. Images are a random sample of the dataset.

it is difficult to determine a catalyst's active site and there are often multiple competing mechanisms to consider [83]. To add to these challenges, computational chemistry methods such as the widely-used Generalized Gradient Approximation (GGA) are less accurate for oxide materials due to the strong electron correlation and complicated electronic structure. Large system sizes and the likelihood of long-range electrostatic or magnetic interactions also result in slower convergence. These additional configurational and computational complexities make the creation of datasets and machine learning models for oxides significantly more expensive and challenging, leading to much fewer and smaller datasets than for metal systems (see [6] for a sample of representative datasets in catalysis).

To address these challenges, we propose training ML models to enable the efficient search of new materials. The training of accurate ML models requires the creation of a large training dataset. For example, the OC20 dataset [40] (ca. 250 million single-point calculations) considered different adsorbates (small adsorbates,

C1/C2 compounds, and N/O-containing intermediates) on top of randomly sampled low Miller index facets of stable materials from the Materials Project[114], but did not include metal oxide materials due to the complexities above. The release of the OC20 dataset helped enable rapid advances in the accuracy and generalizability of Graph Neural Network (GNN) models [136], with decreases of 55+% in the key S2EF metrics in the first two years. Initial baseline models like CGCNN[286] and SchNet[232] focused on local environment representations. Key advances since then include invariant angular interactions (DimeNet/DimeNet++ [133, 131]), faster and more accurate but non-energy conserving models (ForceNet[108] and SpinConv[241]), and triple/quadruplet interactions (GemNet-dT[75], GemNet-XL[250], and GemNet-OC[76]). Other approaches include the use of transformers (3D-Graphormer[290]) and more effective augmentation and learning strategies (Noisy-Nodes[80]). These and further advances are necessary to accurately predict properties of complex structures such as oxide systems.

In this work, we present the Open Catalyst 2022 (OC22) dataset (Figure 5-1) for the oxygen evolution reaction and oxide electrocatalysis more generally, as well as accompanying tasks and GNN baseline models. OC22 is meant to complement OC20, which did not contain any oxide materials. This dataset spans the configurational complexity for oxide surfaces described above, including varying surface terminations, adsorbate+slab configurations and coverage, and non-stoichiometric substitutions and vacancies. To encompass the additional complexities in this dataset, we also expand on the primary tasks in OC20 to include the DFT total energy as a target. A more general property, DFT total energy offers the ability to address potential applications beyond those that just require simple adsorption energies.

With the creation of new datasets, the question arises of whether the data in them is complementary to other datasets for training ML models. For instance, models can be trained jointly using multiple datasets, or transfer learning may be used to train a model on a larger dataset and fine-tuned on a smaller dataset. Recently, the OC20 dataset enabled the catalysis community to use transfer learning to improve model performance [135] on other smaller datasets. The small molecules and drug discovery

communities have seen success in using transfer learning to transfer between varying levels of electronic structure calculations [246] or between related tasks[52, 219, 268]. In this work, we explore the extent OC20 can aid OC22 via transfer learning or by jointly training on both datasets.

We train a variety of leading GNN models on two related proposed community challenges for OC22: (1) predict the DFT total energy and force for a given structure and (2) predict the DFT relaxed total energy given an initial structure. We also evaluate our models' performance on the established task of predicting the relaxed structure given an initial structure. The dataset is split into train/validation/test splits indicative of the situation commonly found in catalysis where the properties of unseen crystal compositions need to be predicted. Splits contain a combination of adsorbate + catalyst and clean catalyst (no adsorbate present) systems. All baseline models, data loaders and training scripts for each of these tasks are available at `https://github.com/Open-Catalyst-Project/ocp`. While we focus on a subset of tasks, models capable of solving these tasks on the OC22 dataset will likely be able to address numerous related catalysis problems.

5.3 The OC22 Dataset

OC22 is designed to provide a training dataset for constructing generalized models to aid in predicting catalytic reactions on oxide surfaces. To achieve this, we built the dataset in four stages: (1) bulk selection, (2) surface selection, (3) initial structure generation, and (4) structure relaxation. The dataset contains isolated surfaces (a.k.a slabs) and surface and adsorbate combinations (a.k.a adsorbate+slabs), 19,142 and 43,189 systems, respectively. This resulted in 9,854,504 single point step calculations, each of which yielded forces and energies which were later partitioned into suitable train, validation, and test validation splits. We prioritized diversity in composition, surface termination, and adsorbate configurations in constructing our dataset to ensure that our models can generalize well. As such the structures in our dataset are not always the most thermodynamically stable. All source code used to generate the

adsorbate configurations will be provided in the Open Catalyst Dataset repository at `https://github.com/Open-Catalyst-Project/Open-Catalyst-Dataset`.

5.3.1 Bulk selection

We begin by confining our set of bulk oxide materials to 4,728 unary (A_xO_y) and binary ($A_xB_yO_z$) metal-oxides from the Materials Project[114] where A and B are metals. These oxides can be composed of any combination of metals or semi-metals listed in the Supplementary Information (SI). In our list of 51 metals, Ce was the only lanthanide considered due to the utility of its oxide compounds in catalytic reactions[248, 57]. For each chemical system, we considered bulk materials with the top five lowest energies above hull with less than 150 atoms to provide the most chemically diverse set of oxides. We also considered 173 unary and binary rutile structures.

Our selection criteria for bulk oxides prioritized chemical diversity over stability. We acknowledge that many of the materials we selected are not electrochemically stable which is a prerequisite for viable electrocatalytic materials. Pourbaix analysis have previously demonstrated that only oxides composed of 26 of the 51 elements we considered are relatively stable under aqueous conditions[279].

We also ignored the fact that certain chemical systems have a far greater set of distinct bulk structures than others. For instance, the Materials Project database has reported over 300 entries for chemical systems such as Ti-O and Mn-Li-O while no entries were reported for 200 chemical systems (see the SI). Other databases such as the Automatic-Flow[51] and Open Quantum Materials Database[223] have also made significant efforts in exploring oxides and contain chemical systems unexplored in the Materials Project. However, to ensure all oxides were obtained using a consistent methodology and open source licensing, we extracted entries from the Materials Project only.

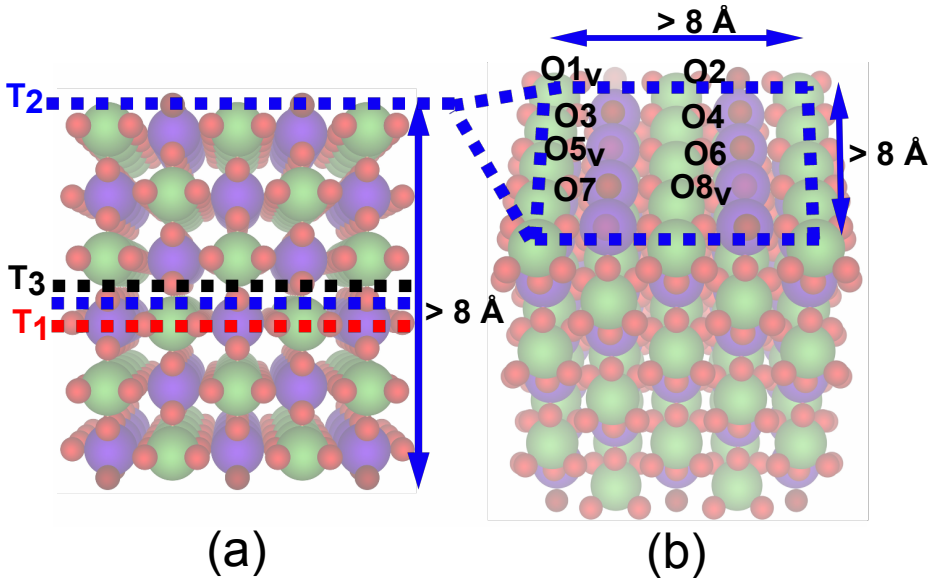


Figure 5-2: Construction of rutile (110) slabs and adsorbate+slabs. (a) Dashed lines indicate the different possible terminations (T_1, T_2 and T_3). The slab is symmetric about T_3. (b) The T_2 terminated surface with its periodic boundary (blue dashed lines) contains 8 oxygen sites. Random removal of 3 surface oxygen (dark red) creates vacancy defects (transparent).

5.3.2 Surface selection

We constructed our dataset by first randomly sampling 4,286 bulk oxides from our original bulk oxide set of 4,728. We limited our dataset to slabs of less than 250 atoms. We construct each slab and adsorbate+slab using the process shown in Figure 5-2. Given a random oxide selected from our bulk dataset, we enumerate through all possible surface terminations with a maximum Miller index less than or equal to 3. As with Figure 5-2(a) all slabs are capped with the same terminating surface regardless of stoichiometry. We randomly select one termination which we replicated to a depth of at least 8 Å and a width in each cross-sectional direction of at least 8 Å.

Next we decorated the surface of the slab with a random number of oxygen vacancies which can act as active sites for reactions such as CO_2 capture[153] and OER[12, 157]. To do so, we first identify all existing oxygen lattice sites on the surface as with Figure 5-2(b). We then select a random number of surface oxygen to remove ranging from 0 (no vacancies) to all surface oxygen. We do the same on the

other surface to maintain charge balance throughout the slab. This is done to avoid the manifestation of non-physical dipole moments which can lead to diverging DFT energies.

The SI provides the chemical space distribution of all slabs and adsorbate+slabs successfully calculated in the dataset. Table 5.1 summarizes the distribution of elemental composition, crystal structures, and number of components of the entire dataset of slabs and adsorbate+slabs.

Table 5.1: Overview of the chemical, structural and adsorbate composition of the entire dataset of slabs and adsorbate+slabs.

Chemical formula	
Unary (A_xO_y)	6,190
Binary ($A_xB_yO_z$)	56,141
Elements sampled	
Alkali	13,541
Alkaline	13, 974
p-block metals	14,029
Metalloids	8,292
Transition metals	48,561
Crystal structures	
Triclinic	6,214
Monoclinic	16,294
Orthorhombic	7,258
Tetragonal (Rutile)	11,550 (4,318)
Trigonal	4,411
Hexagonal	2,680
Cubic	9,606
Adsorbates	
O	10,816
H	5,298
N	4,000
C	3,905
OH	4,092
OOH	4,424
H_2O	4,846
CO	3,994
O_2	1,814
Calc. with PBE+U: 20,812	

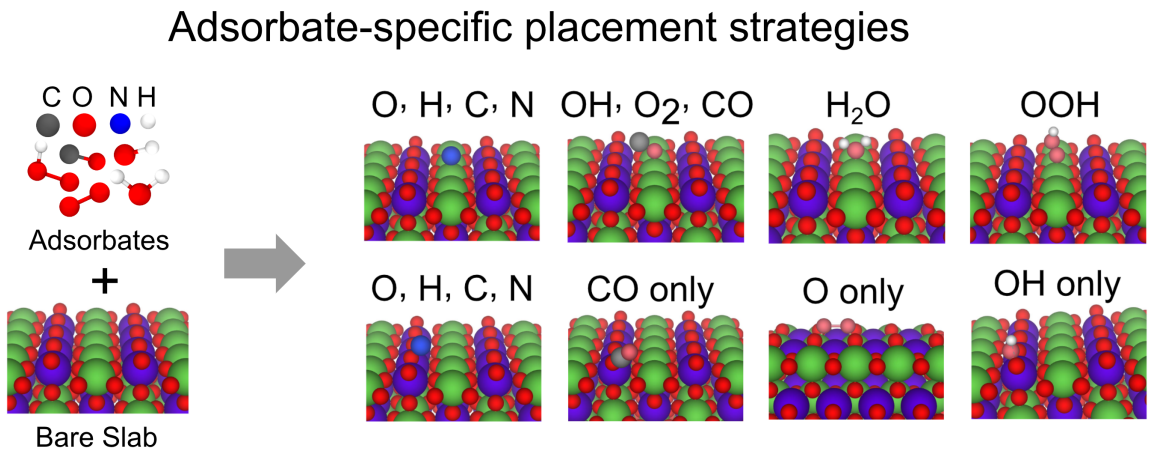


Figure 5-3: Overview of the adsorbate specific placement strategies. Adsorbates include C, O, N, H, OOH, CO, OH, O_2, and H_2O (left). Adsorbates can either bind to undercoordinated surface metals (first row of strategies) or to surface oxygen to form new intermediates (second row).

5.3.3 Initial Structure Generation

To construct our adsorbate+slab, we first randomly sample one adsorbate from the set shown in Figure 5-3. This adsorbate set includes O, *OH, *OH_2, *OOH, and *O_2 which are the intermediates in the proposed reaction mechanisms of OER. To expand the possible chemistry of adsorbates on oxides beyond OER, we also included monatomic *O, *H, *C and *N, as well as *CO. Table 5.1 shows the distribution of the 9 sampled adsorbates across the dataset.

We then determine the coverage of our random adsorbate on our randomly constructed slab. In contrast to the OC20 dataset, here we allow for more than one adsorbate of the same type to bind to the surface. The adsorbate can bind to three types of sites: the surface oxygen, the under-coordinated surface metal, or an oxygen vacancy. The maximum number of adsorbates allowed on the surface is limited by the sum of these three types of sites. However we also ensure that all adsorbates are always separated by a distance greater than the M-O bond of the host material to avoid adsorbate overcrowding.

In this effort, we implemented specific strategies for placing adsorbates on the aforementioned surface sites as shown in Figure 5-3. The first row of placement strategies demonstrates that all adsorbates are able to bind to any undercoordinated

surface metal at the lattice position of oxygen. This includes lattice positions of vacancies introduced during slab generation. An adsorbate containing oxygen will always bind to the metal via the oxygen atom as shown for *OH, $*O_2$, *CO, $*H_2O$ and *OOH. We also considered intermediates that arise due to formation of oxygen dimers which play a role in one of the possible mechanisms of OER[83, 53]. In this configuration, a pair of monatomic oxygen atoms can adsorb on to adjacent undercoordinated metals to form a dimer of 1.68 Å which is longer than the bond length of $*O_2$.

The second and third rows demonstrate how specific molecules that are able to form new molecules with the addition of oxygen can also bind to existing surface oxygen. For example, binding to a surface oxygen with the monatomic adsorbates will form a dimer molecule whereas *CO and *OH can bind to form $*CO_2$ and *OOH respectively. Incorporating these reactions in the dataset will allow for the exploration of intermediate surface reactions that are only possible on oxides.

Lastly, we also allowed for a four-fold rotational degree of freedom about about the normal of the surface for all adsorbates. We randomly select the degree of rotation for each adsorbate on the surface after identifying the adsorbate sites.

5.3.4 Structure Relaxation

The OC22 dataset uses different computational settings than those used for the OC20 dataset. The OC22 dataset models the exchange-correlation effects with the Perdew-Berke-Ernzerhof (PBE), generalized gradient approximation (GGA) [197] which is generally accepted for modeling surface reactions on oxides[83, 104, 271]. In contrast, the OC20 dataset utilizes the RPBE DFT functional. We also accounted for strong electron correlations in some transition metal oxides by applying the Hubbard U correction in accordance to the suggestions made by the Materials Project[114]. The last row of Table 5.1 shows the total number of slabs and adsorbate+slabs calculated using Hubbard U corrections. Although higher-level theory single-point calculations (e.g. hybrid functionals[221]) are often used to verify the final electronic structure and energy of a surface, they still use a scheme similar to the one here to obtain the

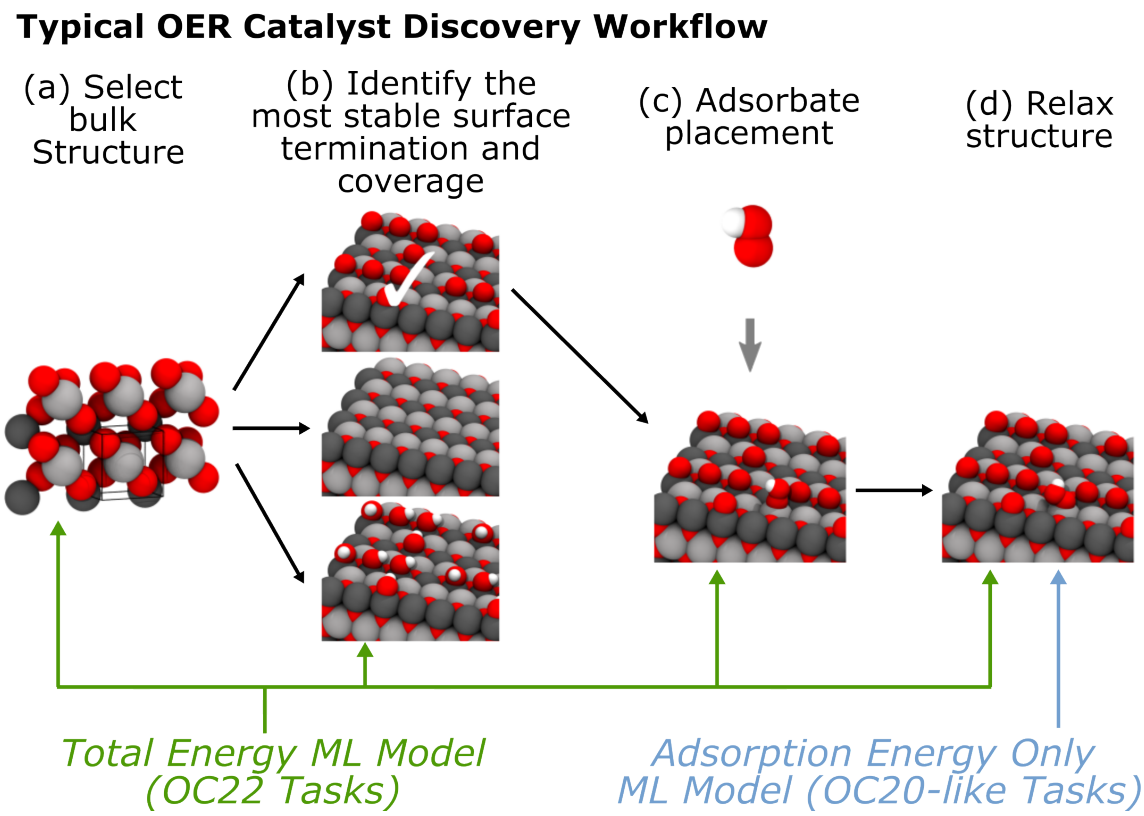


Figure 5-4: A typical OER workflow, motivating the need for total energy models beyond adsorption energies. Total energy models would allow one to study all parts of this workflow, and not just the final relaxation like adsorption energy models. (a) A bulk structure is selected from material datasets like the Materials Project[114] and a surface is created. (b) Surface terminations are enumerated and studied with DFT to identify the most stable termination. Surface Pourbaix diagrams are created and used to make this decision. (c) Only after the most stable termination is identified, an adsorbate is placed and (d) The adsorbate+slab system is relaxed and the referenced adsorption energy is computed.

optimized structure. Models developed for this dataset will greatly accelerate more accurate workflows by focusing expensive calculations on the most stable and relevant structures.

All calculations were performed with spin-polarization to account for the significant spin states in metal oxides. Although some oxide materials can have several magnetic configurations, including antiferromagnetism, we only considered one configuration for each slab with all slabs being initialized with ferromagnetic or nonmagnetic configurations. These different magnetic states for a single crystal structure can significantly change thermodynamic properties at the surface. For example, rutile

VO_2 has been demonstrated to have several different spin states with nonmagnetic surfaces yielding significantly lower surface energies than ferromagnetic surfaces for the same slab[277]. For further details regarding the computational settings, we refer the reader to the SI.

We allowed all atoms of the slab and adsorbate+slab to be relaxed. This will not only yield a lower DFT energy, but also allows for more accurate calculations of the surface energy by ensuring both surfaces are relaxed. This is in contrast to the OC20 dataset where only the adsorbates and the surface atoms were relaxed.

Systems that did not converge ionically were set aside for use in alternative tasks. All intermediate structures, energies, and forces are stored for future training and evaluation. All input structures were constructed with the aid of Python Materials Genomics (pymatgen)[187] and all calculations are performed using the Vienna ab initio simulation package (VASP) [140, 138, 139, 273, 141]. In total, we used over 20 million compute hours to create this dataset.

5.4 Tasks

The goal of the OC22 dataset is to efficiently simulate atomic systems with practical relevance to OER and other oxide applications. Similar to the OC20 dataset, the primary bottleneck to doing so are computationally expensive DFT calculations. Calculations are further exacerbated for OC22 as its systems are larger and more complex than that of OC20. Again, we focus on structure relaxations as they have been a useful means to informing catalyst activity for a broad range of applications[110, 180, 31, 97, 236, 182]. Models developed for OC20 have shown great progress on their proposed tasks[75, 290, 241, 108, 76, 136]. In all of OC20's tasks, energies were referenced to represent adsorption energy. While advantageous for screening purposes, this referencing, however, implicitly limited models to only studying adsorbate+slab combinations and not any one in isolation. In the context of OER, this is especially problematic as typical discovery pipelines require exploring different coverages and configurations of the surface [93, 14, 279, 195, 294, 275, 68].

Figure 5-4 illustrates a typical workflow for OER where studying different surface terminations are necessary before running an adsorption calculation. Here, we propose modified variations of OC20's tasks that would enable models to study surfaces with and without the presence of an adsorbate.

In all tasks, structures can contain a surface and adsorbate combination or just an isolated surface (a.k.a slab). The surface is defined by a unit cell periodic in all directions with a vacuum layer at least 12Å. All ground truth targets are computed using DFT.

We briefly summarize the OC20 tasks below. For all tasks, energy is referenced to correspond to adsorption energy. See the original OC20 manuscript for more details [40]. ***S2EF*** takes a given structure and predicts the energy and per-atom forces. ***IS2RE*** takes an initial structure and predicts the relaxed energy. ***IS2RS*** takes an initial structure and predicts the relaxed structure.

In the curation of both OC20 and OC22, slabs and adsorbate+slabs were relaxed in parallel, with adsorbates being placed on unrelaxed slabs. OC20 makes an assumption in computing an adsorption energy such that the corresponding relaxed slab reference is comparable to that of the adsorbate+slab combination. This assumption was feasible given that the majority of the surface was constrained.

Unlike OC20 where surface atoms are constrained, all atoms in OC22 are unconstrained. While this enables the community to study other surface properties like surface energy, the assumption that the relaxed clean surface and adsorbate+slab surface are comparable no longer holds. Computing an adsorption energy in the same manner of OC20 would correspond to an incorrect reference, resulting in an ill-posed, noisy target (see SI for more details). Instead, we modify OC20's *S2EF* and *IS2RE* tasks to target DFT total energy rather than adsorption energy. We use the *IS2RS* task as is with no modifications.

S2EF-Total takes a given structure and predicts the DFT total energy and per-atoms forces. Compared to *S2EF*, *S2EF-Total* differs only in its energy prediction. *S2EF* takes the DFT total energy and references it by subtracting off a clean surface and gas phase adsorbate energy. *S2EF-Total* is only interested with the DFT total

energy. The two tasks are related as follows:

$$\hat{E}_{S2EF} = \hat{E}_{S2EF\text{-}Total} - E^{DFT}_{slab} - E^{DFT}_{gas} \tag{5.1}$$

IS2RE-Total takes a given structure and predicts the relaxed DFT total energy. Similar to *S2EF-Total*, *IS2RE-Total* is related to *IS2RE* as follows:

$$\hat{E}_{IS2RE} = \hat{E}_{IS2RE\text{-}Total} - E^{DFT}_{slab} - E^{DFT}_{gas} \tag{5.2}$$

DFT total energies are not meaningful on their own. Physically relevant properties like adsorption energy include some reference. A model that can predict a DFT total energy, however, gives the flexibility to reference to whatever is desired. Adsorption energy in this context would involve two predictions - one of the adsorbate+slab and one of the clean surface. For OER this is particularly important to identify the most stable surface coverage (or termination). While this problem is also important for OC20, those systems were much less complicated and the proposed adsorption energy tasks are typically sufficient.

Of the proposed tasks, *S2EF-Total* is the most general and closest to a DFT surrogate. Models trained for this task would enable researchers to also study isolated surfaces, a necessary and important step in the catalyst discovery pipeline. Total energies also allows us to leverage surface trajectories and their energies for training, data that was previously unusable in OC20 using the specified bare slab energy reference.

5.5 Dataset Splits

Similar to OC20, we split our dataset into training, validation, and test splits. Training and validation splits are used to optimize and tune hyperparameters and the test set to report performance.

To explore the extrapolative ability of our models, we split the validation and test sets along the catalyst composition dimension. Unlike OC20, we exclude adsorbate

Table 5.2: Size of train and validation splits. *S2EF-Total* structures come from a superset of *IS2RE-Total* systems, including unrelaxed systems (e.g. 50,810 train systems). Splits are sampled based on catalyst composition, ID for those from the same distribution as training, OOD for unseen catalyst compositions. Splits consist of both adsorbate+slab (adslabs) and slab systems. Validation and test splits are similar in size with exclusive compositions.

Task	Train			ID			OOD		
	Adslabs	Slabs	**Total**	Adslabs	Slabs	**Total**	Adslabs	Slabs	**Total**
S2EF-Total	6,642,168	1,583,125	8,225,293	313,238	81,489	394,727	356,633	94,036	450,669
IS2RE-Total	31,244	14,646	45,890	1,701	923	2,624	1,862	918	2,780
IS2RS	31,244	14,646	45,890	1,701	923	2,624	1,862	918	2,780

extrapolation given the low number of adsorbates present in OC22. Splits were created by first enumerating all possible catalyst compositions in the dataset. From the list of available compositions, a fraction is held out from the training set. Samples were selected from a total of 1,138 unique compositions.

We split our validation and test set into two subsets: ID (sampled from the same catalyst composition training distribution) and OOD. Subsplit sizes for all tasks are given in Table 5.2, with adsorbate+slab (adslabs) and slab counts also shown. Extrapolative subsplits of the validation and test sets are exclusive from one another, e.g. the catalyst compositions held out for OOD are different for the validation and test sets.

5.6 Baseline GNN Models

A wide range of models for catalyst and molecular applications have been proposed [75, 76, 241, 108, 290, 80, 155, 250]. We evaluate our tasks using the latest state of the art models. Additionally, we baseline alternative model architectures including equivariant and (non)energy-conserving models. Code for all baseline models are implemented in PyTorch[194] and PyTorch Geometric[65], and are publicly available in our open source repository at `https://github.com/Open-Catalyst-Project/ocp`.

GNNs have continued to grow in popularity as an efficient and accurate architec-

ture for modeling atomic interactions. Unlike descriptor based models [27, 29, 48, 20], where hand crafted representations are used to describe atomic environments, GNNs learn atomic representations through several message passing steps [79]. Consistent with related work[40, 232, 133], graphs are constructed with atoms treated as nodes and interactions between atoms as edges. Periodic boundary conditions are accounted for in graph construction consistent with OC20. A cutoff radius is introduced for computational tractability.

We benchmark GNNs that have either performed well on OC20 or other molecular datasets. For *S2EF-Total*, we benchmark a larger sample of models including SchNet[232], DimeNet++[131], ForceNet[108], SpinConv[241], PaiNN[233], GemNet-dT[75], and GemNet-OC[76]. *IS2RS* baselines are limited to the top performing models - SpinConv, GemNet-dT, and GemNet-OC. *IS2RE-Total* baselines include SchNet, PaiNN, DimeNet++, and GemNet-dT. Top performing *S2EF-Total* models were also evaluated for *IS2RE-Total* via an iterative relaxations approach[40].

SchNet and DimeNet++ proposed continuous edge filters and directional message passing, respectively. ForceNet and SpinConv proposed architectures with direct force predictions in place of using energy derivatives with respect to atomic positions. PaiNN is an equivariant model with spherical harmonics up to order $l = 1$. We modify PaiNN's original architecture to make direct force predictions as our experiments showed a boost in performance. GemNet-dT incorporates symmetric message passing, scaling factors, equivariant predictions, and several efficient architecture improvements over the similar DimeNet++. GemNet-OC expands on GemNet-dT to efficiently capture quadruplet interactions, the current state of the art model across all tasks for OC20.

Unless otherwise noted, graph edges were computed on-the-fly via a nearest neighbor search for a cutoff radius of 6Å and a maximum of 50 neighbors per atom. GemNet-OC uses different cutoffs for the type of interaction, e.g. triplets and quadruplets. Initial model sizes were taken directly from corresponding OC20 configurations. To accommodate for the fact OC22 has 16x less data, a light hyperparameter sweep was done for all models, with particular emphasis on learning rates, schedulers, and

batch sizes. Effective batch sizes were set to ∼192-256 for *S2EF* and ∼4-64 for *IS2RE*. *S2EF* models used identical learning rate schedulers to more fairly compare baselines, decaying the learning rate at epochs 2, 3, 4, 5, and 6. *IS2RE* used a reduce on plateau learning rate scheduler. Full details on model hyperparameters and training configurations can be found in the SI.

All experiments used the following loss function[40] to balance energy and force predictions:

$$\mathcal{L} = \lambda_E \sum_i |E_i - E_i^{DFT}| \tag{5.3}$$
$$+\lambda_F \sum_{i,j} \frac{1}{3N_i} |F_{ij} - F_{ij}^{DFT}|^p$$

where λ_E and λ_F are emperical parameters, E_i is the energy of system i, F_{ij} is the force on the jth atom in system i, N_i is the number of atoms in system i, and p is the norm order. With the exception of GemNet-dT and GemNet-OC which used $p = 2$, all *S2EF-Total* models used $p = 1$. For *IS2RE-Total* only the energy term is evaluated, i.e $\lambda_F = 0$. Baseline *S2EF-Total* models were trained with $\lambda_E = 1$ and $\lambda_F = N_{atoms}^2$ to insure size invaraince, as detailed by Batzner, et al. [24, 171]

5.7 Experiments

Here we describe the evaluation metrics, training experiments, and share results for our baseline models.

5.7.1 Evaluation Metrics

All our tasks use the same evaluation metrics proposed by OC20. The only difference is rather than ground truth values being DFT adsorption energies, we use DFT total energies for OC22. We briefly mention the metrics below but refer readers to the OC20 manuscript[40] for a more detailed description.

S2EF-Total: The *S2EF-Total* task uses the same metrics as OC20's *S2EF* task. Metrics include Energy Mean Absolute Error (MAE), Force MAE, Force cosine, and Energy Forces within Threshold (EFwT). Ground truth targets correspond to DFT total energy and per-atom forces.

IS2RE-Total: Similarly, *IS2RE-Total* uses the same metrics as OC20's *IS2RE* task. Metrics include Energy MAE and Energy within Threshold (EwT). Ground truth targets correspond to the DFT total energy of the relaxed structure.

IS2RS: *IS2RS* metrics here are identical to that of OC20. Metrics include Average Distance within Threshold (ADwT), Forces below Threshold (FbT), and Average Force below Threshold (AFbT). Ground truth targets are the relaxed structure. DFT is also used to evaluate predicted relaxed structures.

Consistent with OC20, our evaluation metrics still focus on accuracy. Given the complexity of OC22, we are interested in how previously successful models will perform on larger more intricate systems. In addition, we focus on models that are significantly faster than traditional DFT-based techniques. Models that can calculate energy and force estimates in under 10ms would significantly aid oxide-related research.

5.7.2 Training Experiments

The availability of large, diverse datasets like OC20 allows us to explore more interesting experiments alongside the OC22 dataset. In addition to training our baseline models on just OC22 we examine the extent the OC20 dataset and its pretrained models can benefit OC22 performance, and vice-versa.

The varied training strategies are summarized in Figure 5-5. For each task we first study the performance using baseline models just trained on OC22 (**OC22-only**). This is the standard strategy when introducing a new dataset. Next, we leverage both OC20 and OC22 via **joint training** (OC20+OC22). In joint training we train a combined dataset of OC20 and OC22 systems. For *S2EF-Total*, we explore combined datasets with different sizes of OC20 - 2M, 20M, and All. While OC20's energies were originally expressed as adsorption energy, for these ex-

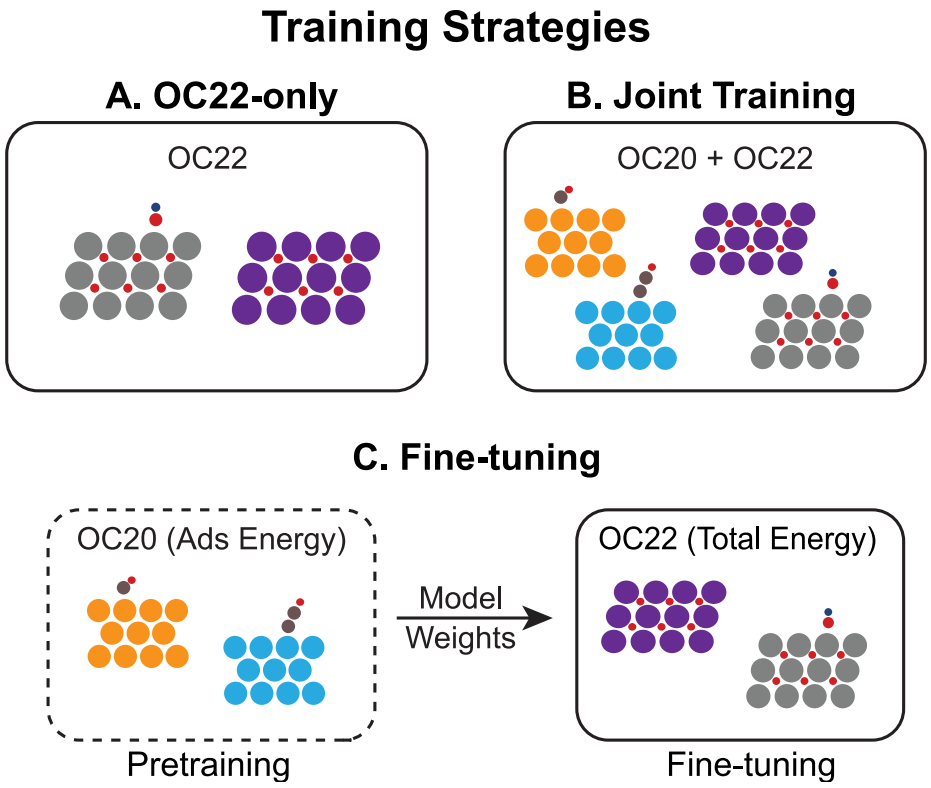


Figure 5-5: The various training strategies explored in OC22. **A.** The OC22-only strategy involves just using OC22 for the proposed tasks. **B.** Joint training refers to models trained on both OC20 and OC22 simultaneously. **C.** In fine-tuning, pretrained models for OC20 are used as starting points to train on just OC22.

Table 5.3: Predicting total energy and force from a structure (*S2EF-Total*). Results are shared for the OC22-only, joint, and fine-tuning training strategies. Experiments are evaluated on the test set.

		S2EF-Total Test							
Training	Model	Energy MAE [eV] ↓		Force MAE [eV/Å] ↓		Force Cosine ↑		EFwT [%] ↑	
		ID	OOD	ID	OOD	ID	OOD	ID	OOD
OC22-only	Median Baseline	163.424	160.455	0.075	0.073	0.002	0.002	0.00	0.00
	SchNet [232]	7.926	7.924	0.060	0.082	0.363	0.220	0.00	0.00
	DimeNet++ [133, 131]	2.098	2.476	0.043	0.059	0.606	0.436	0.00	0.00
	ForceNet [108]	-	-	0.057	0.062	0.349	0.277	0.00	0.00
	SpinConv [241]	1.101	1.981	0.048	0.070	0.514	0.386	0.00	0.00
	PaiNN [233]	0.956	2.632	0.045	0.058	0.504	0.367	0.00	0.00
	GemNet-dT [75]	0.938	1.272	0.032	0.041	0.665	0.530	0.01	0.00
	GemNet-OC [76]	0.383	0.833	0.029	0.040	0.690	0.554	0.03	0.00
OC20-2M + OC22	PaiNN[233]	0.412	1.532	0.048	0.064	0.484	0.357	0.00	0.00
	SpinConv[241]	0.875	1.722	0.035	0.054	0.720	0.550	0.00	0.00
	GemNet-OC [76]	0.426	0.918	0.029	0.037	0.698	0.553	0.01	0.01
OC20-20M + OC22	PaiNN[233]	0.375	1.456	0.046	0.061	0.496	0.358	0.00	0.00
	SpinConv[241]	0.929	1.512	0.035	0.052	0.629	0.470	0.00	0.00
	GemNet-OC [76]	0.320	0.832	0.027	0.037	0.722	0.593	0.06	0.01
OC20-All + OC22	SpinConv[241]	1.217	1.625	0.039	0.046	0.564	0.452	0.00	0.00
	GemNet-OC [76]	0.323	0.695	0.027	0.034	0.698	0.589	0.06	0.00
OC20→OC22	SpinConv[241]	0.978	1.877	0.035	0.049	0.620	0.463	0.01	0.00
	GemNet-dT [75]	0.729	1.327	0.031	0.041	0.667	0.534	0.01	0.00
	GemNet-OC [76]	0.260	0.943	0.030	0.041	0.679	0.546	0.08	0.01
	GemNet-OC-Large [76]	0.458	1.238	0.028	0.040	0.724	0.573	0.04	0.00

periments we use the DFT total energy which is also publicly accessible. One of the limitations to joint training is the need to train on a larger combined dataset, which can significantly increase training time. To address this, we additionally explore **fine-tuning** (OC20 → OC22) experiments. In fine-tuning, models are initialized with pretrained weights learned from training on OC20. The pretrained models are then fine-tuned by training on just OC22. While approaches to fine-tuning vary in which portion of the network's weights are updated, we limit our experiments to updating all the weights and leave more rigorous strategies as future work for the community [135]. For *S2EF-Total*, we experiment with fine-tuning using different fractions of the OC22 dataset. All fine-tuning experiments are performed using public OC20 adsorption-energy model checkpoints found at `https://github.com/Open-Catalyst-Project/ocp/blob/main/MODELS.md`.

Through these experiments we hope to share results that provide insights beyond just performance on OC22. Building a dataset that spans all possible applications, chemical diversity, and level of DFT theory is not computationally feasible. However,

Table 5.4: *S2EF-Total* fine-tuning results trained on various fractions of the OC22 dataset. GemNet-OC[76] was used for all experiments. Note, a fraction of 0% for OC22 corresponds to the baseline of directly evaluating a pretrained checkpoint from OC20 on OC22, with no additional training. All experiments are evaluated on the test set.

		S2EF-Total Test							
Training	Fraction of OC22	Energy MAE [eV] ↓		Force MAE [eV/Å] ↓		Force Cosine ↑		EFwT [%] ↑	
		ID	OOD	ID	OOD	ID	OOD	ID	OOD
OC22-only	5%	0.593	1.800	0.043	0.048	0.497	0.410	0.00	0.00
	15%	0.385	1.468	0.036	0.046	0.612	0.481	0.02	0.00
	30%	0.368	1.326	0.033	0.045	0.661	0.511	0.03	0.00
	50%	0.381	1.209	0.032	0.044	0.659	0.517	0.05	0.00
	100%	0.383	0.833	0.029	0.040	0.690	0.554	0.03	0.00
OC20→OC22	0%	487.121	434.690	0.365	0.362	0.194	0.195	0.00	0.00
	5%	0.559	1.397	0.037	0.039	0.549	0.475	0.00	0.00
	15%	0.326	1.038	0.033	0.038	0.622	0.516	0.03	0.00
	30%	0.270	0.984	0.031	0.038	0.654	0.538	0.06	0.01
	50%	0.254	0.921	0.029	0.039	0.679	0.550	0.08	0.00
	100%	0.260	0.943	0.030	0.041	0.679	0.546	0.08	0.01

as we demonstrate with OC22, by leveraging large datasets, such as OC20, we may be able to train effective models with much smaller datasets for specific domains; even if they contain critical differences like DFT theory and material compositions.

5.7.3 Results

We report results for all baseline models and tasks below. All validation results can be found in the SI.

S2EF-Total: Results on SchNet[232], DimeNet++[131], ForceNet[108], SpinConv[241], PaiNN[233], GemNet-dT[75], and GemNet-OC[76] are shown in Table 5.3(top). All models make energy and per-atom force predictions. SchNet and DimeNet++ make force predictions via a gradient of energy with respect to atomic positions, while all other models make direct force predictions. Across all metrics, GemNet-OC performs the best. While GemNet-dT also demonstrates competitive force metrics, GemNet-OC significantly outperforms all models on energy based metrics. This may be due to GemNet-OC's large receptive field (cutoff=12Å) better capturing long-range interactions and its unique ability to explicitly capture quadruplet interactions.

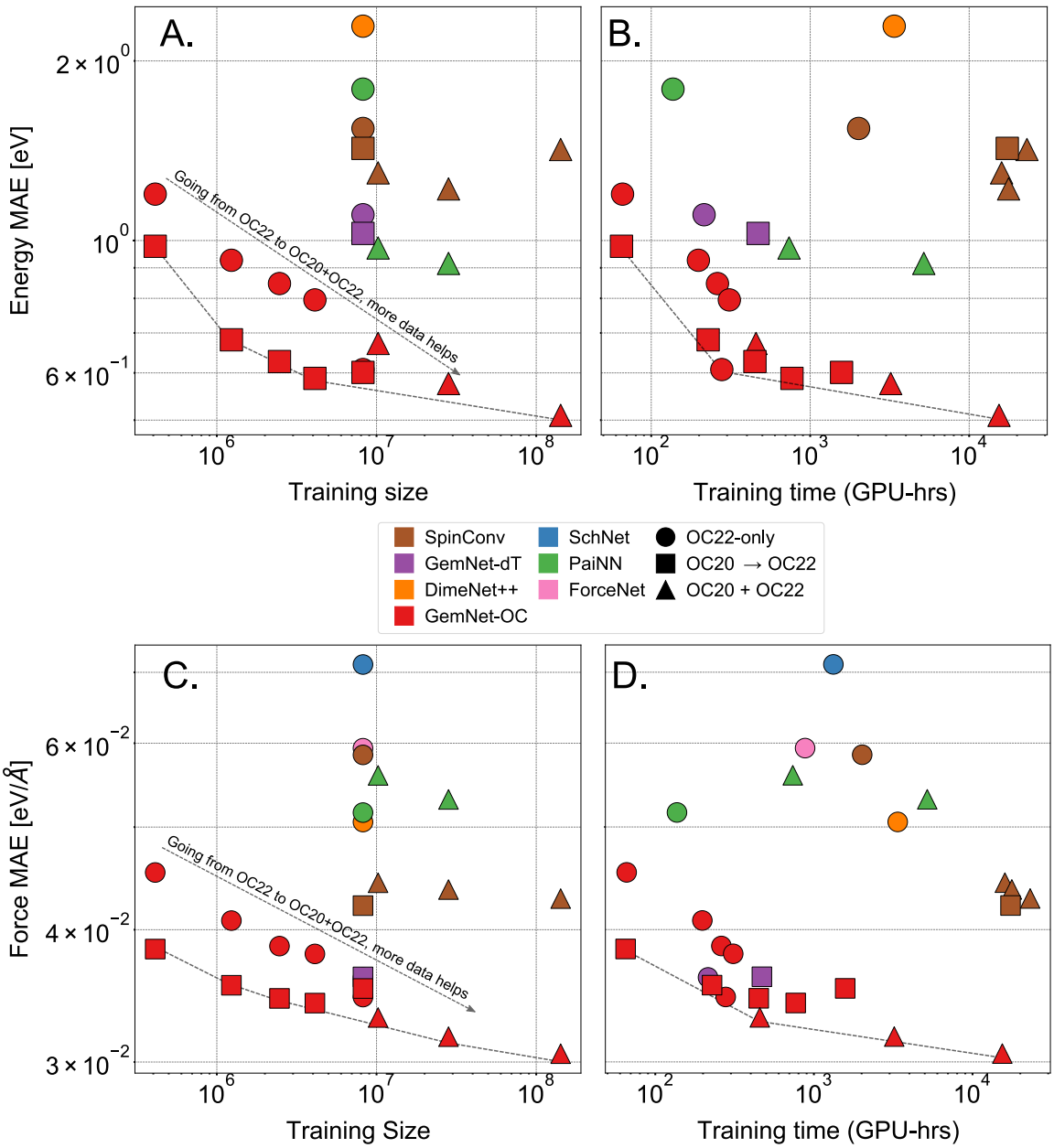


Figure 5-6: Summary of *S2EF-Total* test results as a function of training size (A,C) and training time (B,D). Models are color coded and the respective training strategy is indicated by different shapes. For fixed dataset sizes, fine-tuning experiments see improvements in both energy and force predictions. Increasing data consistently helps performance when moving from OC22 to OC20+OC22. Pareto fronts are provided for current optimums across training sizes and times. Fine-tuning experiments do not consider the dataset sizes and training times used during pretraining. Results are averaged across both ID and OOD splits.

Results across the two test subsplits, In Domain (ID) and Out of Domain (OOD), are shown in Table 5.3. As expected, ID metrics are better than OOD. Unlike OC20

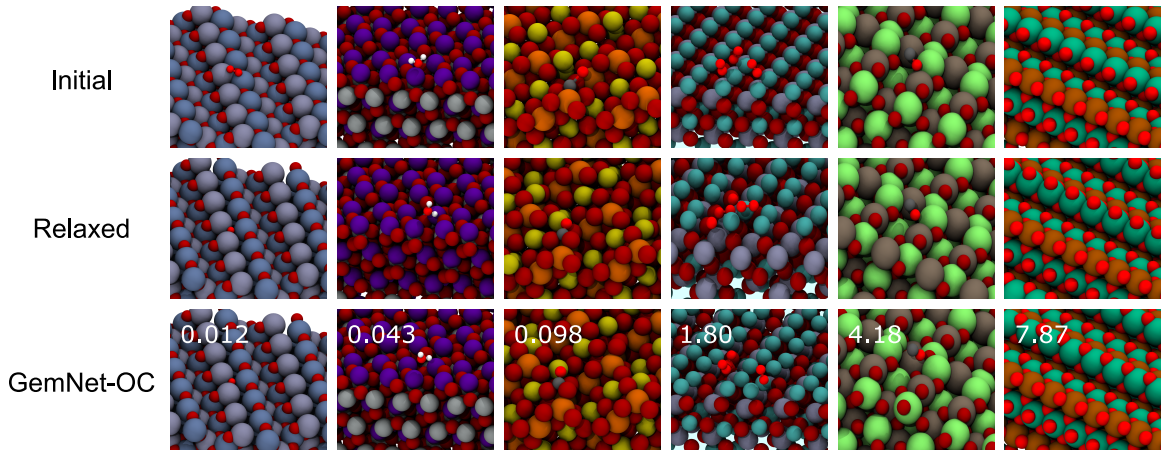


Figure 5-7: Demonstration of GemNet-OC[76] solving the *IS2RS* and *IS2RE-Total* tasks via the relaxation approach. Initial, DFT Relaxed, and the ML predicted relaxed structures are shown for each system. The first three columns were randomly sampled from "successful" cases in which *IS2RE-Total* energy MAE was less than 0.1 eV, while the latter columns are "failure" cases, with energy MAEs greater than 0.5 eV. Oxygen found in the adsorbate is illustrated with a high contrast red and made smaller to distinguish it from oxygen in the catalyst material.

where ID and OOD-based splits had fairly close metrics, OC22 OOD metrics are substantially higher than ID. By definition, OOD contains combinations of material species not seen in the training set, i.e., if Ag-Cu is OOD, then a Ag-Cu only interaction has never been seen during training. This suggests generalization in the context of total energy predictions is more challenging than a referenced adsorption energy. Although physically motivated, OC20's adsorption energy target can also be thought of as a form of Δ-learning [242, 213, 297], simplifying the complexity of the problem to learning a correction to some base property. To explore this in the context of OC22, we report results on a per-element linearly fit reference in the SI that helps improve performance. We refrained from making this the base task for OC22 in order to encourage alternative schemes or approaches to target normalization. OC20 results on the proposed tasks are also available in the SI, with similar poor performance suggesting *S2EF-Total* to be a generally more challenging task.

Joint training experiments on OC20 and OC22 are conducted for the top performing models, GemNet-OC, PaiNN, and SpinConv. Table 5.3 additionally contains results of different sizes of OC20 combined with OC22. To stay consistent

Table 5.5: Predicting total relaxed energy from an initial structure (*IS2RE-Total*). Results are shared for the OC22-only, joint, and fine-tuning training strategies. Experiments are evaluated on the test set.

IS2RE-Total Test

Approach	Training	Model	Energy MAE [eV] ↓		EwT [%] ↑	
			ID	OOD	ID	OOD
Direct	OC22-only	Median Baseline	176.256	171.854	0.00	0.00
		SchNet	2.001	4.847	1.03	0.45
		DimeNet++	1.960	3.519	0.65	0.38
		PaiNN	1.716	3.684	0.88	0.38
		GemNet-dT	1.677	3.084	1.49	0.45
	OC20+OC22	SchNet	3.038	4.300	0.38	0.53
		DimeNet++	1.961	3.461	1.18	0.42
		PaiNN	1.733	3.752	0.76	0.49
		GemNet-dT	2.523	4.229	0.80	0.60
	OC20→OC22	GemNet-OC*	1.153	1.748	3.66	0.98
Relaxation	OC22-only	SpinConv	1.948	2.696	1.11	0.64
		GemNet-dT	1.813	2.044	1.64	0.83
		GemNet-OC	1.329	1.584	2.02	1.40
	OC20+OC22	SpinConv	2.296	2.590	1.26	0.68
		GemNet-OC	1.201	1.534	2.63	2.15
	OC20→OC22	SpinConv	1.800	2.888	1.41	0.57
		GemNet-OC	1.120	1.849	3.89	1.77
		GemNet-OC-Large	1.253	2.115	1.60	0.98

*GemNet-OC pretrained on OC20+OC22 *S2EF-Total*

with OC22, DFT total energy targets were used for OC20. With the addition of OC20 training data, GemNet-OC and SpinConv saw improvements in both energy and force predictions while PaiNN only saw improvements to energy. This suggests that despite the differences in DFT theory, the additional data is still meaningful in improving model predictions. However, increasing the amount of OC20 data had mixed results. GemNet-OC generally saw improvements across all metrics while SpinConv and PaiNN saw either minor improvements or worse performance. We note that training samples were randomly drawn, i.e., experiments with a larger

proportion of OC20 would have seen fewer samples of OC22 during training. The differences in trends could be a result of model data efficiency and capacity. Exploring alternative sampling strategies to joint training could aid models and improve trends further. For our fine-tuning experiments, we evalute GemNet-OC, SpinConv, and GemNet-dT models. Fine-tuning is performed by first training a model on OC20. This pre- trained model is then fine-tuned by training on only OC22. Trained OC20 models are publicly available and were directly obtained from `https://github.com/Open-Catalyst-Project/ocp`. SpinConv saw improvements across all metrics. GemNet-dT generally saw minor improvements or in the case of some OOD performance, worse results. Similarly, GemNet-OC saw significant improvements to energy MAE for ID data, but saw minor changes and worse OOD results, respectively. To drive performance further, we trained GemNet-OC-Large, a larger, more parameterized version of GemNet-OC. The large variant resulted in improved force metrics, but at the cost of worse energy metrics. Fine-tuning experiments were extremely delicate and required careful tuning, details are highlighted in the SI. While our initial fine-tuning results were limited to energy improvements, we hope the future development of more rigorous methods could lead to better performance across all metrics.

A potential benefit accompanying pretraining and fine-tuning is the need for less training data. A model initialized with meaningful weights could simplify the need to learn interactions and representations from scratch by utilizing an alternative dataset. To explore this, we evaluated the performance of a pretrained GemNet-OC model fine-tuned on various fractions of OC22. As shown in Figure 5-8, a fine-tuned GemNet-OC consistently outperforms its OC22-only variant across all data sizes for the ID split, with diminishing returns for both strategies around $\sim$50%. On OOD, energy performance continues to improve with data size. In Table 5.4, we additionally show the performance of a pretrained OC20 GemNet-OC used to directly evaluate OC22 (Fraction = 0%). As expected, energy metrics are extremely poor given OC20's original target is adsorption energy. Force metrics are also extremely poor, suggesting the fine-tuning performance is not merely a result of a good pretrained model, but an

actual transfer of knowledge from the two datasets. Figure 5-6 illustrates the various models and approaches as a function of training size and time. Notably, we see a strong linear trend in performance with data size. With saturation yet in sight, we expect more joint dataset efforts to continue to aid in performance. While for a fixed dataset size, fine-tuning efforts improved performance, they were often more costly in training time (Figure 5-6 B/D). We anticipate future fine-tuning developments to be not only more accurate, but efficient as well. Similar fine-tuning experiments with OC20 models trained on DFT total energy targets were also performed. Results were consistent with those shared above, suggesting that despite a difference in targets, models are learning a similar underlying representation that is being transferred to OC22.

IS2RE-Total: We explore two approaches for predicting relaxed energies from initial structures - “Direct" and “Relaxation" [40]. The first directly predicts the relaxed energy with a single call to the model. The relaxation approach uses a *S2EF-Total* model to run a structural relaxation - iteratively predicting forces and updating atomic positions until a relaxed structure and its corresponding energy is reached. While OC20 has shown relaxation based approaches to be superior to direct, they are 200-300x slower, motivating the potential benefits of direct models.

Table 5.5 presents *IS2RE-Total* results on both direct and relaxation approaches under the different training scenarios. Whereas OC20 saw relaxation based approaches to consistently perform better, we see mixed results here. The best relaxation-based approach, GemNet-OC, achieves an EwT of 3.89% indicating models have significant room for improvement. For the relaxation approach, fine-tuning consistently outperforms OC22-only. The best direct approach, GemNet-OC, also only achieves an EwT of 3.66%. Here, joint training consistently hurts performance. Following literature efforts[250], fine-tuning was done from the top performing *S2EF-Total* checkpoint - GemNet-OC OC20-All+OC22. While the best performing ID results come from a direct approach, OOD metrics are considerably better via the relaxation method, indicating their ability to better generalize. We evaluate OC20 *IS2RE-Total* performance in the SI and observe similar poor results, suggesting *IS2RE-Total* to be a

Table 5.6: Predicting relaxed structures from initial structures (*IS2RS*). All models predicted relaxed structures through an iterative relaxation approach. The initial structure was used as a naive baseline (IS baseline). Experiments are evaluated on the test set.

IS2RS Test

Training	Model	ADwT [%] ↑		FbT [%] ↑		AFbT [%] ↑	
		ID	OOD	ID	OOD	ID	OOD
OC22-only	IS baseline	43.39	45.26	0.00	0.00	0.03	0.10
	SpinConv	56.47	50.60	0.00	0.00	2.77	1.18
	GemNet-dT	57.84	54.17	0.00	0.00	4.16	3.54
	GemNet-OC	59.47	55.72	0.00	0.00	5.49	4.45
OC20+OC22	SpinConv	53.99	52.39	0.00	0.00	2.64	2.38
	GemNet-OC	58.55	58.44	0.00	0.00	8.01	6.58
OC20→OC22	SpinConv	54.21	51.42	0.08	0.00	6.31	3.24
	GemNet-OC	55.55	50.50	0.08	0.00	9.02	6.59
	GemNet-OC-Large	57.23	54.63	0.00	0.00	10.41	8.09

considerably more challenging variation.

IS2RS: To evaluate the prediction of relaxed structures from initial structures, we select the top performing *S2EF-Total* models GemNet-dT, SpinConv, and GemNet-OC. Similar to OC20, we use these models to run ML driven structure relaxations (Figure 5-7). Relaxed structures were then evaluated with DFT to determine whether the predicted relaxed structures are valid. Table 5.6 shows GemNet-OC outperforming all other models across all metrics. Joint training and fine-tuning approaches both improve DFT force based metrics over OC22-only. GemNet-OC-Large fine-tuned achieves the best force metrics. Pursuant to OC20, non-DFT distance based metrics like ADwT struggle to correlate well with the practical DFT metrics [136]. Both FbT and AFbT results indicate the models need significant improvement to achieve the level of accuracy needed for practical applications.

Does OC22 benefit OC20? Alongside developing more accurate models, exploring augmentation strategies is another opportunity to improve performance on existing datasets like OC20[136]. An interesting question is whether OC22 data may

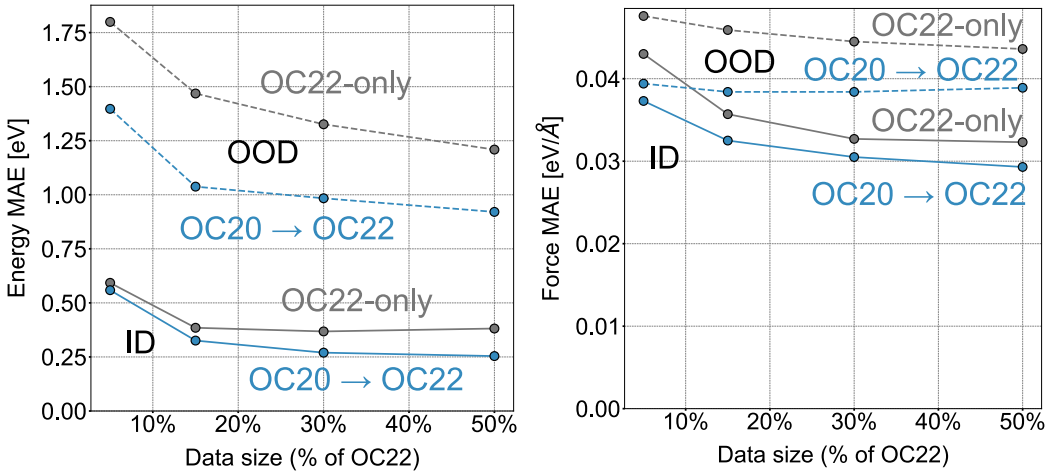


Figure 5-8: Results of GemNet-OC on *S2EF-Total* across different training data sizes. Two strategies are compared here - OC22-only and fine-tuning. Results are reported for both ID (solid) and OOD (dashed) on the test set.

Table 5.7: GemNet-OC results trained on either OC20 or both OC20+OC22 and evaluated on OC20 and OC22. Results are averaged across all ID/OOD validation splits. Total energies are used for all dataset targets.

Training Data	Energy MAE [eV] ↓	Force MAE [eV/] ↓	Force Cosine ↑
	OC22 evaluation		
OC20	55.900	0.384	0.167
OC20+OC22	0.661	0.031	0.657
	OC20 evaluation		
OC20	0.394	0.022	0.651
OC20+OC22	0.317	0.023	0.649

improve model performance on OC20. It has already been shown that the use of auxiliary data such as off-equilibrium MD or rattled data can lead to state-of-the-art results on OC20[76].

To that end, we explore the impact that jointly training with OC22 and OC20 has on OC20 performance. Note OC22 is a significantly smaller and more limited dataset. OC20 contains ∼134M training data points and spans a large swath of materials. OC22 on the other hand is only ∼6% of the size of OC20, limited to only oxide materials, and places no constraints on atoms in the systems. Table 5.7 compares the performance of GemNet-OC trained on OC20 and OC20+OC22 as

evaluated on both OC20 and OC22 seperately. As expected, when trained on only OC20, OC22 metrics are poor - attributed to the the lack of oxides in OC20 and the difference in DFT theories. When trained on OC20+OC22, however, we see a significant improvement in energy MAE (∼20%). Force based metrics are either no different or slightly worse. Despite the joint dataset containing only a small fraction of OC22, it aided by a margin larger than any of the previous MD or rattled data efforts. Exploring in more detail as to how and why such improvements were observed could aid in systematically curating datasets to further improve OC20 performance.

5.8 Outlook and Future Directions

There are many challenges to building large datasets and fitting generalizable models in computational catalysis, some of which were recently summarized [136]. All of the challenges described also apply to the OC22 dataset - model performance varies across adsorbates and materials, direct force predictions tend to perform the best despite breaking energy conservation, developing helpful metrics for common tasks like local relaxations is difficult, and choosing the right calculations to improve the performance and generalizability of models is challenging. This work adds to these difficulties by highlighting additional challenges in capturing long-range interactions, developing models that go beyond adsorption energy, and fitting models with multiple datasets and levels of theory.

The performance of baseline models in this work is impressive given the difficulty of predicting the total system energy of complex oxide surfaces, but challenges still remain. The best results on the most general *S2EF-Total* task using a transfer learning approach from OC20 has an energy MAE of 0.26 eV for ID performance and 0.94 eV for OOD performance. Using that same model to predict relaxed total energies yields energy MAEs of 1.12 eV for ID and 1.85 eV for OOD predictions. These results are somewhat more impressive on a per-atom basis as is common for formation energy estimates of materials. However, for predicting experimentally-relevant properties like the overpotential for the OER, these results are far from sufficient. We note that

the initial baseline models for OC20 were similarly unhelpful for catalyst activity predictions, but rapid contributions from the broader community greatly improved their accuracy and predictive power. We hope that similar progress is seen for the tasks here. We also expect that the current models may already be helpful for certain more limited tasks, such as accelerating future oxide calculations with the use of online fine-tuning [172].

The OC22 dataset contains long-range interactions that are likely difficult to capture in exiting GNN models. Unlike metal surfaces which have a sea of electrons that can screen interactions, many of the oxides in OC22 are semiconductors with considerable partial charges (especially on the oxygen atoms). Electrostatics have very long range effects (energy decaying as $1/r$), and the partial charges can vary from system to system. The interaction of magnetic spins in systems with spin polarization is also long-ranged. This poses a challenge for the GNNs used in this work, which are often developed under the assumption that local interactions dominate. The use of several message passing steps or long-range local cutoffs may allow for these long-range interactions to be captured. There has been considerable effort in developing ML models that include long range interactions[134, 299, 28], and we expect those approaches to be very useful in improving predictions for OC22.

The tasks proposed in this work aim to push the community more in the direction of a general purpose potential, rather than separate models for each specific property. As an example, the tasks in OC20 were limited to the prediction of a specific property - the adsorption energy. This was a reasonable choice as the adsorption energy was the primary consideration for their application, and the adsorption energy itself was thought to be easier to fit than the DFT total energies. However, defining the tasks in this way meant that resulting models could only predict the adsorption energy and were unhelpful for predicting other surface properties like the surface energy. These limitations are highlighted in oxide catalysis where the stability of various surface terminations is needed. The total energy tasks in this work should encourage models that serve as general DFT surrogates - making predictions on a much wider range of properties.

The OC22 dataset also highlights the challenges of requiring multiple levels of theory for varying properties and materials. The OC20 dataset was constructed with the RPBE functional and neglected spin polarization, which represented a good trade-off between accuracy and computational cost for adsorption energies. However, most oxide surface calculations use or start with the PBE+U functional, and spin polarization tends to be more important. Combining datasets with multiple levels of theory, or upgrading datasets from less accurate to more accurate methods are popular questions in the small molecule community [86, 60], but applying these ideas to OC20/OC22 will require extending these approaches to large datasets and inorganic materials, and we hope the community rises to this challenge. An obvious future direction is to improve the data quality with far more expensive hybrid functionals on the relaxed structures here. Another future direction is the incorporation of magnetic configurations as variables in our models. Many oxides exhibit different magnetic configurations for the same structure. These magnetic polymorphs have significant consequences in electronic and thermodynamic properties. Including magnetic polymorphs in the future will allow for more general models.

Joint training on both the OC20 and OC22 datasets leads to several unexpected results. Surprisingly, naively fitting on both OC20 and OC22 (much smaller dataset) leads to large accuracy improvements for predicting OC20 energies, as shown in Table 5.7. In addition, models trained on either OC22 or OC20+OC22 both appear to follow the same log-log scaling for energy MAE (Figure 5-6). These observations open the door to using a wide array of existing large datasets (NOMAD[59], Materials Project[114], OQMD[223]) that although different, could aid in model development. These ideas can be rationalized if all of these datasets together can help learn more flexible and useful representations, regardless of their specific tasks or details.

Fine-tuning and transfer learning baselines were investigated as potential routes to improve accuracy across both OC20 and OC22 and reduce the computational intensity of training GNNs for these tasks. The most accurate models for both OC20 and OC22 were models trained on both datasets simultaneously, which indicates that a common representation can be learned and shared by both datasets. Surprisingly,

the limited fine-tuning experiments in this work did not improve substantially on the accuracy/cost Pareto front (Table 5-6). However, there are many possible fine-tuning strategies and a large number of variations (e.g. which sections of the GNN to freeze or fit, or leaving this decision to an attention block [135]), and we expect more progress from the community in this area. These approaches are necessary to encourage the re-use of large models, and to reduce the computational cost of obtaining state-of-the-art models for future small datasets.

Models trained on OC22 could predict the total energies for any clean or adsorbate+slab which ultimately allows us to determine any thermodynamic quantity including adsorption energy, surface energy, and reaction energy. Adsorption and reaction energies are useful for identifying viable catalysts. We can also predict the surface energy in order to construct elaborate phase diagrams which can be used to assess the thermodynamic stability of a surface at varying adsorbate coverages. Pourbaix diagrams (applied potential vs pH) are especially important for determining the thermodynamic viability of electrocatalysts. The surface energy can also be used to model the equilibrium crystal structure or Wulff shape. With a predictive model that circumvents DFT calculations, all these applications, which ordinarily require hundreds of DFT calculations, are possible with little to no computational cost.

This dataset will have broad impact in discovering oxide catalysts for a variety of reaction families and unraveling complex reaction mechanisms in these systems. Oxide materials are likely present in any reaction under strong oxidative conditions, such as the accelerated degradation of long-lived contaminants like PFOA[152] or systematically upgrading chemical building blocks [274]. Photocatalysis, which directly uses available sunlight to drive chemical reactions also relies heavily on oxides such as TiO_2 due to their desirable optical properties [50] and could benefit from this dataset. One example which is currently computationally expensive to study is the Mars-van Krevelen (MvK) mechanism, which is one of the most common catalytic mechanisms in ionic crystals[162, 106]. In the MvK, an adsorbate binds to a surface oxygen to form a new intermediate which desorbs to leave behind an oxygen vacancy, which can later be replenished by oxygen atoms from incoming adsorbates. By explicitly

including oxygen defects and vacancies in the dataset generation process, we hope the resulting models will be helpful for accelerating these studies. Similar reactions that could benefit from these approaches are CO_2 capture on carbides [85] or nitrate reduction on nitrides[1].

5.9 Supporting Information Available

The supporting information contains details on OC20 *S2EF-Total* and *IS2RE-Total* results, results using an alternative reference scheme, a discussion on adsorption energy for OC22, performance on OC22 adslabs and slabs, independently, training and hyperparameters for baseline models, full results on the validation splits, and the Hubbard U corrections used. The full dataset is provided at `http://opencatalystproject.org` and available in an ASE[143] trajectory or model-ready LMDB format. Baseline models, dataloaders, and trainers are provided in the open source repository `https://github.com/Open-Catalyst-Project/ocp`.

Chapter 6

Open Challenges in Developing Generalizable Large Scale Machine Learning Models for Catalyst Discovery

This work originally appeared as: Kolluru, A., Shuaibi, M.*, Palizhati, A., Shoghi, N., Das, A., Wood, B., Zitnick, C.L., Kitchin, J., Ulissi, Z., 2022. Open Challenges in Developing Generalizable Large Scale Machine Learning Models for Catalyst Discovery. ACS Catalysis, 12(14), pp.8572-8581.*These authors contributed equally.*
My contribution in this work included model training and evaluation, identifying and gathering community challenges, providing my perspective and outlook on the topic, and manuscript writing.

6.1 Abstract

The development of machine learned potentials for catalyst discovery has predominantly been focused on very specific chemistries and material compositions. While effective in interpolating between available materials, these approaches struggle to generalize across chemical space. The recent curation of large-scale catalyst datasets

has offered the opportunity to build a universal machine learning potential, spanning chemical and composition space. If accomplished, said potential could accelerate the catalyst discovery process across a variety of applications (CO_2 reduction, NH_3 production, etc.) without additional specialized training efforts that are currently required. The release of the OC20[40] has begun just that, pushing the heterogeneous catalysis and machine learning communities towards building more accurate and robust models. In this perspective, we discuss some of the challenges and findings of recent developments on OC20. We examine the performance of current models across different materials and adsorbates to identify notably underperforming subsets. We then discuss some of the modeling efforts surrounding energy-conservation, approaches to finding and evaluating the local minima, and augmentation of off-equilibrium data. To complement the community's ongoing developments, we end with an outlook to some of the important challenges that have yet to be thoroughly explored for large-scale catalyst discovery.

6.2 Introduction

Catalysts have played a key role in the synthesis of everyday chemicals and fuels necessary for a 21st century society. As renewable energy prices continue to decrease, traditional chemical synthesis processes are being revisited for more sustainable alternatives. At the center of this, catalyst discovery plays a key role in the advancement of renewable energy processes and sustainable chemical production, i.e. ammonia for fertilizer and hydrogen production. Unfortunately, the search space for catalyst materials is enormous for even high-throughput experiments [214]. This presents a need for computational tools to simulate systems through quantum mechanical (QM) models like DFT. QM approaches have made notable advancements in bridging computational results to experimental findings [120, 210, 288, 142, 215, 196]. While effective, QM tools scale very poorly, $O(N^3)$ or worse in the number of electrons. The computational cost associated with QM tools render them infeasible to the scale of the systems and search space desired for catalyst discovery. As a result, the catalysis com-

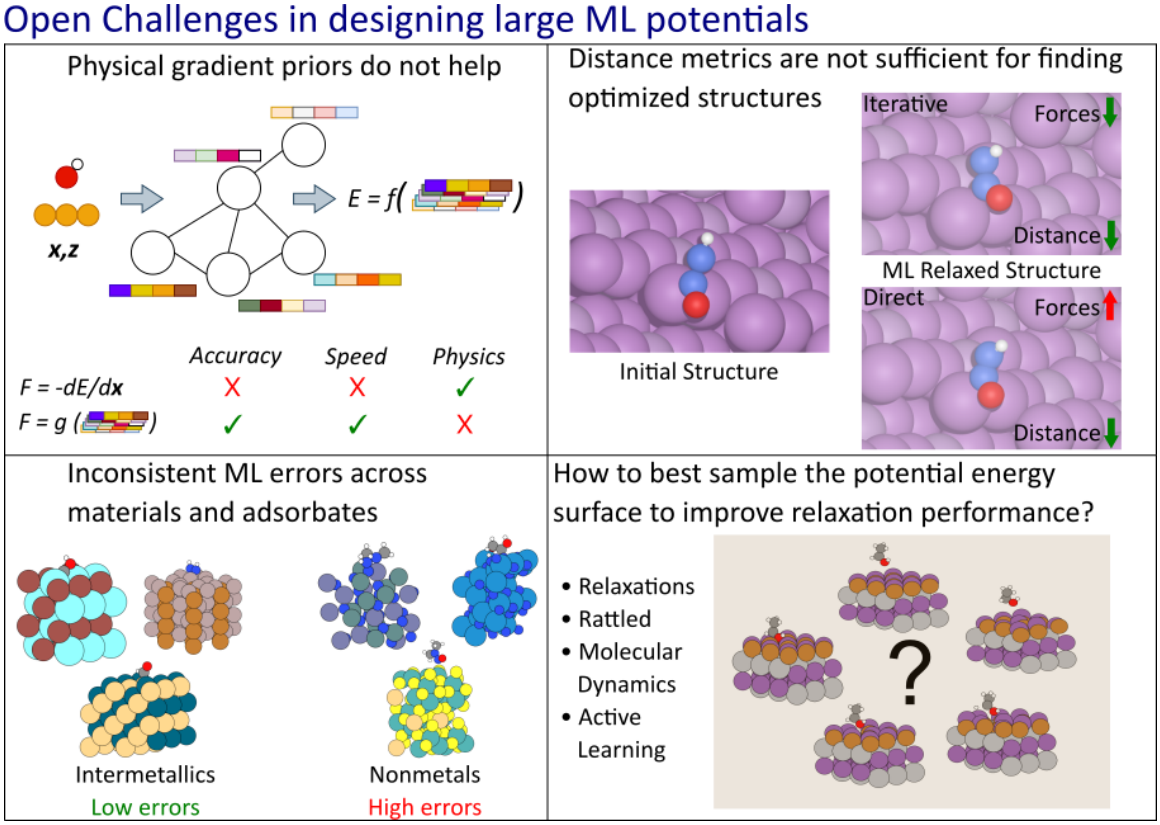


Figure 6-1: Summary of challenges associated with training on large dataset with large ML potentials discussed in the paper. *Top left* Trade offs in direct and gradient GNN force predictions. *Top right* An example system for a case where the distance metrics are relatively good for the direct approach but the force metrics are worse. *Bottom left* Demonstration of inconsistent error across a metallic surface and a non-metal through an example. *Bottom right* Augmenting existing relaxation datasets with off-equlibrium data can aid in relaxation performance.

munity has moved towards a more data driven approach [3, 92, 218, 105, 287]. With the QM data available, researchers are often interested in building machine learning surrogates for a particular chemical property [266, 13, 291, 77]. Such efforts, however, were limited to the finite data available, often for a very specific chemistry or system, limiting the generalizability ability of such models [92, 262]. Fortunately, as the community continues to curate larger, and more diverse datasets, machine learning models will continue to improve as they move towards larger, and more sophisticated architectures.

In the field of small molecules, a vast collection of datasets have been developed for varying use cases, including molecular dynamics simulations (MD17[48], ANI-1[245],

COLL[131]) and quantum mechanical properties (QM9[212] Alchemy[45]). These datasets are often limited to a few (5-10) unique elements, on average 10-20 atoms per system, and training set sizes in the range of 10k-1M samples. In the field of heterogeneous catalysis, datasets are often much more limited with training set sizes between 100 - 50k [5, 2, 160, 179]. These datasets were often created for very specific applications involving a handful of small adsorbates (i.e. hydrogen containing adsorbates on transition metal surfaces, CO_2 reduction catalysts, etc.). The release of OC20 marks a push towards a large, sparse collection of the material space. OC20 spans 55 unique elements, 82 adsorbates and includes a collection of unary, binary and ternary materials. A total of 1.28 million DFT relaxations were performed, comprising $\sim$260M single point evaluations of system energy and per-atom forces.

OC20 presented several practical tasks for the community to work towards. The most general of the tasks, *S2EF* evaluates a model's ability to serve as a surrogate to DFT - predicting a configuration's energy and per-atom forces. *IS2RE* asks to predict the relaxed state energy, given only the initial structure. *IS2RS* explores how well the relaxed structure can be predicted given only the initial configuration. In the scope of OC20, all energies were referenced to represent adsorption energy. For more details, we refer readers to the original manuscript [40].

In this perspective we shed light on the challenges of training Graph Neural Networks (GNNs) on large-scale datasets spanning material and composition space, illustrated in Figure 6-1. We begin with a quick overview on the current state of the community's progress and share some takeaways from what we have observed. We then discuss some telling trends on the performance of models across different adsorbates and material types. We discuss how different approaches and modeling decisions impact the prediction tasks and highlight the challenges associated with each. Further, we explain what the accuracies in various proposed metrics mean and some of the challenges in analyzing them. Finally, we share our outlook on the direction the community is headed and what still remains to achieve a large scale, generalizable potential for catalyst discovery.

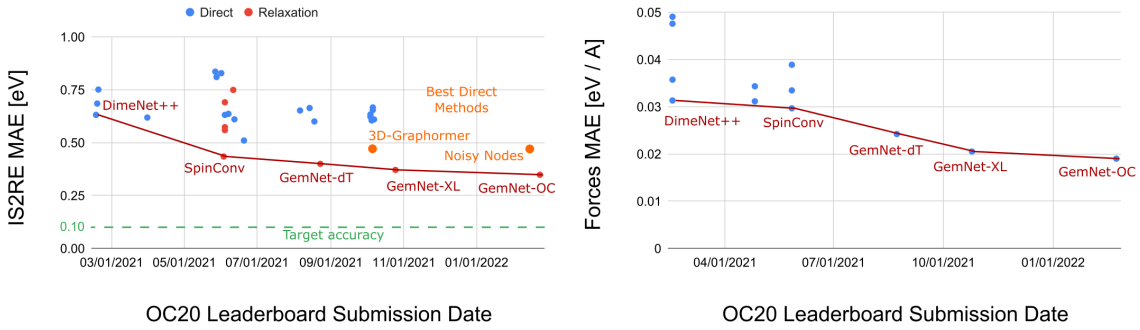


Figure 6-2: Community progress on the OC20 dataset since release. Left: *IS2RE* performance for both direct and relaxation based approaches. The current error target of 0.10eV would make these models more practically useful for researchers' applications. Right: *S2EF* performance as evaluated by mean absolute error of the forces. *IS2RE* and *S2EF* MAEs for their median baselines are 1.756 eV and 0.084 eV/Å, respectively.

6.3 Community progress in developing ML models for catalysis

Molecular modeling has progressed at an incredible rate over the past few decades. Simple linear models, neural networks, and kernel methods were originally developed relying on hand-crafted atomic representations, or descriptors [29, 158, 46, 19, 21] as inputs to the models. Descriptors capture invariant geometric information in the form of bonds and angles of an atoms local environment. While effective, the parameterization of such descriptors has been a challenging and non-trivial task. The past few years has seen a shift towards deep learning approaches. Rather than relying on hand crafted representations, models are being developed to learn similar or more expressive representations, specifically by exploiting the graphical nature of molecules using GNNs [232, 133, 130, 25, 233, 155]. Such models only take in 3D atomic coordinates and atomic numbers. A graph is then generated, where atoms are treated as nodes, and the distance between them as edges. Once a graph has been constructed, GNNs will undergo several rounds of message passing in which node representations are updated based off messages sent between neighboring nodes. While models may differ in their exact architecture, the update and message functions of-

ten include a series of multi-layer perceptrons and nonlinearities. Unlike traditional descriptor based models, GNNs end up learning node representations as part of the training process. Learned representations proceed through a final output block where a final prediction is made. In recent years, GNNs have come to surpass traditional descriptor based models [232, 133, 130, 25, 233, 155]. While typically data hungry, recent models like NequIP [25] are demonstrating great performance with as little as 100 samples. GNNs continue to gain traction as models continue to demonstrate state of the art performance on molecular datasets.

Since the release of OC20, the community has been rapidly developing new approaches to improve existing baselines. Models being developed range from traditional descriptor-style models [144] to complex and large GNN architectures [130, 241, 250, 290, 80]. Godwin, et al. present a simple, but effective GNN regularization technique to improve graph-level predictions, namely *IS2RE*. Liu, et al. use a similar technique in addition to a graph-based transformer to win 1^{st} place in the NeurIPS 2021 Open Catalyst Challenge [188] for direct *IS2RE* predictions. Klicpera, et al.[130, 76] and Shuaibi, et al.[241] explore various higher order representations (i.e., triplets and quadruplets) and leverage training on the entire OC20 to achieve impressive performance on the *S2EF* task, with GemNet-OC[76] holding the current state of the art across all tasks. Sriram, et al.[250] introduces Graph Parallelism, allowing them to scale GemNet to nearly a billion parameters across multiple GPUs. The scale and diversity of OC20 has additionally enabled transfer learning approaches to smaller datasets. Kolluru, et al. [135] propose a transfer learning technique to use OC20 pretrained models to improve performance on smaller, out-of-distribution datasets. Similar work has also been demonstrated for other big material datasets [43].

As the community continues to improve performance (Figure 6-2), it's important to understand some of the challenges, trends, and pitfalls in developing a generalizable potential.

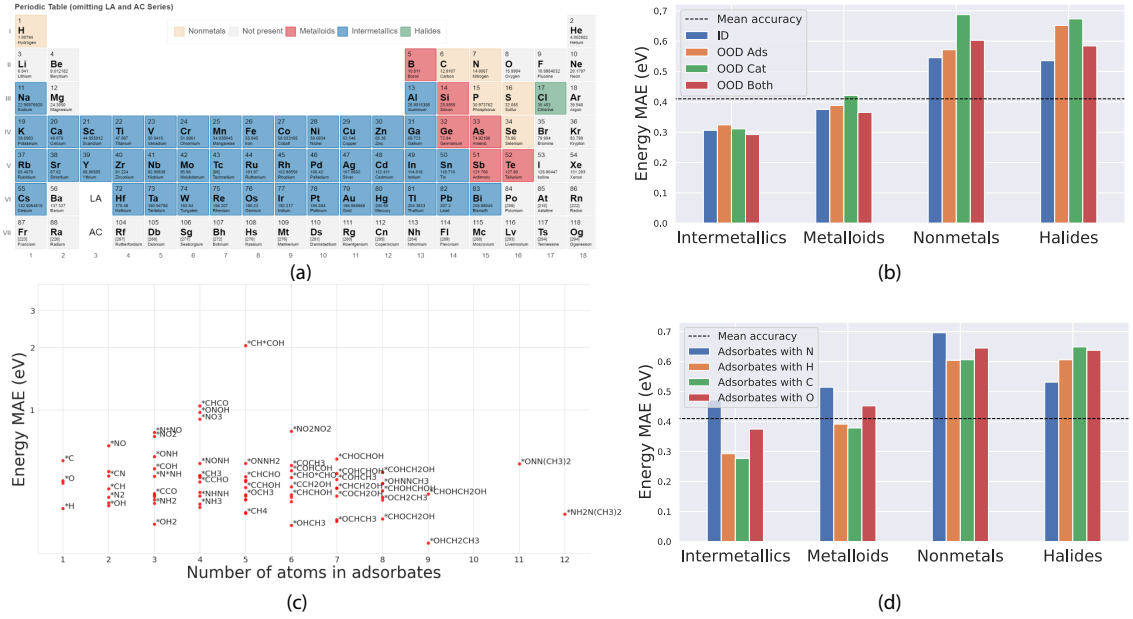


Figure 6-3: Analysis of GemNet-dT errors on the OC20 validation sets. (a) The categorization of OC20 elements into intermetallics, nonmetals, metalloids and halides for analysis. (b) Model performance across the different distributions and material types. (c) Errors averaged across all validation splits for specific adsorbate containing systems. (d) Errors averaged across all validation splits for adsorabtes containing certain elements.

6.4 Where are molecular GNNs still erroneous?

Most of the independent work done in developing ML potentials has been confined to datasets built for certain applications. For example, ML potentials for the applications of CO2RR are usually just trained with CO and H adsorbates [15, 179, 292, 41]. While this approach might interpolate well across materials, extrapolation to different adsorbates or more complicated materials will likely suffer in performance. A universal ML potential, if possible, would first require a large, diverse dataset that spans material and chemical space. OC20 dataset was created to build ML potentials that cover a large and diverse space of heterogeneous catalysts.

Errors across material types: With over 300k unique surfaces, OC20 spans a vast range of material compositions. When training large GNNs on the entire OC20 dataset, we observe that the accuracies are not uniform across element and adsorbate types. To analyze this, we divide the validation set into four different material types: intermetallics, metalloids, non-metals and halides, Figure 6-3(a). The distribution of data across these classes of materials is not the same, we have significantly more intermetallics and relatively fewer halides. We observe that the performance on non-metals is significantly worse, although both nonmetals and metalloids contribute to similar percentage of training data (Figure 6-3(b)). On the other hand, models tend to do much better across the board for intermetallics. Inaccuracies coming from non-metals disproportionately contribute to the overall errors, leading to worse performance for both force and energy predictions.

Errors across adsorbates: Large adsorbates are inherently more complicated as the degrees of freedom increases with the number of atoms. However, we observe no correlation with our model's performance and the size of the adsorbate. Model accuracies are poor for bidentate adsorbates like *CH*COH, *N*NO, *CH2*O, shown in Figure 6-3(c). Figure 6-3(d) also shows that adsorbates with N and O are generally more erroneous.

Model	Energy MAE (eV) ↓				Force MAE (eV/Å) ↓			
	ID	OOD Ads.	OOD Cat.	OOD Both	ID	OOD Ads.	OOD Cat.	OOD Both
Median	2.04	2.42	1.99	2.58	0.081	0.080	0.079	0.098
Gradient forces								
SpinConv[241]	-	-	-	-	0.031	0.035	0.032	0.042
GemNet-dT[130]	0.36	0.39	0.48	0.58	0.030	0.034	0.033	0.042
Direct forces								
SpinConv[241]	0.26	0.29	0.38	0.47	0.027	0.030	0.029	0.037
GemNet-dT[130]	0.23	0.25	0.35	0.41	0.021	0.024	0.025	0.032

Table 6.1: Results on the OC20 S2EF task via gradient-derived or direct force predictions. All models were trained on the OC20 S2EF All dataset. Results reported for the validation set. Energy metrics are unavailable for the gradient based SpinConv model due to being optimized only on forces.

6.5 Modeling trade-offs

6.5.1 Energy-conserving forces

Force predictions play an important role in the applications of ML models for catalyst discovery. While some tasks may only be interested in property predictions like adsorption or formation energy [15, 150, 161], forces are necessary to study dynamics such as structural relaxations, molecular dynamics, and transition state calculations [40, 25, 232, 198].

Physically, energy-conserving forces are derived as the gradient of energy with respect to atomic positions:

$$\boldsymbol{F_i} = -\frac{dE}{dx_i} \tag{6.1}$$

Energy-conservation is critical in studying molecular dynamics accurately. ML models estimating energy-conserving forces must ensure the architecture is continuous and differentiable, often satisfied by appropriate non-linear activation functions [232, 133, 130]. Geometrically, forces derived in an energy-conserving manner ensures forces are rotationally equivariant, a necessary physical relation of molecular systems [49]. Unfortunately, a gradient calculation increases model overhead in both memory usage and computational time by a factor of 2-4 [241, 108]. For datasets like MD17, calculating forces as a gradient is known to help in model accuracies as that is an important physical prior to the model [25, 130, 241]. Models trained on

			Energy MAE [eV] ↓				EwT ↑			
Model	Approach	Dataset Size	ID	OOD Ads	OOD Cat	OOD Both	ID	OOD Ads	OOD Cat	OOD Both
Median baseline	-	-	1.75	1.88	1.71	1.66	0.71%	0.72%	0.89%	0.74%
DimeNet++ [133]	Direct	460,328	0.56	0.73	0.58	0.66	4.25%	2.07%	4.10%	2.41%
SpinConv[241]	Direct	460,328	0.56	0.72	0.57	0.67	4.08%	2.26%	3.82%	2.33%
NoisyNodes[80]	Direct	460,328	0.42	**0.57**	0.44	**0.47**	**9.12%**	**3.49%**	8.01%	**4.64%**
Graphormer[290]	Direct	460,328	**0.40**	0.57	**0.42**	0.50	8.97%	3.45%	**8.18%**	3.79%
DimeNet++ – LF + LE[133, 40, 189]	Relaxation	2,000,000	0.53	0.57	0.56	0.52	6.79%	4.71%	6.49%	4.54%
SpinConv[241, 189]	Relaxation	2,000,000	0.46	0.51	0.47	0.44	7.38%	4.82%	7.05%	5.31%
GemNet-dT[130]	Relaxation	2,000,000	0.44	0.44	0.45	0.42	9.37%	6.59%	8.42%	6.40%
GemNet-OC[76]	Relaxation	2,000,000	**0.41**	**0.42**	**0.42**	**0.39**	**11.02%**	**8.68%**	**10.10%**	**7.82%**
DimeNet++ – LF + LE [133, 40]	Relaxation	133,934,018	0.50	0.54	0.58	0.61	6.57%	4.34%	5.09%	3.93%
SpinConv[241]	Relaxation	133,934,018	0.42	0.44	0.46	0.42	9.37%	7.47%	8.16%	6.56%
GemNet-dT[130]	Relaxation	133,934,018	0.39	0.39	0.43	0.38	12.37%	9.11%	10.09%	7.87%
GemNet-OC[76]	Relaxation	133,934,018	**0.35**	**0.35**	**0.38**	**0.34**	**16.06%**	**12.62%**	**13.17%**	**11.06%**

Table 6.2: Results on the OC20 *IS2RE* task using one of two approaches. **Direct** Directly predicting the relaxed state energy and **Relaxation** Training a model for energy and force predictions, followed by an iterative ML-based geometry optimization to arrive at a relaxed structure and energy. Relaxation results on the 2M subset suggest that competitive results are still possible with a limited compute budget. Results reported for the test set.

MD17 are often used to run molecular dynamics, further necessitating the need for energy-conservation [25]. However, for the OC20 dataset, particularly in the task of geometric optimization, we observe that the gradient approach for calculating forces to perform worse than direct prediction of forces for GemNet-dT [130] and Spinconv [241]. Dimenet ++ [133] and ForceNet [108] were built for gradient and direct approach respectively. The gradient approach could also make the training unstable in certain cases, which has been observed for ForceNet[108] and GemNet-OC[76]. Table 6.1 compares performance on the S2EF task for two recent top performing models, GemNet-dT [130] and SpinConv [241]. Not only are the force accuracies worse for the gradient approach, but the corresponding relaxed structure and relaxed energy metrics calculated via optimization are also significantly worse [241].

While energy-conservation plays a critical role in many molecular applications, we observe that direct force computations brings efficiency and performance advantages [108, 241]. Models trained for direct force predictions are limited to applications where strict enforcement of energy-conservation can reasonably be ignored, i.e. OC20's structural relaxations. Here, atomic positions are updated solely from force estimates [40, 54]. If necessary, DFT, or a subsequent ML model, can then be used to make

reliable energy predictions on the ML optimized structure. Similarly, transition states or saddle points can be derived in a similar manner with direct-force models. We want to emphasize that although unorthodox, direct-force models still prove to be useful in certain catalyst applications, i.e. OC20-like tasks.

6.5.2 Prediction of relaxed energy and structure

Adsorption energy is one of many properties that helps inform catalyst performance[278]. Computationally, this is computed via a series of QM structural relaxations. The relaxed energy is then referenced to represent the adsorption energy, see Chanussot et al.[40], García-Muelas et al.[71] for more details. From a data-driven approach, we can predict the relaxed energy or the relaxed structure of an atomic system usually via two methods. First, we can build a surrogate to DFT, approximating system energy and per-atom structures, and running ML optimizations to find the minimum energy, a common approach within the field. Alternatively, given a large enough dataset of relaxed structures and energies, we can try to predict these properties directly using a ML model instead of optimizing via an iterative loop. The advantage of the direct method over the relaxation approach is that it requires only a single call to the ML model, whereas the relaxation approach could require on average 200-300 calls for a single relaxation. Direct approaches are particularly advantageous when we talk about the computational cost of approaching large scale inference on the order of hundreds of millions to billions of systems.

The community has made tremendous progress in predicting adsorption energy as evaluated by the OC20 IS2RE task (Figure 6-2). Direct approaches, despite using 300x less data, are approaching the competitive relaxation based approaches of GemNet-XL and GemNet-OC. Inference time aside, models trained on the full 133M dataset for the relaxation based approaches are typically compute intensive, using between 128-512 GPUs [241, 108, 130, 40]. While this is certainly a small price to pay if the models developed accelerate the discovery process, it does make it difficult for the community to engage in and aid in development. This has been particularly observed in the NeurIPS 2021 Open Catalyst Challenge [188], where of the 30 submissions,

0 were made via the relaxation approach. Here, we show that models trained on a 2M subset of the full dataset are still able to provide competitive results and even, averaged across all splits, out perform direct approaches. Given the trends in the 2M dataset correlate well with the full 133M dataset [76], this should help incentivize the community to explore other approaches even with resource limitations. Although the relaxation approach is computationally expensive for both training and inference, we have observed that the models trained through this approach tend to generalize better on out-of-distribution (OOD) data, Table 6.2.

Direct relaxed energy predictions are an easier ML problem than direct structure predictions. For a system of size N, energy predictions require a single scalar output, while structure predictions require 3N components. We find that for relaxed energy prediction tasks, metrics are closer for direct and relaxation approach whereas for structure prediction task the metrics are worse. The OC20 paper provides a baseline for relaxed structure prediction only via the relaxation approach [40]. In Table 6.3 we provide baselines for direct relaxed structure prediction. A considerable gap exists between the direct and relaxation based approaches (especially in the DFT based metrics).

6.5.3 Metrics for finding local minima

Relaxed structure prediction is less straightforward than some of the other common energy and force prediction tasks. Given a dataset like OC20 where relaxed structures are not necessarily global minima, a model trained on such a dataset could either (1) predict and arrive at the same local minima, (2) arrive at a different, but still suitable minima, or (3) fail to arrive at any sort of minima.

To account for this, two main metrics have been presented in the OC20 paper. ADwT is a distance based metric and measures how close the predicted structure compares to the actual structure. This is similar to the Global Distance Test (GDT) metric in the protein folding task [122, 295]. ADwT takes an average across different thresholds varying from 0.1 to 0.5Åto ensure a signal is captured. For the OC20 dataset, we evaluate this metric for the input initial structures for an accuracy of

Table 6.3: Baseline metrics for IS2RS direct task in comparison with the relaxation approach. Metrics are reported on a 2k subset of the validation set, across all splits. DwT is evaluated at a threshold of 0.04 Å. For compute reasons, DFT-based metrics were evaluated on a 200 system subset of the 2k, 50 systems from each split.

	Model	DwT (at 0.04 Å) ↑	ADwT ↑	FbT* ↑	AFbT* ↑
Direct	ForceNet[108]	0.70	45.69%	0.00%	0.00%
	SpinConv[241]	1.05	47.76%	0.00%	0.00%
	GemNet-dT[130]	1.75	45.87%	0.00%	0.08%
Relaxation	ForceNet[108]	1.45	46.51%	0.00%	7.64%
	SpinConv[241]	8.20	55.81%	0.00%	12.55%
	GemNet-dT[130]	13.95	60.88%	0.00%	20.35%

21.18% on the in-domain validation set [40]. Models, at the bare minimum, should perform better than this baseline. To ensure invariance to arbitrary coordinate reference frames, we predict the difference between initial and final positions instead of the final position Cartesian coordinates. Predicting the delta difference helps simplify this task and results in improved ADwT accuracies.

A model that predicts a relaxed structure that is not identical to its DFT reference may still be considered successful for two reasons. (1) the model could have predicted a symmetrically identical site on the surface and (2) the model predicted a different, but still suitable local minima. The former is more a concern surrounding the distance-based metric, as ADwT, although accounts for periodic-boundary conditions, does not consider symmetrically identical sites. While it is rather unlikely an adsorbate initialized over a particular site will hop several sites over to a symmetrically identical site, it is worth raising awareness to the possibility. On the other hand, a model that arrives at a different relaxed structure entirely will fail according to ADwT. However, to verify whether the model has predicted a different suitable minima, we can evaluate the DFT forces corresponding to the ML predicted structures. This metric is called AFbT and it measures the percent of structures having their forces close to zero [40]. Since models are expected to predict relaxed structures, DFT forces should be close to zero. This is a stricter metric as compared to ADwT. However, this is far more expensive due to the additional DFT calculations. A more

practically useful metric would be number of DFT calculations required to find the relaxed structure starting from the ML relaxed structure. This would give us an idea of the percent of DFT calculations that the current ML models can reduce. Although useful, this is a significantly more expensive metric than AFbT calculations. While it is not something Open Catalyst Project's (OCP) tracks on their public leaderboard, we bring awareness to it as there could be instances where models do poorly on ADwT and AFbT but resulting structures are only a few DFT steps away from the relaxed structure.

In Table 6.3 we compare relaxed structure prediction via a direct and relaxation approach. We observe that direct methods, although having competitive ADwT metrics, have AFbT metrics that are significantly worse. This suggests that direct models do a reasonable job at getting close to the relaxed structure but are in high-force configurations, failing to capture repulsive physical interactions [80]. We speculate models struggle with this since small perturbations distances can have large consequences on forces, e.g. moving two atoms at an equilibrium bond length fractions of an angstrom towards each other. Relaxed structure prediction via the relaxation approach avoids this issue by using ML forces to drive a geometric optimizer.

We observe that distance metrics at tighter thresholds correlate better with force based metrics, however, going below 0.04 Å does not give sufficient signal and the accuracies for most systems fall to zero. Moreover, the Distance within Threshold (DwT) at 0.04 Å isn't a good enough signal that can replace AFbT. For example, DwT (at 0.04Å) for ForceNet relaxation approach and GemNet-dT direct approach are similar, however, the AFbT metrics still differ by 7.56% (as shown in Table 6.3). We believe that finding non DFT-based metrics that correlate well with DFT-based metrics is still an open and important question in the community which would make model evaluation computationally less expensive.

6.5.4 Additional data

The OC20 paper [40] released two additional data subsets generated with ab-initio molecular dynamics ('MD') and structural perturbations ('Rattled'). These provide

$38M$ and $17M$ additional *S2EF* training data points respectively.

Table 6.4 presents results for GemNet-OC[76] models trained on *S2EF*, Rattled, and MD data compared against similar analysis from the OC20 paper for DimeNet++ [133, 131]. First, on the force MAE metric, addition of MD data hurts DimeNet++ while it improves GemNet-OC. We speculate this to be another artifact of modeling forces as negative gradients of energy (as in DimeNet++) *vs.* direct prediction (as in GemNet-OC). Second, consistent with the OC20 paper, adding MD data to the training set provides a useful signal for *IS2RS* structure relaxations as per the AFbT metric. Finally, adding Rattled data helps with *IS2RS* metrics, but did not help or marginally hurt the *S2EF* force MAE. This could be due to a variety of reasons – random perturbations being too large / small to be useful, intermediate structures along a trajectory being less useful compared to closer to the local minimum (as in MD initial structures), etc. A promising direction here could be active learning approaches to optimally query additional training data points.

		S2EF Val ID	IS2RS Test	
	Training Data (# samples)	Force MAE ↓	ADwT ↑	AFbT ↑
DN++	$20M$ ($20M$)	0.0511	34.37%	2.67%
DN++	$20M$ + MD ($58M$)	0.0594	47.69%	17.09%
DN++	$20M$ + Rattled ($37M$)	0.0614	43.94%	12.51%
GN-OC	All ($133M$)	0.0179	60.33%	35.27%
GN-OC	All + MD ($172M$)	0.0173	60.77%	38.05%
GN-OC	All + MD + Rattled ($189M$)	0.0174	-	-

Table 6.4: Results with DimeNet++ (DN++) and GemNet-OC (GN-OC) trained on MD and Rattled. S2EF results reported for the validation in-distribution set. IS2RS results reported on the test set.

6.6 Summary and Outlook

The development of generalizable or universal ML models has only recently been seriously considered with the emergence of large-scale datasets like OC20 [40]. Since its release, the catalysis and ML communities have both made tremendous progres-

sive in developing models for catalyst applications. As the community continues to grow and as more datasets emerge that span material and composition space, the prospect of large-scale generalizable models is within reason. Progress thus far has demonstrated several challenges in accomplishing this feat: classes of materials and adsorbates with inconsistent errors, energy-conserving forces, relaxed vs direct approaches, DFT metrics, and data augmentation strategies. In this perspective, we discussed these challenges in detail and provided some insights as to how and why they are important. Although these challenges were discussed in the context of OC20, we anticipate similar challenges to future datasets of its kind.

Datasets like OC20 has offered new ways to how we think about building large, generalizable, and reliable models. While model performance has been the focal point of community progress thus far, we provide an outlook of other important challenges that we hope the community to engage in.

Training strategies. OC20 was released with predefined training, validation, and test sets. Its splits were curated in a manner to tackle the problem of building a single generalizable model for catalysis. However, it could be the case that multiple models for different subsets of the data, e.g. adsorbates, compositions, materials, do better. In the case of nonmetals, for instance, we have shown that this actually hurts performance - a possible consequence of the reduced dataset size.

Uncertainty and active learning. While model performance is a necessary step for the discovery process, it is not always sufficient. A practical ML-aided catalyst discovery pipeline will ultimately turn to experiments to validate whether the ML predicted “great" catalyst is at all effective. Having confidence in these predictions is particularly important to avoid wasted expensive experiments. Uncertainty quantification has been a particularly popular topic within the catalysis community, often focused on the small data regime and active learning [265, 36, 242, 272, 116, 263, 226]. The effectiveness of traditional uncertainty estimation techniques on large datasets like OC20 is a necessary and important step for the future of this work. Similarly, how to best leverage active learning for either dataset generation and/or augmentation [247] or online active learning [272, 242] at the scale of OC20 will be an exciting

future direction.

Model efficiency. In addition to model performance and reliability, model efficiency will continue to be critical for all applications. For training, faster, more data efficient models can help attract the community to tackle some of the bigger challenges like a surrogate to DFT, i.e. OC20's S2EF task. Progress so far has shown that the best models are also the largest models. From an inference perspective, this poses obvious challenges of slower speeds and ultimately reduced screening throughput. While models still remain orders of magnitude faster than DFT, when considering the possibility of screening billions of systems, computational costs add up. Recent models encoding equivariant representations [25, 171] have shown incredible scaling and efficiency gains that could be promising to explore. Moving forward, efficient architectures and model distillation[69] will be an important contribution to reduce the computational cost of large-scale inference, even if it means sacrificing some accuracy.

Data augmentation. The scale of OC20 makes data augmentation a non-trivial challenge. With 130M+ training data points, randomly adding 10-100k data points will likely have negligible impact on the models. We observed that models using the additional MD data are able to perform the best, while the rattled data has little impact. Identifying strategies to combine and train large molecular and material datasets like ANI-1[245] and OQMD[128] with OC20 could help improve models even further. The biggest challenge surrounding this comes from combining datasets of varying levels of DFT theory.

Energy-conserving forces. In the context of OC20, we have observed that the best performing models make a direct force-prediction. While this may be suitable for some applications, the more physically motivated gradient approach to force prediction is desired for other applications like MD. The same direct models applied to MD17 observe the opposite effect, better performance via the gradient method [76]. It remains an open question why this is the case, and we encourage others to investigate this observation.

Physics-based modeling. The majority of models submitted to OC20 have followed a purely data-driven approach, only taking in atomic numbers and positions

as inputs. Exploring ways to leverage OC20 charge density or Bader charge data [1] could prove useful, particularly in the low data regime. Additionally, models like UNiTE [205] or OrbNet[206] that leverage tight binding DFT[18] for featurization could be interesting to explore for catalyst applications.

[1] To be made publically available at https://github.com/Open-Catalyst-Project/ocp/blob/main/DATASET.md

Chapter 7

Conclusions and Outlook

The ability to screen billions of catalyst materials accurately would accelerate the discovery of efficient, low-cost catalysts that are necessary for renewable energy technologies. Through this thesis, we contribute to the development of generalizable machine learning models and methods to accelerate the catalyst discovery process across chemical and material space.

7.1 Contributions

In Chapter 2, we focus our attention on the low data regime and how we can build more accurate methods and frameworks for accelerating atomic simulations. Specifically, we show how incorporating physical priors can aid neural network based architectures in online and offline active learning frameworks. We developed `amptorch` to efficiently train, optimize, and take advantage of modern machine learning methods for descriptor based models. We also developed a modular active learning package, *almlp* (now, *FINETUNA*), to conveniently explore alternative models and querying strategies. Through this work we were able to accelerate simulations of small molecules on catalyst materials anywhere from 4-10x.

We then tackled the task of developing a general-purpose model that could screen across chemical space without the reliance on any DFT. In Chapter 3, we constructed the world's largest catalyst dataset - OC20 to enable the desired model development.

OC20 consists of over 1M DFT simulations (200M+ data points) spanning a swath of different materials, surfaces, and adsorbates relevant for renewable energy technologies. To garner attention from the ML and catalysis communities, we formulated well-defined tasks and challenges relevant to every-day catalyst applications. Baseline GNNs were released and a public leaderboard was established to encourage developments and track progress. Accompanying this work, we developed the `ocp` codebase, which includes all baseline models, data loaders, evaluators, and tools necessary to bootstrap research. While baseline models were far from practical applications, the community has made significant developments with new models currently being applied to catalyst discovery pipelines.

The size and uniqueness of OC20 brought forth a wave of model development for catalysis from ourselves and the community at large. In Chapter 4, we present one of these efforts to more accurately model catalysts - SpinConv. SpinConv is a GNN that explicitly tries to capture the 3D environment of an atom through projections of the local environment onto a grid. At the time of its release, SpinConv achieved state-of-the-art performance over existing baselines by $\sim$15% across key metrics. Since then, model development has progressed tremendously and, although no longer state-of-the-art, some of its design decisions have aided or inspired newer developments.

While OC20's size and diversity spans a large material space, it is still a sparse sampling of the billions of possible combinations and naturally has its limitations. In Chapter 5, we constructed the largest oxide catalyst dataset - OC22. OC22 consists of over 60,000 DFT relaxations across a range of oxide surface combinations and adsorbates important for green hydrogen storage, and other oxide applications. Here we expand on the OC20 tasks to focus on total energy predictions, a more general property for catalyst applications. We show how models trained on total energies enables the exploration of clean surface configurations, an important step in downstream discovery workflows. We also demonstrate how datasets like OC22 can complement existing datasets like OC20, and vice-versa, through alternative training strategies like joint training and transfer learning. This work offers a precedent to future catalyst datasets that seek to improve model performance by leveraging larger

datasets like OC20 and OC22.

In Chapter 6, we share our perspective on some of the findings and challenges in building a generalizable machine learning model for catalyst discovery. We show that (1) models developed with direct-force predictions are more accurate and faster than physically motivated gradient-based predictions; (2) when tasked with predicting relaxed energies, approaches that run ML-driven structure optimizations are more accurate and generalize better than models that directly predict relaxed energies; and (3) augmenting off-equilibrium data, while may not improve training metrics, helps improve downstream DFT evaluations. Some of the open challenges that still remain include (1) poor model performance on material subsets - specifically nonmetals and halides; and (2) identifying metrics that can efficiently and accurately evaluate model relaxations without relying on costly DFT. While the points raised in this work preceded OC22, the same trends and challenges are observed in OC22.

7.2 Outlook

The development of generalizable machine learning models for catalysis has only recently been more seriously discussed in light of large dataset efforts like OC20 and OC22, among others. We recognize that as a result of its infancy, a multitude of future directions may stem from this research. Future modeling efforts may consider how to more efficiently capture and represent the atomic environment. Additionally, models here were purely data-driven. Exploring the extent physical biases or features can be incorporated has yet to be properly explored in depth. As models and datasets continue to grow, more efficient architectures and model distillation strategies will be increasingly important contributions to reduce computational costs.

While improving model performance will continue to be a goal, uncertainty estimation will be a critical component in practical discovery applications. GNNs discussed in this work provide no uncertainty estimates associated with their predictions. A set of baselines for GNN uncertainty prediction followed by more accurate strategies, if necessary, will be an important next step for downstream catalyst applications like

active learning or experimental collaborations. Future work may also consider developing better training and augmentation strategies. Transfer learning, in particular, has been shown in this work to be a delicate and challenging task, but can provide meaningful improvements. Exploring how to more carefully and systematically transfer knowledge between datasets for any new, arbitrary dataset will be important for community adoption. This work, unsurprisingly, has shown that more data continues to help model performance. Future work may explore how to generate more meaningful training data, whether it be through off-equilibrium simulations, active learning, or new datasets entirely.

The tasks, challenges, and datasets presented in this work serve to accelerate the development of models and methods for catalyst discovery. However, when it comes time for models to suggest candidate materials, we will need to evaluate their performance based off experimental feedback. Future work will need to evaluate model performance beyond just the tasks and metrics proposed here. Getting experimental feedback early will play a critical role in the direction model development should take - are pushing these task metrics sufficient or are more complex tasks necessary? Exploring the full pipeline of ML predictions to experimental results will ultimately be vital to the success of large scale catalyst screening.

Bibliography

[1] Younes Abghoui and Egill Skúlason. Electrochemical synthesis of ammonia via mars-van krevelen mechanism on the (111) facets of group iii–vii transition metal mononitrides. *Catalysis Today*, 286:78–84, 2017. 148

[2] Frank Abild-Pedersen, Jeff Greeley, Felix Studt, Jan Rossmeisl, TR Munter, Poul Georg Moses, Egill Skulason, Thomas Bligaard, and Jens Kehlet Nørskov. Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces. *Phys. Rev. Lett.*, 99(1):016105, 2007. 44, 68, 152

[3] Ankit Agrawal and Alok Choudhary. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *Apl Materials*, 4(5):053208, 2016. 151

[4] Zeynep Aksoz and Clemens Preisinger. An Interactive Structural Optimization of Space Frame Structures Using Machine Learning. In *Impact: Design With All Senses*, pages 18–31. Springer International Publishing, sep 2020. 67

[5] Mie Andersen, Sergey V. Levchenko, Matthias Scheffler, and Karsten Reuter. Beyond scaling relations for the description of catalytic materials. *ACS Catalysis*, 9(4):2752–2759, 2019. 44, 68, 152

[6] Mie Andersen and Karsten Reuter. Adsorption enthalpies for catalysis modeling through machine-learned descriptors. *Accounts of Chemical Research*, 54(12):2741–2749, 2021. 118

[7] Brandon Anderson, Truong Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. In *Advances in Neural Information Processing Systems*, pages 14537–14546, 2019. 92, 96, 110, 112, 113

[8] Shi Jun Ang, Wujie Wang, Daniel Schwalbe-Koda, Simon Axelrod, and Rafael Gomez-Bombarelli. Active Learning Accelerates Ab Initio Molecular Dynamics on Pericyclic Reactive Energy Surfaces. *ChemRxiv*, 4 2020. 57

[9] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 68

[10] Nongnuch Artrith. Machine learning for the modeling of interfaces in energy storage and conversion materials. *Journal of Physics: Energy*, 1(3):032002, 2019. 67

[11] Nongnuch Artrith and Alexie M. Kolpak. Understanding the composition and activity of electrocatalytic nanoalloys in aqueous solvents: A combination of dft and accurate neural network potentials. *Nano Letters*, 14(5):2670–2676, 2014. PMID: 24742028. 39, 48

[12] Majid Asnavandi, Yichun Yin, Yibing Li, Chenghua Sun, and Chuan Zhao. Promoting Oxygen Evolution Reactions through Introduction of Oxygen Vacancies to Benchmark NiFe-OOH Catalysts. *ACS Energy Letters*, 3(7):1515–1520, 2018. 122

[13] Seoin Back, Kevin Tran, and Zachary W Ulissi. Toward a design of active oxygen evolution catalysts: insights from automated density functional theory calculations and machine learning. *Acs Catalysis*, 9(9):7651–7659, 2019. 151

[14] Seoin Back, Kevin Tran, and Zachary W Ulissi. Discovery of acid-stable oxygen evolution catalysts: high-throughput computational screening of equimolar bimetallic oxides. *ACS Applied Materials & Interfaces*, 12(34):38256–38265, 2020. 127

[15] Seoin Back, Junwoong Yoon, Nianhan Tian, Wen Zhong, Kevin Tran, and Zachary W Ulissi. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *The Journal of Physical Chemistry Letters*, 10(15):4401–4408, 2019. 34, 67, 156, 157, 213

[16] R.F.W. Bader and R.F. Bader. *Atoms in Molecules: A Quantum Theory*. International series of monographs on chemistry. Clarendon Press, 1994. 69, 77

[17] Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. Extended tight-binding quantum chemistry methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2020. 81

[18] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15(3):1652–1671, 2019. 166

[19] Albert P Bartók and Gábor Csányi. G aussian approximation potentials: A brief tutorial introduction. *International Journal of Quantum Chemistry*, 115(16):1051–1057, 2015. 153

[20] Albert P. Bartók, James Kermode, Noam Bernstein, and Gábor Csányi. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X*, 8:041048, Dec 2018. 131

[21] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104:136403, Apr 2010. 39, 44, 46, 48, 80, 153

[22] Albert P. Bartók and Gábor Csányi. Gaussian approximation potentials: A brief tutorial introduction, 2015. 46, 51

[23] Thomas AA Batchelor, Jack K Pedersen, Simon H Winther, Ivano E Castelli, Karsten W Jacobsen, and Jan Rossmeisl. High-entropy alloys as a discovery platform for electrocatalysis. *Joule*, 3(3):834–845, 2019. 68

[24] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):1–11, 2022. 132, 229

[25] Simon Batzner, Tess E Smidt, Lixin Sun, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, and Boris Kozinsky. Se (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *arXiv preprint arXiv:2101.03164*, 2021. 41, 43, 96, 112, 113, 153, 154, 157, 158, 165

[26] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *Journal of Chemical Physics*, 2011. 15, 40

[27] Jörg Behler. Perspective: Machine learning potentials for atomistic simulations. *The Journal of Chemical Physics*, 145(17):170901, 2016. 39, 48, 50, 80, 131

[28] Jörg Behler and Gábor Csányi. Machine learning potentials for extended systems: a perspective. *The European Physical Journal B*, 94(7):1–11, 2021. 145

[29] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, Apr 2007. 15, 40, 50, 131, 153

[30] Noam Bernstein, Gábor Csányi, and Volker L. Deringer. De novo exploration and self-guided learning of potential-energy surfaces. *npj Computational Materials*, 2019. 45

[31] T Bligaard and Jens Kehlet Nørskov. Ligand effects in heterogeneous catalysis and electrochemistry. *Electrochimica Acta*, 52(18):5512–5516, 2007. 127

[32] Peter E Blöchl. Projector augmented-wave method. *Physical Review B*, 50(24):17953, 1994. 205

[33] Jacob R Boes and John R Kitchin. Neural network predictions of oxygen interactions on a dynamic pd surface. *Molecular Simulation*, 43(5-6):346–354, 2017. 67

[34] Jacob R Boes, Osman Mamun, Kirsten Winther, and Thomas Bligaard. Graph theory approach to high-throughput surface adsorption structure generation. *The Journal of Physical Chemistry A*, 123(11):2281–2285, 2019. 74

[35] Lars A Bratholm, Will Gerrard, Brandon Anderson, Shaojie Bai, Sunghwan Choi, Lam Dang, Pavel Hanchar, Addison Howard, Guillaume Huard, Sanghoon Kim, et al. A community-powered search of machine learning strategy space to find nmr property prediction models. *arXiv preprint arXiv:2008.05994*, 2020. 92

[36] Jonas Busk, Peter Bjørn Jørgensen, Arghya Bhowmik, Mikkel N Schmidt, Ole Winther, and Tejs Vegge. Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks. *Machine Learning: Science and Technology*, 3(1):015012, 2021. 164

[37] CAISO. Current and Forecasted Demand, 2020. 15, 38

[38] Federico Calle-Vallejo, José I Martínez, Juan M García-Lastra, Philippe Sautet, and David Loffreda. Fast prediction of adsorption properties for platinum nanocatalysts with generalized coordination numbers. *Angew. Chem. Int. Ed.*, 53(32):8316–8319, 2014. 44, 68

[39] Anand Chandrasekaran, Deepak Kamal, Rohit Batra, Chiho Kim, Lihua Chen, and Rampi Ramprasad. Solving the electronic structure problem with machine learning. *npj Computational Materials*, 5(1):1–7, 2019. 77

[40] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021. 31, 39, 62, 96, 97, 104, 105, 108, 109, 118, 128, 131, 132, 141, 150, 152, 157, 158, 159, 160, 161, 162, 163

[41] An Chen, Xu Zhang, Letian Chen, Sai Yao, and Zhen Zhou. A machine learning model on simple features for CO_2 reduction electrocatalysts. *The Journal of Physical Chemistry C*, 124(41):22471–22478, 2020. 156

[42] Benjamin W. J. Chen, Lang Xu, and Manos Mavrikakis. Computational methods in heterogeneous catalysis. *Chemical Reviews*, 121(2):1007–1048, 2021. PMID: 33350813. 117

[43] Chi Chen and Shyue Ping Ong. Atomsets as a hierarchical transfer learning framework for small and large materials datasets. *npj Computational Materials*, 7(1):1–9, 2021. 154

[44] Chi Chen, Yunxing Zuo, Weike Ye, Xiangguo Li, Zhi Deng, and Shyue Ping Ong. A critical review of machine learning of energy materials. *Advanced Energy Materials*, 10(8):1903242, 2020. 48, 67

[45] Guangyong Chen, Pengfei Chen, Chang-Yu Hsieh, Chee-Kong Lee, Benben Liao, Renjie Liao, Weiwen Liu, Jiezhong Qiu, Qiming Sun, Jie Tang, et al. Alchemy: A quantum chemistry dataset for benchmarking ai models. *arXiv preprint arXiv:1906.09427*, 2019. 152

[46] Jun Chen, Xin Xu, Xin Xu, and Dong H Zhang. A global potential energy surface for the $H_2 + OH \longrightarrow H_2O + H$ reaction using neural networks. *The Journal of Chemical Physics*, 138(15):154301, 2013. 153

[47] Stefan Chmiela, Huziel E Sauceda, Klaus-Robert Müller, and Alexandre Tkatchenko. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature communications*, 9(1):1–10, 2018. 98, 104, 110

[48] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017. 43, 44, 97, 98, 99, 104, 110, 111, 112, 113, 131, 151

[49] Anders S Christensen and O Anatole von Lilienfeld. On the role of gradients for machine learning of molecular energies and forces. *Machine Learning: Science and Technology*, 1(4):045018, 2020. 111, 157

[50] Benjamin M Comer and Andrew J Medford. Analysis of photocatalytic nitrogen fixation on rutile tio2 (110). *ACS Sustainable Chemistry & Engineering*, 6(4):4648–4660, 2018. 147

[51] Stefano Curtarolo, Wahyu Setyawan, Shidong Wang, Junkai Xue, Kesong Yang, Richard H. Taylor, Lance J. Nelson, Gus L W Hart, Stefano Sanvito, Marco Buongiorno-Nardelli, Natalio Mingo, and Ohad Levy. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science*, 58(N/A):227–235, 2012. 67, 121

[52] Wenyuan Dai, Ou Jin, Gui-Rong Xue, Qiang Yang, and Yong Yu. Eigentransfer: a unified framework for transfer learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 193–200, 2009. 120

[53] Holger Dau, Christian Limberg, Tobias Reier, Marcel Risch, Stefan Roggan, and Peter Strasser. The Mechanism of Water Oxidation: From Electrolysis via Homogeneous to Biological Catalysis. *ChemCatChem*, 2(7):724–761, 2010. 125

[54] Estefanía Garijo del Río, Jens Jørgen Mortensen, and Karsten Wedel Jacobsen. Local bayesian optimizer for atomic structures. *Physical Review B*, 100(10):104103, 2019. 158

[55] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 68

[56] Volker L Deringer, Andrei L Tchougréeff, and Richard Dronskowski. Crystal orbital hamilton population (cohp) analysis as projected from plane-wave basis sets. *The Journal of Physical Chemistry A*, 115(21):5461–5466, 2011. 77

[57] Subhashish Dey and Ganesh Chandra Dhal. Cerium catalysts applications in carbon monoxide oxidations. *Materials Science for Energy Technologies*, 3:6–24, 2020. 121

[58] Colin F Dickens, Joseph H Montoya, Ambarish R Kulkarni, Michal Bajdich, and Jens K Nørskov. An electronic structure descriptor for oxygen reactivity at metal and metal-oxide surfaces. *Surf. Sci.*, 681(N/A):122–129, 2019. 68

[59] Claudia Draxl and Matthias Scheffler. Nomad: The fair concept for big-data-driven materials science. *arXiv preprint arXiv:1805.05039*, page p. N/A, 2018. 146

[60] Chenru Duan, Daniel BK Chu, Aditya Nandy, and Heather J Kulik. Two wrongs can make a right: A transfer learning approach for chemical discovery with chemical accuracy. *arXiv preprint arXiv:2201.04243*, page p. N/A, 2022. 146

[61] Annual Energy Outlook 2020, 2020. 66

[62] U.S. Energy Information Administration (EIA), 2020. 37

[63] Jacques A. Esterhuizen, Bryan R. Goldsmith, and Suljo Linic. Theory-guided machine learning finds geometric structure-property relationships for chemisorption on subsurface alloys. *Chem*, 6(11):3100–3117, 2020. 67

[64] Jiameng Fan and Wenchao Li. Adversarial training and provable robustness: A tale of two objectives, 2020. 63

[65] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019. 50, 79, 130

[66] Richard Phillips Feynman. Forces in molecules. *Physical Review*, 56(4):340, 1939. 206

[67] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pages 3165–3176. PMLR, 2020. 110

[68] Raul A Flores, Christopher Paolucci, Kirsten T Winther, Ankit Jain, Jose Antonio Garrido Torres, Muratahan Aykol, Joseph Montoya, Jens K Nørskov, Michal Bajdich, and Thomas Bligaard. Active learning accelerated discovery of stable iridium oxide polymorphs for the oxygen evolution reaction. *Chemistry of Materials*, 32(13):5854–5863, 2020. 117, 127

[69] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018. 165

[70] Fabian B Fuchs, Daniel E Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *arXiv preprint arXiv:2006.10503*, 2020. 110

[71] Rodrigo García-Muelas and Núria López. Statistical learning goes beyond the d-band model providing the thermochemistry of adsorbates on transition metals. *Nature communications*, 10(1):1–7, 2019. 159

[72] Estefanía Garijo del Río, Jens Jørgen Mortensen, and Karsten Wedel Jacobsen. Local bayesian optimizer for atomic structures. *Phys. Rev. B*, 100:104103, Sep 2019. 48

[73] José A. Garrido Torres, Paul C. Jennings, Martin H. Hansen, Jacob R. Boes, and Thomas Bligaard. Low-scaling algorithm for nudged elastic band calculations using a surrogate machine learning model. *Phys. Rev. Lett.*, 122:156001, Apr 2019. 48, 49

[74] Kevin F. Garrity, Joseph W. Bennett, Karin M. Rabe, and David Vanderbilt. Pseudopotentials for high-throughput dft calculations. *Computational Materials Science*, 81:446–452, 2014. 63

[75] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34:6790–6802, 2021. 119, 127, 130, 131, 135, 136, 224, 232

[76] Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ulissi, C Lawrence Zitnick, and Abhishek Das. How do graph networks generalize to large and diverse molecular systems? *arXiv preprint arXiv:2204.02782*, 2022. 24, 29, 32, 44, 119, 127, 130, 131, 135, 136, 138, 143, 154, 158, 160, 163, 165, 224, 232

[77] Lei Ge, Hao Yuan, Yuxiang Min, Li Li, Shiqian Chen, Lai Xu, and William A Goddard III. Predicted optimal bifunctional electrocatalysts for the hydrogen evolution reaction and the oxygen evolution reaction using chalcogenide heterostructures based on machine learning analysis of in silico quantum mechanics based high throughput screening. *The Journal of Physical Chemistry Letters*, 11(3):869–876, 2020. 151

[78] Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, Andrea Dal Corso, Stefano de Gironcoli, Stefano Fabris, Guido Fratesi, Ralph Gebauer, Uwe Gerstmann, Christos Gougoussis, Anton Kokalj, Michele Lazzeri, Layla Martin-Samos, Nicola Marzari, Francesco Mauri, Riccardo Mazzarello, Stefano Paolini, Alfredo Pasquarello, Lorenzo

Paulatto, Carlo Sbraccia, Sandro Scandolo, Gabriele Sclauzero, Ari P Seitsonen, Alexander Smogunov, Paolo Umari, and Renata M Wentzcovitch. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter*, 21(39):395502, sep 2009. 54, 63

[79] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning (ICML)*, pages 1273–1272, 2017. 41, 96, 110, 111, 112, 131

[80] Jonathan Godwin, Michael Schaarschmidt, Alexander L Gaunt, Alvaro Sanchez-Gonzalez, Yulia Rubanova, Petar Veličković, James Kirkpatrick, and Peter Battaglia. Simple gnn regularisation for 3d molecular property prediction and beyond. In *International Conference on Learning Representations*, 2021. 119, 130, 154, 158, 162

[81] Bryan R. Goldsmith, Jacques Esterhuizen, Jin-Xun Liu, Christopher J. Bartel, and Christopher Sutton. Machine learning for heterogeneous catalyst design and discovery. *AIChE Journal*, 64(7):2311–2323, 2018. 67

[82] Sheng Gong, Tian Xie, Taishan Zhu, Shuo Wang, Eric R Fadel, Yawei Li, and Jeffrey C Grossman. Predicting charge density distribution of materials using a local-environment-based graph convolutional network. *Physical Review B*, 100(18):184103, 2019. 77

[83] Danilo González, Javier Heras-Domingo, Mariona Sodupe, Luis Rodríguez-Santiago, and Xavier Solans-Monfort. Importance of the oxyl character on the IrO2 surface dependent catalytic activity for the oxygen evolution reaction. *Journal of Catalysis*, 396:192–201, 2021. 118, 125, 234

[84] Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734. IEEE, 2005. 96, 98

[85] Jose M Gracia, Frans F Prinsloo, and JW Niemantsverdriet. Mars-van krevelen-like mechanism of co hydrogenation on an iron carbide surface. *Catalysis Letters*, 133(3):257–261, 2009. 148

[86] Colin A Grambow, Yi-Pei Li, and William H Green. Accurate thermochemistry with small data sets: A bond additivity correction and transfer learning approach. *The Journal of Physical Chemistry A*, 123(27):5826–5835, 2019. 146

[87] Geun Ho Gu, Changhyeok Choi, Yeunhee Lee, Andres B. Situmorang, Juhwan Noh, Yong-Hyun Kim, and Yousung Jung. Progress in computational and machine-learning methods for heterogeneous small-molecule activation. *Advanced Materials*, 32(35):1907865, 2020. 67

[88] Geun Ho Gu, Juhwan Noh, Inkyung Kim, and Yousung Jung. Machine learning for renewable energy materials. *Journal of Materials Chemistry A*, 7(29):17096–17117, 2019. 67

[89] Geun Ho Gu, Juhwan Noh, Sungwon Kim, Seoin Back, Zachary Ulissi, and Yousung Jung. Practical deep-learning representation for fast heterogeneous catalyst screening. *The Journal of Physical Chemistry Letters*, 11(9):3185–3191, 2020. PMID: 32191473. 67, 213

[90] Geun Ho Gu, Petr Plechac, and Dionisios G Vlachos. Thermochemistry of gas-phase and surface species via lasso-assisted subgraph selection. *Reaction Chemistry & Engineering*, 3(4):454–466, 2018. 67

[91] Shuang Gu, Bingjun Xu, and Yushan Yan. Electrochemical energy engineering: a new frontier of chemical engineering innovation. *Annual Review of Chemical and Biomolecular Engineering*, 5(1):429–454, 2014. 37

[92] Yani Guan, Donovan Chaffart, Guihua Liu, Zhaoyang Tan, Dongsheng Zhang, Yanji Wang, Jingde Li, and Luis Ricardez-Sandoval. Machine learning in solid heterogeneous catalysis: Recent developments, challenges and perspectives. *Chemical Engineering Science*, 248:117224, 2022. 151

[93] GT Kasun Kalhara Gunasooriya and Jens K Nørskov. Analysis of acid-stable and active oxides for the oxygen evolution reaction. *ACS Energy Letters*, 5(12):3778–3787, 2020. 127

[94] Thomas Hager. *The alchemy of air: a Jewish genius, a doomed tycoon, and the scientific discovery that fed the world but fueled the rise of Hitler*. Broadway Books, 2009. 114

[95] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 2017. 80

[96] Bjørk Hammer, Lars Bruno Hansen, and Jens Kehlet Nørskov. Improved adsorption energetics within density-functional theory using revised perdew-burke-ernzerhof functionals. *Physical Review B*, 59(11):7413, 1999. 205

[97] Bjørk Hammer and Jens Kehlet Nørskov. Theoretical surface science and catalysis—calculations and concepts. In *Advances in Catalysis*, volume 45, pages 71–129. Elsevier, 2000. 127

[98] Kathleen Hancock and Juliann Allison. *Renewable Power Generation Costs in 2018*. International Renewable Energy Agency, Abu Dhabi, 2019. 37

[99] Kazuki Hayashi and Makoto Ohsaki. Reinforcement Learning and Graph Embedding for Binary Truss Topology Optimization Under Stress and Displacement Constraints. *Frontiers in Built Environment*, 6(N/A):59, apr 2020. 67

[100] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 92

[101] Graeme Henkelman, Andri Arnaldsson, and Hannes Jónsson. A fast and robust algorithm for bader decomposition of charge density. *Computational Materials Science*, 36(3):354–360, 2006. 69, 77

[102] Graeme Henkelman and Hannes Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *The Journal of Chemical Physics*, 113(22):9978–9985, 2000. 57

[103] Graeme Henkelman, Blas P. Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics*, 113(22):9901–9904, 2000. 57

[104] Javier Heras-Domingo, Mariona Sodupe, and Xavier Solans-Monfort. Interaction between Ruthenium Oxide Surfaces and Water Molecules. Effect of Surface Morphology and Water Coverage. *Journal of Physical Chemistry C*, 123(13):7786–7798, 2019. 125, 234

[105] Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. Data-driven materials science: status, challenges, and perspectives. *Advanced Science*, 6(21):1900808, 2019. 151

[106] Yoyo Hinuma, Takashi Toyao, Takashi Kamachi, Zen Maeno, Satoru Takakusagi, Shinya Furukawa, Ichigaku Takigawa, and Ken Ichi Shimizu. Density Functional Theory Calculations of Oxygen Vacancy Formation and Subsequent Molecular Adsorption on Oxide Surfaces. *Journal of Physical Chemistry C*, 122(51):29435–29444, 2018. 147

[107] J Hoja, LM Sandonas, B Ernst, A Vazquez-Mayagoitia, RAJ DiStasio, and A Tkatchenko. Qm7-x: a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules (version 2.0). *ZENODO https://doi. org/10.5281/zenodo*, 4288677, 2020. 44

[108] Weihua Hu, Muhammed Shuaibi, Abhishek Das, Siddharth Goyal, Anuroop Sriram, Jure Leskovec, Devi Parikh, and C Lawrence Zitnick. Forcenet: A graph neural network for large-scale quantum calculations. *arXiv preprint arXiv:2103.01436*, 2021. 105, 106, 108, 112, 113, 119, 127, 130, 131, 135, 136, 157, 158, 159, 161, 232

[109] Bing Huang, Nadine O Symonds, and O Anatole von Lilienfeld. The fundamentals of quantum machine learning. *arXiv preprint arXiv:1807.04259*, 2018. 90

[110] Hai-Cai Huang, Jun Li, Yang Zhao, Jing Chen, Yu-Xiang Bu, and Shi-Bo Cheng. Adsorption energy as a promising single-parameter descriptor for single

atom catalysis in the oxygen evolution reaction. *Journal of Materials Chemistry A*, 9(10):6442–6450, 2021. 127

[111] Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Normalization techniques in training dnns: Methodology, analysis and application. *arXiv preprint arXiv:2009.12836*, page p. N/A, 2020. 224

[112] Jens S. Hummelshøj, Frank Abild-Pedersen, Felix Studt, Thomas Bligaard, and Jens K. Nørskov. CatApp: A web application for surface chemistry and heterogeneous catalysis. *Angewandte Chemie - International Edition*, 51(1):272–274, 2012. 92

[113] Anubhav Jain, Geoffroy Hautier, Shyue Ping Ong, Charles J Moore, Christopher C Fischer, Kristin A Persson, and Gerbrand Ceder. Formation enthalpies by mixing GGA and GGA+U calculations. *Physical Review B*, 84(4):045115, jul 2011. 235

[114] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. 23, 26, 69, 73, 119, 121, 125, 126, 146, 208

[115] Mohammed Ibrahim Jamesh and Xiaoming Sun. Recent progress on earth abundant electrocatalysts for oxygen evolution reaction (OER) in alkaline medium to achieve efficient water splitting – A review. *Journal of Power Sources*, 400(February):31–68, 2018. 117

[116] Ryosuke Jinnouchi, Jonathan Lahnsteiner, Ferenc Karsai, Georg Kresse, and Menno Bokdam. Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with bayesian inference. *Phys. Rev. Lett.*, 122:225701, Jun 2019. 45, 48, 53, 58, 164

[117] Fabian Jirasek, Rodrigo A.S. Alves, Julie Damay, Robert A. Vandermeulen, Robert Bamler, Michael Bortz, Stephan Mandt, Marius Kloft, and Hans Hasse. Machine Learning in Thermodynamics: Prediction of Activity Coefficients by Matrix Completion. *Journal of Physical Chemistry Letters*, 11(3):981–985, feb 2020. 67

[118] Steven Johnson and John Joannopoulos. Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis. *Optics Express*, 8(3):173, 2001. 234

[119] Peter Bjørn Jørgensen, Karsten Wedel Jacobsen, and Mikkel N Schmidt. Neural message passing with edge updates for predicting properties of molecules and materials. *arXiv preprint arXiv:1806.03146*, 2018. 96, 111, 112

[120] Rajan Jose, Nurbosyn U Zhanpeisov, Hiroshi Fukumura, Yoshinobu Baba, and Mitsuru Ishikawa. Structure- property correlation of cdse clusters using experimental results and first-principles dft calculations. *Journal of the American Chemical Society*, 128(2):629–636, 2006. 39, 150

[121] Matthew Jouny, Wesley Luc, and Feng Jiao. High-rate electroreduction of carbon monoxide to multi-carbon products. *Nature Catalysis*, 2018. 37

[122] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. 160

[123] Anuj Karpatne, William Watkins, Jordan Read, and Vipin Kumar. Physics-guided neural networks (pgnn): An application in lake temperature modeling. *arXiv preprint arXiv:1710.11431*, 2017. 51

[124] Steven K Kauwe, Jake Graser, Antonio Vazquez, and Taylor D Sparks. Machine learning prediction of heat capacity for solid inorganics. *Integrating Materials and Manufacturing Innovation*, 7(2):43–51, 2018. 67

[125] Alireza Khorshidi and Andrew A. Peterson. Amp: A modular approach to machine learning in atomistic simulations. *Computer Physics Communications*, 207:310–324, 2016. 39, 40, 44, 48, 50, 51, 67, 81

[126] Myungjoon Kim, Byung Chul Yeo, Youngtae Park, Hyuck Mo Lee, Sang Soo Han, and Donghun Kim. Artificial intelligence to accelerate the discovery of n_2 electroreduction catalysts. *Chemistry of Materials*, 32(2):709–720, 2019. 67

[127] Yoolhee Kim, Edward Kim, Erin Antono, Bryce Meredig, and Julia Ling. Machine-learned metrics for predicting thelikelihood of success in materials discovery. *arxiv.1911.11201*, pages 1–13, 2019. 78

[128] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1(1):1–15, 2015. 67, 165

[129] John R Kitchin. Machine learning in catalysis. *Nature Catalysis*, 1(4):230–232, 2018. 67

[130] Johannes Klicpera, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *arXiv preprint arXiv:2106.08903*, 2021. 41, 43, 153, 154, 157, 158, 159, 161

[131] Johannes Klicpera, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020. 34, 44, 80, 85, 86, 87, 88, 119, 131, 135, 136, 152, 163, 212, 213, 217, 218, 219, 220, 232

[132] Johannes Klicpera, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. In *NeurIPS-W*, 2020. 43, 96, 97, 105, 107, 108, 110, 112, 113

[133] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations (ICLR)*, 2020. 16, 34, 40, 41, 42, 43, 80, 81, 85, 86, 87, 88, 96, 97, 108, 111, 112, 113, 119, 131, 135, 153, 154, 157, 158, 163, 212, 213, 217, 218, 219, 220, 232

[134] Tsz Wai Ko, Jonas A Finkler, Stefan Goedecker, and Jörg Behler. General-purpose machine learning potentials capturing nonlocal charge transfer. *Accounts of Chemical Research*, 54(4):808–817, 2021. 145

[135] Adeesh Kolluru, Nima Shoghi, Muhammed Shuaibi, Siddharth Goyal, Abhishek Das, Lawrence Zitnick, and Zachary W Ulissi. Transfer learning using attentions across atomic systems with graph neural networks (taag). *The Journal of Chemical Physics*, 2022. 119, 135, 147, 154

[136] Adeesh Kolluru, Muhammed Shuaibi, Aini Palizhati, Nima Shoghi, Abhishek Das, Brandon Wood, C Lawrence Zitnick, John R Kitchin, and Zachary W Ulissi. Open challenges in developing generalizable large scale machine learning models for catalyst discovery. *arXiv preprint arXiv:2206.02005*, 2022. 119, 127, 142, 144

[137] G. Kresse and J. Hafner. Ab initio molecular dynamics for open-shell transition metals. *Phys. Rev. B*, 48:13115–13118, Nov 1993. 54, 63

[138] Georg Kresse and Jürgen Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6(1):15–50, 1996. 54, 63, 75, 76, 127, 205, 234

[139] Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169–11186, 1996. 75, 76, 127, 205, 234

[140] Georg Kresse and Jürgen Hafner. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Physical Review B*, 49(20):14251–14269, 1994. 75, 76, 127, 205, 234

[141] Georg Kresse and Daniel Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B*, 59(3):1758, 1999. 75, 76, 127, 205, 234

[142] N Ktari, N Fourati, C Zerrouki, M Ruan, M Seydou, F Barbaut, F Nal, N Yaakoubi, MM Chehimi, and R Kalfat. Design of a polypyrrole mip-saw sensor for

selective detection of flumequine in aqueous media. correlation between experimental results and dft calculations. *RSC advances*, 5(108):88666–88674, 2015. 39, 150

[143] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, jun 2017. 63, 74, 76, 148, 216

[144] Xiangyun Lei and Andrew J Medford. A universal framework for featurization of atomistic systems. *arXiv preprint arXiv:2102.02390*, 2021. 154

[145] Baotong Li, Congjia Huang, Xin Li, Shuai Zheng, and Jun Hong. Non-iterative structural topology optimization using deep learning. *CAD Computer Aided Design*, 115(N/A):172–180, oct 2019. 67

[146] Bowen Li and Srinivas Rangarajan. Designing compact training sets for data-driven molecular property prediction through optimal exploitation and exploration. *Molecular Systems Design & Engineering*, 4(5):1048–1057, 2019. 67

[147] Hao Li, Zhien Zhang, and Zhijian Liu. Application of artificial neural networks for catalysis: a review. *Catalysts*, 7(10):306, 2017. 67

[148] Shunning Li, Yuanji Liu, Dong Chen, Yi Jiang, Zhiwei Nie, and Feng Pan. Encoding the atomic structure for machine learning in materials science. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(1):e1558, 2022. 43

[149] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. In *International Conference on Learning Representations (ICLR)*, 2016. 98

[150] Zheng Li, Siwen Wang, Wei Shan Chin, Luke E Achenie, and Hongliang Xin. High-throughput screening of bimetallic catalysts enabled by machine learning. *J. Mater. Chem. A*, 5(46):24131–24138, 2017. 68, 157

[151] Zheng Li, Siwen Wang, and Hongliang Xin. Toward artificial intelligence in catalysis. *Nature Catalysis*, 1(9):641–642, 2018. 67

[152] Shangtao Liang, Randall" David" Pierce Jr, Hui Lin, Sheau-Yun Chiang, and Qingguo" Jack" Huang. Electrochemical oxidation of pfoa and pfos in concentrated waste streams. *Remediation Journal*, 28(2):127–134, 2018. 147

[153] Bin Liu, Congming Li, Guoqiang Zhang, Xuesi Yao, Steven S.C. Chuang, and Zhong Li. Oxygen Vacancy Promoting Dimethyl Carbonate Synthesis from CO2 and Methanol over Zr-Doped CeO2 Nanorods. *ACS Catalysis*, 8(11):10446–10456, 2018. 122

[154] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989. 89

[155] Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013*, 2021. 110, 111, 112, 113, 130, 153, 154

[156] Yuanbin Liu, Weixiang Hong, and Bingyang Cao. Machine learning for predicting thermodynamic properties of pure fluids and their mixtures. *Energy*, 188(N/A):116091, dec 2019. 67

[157] Pietro P. Lopes, Dong Young Chung, Xue Rui, Hong Zheng, Haiying He, Pedro Farinazzo Bergamo Dias Martins, Dusan Strmcnik, Vojislav R. Stamenkovic, Peter Zapol, J. F. Mitchell, Robert F. Klie, and Nenad M. Markovic. Dynamically stable active sites from surface evolution of perovskite materials during the oxygen evolution reaction. *Journal of the American Chemical Society*, 143(7):2741–2750, 2021. 122

[158] Sönke Lorenz, Axel Groß, and Matthias Scheffler. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chemical Physics Letters*, 395(4-6):210–215, 2004. 153

[159] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2016. 50, 53, 202

[160] Xianfeng Ma and Hongliang Xin. Orbitalwise coordination number for predicting adsorption properties of metal nanocatalysts. *Phys. Rev. Lett.*, 118(3):036101, 2017. 44, 68, 152

[161] Osman Mamun, Kirsten T. Winther, Jacob R. Boes, and Thomas Bligaard. High-throughput calculations of catalytic properties of bimetallic alloy surfaces. *Scientific data*, 6(1):76, 2019. 44, 68, 157

[162] P. Mars and D. W. van Krevelen. Oxidations carried out by means of vanadium oxide catalysts. *Chemical Engineering Science*, 3:41–59, 1954. 147

[163] Antonio J. Martín, Gastón O. Larrazábal, and Javier Pérez-Ramírez, 2015. 37

[164] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293, 2014. 103

[165] Sebastian Matera, William F. Schneider, Andreas Heyden, and Aditya Savara. Progress in accurate chemical kinetic modeling, simulations, and parameter estimation for heterogeneous catalysis. *ACS Catalysis*, 9(8):6624–6647, 2019. 67

[166] Andrew J Medford, M Ross Kunz, Sarah M Ewing, Tammie Borders, and Rebecca Fushimi. Extracting knowledge from data through catalysis informatics. *ACS Catalysis*, 8(8):7403–7429, 2018. 67

[167] Bryce Meredig, Erin Antono, Carena Church, Maxwell Hutchinson, Julia Ling, Sean Paradiso, Ben Blaiszik, Ian Foster, Brenna Gibbons, Jason Hattrick-Simpers, Apurva Mehta, and Logan Ward. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Molecular Systems Design and Engineering*, 3(5):819–825, 2018. 78

[168] Benjamin Kurt Miller, Mario Geiger, Tess E Smidt, and Frank Noé. Relevance of rotationally equivariant convolutions for predicting molecular properties. *arXiv preprint arXiv:2008.08461*, 2020. 92, 110

[169] Hendrik J Monkhorst and James D Pack. Special points for brillouin-zone integrations. *Physical Review B*, 13(12):5188, 1976. 206

[170] Tim Mueller, Alberto Hernandez, and Chuhong Wang. Machine learning for interatomic potential models. *The Journal of Chemical Physics*, 152(5):050902, 2020. 48

[171] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *arXiv preprint arXiv:2204.05249*, 2022. 132, 165, 229

[172] Joseph Musielewicz, Xiaoxiao Wang, Tian Tian, and Zachary Ulissi. Finetuna: Fine-tuning accelerated molecular simulations. *arXiv preprint arXiv:2205.01223*, page p. N/A, 2022. 145

[173] Ryo Nagai, Ryosuke Akashi, and Osamu Sugino. Completing density functional theory by machine learning hidden messages from molecules. *npj Computational Materials*, 6(1):1–8, 2020. 77

[174] Suresh Kondati Natarajan and Jörg Behler. Neural network molecular dynamics simulations of solid–liquid interfaces: water at low-index copper surfaces. *Phys. Chem. Chem. Phys.*, 18:28704–28725, 2016. 39, 48

[175] United Nations. Sustainable development goals. `https://www.un.org/sustainabledevelopment/climate-change/`. Accessed: 2019-07. 37

[176] Ryky Nelson, Christina Ertural, Janine George, Volker L. Deringer, Geoffroy Hautier, and Richard Dronskowski. Lobster: Local orbital projections, atomic charges, and chemical-bonding analysis from projector-augmented-wave-based density-functional theory. *Journal of Computational Chemistry*, 41(21):1931–1940, 2020. 77

[177] Richard G Newell, Daniel Raimi, Seth Villanueva, and Brian Prest. Global Energy Outlook 2020: Energy Transition or Energy Addition? With Commentary on Implications of the COVID-19 Pandemic. Technical report, Resources for the future, 2020. 66

[178] Jigyasa Nigam, Sergey Pozdnyakov, and Michele Ceriotti. Recursive evaluation and iterative contraction of n-body equivariant features. *The Journal of Chemical Physics*, 153(12):121101, 2020. 92

[179] Juhwan Noh, Seoin Back, Jaehoon Kim, and Yousung Jung. Active learning with non-ab initio input features toward efficient co_2 reduction catalysts. *Chem. Sci.*, 9(23):5152–5159, 2018. 44, 68, 152, 156

[180] Jens K Nørskov, Frank Abild-Pedersen, Felix Studt, and Thomas Bligaard. Density functional theory in surface chemistry and catalysis. *Proceedings of the National Academy of Sciences*, 108(3):937–943, 2011. 127

[181] Jens K. Nørskov and Thomas Bligaard. The catalyst genome, 2013. 66

[182] Jens K Nørskov, Thomas Bligaard, Ashildur Logadottir, S Bahn, Lars B Hansen, Mikkel Bollinger, H Bengaard, Bjørk Hammer, Z Sljivancanin, Manos Mavrikakis, et al. Universality in heterogeneous catalysis. *Journal of catalysis*, 209(2):275–278, 2002. 127

[183] Jens K Nørskov, Felix Studt, Frank Abild-Pedersen, and Thomas Bligaard. *Fundamental concepts in heterogeneous catalysis*. John Wiley & Sons, 2014. 39, 66

[184] Ivan S. Novikov, Konstantin Gubaev, Evgeny V. Podryabinkin, and Alexander V. Shapeev. The mlip package: Moment tensor potentials with mpi and active learning, 2020. 53, 54

[185] J. K. Nørskov, J. Rossmeisl, A. Logadottir, L. Lindqvist, J. R. Kitchin, T. Bligaard, and H. Jónsson. Origin of the overpotential for oxygen reduction at a fuel-cell cathode. *The Journal of Physical Chemistry B*, 108(46):17886–17892, 2004. 39

[186] Department of Energy. Department of Energy, Accessed: 2019-07. 37

[187] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68(N/A):314–319, 2013. 73, 74, 127

[188] Open catalyst project challenge. `https://opencatalystproject.org/challenge.html`, 2021. 154, 159

[189] Is2re leaderboard concerns. https://discuss.opencatalystproject.org/t/is2re-leaderboard-concerns/66, 06 2021. 158

[190] Aini Palizhati, Wen Zhong, Kevin Tran, Seoin Back, and Zachary W Ulissi. Toward predicting intermetallics surface properties with high-throughput dft and convolutional neural networks. *Journal of chemical information and modeling*, 59(11):4742–4749, 2019. 43

[191] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 68

[192] Robert G Parr. Density functional theory of atoms and molecules. In *Horizons of quantum chemistry*, pages 5–15. Springer, 1980. 39, 43

[193] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary Devito Facebook, A I Research, Zeming Lin, Alban Desmaison, Luca Antiga, Orobix Srl, and Adam Lerer. Automatic differentiation in PyTorch. In *Advances in Neural Information Processing Systems 32*, 2017. 50

[194] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 79, 130

[195] Tipaporn Patniboon and Heine Anton Hansen. Acid-stable and active m–n–c catalysts for the oxygen reduction reaction: The role of local structure. *ACS Catalysis*, 11(21):13102–13118, 2021. 127

[196] Laurence Pause, Marc Robert, Joachim Heinicke, and Olaf Kühl. Radical anions of carbenes and carbene homologues. dft study and preliminary experimental results. *Journal of the Chemical Society, Perkin Transactions 2*, pages 1383–1388, 2001. 39, 150

[197] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865, 1996. 63, 75, 125, 205, 234

[198] Andrew A. Peterson. Acceleration of saddle-point searches with machine learning. *The Journal of Chemical Physics*, 145(7):074106, 2016. 39, 48, 51, 67, 157

[199] Andrew A. Peterson, Rune Christensen, and Alireza Khorshidi. Addressing uncertainty in atomistic machine learning. *Phys. Chem. Chem. Phys.*, 19:10978–10985, 2017. 53

[200] Dirk Porezag, Th Frauenheim, Th Köhler, Gotthard Seifert, and R Kaschner. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Physical Review B*, 51(19):12947, 1995. 43

[201] Philipp Pracht, Eike Caldeweyher, Sebastian Ehlert, and Stefan Grimme. A robust non-self-consistent tight-binding quantum chemistry method for large molecules. *ChemRxiv*, 2019. 81, 216

[202] William H Press, Brian P Flannery, Saul A Teukolsky, and William T Vetterling. Numerical recipes, cambridge: Cambridge univ, 1986. 206

[203] A Pukrittayakamee, M Malshe, M Hagan, LM Raff, R Narulkar, S Bukkapatnum, and R Komanduri. Simultaneous fitting of a potential-energy surface and its corresponding force fields using feedforward neural networks. *The Journal of chemical physics*, 130(13):134101, 2009. 88

[204] Péter Pulay. Convergence acceleration of iterative sequences. the case of scf iteration. *Chemical Physics Letters*, 73(2):393–398, 1980. 234

[205] Zhuoran Qiao, Anders S Christensen, Matthew Welborn, Frederick R Manby, Anima Anandkumar, and Thomas F Miller III. Unite: Unitary n-body tensor equivariant network with applications to quantum chemistry. *arXiv preprint arXiv:2105.14655*, 2021. 166

[206] Zhuoran Qiao, Matthew Welborn, Animashree Anandkumar, Frederick R Manby, and Thomas F Miller III. Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *arXiv preprint arXiv:2007.08026*, 2020. 43, 96, 111, 112, 113, 166

[207] David Raciti and Chao Wang. Electrochemical alternative to Fischer–Tropsch, 2018. 37

[208] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019. 92

[209] Krishnan Raghavachari, Gary W Trucks, John A Pople, and Martin Head-Gordon. A fifth-order perturbation comparison of electron correlation theories. *Chemical Physics Letters*, 157(6):479–483, 1989. 43

[210] Seyfeddine Rahali, Mohamed Ali Ben Aissa, Lotfi Khezami, Nuha Elamin, Mahamadou Seydou, and Abueliz Modwi. Adsorption behavior of congo red onto barium-doped zno nanoparticles: correlation between experimental results and dft calculations. *Langmuir*, 37(24):7285–7294, 2021. 39, 150

[211] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 105

[212] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014. 32, 43, 44, 51, 80, 96, 98, 104, 110, 111, 152

[213] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Big data meets quantum chemistry approximations: the δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015. 138

[214] Fang Ren, Logan Ward, Travis Williams, Kevin J Laws, Christopher Wolverton, Jason Hattrick-Simpers, and Apurva Mehta. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Science advances*, 4(4):eaaq1566, 2018. 150

[215] Navnath D Rode, Issam Abdalghani, Antonio Arcadi, Massimiliano Aschi, Marco Chiarini, and Fabio Marinelli. Synthesis of 2-acylindoles via ag-and cu-catalyzed anti-michael hydroamination of β-(2-aminophenyl)-α, β-ynones: Experimental results and dft calculations. *The Journal of organic chemistry*, 83(12):6354–6362, 2018. 39, 150

[216] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*, 2019. 96

[217] Andrew S Rosen, Shaelyn M Iyer, Debmalya Ray, Zhenpeng Yao, Alan Aspuru-Guzik, Laura Gagliardi, Justin M Notestein, and Randall Q Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021. 43

[218] Andrew S Rosen, Justin M Notestein, and Randall Q Snurr. Realizing the data-driven, computational discovery of metal-organic framework catalysts. *Current Opinion in Chemical Engineering*, 35:100760, 2022. 151

[219] Lars Rosenbaum, Alexander Dörr, Matthias R Bauer, Frank M Boeckler, and Andreas Zell. Inferring multi-target qsar models with taxonomy-based multi-task learning. *Journal of Cheminformatics*, 5(1):1–20, 2013. 120

[220] Kevin Rossi, Veronika Jurásková, Raphael Wischert, Laurent Garel, Clémence Corminbœuf, and Michele Ceriotti. Simulating solvation and acidity in complex mixtures with first-principles accuracy: The case of ch3so3h and h2o2 in phenol. *Journal of Chemical Theory and Computation*, 16(8):5139–5149, 2020. PMID: 32567854. 53

[221] Roger Rousseau, Vassiliki-Alexandra Glezakou, and Annabella Selloni. Theoretical insights into the surface physics and chemistry of redox-active oxides. *Nature Reviews Materials*, 5(6):460–475, 2020. 125

[222] Matthias Rupp, Raghunathan Ramakrishnan, and O. Anatole von Lilienfeld. Machine learning for quantum mechanical properties of atoms in molecules. *The Journal of Physical Chemistry Letters*, 6(16):3309–3313, 2015. 39, 48

[223] James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *JOM*, 65(11):1501–1509, 2013. 121, 146

[224] Edward Sanville, Steven D Kenny, Roger Smith, and Graeme Henkelman. Improved grid-based algorithm for bader charge allocation. *Journal of Computational Chemistry*, 28(5):899–908, 2007. 77

[225] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. *arXiv preprint arXiv:2102.09844*, 2021. 110, 112, 113

[226] Gabriele Scalia, Colin A Grambow, Barbara Pernici, Yi-Pei Li, and William H Green. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of chemical information and modeling*, 60(6):2697–2717, 2020. 164

[227] Gabriel R Schleder, Antonio C M Padilha, Carlos Mera Acosta, Marcio Costa, and Adalberto Fazzio. From DFT to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, 2(3):032001, may 2019. 48

[228] Philomena Schlexer Lamoureux, Kirsten T Winther, Jose Antonio Garrido Torres, Verena Streibel, Meng Zhao, Michal Bajdich, Frank Abild-Pedersen, and Thomas Bligaard. Machine learning for computational heterogeneous catalysis. *ChemCatChem*, 11(16):3581–3601, 2019. 67

[229] Jonathan Schmidt, Mário RG Marques, Silvana Botti, and Miguel AL Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):1–36, 2019. 67

[230] Christoph Schran, Jörg Behler, and Dominik Marx. Automated fitting of neural network potentials at coupled cluster accuracy: Protonated water clusters as testing ground. *Journal of Chemical Theory and Computation*, 16(1):88–99, 2020. PMID: 31743025. 51

[231] K. T. Schütt, P. J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K. R. Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, 2017. 43

[232] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A

continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, pages 991–1001, 2017. 16, 21, 34, 40, 41, 42, 43, 44, 80, 85, 86, 87, 88, 91, 96, 99, 108, 110, 111, 112, 113, 119, 131, 135, 136, 153, 154, 157, 212, 213, 217, 218, 219, 220, 232

[233] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021. 131, 135, 136, 153, 154, 232

[234] Kristof T Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8:13890, 2017. 44, 96, 111, 112

[235] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. 96, 99, 105, 106, 111, 112

[236] Zhi Wei Seh, Jakob Kibsgaard, Colin F Dickens, Ib Chorkendorff, Jens K Nørskov, and Thomas F Jaramillo. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science*, 355(6321), 2017. 37, 39, 66, 127

[237] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020. 96

[238] Burr Settles. Active Learning Literature Survey. *Machine Learning*, 2010. 16, 45, 48, 54

[239] David S. Sholl and Janice A. Steckel. *Density Functional Theory*. John Wiley & Sons, Inc., Hoboken, NJ, USA, mar 2009. 66

[240] Muhammed Shuaibi, Ben Comer, and Xiangyun Lei. Amptorch: Atomistic machine-learning package - pytorch. `https://github.com/ulissigroup/amptorch`, 2020. 50, 63

[241] Muhammed Shuaibi, Adeesh Kolluru, Abhishek Das, Aditya Grover, Anuroop Sriram, Zachary Ulissi, and C. Lawrence Zitnick. Rotation invariant graph neural networks using spin convolutions. *arXiv preprint arXiv:2106.09575*, 2021. 119, 127, 130, 131, 135, 136, 154, 157, 158, 159, 161, 224, 232

[242] Muhammed Shuaibi, Saurabh Sivakumar, Rui Qi Chen, and Zachary W Ulissi. Enabling robust offline active learning for machine learning potentials using simple physics-based priors. *Machine Learning: Science and Technology*, 2(2):025007, 2020. 138, 164

[243] Muhammed Shuaibi, Saurabh Sivakumar, Rui Qi Chen, and Zachary W Ulissi. Offlineal for mlps manuscript. `https://github.com/ulissigroup/OfflineAL-for-MLPs-manuscript`, 2020. 50, 200

[244] Ganesh Sivaraman, Anand Narayanan Krishnamoorthy, Matthias Baur, Christian Holm, Marius Stan, Gábor Csányi, Chris Benmore, and Álvaro Vázquez-Mayagoitia. Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Computational Materials*, 6(104), Jul 2020. 53

[245] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chemical science*, 8(4):3192–3203, 2017. 44, 151, 165

[246] Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature communications*, 10(1):1–8, 2019. 120

[247] Justin S Smith, Roman Zubatyuk, Benjamin Nebgen, Nicholas Lubbers, Kipton Barros, Adrian E Roitberg, Olexandr Isayev, and Sergei Tretiak. The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific data*, 7(1):1–10, 2020. 164

[248] Xue Zhi Song, Wen Yu Zhu, Xiao Feng Wang, and Zhenquan Tan. Recent Advances of CeO2-Based Electrocatalysts for Oxygen and Hydrogen Evolution as well as Nitrogen Reduction. *ChemElectroChem*, 8(6):996–1020, 2021. 121

[249] Joshua M. Spurgeon and Bijandra Kumar. A comparative technoeconomic analysis of pathways for commercial electrochemical co2 reduction to liquid products. *Energy and Environmental Science*, 2018. 37

[250] Anuroop Sriram, Abhishek Das, Brandon M Wood, Siddharth Goyal, and C Lawrence Zitnick. Towards training billion parameter graph neural networks for atomic simulations. In *International Conference on Learning Representations*, 2022. 119, 130, 141, 154

[251] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks, 2017. 63

[252] Geng Sun and Philippe Sautet. Metastable Structures in Cluster Catalysis from First-Principles: Structural Ensemble in Reaction Conditions and Metastability Triggered Reactivity. *Journal of the American Chemical Society*, 140(8):2812–2820, feb 2018. 67

[253] Geng Sun and Philippe Sautet. Toward fast and reliable potential energy surfaces for metallic pt clusters by hierarchical delta neural networks. *Journal of Chemical Theory and Computation*, 15(10):5614–5627, 2019. 67

[254] Daniel P Tabor, Loïc M Roch, Semion K Saikin, Christoph Kreisbeck, Dennis Sheberla, Joseph H Montoya, Shyam Dwaraknath, Muratahan Aykol, Carlos Ortiz, Hermann Tribukait, et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature Reviews Materials*, 3(5):5–20, 2018. 67

[255] Keisuke Takahashi, Lauren Takahashi, Itsuki Miyazato, Jun Fujima, Yuzuru Tanaka, Takeaki Uno, Hiroko Satoh, Koichi Ohno, Mayumi Nishida, Kenji Hirai, Junya Ohyama, Thanh Nhat Nguyen, Shun Nishimura, and Toshiaki Taniike. The Rise of Catalyst Informatics: Towards Catalyst Genomics, 2019. 67

[256] W Tang, E Sanville, and G Henkelman. A grid-based bader analysis algorithm without lattice bias. *Journal of Physics: Condensed Matter*, 21(8):084204, 2009. 77

[257] Y Tang, O Selvitopi, D T Popovici, and A Buluç. A High-Throughput Solver for Marginalized Graph Kernels on GPU. In *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 728–738, 2020. 21, 91

[258] Yu-Hang Tang and Wibe A. de Jong. Prediction of atomization energy using graph kernel and active learning. *The Journal of Chemical Physics*, 150(4):044107, 2019. 21, 91, 209

[259] Michael P Teter, Michael C Payne, and Douglas C Allan. Solution of schrödinger's equation for large systems. *Physical Review B*, 40(18):12255, 1989. 206

[260] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 96, 110, 112, 113

[261] Janis Timoshenko and Anatoly I Frenkel. "inverting" x-ray absorption spectra of catalysts by machine learning in search for activity descriptors. *ACS Catalysis*, 9(11):10192–10211, 2019. 67

[262] Alessandra Toniato, Alain C Vaucher, and Teodoro Laino. Grand challenges on accelerating discovery in catalysis. *Catalysis Today*, 387:140–142, 2022. 151

[263] José A Garrido Torres, Paul C Jennings, Martin H Hansen, Jacob R Boes, and Thomas Bligaard. Low-scaling algorithm for nudged elastic band calculations using a surrogate machine learning model. *Physical review letters*, 122(15):156001, 2019. 164

[264] Takashi Toyao, Zen Maeno, Satoru Takakusagi, Takashi Kamachi, Ichigaku Takigawa, and Ken-ichi Shimizu. Machine learning for catalysis informatics: Recent applications and prospects. *ACS Catalysis*, 10(3):2260–2297, 2019. 67

[265] Kevin Tran, Willie Neiswanger, Junwoong Yoon, Qingyang Zhang, Eric Xing, and Zachary W Ulissi. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, 1(2):025006, 2020. 44, 54, 67, 68, 164, 213

[266] Kevin Tran and Zachary W. Ulissi. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nature Catalysis*, 1:696–703, 2018. 21, 34, 67, 75, 91, 92, 151, 213

[267] Sergio Trasatti. Electrocatalysis by oxides - Attempt at a unifying approach. *Journal of Electroanalytical Chemistry*, 111(1):125–131, 1980. 117

[268] Turki Turki, Zhi Wei, and Jason TL Wang. Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. *IEEE Access*, 5:7381–7393, 2017. 120

[269] Zachary W Ulissi, Andrew J Medford, Thomas Bligaard, and Jens K Nørskov. To address surface reaction network complexity using scaling relations machine learning and dft calculations. *Nature Communications*, 8(1):1–7, 2017. 67

[270] Oliver T Unke and Markus Meuwly. Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of chemical theory and computation*, 15(6):3678–3693, 2019. 44, 111, 112, 113

[271] Maxime Van Den Bossche and Henrik Grönbeck. Adsorbate Pairing on Oxide Surfaces: Influence on Reactivity and Dependence on Oxide, Adsorbate Pair, and Density Functional. *Journal of Physical Chemistry C*, 121(15):8390–8398, 2017. 125, 234

[272] Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Computational Materials*, 6(20), Mar 2020. 45, 46, 48, 53, 58, 164

[273] "the calculations in this work have been performed using the ab-initio total-energy and molecular- dynamics package vasp (vienna ab-initio simulation package) developed at the institut für materialphysik of the universiät wien". 75, 76, 127, 205, 234

[274] Jacques C Védrine. Metal oxides in heterogeneous oxidation catalysis: State of the art and challenges for a more sustainable world. *ChemSusChem*, 12(3):577–588, 2019. 147

[275] Olga Vinogradova, Dilip Krishnamurthy, Vikram Pande, and Venkatasubramanian Viswanathan. Quantifying confidence in dft-predicted surface pourbaix diagrams of transition-metal electrode–electrolyte interfaces. *Langmuir*, 34(41):12259–12269, 2018. 127

[276] O Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry*, pages 1–12, 2020. 21, 71, 91

[277] Matthew J. Wahila, Nicholas F. Quackenbush, Jerzy T. Sadowski, Jon Olaf Krisponeit, Jan Ingo Flege, Richard Tran, Shyue Ping Ong, Christoph Schlueter, Tien Lin Lee, Megan E. Holtz, David A. Muller, Hanjong Paik, Darrell G. Schlom, Wei Cheng Lee, and Louis F.J. Piper. The breakdown of Mott physics at VO2 surfaces. *arXiv*, pages 1–9, 2020. 127

[278] Youwei Wang, Wujie Qiu, Erhong Song, Feng Gu, Zhihui Zheng, Xiaolin Zhao, Yingqin Zhao, Jianjun Liu, and Wenqing Zhang. Adsorption-energy-based activity descriptors for electrocatalysts in energy storage applications. *National Science Review*, 5(3):327–341, 2018. 159

[279] Zhenbin Wang, Ya-Rong Zheng, Ib Chorkendorff, and Jens K Nørskov. Acid-stable oxides for oxygen electrocatalysis. *ACS Energy Letters*, 5(9):2905–2908, 2020. 121, 127

[280] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *arXiv preprint arXiv:1807.02547*, 2018. 96, 112, 113

[281] Devin T Whipple and Paul J A Kenis. Prospects of co2 utilization via direct heterogeneous electrochemical reduction. *The Journal of Physical Chemistry Letters*, 1(24):3451–3458, 2010. 37

[282] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating physics-based modeling with machine learning: A survey, 2020. 51

[283] Kirsten T. Winther, Max J. Hoffmann, Jacob R. Boes, Osman Mamun, Michal Bajdich, and Thomas Bligaard. Catalysis-Hub.org, an open electronic structure database for surface reactions. *Scientific Data*, 2019. 44, 68

[284] D. M. Wood and A. Zunger. A new method for diagonalising large matrices. *Journal of Physics A: General Physics*, 18(9):1343–1359, 1985. 234

[285] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 105

[286] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14):145301, 2018. 16, 21, 34, 41, 42, 43, 80, 85, 87, 88, 91, 92, 96, 108, 111, 112, 113, 119, 211, 213, 218, 219

[287] Jiayan Xu, Xiao-Ming Cao, and P Hu. Perspective on computational reaction prediction using machine learning methods in heterogeneous catalysis. *Physical Chemistry Chemical Physics*, 23(19):11155–11179, 2021. 151

[288] Lejin Xu, Xiang Meng, Ming Li, Wuyang Li, Zengguang Sui, Jianlong Wang, and Jun Yang. Dissolution and degradation of nuclear grade cationic exchange resin by fenton oxidation combining experimental results and dft calculations. *Chemical Engineering Journal*, 361:1511–1523, 2019. 39, 150

[289] Li Yang, Lei Sun, and Wei Qiao Deng. Combination Rules for Morse-Based van der Waals Force Fields. *Journal of Physical Chemistry A*, 2018. 201, 202

[290] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? *Advances in Neural Information Processing Systems*, 34, 2021. 119, 127, 130, 154, 158

[291] Yiran Ying, Ke Fan, Xin Luo, Jinli Qiao, and Haitao Huang. Unravelling the origin of bifunctional oer/orr activity for single-atom catalysts supported on c2n by dft and machine learning. *Journal of Materials Chemistry A*, 9(31):16860–16867, 2021. 151

[292] Junwoong Yoon and Zachary W Ulissi. Differentiable optimization for the prediction of ground state structures (dogss). *Physical Review Letters*, 125(17):173001, 2020. 156

[293] Nannan Yuan, Qianqian Jiang, Jie Li, and Jianguo Tang. A review on non-noble metal based electrocatalysis for the oxygen evolution reaction. *Arabian Journal of Chemistry*, 13(2):4294–4309, 2020. 117

[294] Alexandra Zagalskaya, Iman Evazzade, and Vitaly Alexandrov. Ab initio thermodynamics and kinetics of the lattice oxygen evolution reaction in iridium oxides. *ACS Energy Letters*, 6(3):1124–1133, 2021. 127

[295] Adam Zemla. Lga: a method for finding 3d similarities in protein structures. *Nucleic acids research*, 31(13):3370–3374, 2003. 160

[296] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020. 96, 98

[297] Junmian Zhu, Bobby G Sumpter, Stephan Irle, et al. Artificial neural network correction for density-functional tight-binding molecular dynamics simulations. *MRS Communications*, 9(3):867–873, 2019. 51, 138

[298] C. L. Zitnick, L. Chanussot, A. Das, S. Goyal, J. Heras-Domingo, C. Ho, W. Hu, T. Lavril, A. Palizhati, M. Riviere, M. Shuaibi, A. Sriram, K. Tran, B. Wood, J. Yoon, D. Parikh, and Z. Ulissi. An introduction to electrocatalyst design using machine learning for renewable energy storage. *arXiv preprint arXiv:2010.09435*, 2020. 15, 37, 38, 39, 96, 114

[299] Tetiana Zubatiuk, Benjamin Nebgen, Nicholas Lubbers, Justin S Smith, Roman Zubatyuk, Guoqing Zhou, Christopher Koh, Kipton Barros, Olexandr Isayev, and Sergei Tretiak. Machine learned hückel theory: Interfacing physics and deep neural networks. *The Journal of Chemical Physics*, 154(24):244108, 2021. 145

[300] Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Jörg Behler, Gábor Csányi, Alexander V. Shapeev, Aidan P. Thompson, Mitchell A. Wood, and Shyue Ping Ong. Performance and cost assessment of machine learning interatomic potentials. *The Journal of Physical Chemistry A*, 124(4):731–745, 2020. PMID: 31916773. 48

Appendix A

Supplementary Information for Chapter 2

This work originally appeared as the Supplementary Information for: Shuaibi, M., Sivakumar, S., Chen, R.Q. and Ulissi, Z.W., 2020. Enabling robust offline active learning for machine learning potentials using simple physics-based priors. Machine Learning: Science and Technology, 2(2), p.025007.

A.1 High-temperature MD

In a similar manner to Figure 4 of the main text, we demonstrate our framework's ability to successfully converge to an accurate high-temp MD despite the highly perturbed sampled configurations. This same experiment was unsuccessful without the inclusion of the more potential prior we have introduced in this work. Beginning with a dataset containing a single structure, we run the proposed framework over a 2ps MD simulation of CO on Cu(100) in a 800K NVT ensemble. We illustrate our results in Figure A-1, with good agreement as early as the 3rd iteration - suggesting that the sampled highly-perturbed structures aid in reaching a converged simulation in much fewer iterations.

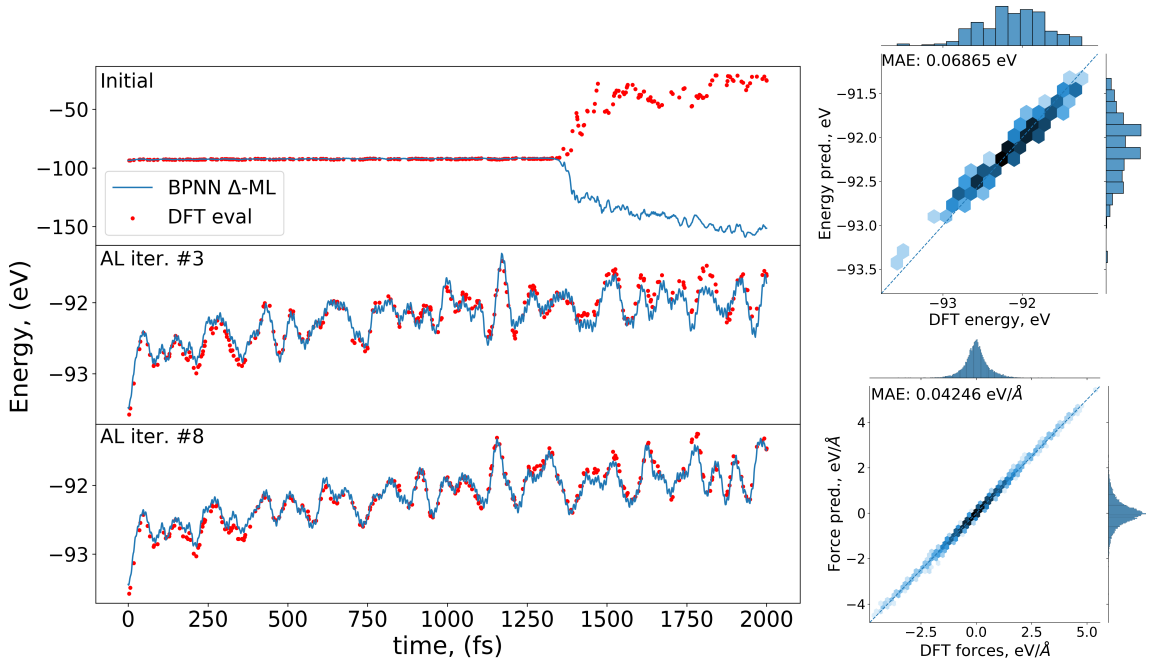


Figure A-1: Offline-AL demonstration to a 2ps MD simulation of CO on Cu(100) at 800K**(a)** Evolution of the MD trajectory over several iterations of the active learning framework. We verify the effectiveness of our framework by randomly sampling configurations and comparing DFT evaluated energy and forces with that of our model's predictions. **(b)** Parity plots associated with the DFT evaluated configurations and our model's predictions on the 8th iteration, demonstrating good agreement. Shading was scaled logarithmically with darker shading corresponding to a higher density of points.

A.2 Interactive examples

Several interactive Google Colab notebooks have been prepared to allow readers to conveniently explore the proposed methods. Accelerated structural relaxations and transition state calculations can be found at Ref [243]. Random query strategies are used to demonstrate the effectiveness of even the simplest of strategies. We encourage users to explore query and termination strategies that best suites their application of choice. DFT calculations are performed directly in the notebook examples via a GPU-enabled Quantum Espresso package.

A.3 Morse parameters fitting

The Morse potential was selected for our primarily bonded, catalytic systems. Parameters of the Morse potential, D_e, r_e, and a, corresponding to well depth, equilibrium distance, and well width were computed in the following manner for a given element, X:

1. Lone atomic energies, E_X, obtained through singlepoint DFT calculations;
2. Diatomic atoms relaxed to obtain a relaxed state energy, E_{X_2}, and equilibrium distance, r_e.
3. Well depth, D_e, is calculated as follows:

$$D_e = -(E_{X_2} - 2 * E_X) \tag{A.1}$$

4. Diatomic bond stretched and corresponding DFT points fit to Morse potential functional form (A.2) to obtain a Figure (A-2).

$$E_{morse} = D_e(e^{-2a(r-r_e)} - 2e^{-a(r-r_e)}) \tag{A.2}$$

To make use of the Morse potential for multi-element systems, linear mixing rules are utilized to compute element pair parameters. Adapted from Yang, et al.[289] the Morse potential is rewritten and parameter combinations applied accordingly (A.3-A.6)

$$E_{morse} = D_e(\exp[-\frac{2C}{\sigma}(r - r_e)] - 2\exp[-\frac{C}{\sigma}(r - r_e))] \tag{A.3}$$

$$D_{AB} = \sqrt{D_A D_B} \tag{A.4}$$

$$r_{e,AB} = \frac{r_{e,A} + r_{e,B}}{2} \tag{A.5}$$

$$\sigma_{AB} = \frac{\sigma_A + \sigma_B}{2} \tag{A.6}$$

Where $C = \ln 2/(r_e - \sigma)$ and σ corresponds to $E_{morse}(\sigma) = 0$. Although more sophis-

ticate combination rules exist [289], the accuracy of our Morse potential is not crucial for the success of our framework as it is meant to provide some guidance to the model.

A.4 Convergence

The convergence of the Offline-AL loop can be accelerated through the use of a learning rate scheduler. Figure A-3 compares the learning curves of AL frameworks with and without a learning rate scheduler, ceteris paribus. We demonstrate that a cosine annealing scheduler with warm restarts [159] was able to assist the convergence by smoothing out the learning curve and requiring fewer training images to reach a similar level or error.

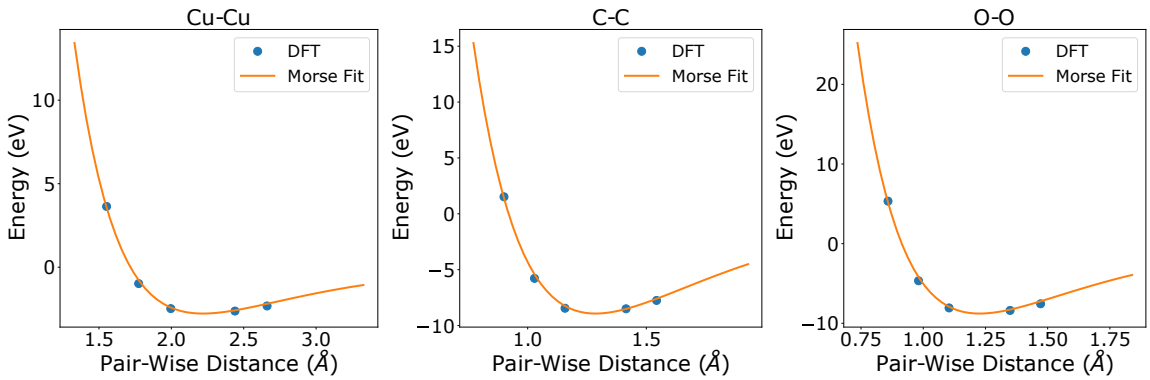


Figure A-2: Morse parameters are obtained by fitting DFT points near the equilibrium distance to equation A.2. Sample fittings are illustrated for copper, carbon, and oxygen.

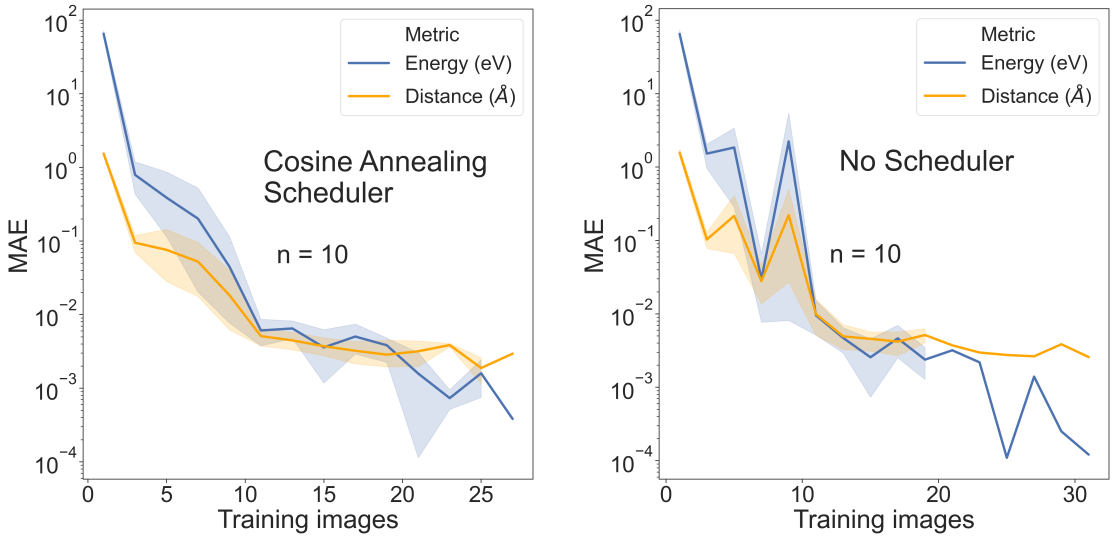


Figure A-3: Offline-AL convergence of our BPNN Δ-ML is compared with and without a learning rate scheduler. The use of a scheduler, particularly with small data, enables our framework to converge more reliably to the local minima.

Appendix B

Supplementary Information for Chapter 3

This work originally appeared as the Supplementary Information for: Chanussot, L., Das, A.*, Goyal, S.*, Lavril, T.*, Shuaibi, M.*, Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., Palizhati, A., Sriram, A., Wood, B., Yoon, J., Parikh, D., Zitnick, C.L., and Ulissi, Z., 2021. Open catalyst 2020 (OC20) dataset and community challenges. ACS Catalysis, 11(10), pp.6059-6072. *These authors contributed equally.*

B.1 DFT Relaxations

DFT calculations were performed with the *Vienna Ab Initio Simulation Package* (VASP)[140, 138, 139, 273, 141] with periodic boundary conditions and the projector augmented wave (PAW) pseudopotentials [32, 141]. 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 [197] and the revised Perdew-Burke-Ernzerhof (RPBE) functional, because of its improved description of the energetics of atomic and molecular bonding to surfaces [96]. 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

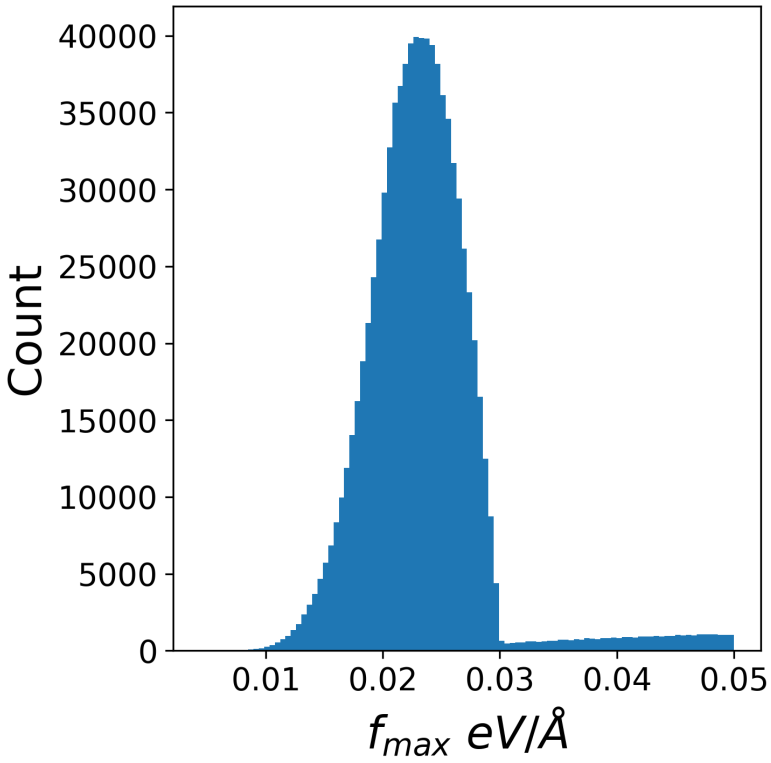


Figure B-1: The distribution of max-absolute forces, f_{max}, for systems that converged and completed successfully. Systems in which $f_{max} > 0.05$ eV/Å were excluded from all tasks except S2EF.

Monkhorst-Pack grid [169]. The ionic degrees of freedom were relaxed using a Conjugate Gradient minimization [259, 202]. The relaxation was terminated when either the Hellmann-Feynman forces [66] 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 B-1 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.

B.2 Adsorption Energy

$$E_{ad} = E_{sys} - E_{slab} - E_{gas}$$

Gas phase references, E_{gas}, for each adsorbate was computed as a linear combination of N_2, H_2O, CO, and H_2 resulting in the atomic energies from Table B.1.

Adsorbate atom	Energy (eV)
H	-3.477
O	-7.204
C	-7.282
N	-8.083

Table B.1: The per atom energy of individual adsorbate atoms used to calculate the gas phase reference energy for an adsorbate molecule

B.3 Computational Workflow

An illustration of the workflow used to sampled from the dataset and perform calculations is show in Figure B-2.

B.4 Graph Construction

Given a set of atoms in the 3D unit cell that is periodically repeated, we construct a radius graph where nodes represent the atoms and edges represent nearby interaction between pairs of atoms. Specifically, we draw a directed edge from atom j to atom i if atom j is within the cutoff distance from atom i, and vice versa. This means that the edges are always bidirectional. Furthermore, since the nodes are periodically repeated, two atoms may have multiple directed edges if they lie within the cutoff distance in multiple repeated cells. If an atom i has more than one edge to an atom j, each edge represents atom j in a different cell, resulting in unique relative distances and edge features, Figure B-3. From the atom-centric view, the above directed multi-

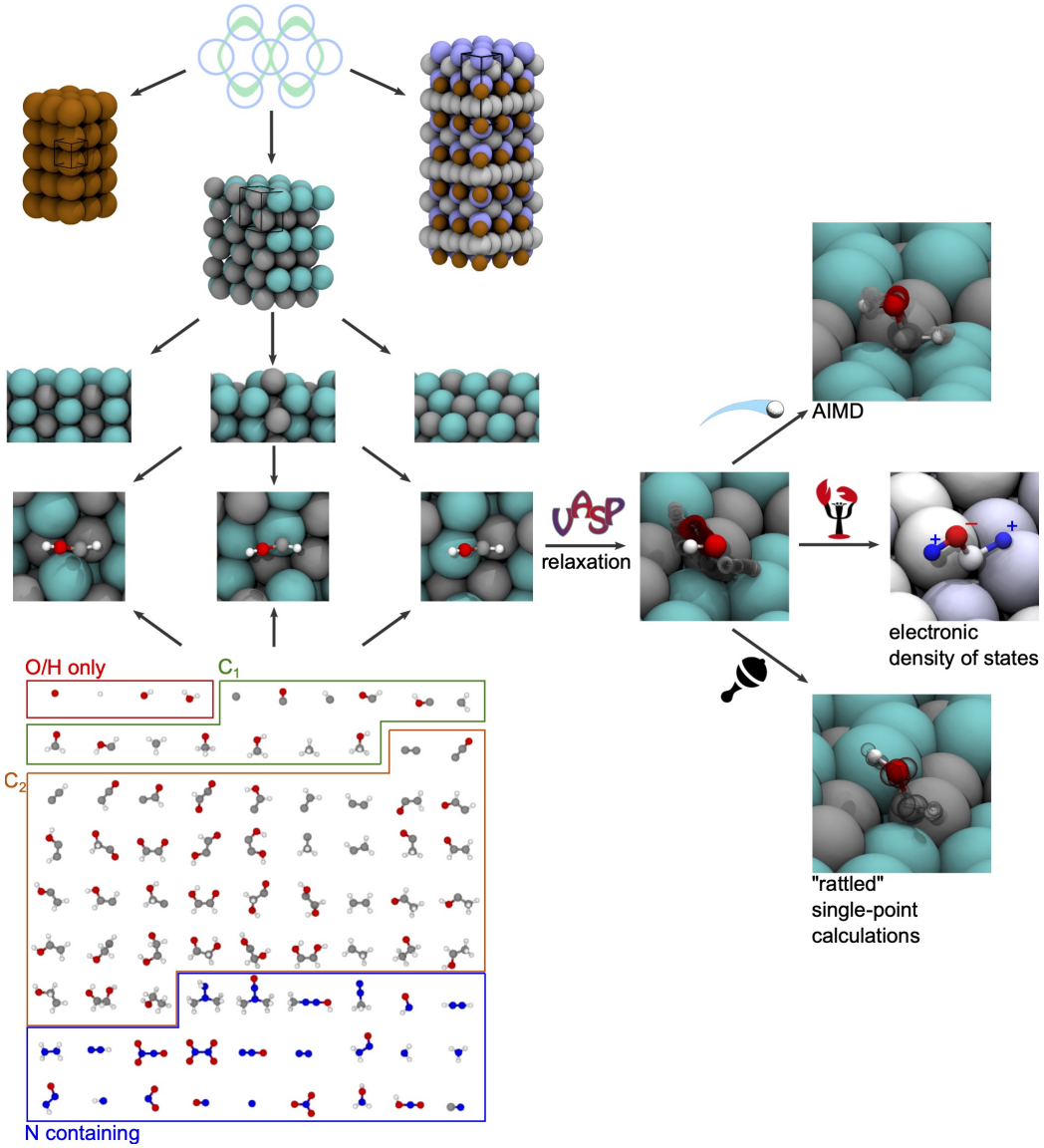


Figure B-2: The workflow used to generate the Open Catalyst Dataset. Stable materials were downloaded from The Materials Project[114] and paired with heuristically chosen adsorbates to create adsorption structures. These structures were randomly sampled for DFT relaxation and then subsequent AIMD, electronic structure analysis, and single-point rattling calculations.

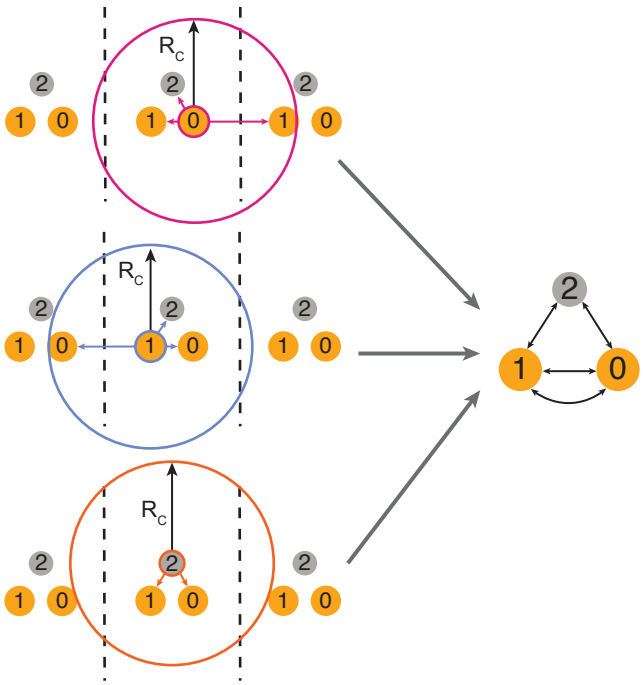


Figure B-3: A simple example of constructing a radius graph with periodic boundary conditions. The graph on the right represents all edges assuming each atom as the center node individually (shown on the left).

graph representation of the atomic system precisely captures the local 3D structure surrounding each atom, taking periodic boundary condition into account.

B.5 Graph Pairwise Similarity

The mean pairwise similarity (mps) between a collection of graphs gives an indication of 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. [258]. 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 6 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 6 Å and the molecular kernel edge length scale was set to 18 Å, nearly identical results were achieved with 2 Å and 5 Å respectively. All other parameters were set to default values.

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

B.6 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++.

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.

B.7 Hyperparameters for Baseline Models

Model hyperparameters for the 'All' splits of the IS2RE and S2EF tasks are provided in Tables B.2, B.3, and B.4. Hyperparameters of the remaining splits can be found in the corresponding repo: `https://github.com/Open-Catalyst-Project/ocp/tree/master/configs`.

Hyperparameters	IS2RE	S2EF
Size of atom embeddings	384	512
Size of fully connected layers	512	128
Number of fully connected layers	4	3
Number of graph convolutional layers	6	3
Number of Gaussians used for smearing	100	100
Cutoff distance for interatomic interactions	6	6
Batch size (per gpu)	16	24
Initial learning rate	0.001	0.0005
Learning rate gamma	0.1	0.1
Learning rate milestones	[5, 9, 13]	[3, 5, 7]
Warmup epochs	3	2
Warmup factor	0.2	0.2
Max epochs	20	20
Force coefficient	N/A	10

Table B.2: CGCNN [286] hyperparameters on the All split of the IS2RE and S2EF tasks.

Hyperparameters	IS2RE	S2EF
Number of hidden channels	384	1024
Number of filters	128	256
Number of interaction blocks	4	5
Number of Gaussians used for smearing	100	200
Cutoff distance for interatomic interactions	6	6
Global aggregation	add	add
Batch size (per gpu)	64	20
Initial learning rate	0.001	0.0001
Learning rate gamma	0.1	0.1
Learning rate milestones	[10, 15, 20]	[3, 5, 7]
Warmup epochs	3	2
Warmup factor	0.2	0.2
Max epochs	30	15
Force coefficient	N/A	30

Table B.3: SchNet [232] hyperparameters on the All split of the IS2RE and S2EF tasks.

Hyperparameters	IS2RE	S2EF
Number of hidden channels	256	192
Output block embedding size	192	192
Number of interaction blocks	3	3
Number of radial basis functions	6	6
Number of spherical harmonics	7	7
Number of residual layers before skip connection	1	1
Number of residual layers after skip connection	2	2
Number of linear layers in output blocks	3	3
Cutoff distance for interatomic interactions	6	6
Batch size (per GPU)	4	8
Initial learning rate	0.0001	0.0001
Learning rate gamma	0.1	0.1
Learning rate milestones	[4, 8, 12]	[2, 3, 4]
Warmup epochs	2	2
Warmup factor	0.2	0.2
Max epochs	20	7
Force coefficient	N/A	50

Table B.4: DimeNet++ [133, 131] hyperparameters on the All split of the IS2RE and S2EF tasks.

B.8 IS2RE Performance of Baseline Models on Previous Datasets

The MAE metrics of the baseline models for the *IS2RE* task are significantly higher than have been reported in recent studies applying ML models to predict adsorption energies[15, 265, 89]. There are three key differences in this work. First, the dataset here is larger, more diverse, sparser, and more uniformly sampled than previous datasets making this task more challenging. Second, we are using a more difficult definition of the *IS2RE* task - predict the final energy directly from the initial structure, rather than a clean representation of the final structure [15]. Finally, the baseline models themselves are somewhat different (both implementation, and details of the training and precise form).

To test that the baselines models were consistent with previous efforts, we applied all three models to the *IS2RE* task for a literature dataset of CO adsorption energies [15, 266], show in Table B.5. Our results are consistent, and often better, than previously reported validation accuracy for a CGCNN-based model at approximately 0.190 eV MAE on the literature dataset. This is far lower than the 0.57 eV MAE for our baseline models trained only on the CO subset of the OC20 dataset. This suggests that the dataset diversity is the dominant factor in this variation, and further emphasizes that a uniformly sampled dataset can be more difficult to fit than one obtained through an active learning process that emphasizes high-performing catalysts.

Model	Validation
	Energy MAE [eV] ↓
Previous Work [15, 266]	0.190
CGCNN [286]	0.174
SchNet [232]	0.170
DimeNet++ [133, 131]	0.149

Table B.5: Benchmark of our baseline models' implementations on a literature CO dataset[15, 266] as evaluated by Energy MAE.

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

Table B.6: Adsorbates considered in OC20

B.9 Adsorbates Included

The full list of adsorbates is indicated in Table B.6. 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 mono-dentate (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 mono-dentate or bi-dentate adsorption location is indicated by *.

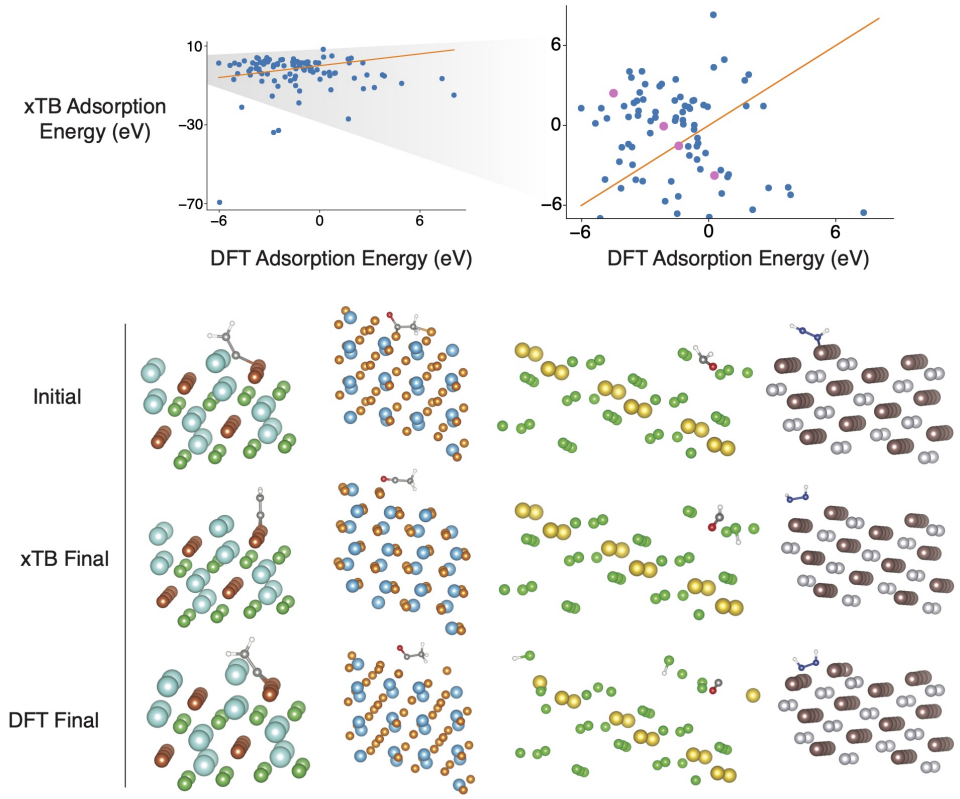


Figure B-4: Top: A parity plot comparing xTB adsorption energies with DFT adsorption energies and an inset that limits xTB values to a range similar to that of DFT. Bottom: Initial and final structures corresponding to the pink markers in the plot above organized from left to right.

B.10 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 *$ONNO_2$.

B.11 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 demon-

strated by SI Figure B-4. We preformed tight binding calculations on 100 random systems from the validation set with extended tight binding (xTB) and the atomic simulation environment (ASE) [143] interface using the GFN0 parameters [201]. All xTB calculations were carried out in accordance to our DFT procedures with a few notable differences. For the combined systems, i.e. an adsorbate on a surface, all surface atoms were fixed during the relaxation. Relaxations with xTB featured a BFGS optimizer instead of conjugated gradient, but the convergence criteria remained the same as other DFT calculations, f_{max} of 0.03 eV/Å or a maximum of 200 steps except for adsorbate references where f_{max} was 0.05 eV/Å. 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 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.

B.12 Additional Data: Rattled & Molecular Dynamics

Off-equilibrium data was additionally generated to diversify the structures in the dataset. Two approaches were use to generate this additional data: structural perturbations ("rattled") and molecular dynamics.

Rattled. Structures along the relaxation path way were sampled, perturbed via random atomic position displacements, and evaluated with DFT. For each relaxation, 20% of the intermediate structures were sampled for rattling. Atomic displacements were sampled from a normal distributions with $\mu = 0$ and $\sigma = 0.05$. Approximately 30 million single-point calculations were carried out. Upon filtering, $17M$ *S2EF* data points were used for training.

Molecular Dynamics. Short time-scale molecular dynamics simulations were per-

formed on previously relaxed structures. Simulations took place at 900K for 80 or 320 fs with an integration step size of 2 fs in the NVE ensemble. Approximately 64 million single-point calculations were carried out. Upon filtering, 38M *S2EF* data points were used for training.

Performance of baseline models. We report *S2EF* and *IS2RS* results for SchNet [232] and DimeNet++ [131] models optimized for force-prediction in Table B.7. Consistent with results in the main paper, we find that DimeNet++ outperforms SchNet (lower Force MAE, higher Force cosine, higher AFbT). Compared to training only on *S2EF* data, training on MD data seems to provide a complementary learning signal and leads to better sample efficiency – both DimeNet++ and SchNet trained on *S2EF*-20M + MD (58M training samples) outperform corresponding models trained on *S2EF*-All (134M training samples) as per AFbT. Finally, *IS2RS* AFbT seems to correlate better with *S2EF* Force cosine than *S2EF* Force MAE, especially when comparing models trained on Rattled or MD data.

			S2EF Test		IS2RS Test		
Model	Training Data	# Samples	Force MAE	Force cosine	ADwT	FbT	AFbT
SchNet [232]	S2EF-20M	20M	0.0535	0.3006	27.68%	0.00%	1.68%
SchNet [232]	S2EF-All	134M	0.0490	0.3417	31.78%	0.00%	3.38%
SchNet [232]	S2EF-20M + Rattled	37M	0.0691	0.3619	36.70%	0.10%	5.14%
SchNet [232]	S2EF-20M + MD	58M	0.0775	0.3885	41.10%	0.15%	8.97%
DimeNet++ [133, 131]	S2EF-20M	20M	0.0509	0.3382	34.37%	0.00%	2.67%
DimeNet++ [133, 131]	S2EF-All	134M	0.0357	0.4787	48.91%	0.25%	15.17%
DimeNet++ [133, 131]	S2EF-20M + Rattled	37M	0.0658	0.4395	43.94%	0.05%	12.51%
DimeNet++ [133, 131]	S2EF-20M + MD	58M	0.0635	0.4644	47.69%	0.15%	17.09%
DimeNet++ [133, 131]-large	S2EF-All	134M	0.0313	0.5443	51.67%	0.40%	21.74%

Table B.7: *S2EF* and IS2RS results of force-only SchNet and DimeNet++ models on *S2EF*, MD, and Rattled data.

B.13 Results on Validation splits

Full results on the validation splits are shown in Tables B.9, B.10, and B.8 for the *S2EF*, *IS2RS*, and *IS2RE* tasks respectively.

		Energy MAE [eV] ↓				EwT ↑			
		Validation							
Model	Approach	ID	OOD Ads	OOD Cat	OOD Both	ID	OOD Ads	OOD Cat	OOD Both
Median baseline	-	1.7466	1.7647	1.7283	1.5640	0.78%	0.80%	0.83%	0.91%
CGCNN [286]	Direct	0.6203	0.7426	0.6001	0.6708	3.36%	2.11%	3.53%	2.29%
SchNet [232]	Direct	0.6465	0.7074	0.6475	0.6626	2.96%	2.22%	3.03%	2.38%
DimeNet++ [133, 131]	Direct	0.5636	0.7127	0.5612	0.6492	4.25%	2.48%	4.40%	2.56%
SchNet [232]	Relaxation	0.7150	0.7395	0.8010	0.8197	4.03%	3.09%	3.87%	2.72%
SchNet [232] – force-only + energy-only	Relaxation	0.7110	0.7574	0.8316	0.8075	4.33%	2.88%	3.63%	2.57%

Table B.8: 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.

	S2EF Validation			
Model	ID	OOD Ads	OOD Cat	OOD Both
	Energy MAE [eV] ↓			
Median baseline	2.0715	2.2275	2.0558	2.3313
CGCNN [286]	0.5041	0.5986	0.5252	0.7308
SchNet [232]	0.4468	0.4973	0.5453	0.7047
SchNet [232] – force-only	34.0183	33.4238	34.2519	38.1693
SchNet [232] – energy-only	0.4011	0.4727	0.5607	0.7165
DimeNet++ [133, 131]	0.4545	0.5093	0.5184	0.6753
DimeNet++ [133, 131] – force-only	28.2095	28.4266	28.8740	35.0468
DimeNet++ [133, 131] – energy-only	0.3599	0.4500	0.5412	0.7108
DimeNet++ [133, 131]-Large – force-only	29.3524	29.4825	29.9799	36.6944
	Force MAE [eV/Å] ↓			
Median baseline	0.0810	0.0799	0.0798	0.0942
CGCNN [286]	0.0684	0.0746	0.0679	0.0852
SchNet [232]	0.0493	0.0574	0.0520	0.0685
SchNet [232] – force-only	0.0442	0.0514	0.0465	0.0618
SchNet [232] – energy-only	0.5810	0.6254	0.5875	0.6562
DimeNet++ [133, 131]	0.0443	0.0508	0.0445	0.0589
DimeNet++ [133, 131] – force-only	0.0331	0.0366	0.0343	0.0436
DimeNet++ [133, 131] – energy-only	0.3410	0.3322	0.3425	0.3502
DimeNet++ [133, 131]-Large – force-only	0.0281	0.0318	0.0315	0.0396
	Force Cosine ↑			
Median baseline	0.0000	0.000	0.000	0.000
CGCNN [286]	0.1550	0.1320	0.1456	0.1338
SchNet [232]	0.3185	0.2862	0.2973	0.2854
SchNet [232] – force-only	0.3604	0.3296	0.3294	0.3266
SchNet [232] – energy-only	0.0841	0.0695	0.0807	0.0699
DimeNet++ [133, 131]	0.3632	0.3401	0.3512	0.3556
DimeNet++ [133, 131] – force-only	0.4877	0.4747	0.4599	0.4849
DimeNet++ [133, 131] – energy-only	0.1064	0.0855	0.1043	0.0880
DimeNet++ [133, 131]-Large – force-only	0.5640	0.5500	0.5106	0.5390
	EFwT ↑			
Median baseline	0.00%	0.01%	0.01%	0.01%
CGCNN [286]	0.01%	0.00%	0.00%	0.01%
SchNet [232]	0.13%	0.00%	0.10%	0.00%
SchNet [232] – force-only	0.00%	0.00%	0.00%	0.00%
SchNet [232] – energy-only	0.00%	0.00%	0.00%	0.00%
DimeNet++ [133, 131]	0.09%	0.00%	0.09%	0.00%
DimeNet++ [133, 131] – force-only	0.00%	0.00%	0.00%	0.00%
DimeNet++ [133, 131] – energy-only	0.00%	0.00%	0.00%	0.00%
DimeNet++ [133, 131]-Large – force-only	0.00%	0.00%	0.00%	0.00%

Table B.9: Predicting energy and forces from a structure (*S2EF*) as evaluated by Mean Absolute Error (MAE) of the energies, force MAE, force cosine, and the percentage of Energies and Forces within Threshold (EFwT). Results reported for models trained on the entire training dataset (S2EF-All).

	IS2RS Validation			
Model	ID	OOD Ads	OOD Cat	OOD Both
	ADwT ↑			
IS baseline	21.18%	23.49%	20.25%	28.29%
SchNet [232]	15.53%	16.57%	14.50%	17.29%
SchNet [232] – force-only	32.41%	33.33%	30.02%	37.48%
DimeNet++ [133, 131]	30.40%	30.77%	29.94%	34.89%
DimeNet++ [133, 131] – force-only	49.05%	46.91%	46.54%	55.23%

Table B.10: Predicting relaxed structure from initial structure (*IS2RS*) as evaluated by Average Distance within Threshold (ADwT). All values in percentages, higher is better. Results reported for structure to energy-force (S2EF) models trained on the All training dataset. The initial structure was used as a naive baseline (IS baseline). Note that metrics requiring expensive DFT calculations – FbT and AFbT – are only computed for test splits, not val.

Appendix C

Supplementary Information for Chapter 5

This work originally appeared as the Supplementary Information for: Tran, R., Lan, J.*, Shuaibi, M.*, Wood, B.M.*, Goyal, S.*, Das, A., Heras-Domingo, J., Kolluru, A., Rizvi, A., Shoghi, N., Sriram, A., Ulissi, Z., Zitnick, C.L. 2022. The Open Catalyst 2022 (OC22) Dataset and Challenges for Oxide Electrocatalysis. arXiv preprint arXiv:2206.08917. ACS Catalysis, under review. *These authors contributed equally.*

C.1 OC20 *S2EF-Total* and *IS2RE-Total* results

To enable the comparison of total energy metrics between the OC22 and OC20 datasets, we trained baseline OC20 models for the proposed *S2EF-Total* and *IS2RE-Total* tasks. Table C.1 shows that across all models, *S2EF-Total* metrics are considerably worse than their *S2EF* counterparts. Similar to OC22, we see OOD metrics to be significantly worse than ID. Table C.2 shows a similar trend of *IS2RE-Total* with worse performance. It is worth noting that total energy based metrics are a more challenging task than their referenced counter parts. A model tasked with predicting total energies is required to capture all subsurface, surface, and adsorbate interactions accurately. In the case of OC20's adsorption reference, because a slab reference energy is subtracted off, models are ultimately focused on the energy associated with

Table C.1: A comparison of OC20 performance on *S2EF* and *S2EF-Total*. Across all models and splits, *S2EF-Total*, results in worse performance.

Task	Model	Energy MAE [eV] ↓	Force MAE [eV/Å] ↓	Force Cosine ↑
		ID		
S2EF	SchNet	0.447	0.049	0.319
	DimeNet++	0.455	0.044	0.363
	GemNet-dT	0.242	0.023	0.613
S2EF-Total	SchNet	3.737	0.047	0.343
	DimeNet++	3.043	0.032	0.515
	GemNet-dT	0.466	0.025	0.586
		OOD Ads		
S2EF	SchNet	0.497	0.057	0.286
	DimeNet++	0.509	0.051	0.340
	GemNet-dT	0.247	0.025	0.605
S2EF-Total	SchNet	3.756	0.053	0.318
	DimeNet++	3.052	0.035	0.503
	GemNet-dT	0.473	0.028	0.586
		OOD Cat		
S2EF	SchNet	0.545	0.052	0.297
	DimeNet++	0.518	0.045	0.351
	GemNet-dT	0.357	0.027	0.561
S2EF-Total	SchNet	3.853	0.049	0.312
	DimeNet++	3.159	0.034	0.472
	GemNet-dT	1.033	0.031	0.526
		OOD Both		
S2EF	SchNet	0.705	0.069	0.285
	DimeNet++	0.675	0.059	0.356
	GemNet-dT	0.415	0.034	0.596
S2EF-Total	SchNet	4.770	0.064	0.313
	DimeNet++	3.946	0.043	0.500
	GemNet-dT	1.263	0.039	0.560

only the adsorbate-surface interface.

Table C.2: A comparison of OC20 performance on *IS2RE* and *IS2RE-Total*. Across all models and splits, *IS2RE-Total* results in worse performance.

Task	Model	Energy MAE [eV] ↓	EwT [%] ↑
		ID	
IS2RE	SchNet	0.637	2.96
	DimeNet++	0.561	4.26
	GemNet-dT	0.526	4.65
IS2RE-Total	SchNet	1.698	0.87
	DimeNet++	1.378	1.35
	GemNet-dT	1.196	1.41
		OOD Ads	
IS2RE	SchNet	0.734	2.33
	DimeNet++	0.725	2.06
	GemNet-dT	0.705	2.21
IS2RE-Total	SchNet	1.683	1.05
	DimeNet++	1.426	1.19
	GemNet-dT	1.218	1.33
		OOD Cat	
IS2RE	SchNet	0.661	2.95
	DimeNet++	0.575	4.10
	GemNet-dT	0.533	4.59
IS2RE-Total	SchNet	2.359	0.64
	DimeNet++	1.766	0.93
	GemNet-dT	2.102	0.77
		OOD Both	
IS2RE	SchNet	0.704	2.22
	DimeNet++	0.661	2.42
	GemNet-dT	0.643	2.31
IS2RE-Total	SchNet	2.576	0.59
	DimeNet++	1.994	0.80
	GemNet-dT	2.331	0.72

C.2 Alternative reference scheme

Whereas OC20 references systems to correspond to an adsorption energy, OC22 is only concerned with making total energy predictions. In the context of model training, an adsorption energy reference modifies the target energy distribution of the dataset. Normalization schemes have been known to aid in accelerating and improving model training, particularly for deep neural networks [111]. We present a "linear referencing" approach as a normalization scheme for OC22. First, we fit a linear regression model to learn per-atom energies, i.e.

$$\boldsymbol{K} = \begin{bmatrix} K_H^1 & K_{He}^1 & K_{Li}^1 & \dots & K_{Fm}^1 \\ K_H^2 & K_{He}^2 & K_{Li}^2 & \dots & K_{Fm}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ K_H^N & K_{He}^N & K_{Li}^N & \dots & K_{Fm}^N \end{bmatrix}, \boldsymbol{P} = \begin{bmatrix} E_H \\ E_{He} \\ \vdots \\ E_{Fm} \end{bmatrix}$$

$$\boldsymbol{KP} = \begin{bmatrix} E_1^{DFT} \\ E_2^{DFT} \\ \vdots \\ E_N^{DFT} \end{bmatrix} \tag{C.1}$$

Where K_j^i corresponds to the count of element j in system i, E_j the per-element energy being fit, and E_i^{DFT} the ground truth DFT total energy. Once fitted, energy targets used for training, E_i^{ML}, are referenced as follows:

$$E_i^{ML} = E_i^{DFT} - \boldsymbol{K^i P} \tag{C.2}$$

Where $\boldsymbol{K^i}$ are the element counts for system i and $\boldsymbol{P}$ is the set of fitted per-element energy coefficients. Table C.3 compares model performance on *S2EF-Total* with and without the proposed linear reference. Across all models, we see a 6.5%, 43.9%, and 48.5% improvement in ID energy metrics for GemNet-OC[76], GemNet-dT[75], and SpinConv[241], respectively. With the execption of GemNet-OC, all models see

Table C.3: OC22 *S2EF-Total* test results for several top performing baseline GNNs, with and without a linear referencing scheme. A linear reference serves as an energy normalization strategy, aiding in overall energy performance across all models.

		ID			OOD		
Reference	Model	Energy MAE [eV] ↓	Force MAE [eV/Å] ↓	Force Cosine ↑	Energy MAE [eV] ↓	Force MAE [eV/Å] ↓	Force Cosine ↑
None	SpinConv	1.101	0.048	0.5140	1.981	0.070	0.386
	GemNet-dT	0.938	0.032	0.6647	1.272	0.041	0.530
	GemNet-OC	0.383	0.029	0.6903	0.833	0.040	0.554
Linear	SpinConv	0.567	0.036	0.6112	1.394	0.067	0.400
	GemNet-dT	0.470	0.032	0.6727	1.091	0.042	0.526
	GemNet-OC	0.357	0.030	0.6914	1.057	0.040	0.550

similar energy improvements for the OOD split. As expected, force metrics across all models see little change. This referencing was omitted from the main paper as to encourage other strategies to energy normalization, particularly for large, diverse datasets like OC22 and OC20.

C.3 Use of total energy models to predict adsorption energies

Total DFT energy predictions, although more general than adsorption energies, are not physically meaningful on their own, as only relative DFT energies are meaningful. One benefit to total energy prediction is that there are many possible catalysis and materials relevant properties that can be computed, only one of which is the adsorption energy. Here we elaborate on adsorption energy predictions via a direct route, as was done in OC20, and via total energy predictions, as is the case for OC22. The first approach, which we refer to as *ads-ref*, directly predicts adsorption energy as:

$$\hat{E}_{ads} = E_{ads}^{ML} \tag{C.3}$$

Where models are trained on adsorption energy targets ($E_{ads} = E_{adslab} - E_{slab} - E_{gas}$) that make use of DFT for the adslab, clean slab, and gas-phase adsorbate reference. The direct approach requires one model forward pass for predictions. The second approach, referred to as *total-ref*, involves making two predictions — a relaxed adslab and clean slab prediction:

$$\hat{E}_{ads} = E_{adslab}^{ML} - E_{slab}^{ML} - E_{adsorbate}^{DFT} \tag{C.4}$$

The *total-ref* approach uses total DFT energy targets for both adslabs and slabs and has the advantage of being more general.

The motivation for the change in referencing is two-fold: it enables adsorption energy predictions that span different surface coverages that are particularly important in oxide catalysts and it opens up the possibility of new property predictions such as surface energies. In order to maximize the properties accessible in the dataset we allowed all the slab atoms to relax. However, this adds the complication of potentially having possible inconsistent slab references. We note that to make an accurate adsorption energy calculation, the corresponding relaxed slab reference needs to be identical or similar to that of the relaxed adslab (Figure C-1a). Since OC22 created

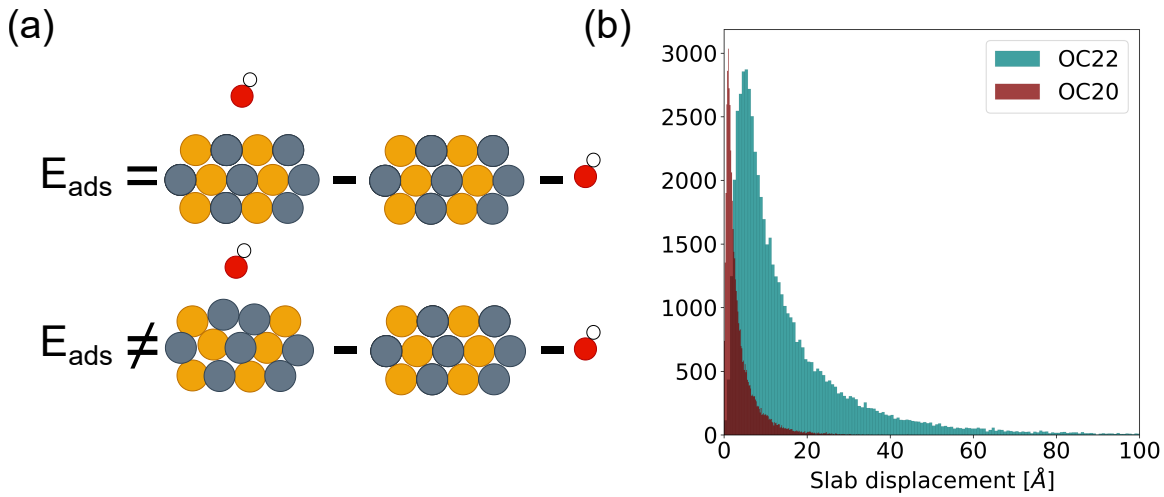


Figure C-1: **(a)** A correct adsorption calculation assumes the relaxed adslab and relaxed clean slab reference are consistent. **(b)** A histogram of cumulative slab displacement between the relaxed adslab and relaxed clean slab. OC22 systems observed a significant amount of movement compared to OC20, a consequence of all OC22 atoms being unconstrained and slabs not being optimized before adsorbate placement.

adslabs from unrelaxed slabs and relaxed corresponding pairs in parallel, it is possible that the slab is no longer a consistent reference. Figure C-1b illustrates the cumulative slab atom displacement between the relaxed adslab and relaxed clean slab for both OC20 and OC22.

Unsurprising, allowing the slab atoms to relax results in more atom movement on average in OC22 slabs compared to OC20. To account for this, we introduce a maximum allowed slab displacement to limit the validity of our systems to those within a tolerable threshold. As reference, OC20 had on average a maximum slab

Table C.4: Predicting OC22 adsorption energies via the proposed *total-ref* approach. Due to OC22 not always having consistent references, results are reported for varying subsets of the validation set in which max cumulative slab displacement is below a specified threshold.

	Energy MAE [eV] (validation size)	
Max slab displacement[A]	ID	OOD
None	0.890 (1097)	1.07 (1750)
10	0.568 (499)	0.71 (683)
6	0.533 (295)	0.65 (369)

displacement of ∼4Å. Table C.4 reports the adsorption energy performance of the total-ref on the OC22 dataset. A max slab displacement of “None" corresponds to the baseline of not filtering any OC22 systems. Without filtering any systems we observe large adsorption energy errors. When we limit our analysis to systems in which slabs are fairly consistent we see significantly better error metrics, suggesting a possible cancellation of errors.

While our analysis here is fairly limited given the size of the OC22 validation sets, it does provide a promising application of OC22 models to adsorption energy. Future work will involve a more rigorous analysis of OC22 models on adsorption energy upon curation of the relevant validation dataset. Regardless, all of the total energy models trained in this work will be relevant to these applications.

Table C.5: *S2EF-Total* results on adslab and slab subsets of the OC22 test splits. Models were trained and evaluated on only that subset. GemNet-OC* corresponds to the baseline model trained on all of OC22 but evaluated on the subsets in isolation.

Subset	Model	Energy MAE [eV] ↓		Force MAE [eV/Å] ↓		Force Cosine ↑	
		ID	OOD	ID	OOD	ID	OOD
Adslabs	SpinConv	0.933	1.955	0.034	0.062	0.601	0.409
	GemNet-dT	0.891	1.404	0.030	0.040	0.659	0.527
	GemNet-OC	0.448	0.871	0.029	0.037	0.674	0.548
	GemNet-OC *	0.358	0.815	0.027	0.037	0.695	0.553
Slabs	SpinConv	1.479	2.147	0.071	0.090	0.696	0.492
	GemNet-dT	1.613	2.033	0.057	0.062	0.490	0.414
	GemNet-OC	1.126	1.660	0.052	0.055	0.531	0.436
	GemNet-OC*	0.482	0.905	0.037	0.050	0.673	0.558

C.4 OC22 adslab and slab only performance

The OC22 dataset contains a combination of both adslab (adsorbate on a slab) and slab systems. While evaluation metrics are averaged across all systems, it may be useful to explore the performance of a particular subset. We trained several models on subsets in isolation and report results in Table C.5. Adslab performance does considerably better than slabs, a possible consequence of dataset size and the nature of adslab relaxations sampling a larger configurational space (e.g. slab relaxations often only require a few dozen DFT calculations). Best adslab and slab performance is achieved when training on both subsplits, suggesting adslabs are useful to improving slab performance, and vice-versa.

C.5 Training and hyperparameters

Baseline models used hyperparameters originally from OC20 or included a light sweep over some of the training settings - learning rate, optimizer, scheduler. For experiments within a particular task, model architectures were fixed. All model hyperparameters will be accessible at `https://github.com/Open-Catalyst-Project/ocp/tree/main/configs/oc22`. *S2EF-Total* models trained only on OC22 used an atom-wise loss function[24, 171] - weighing energy+forces in the loss function by $1 : N_{atoms}^2$. We found this to improve force metrics. A stepwise learning rate scheduler was also

used for these experiments, decaying the learning rate at 2,3,4,5,6 epochs. *S2EF-Total* joint training jobs used the original OC20 loss function and used a reduce-on-plateau learning rate scheduler. *S2EF-Total* fine-tuning experiments all used the original OC20 loss function as we noticed an atomwise loss function overfit very quickly on forces. The OC22-only experiments of the main paper used the OC20 loss function to allow for direct comparisons with the fine-tuning experiments. All joint training experiments involving OC20 used DFT total energies instead of adsorption energies. Unlike OC20, no energy normalization was done for training as we saw it to hurt performance across the board.

Models were trained using anywhere from 4-64 GPUs on 32Gb NVIDIA Volta cards. Learning rates for all fine-tuning experiments were reduced by 5-10x as compared to their base counterparts to ensure stable training. All models were optimized using AMSGrad. We provide the hyperparameters for our best performing *S2EF-Total* model variant, GemNet-OC, for the joint and fine-tuning training strategies in Table C.6.

Table C.6: Model hyperparameters for the top performing GemNet-OC joint and fine-tuning experiments.

Hyperparameters	OC20+OC22	OC20→OC22
No. spherical basis	7	7
No. radial basis	128	128
No. blocks	4	4
Atom embedding size	256	256
Edge embedding size	512	512
Triplet edge embedding input size	64	64
Triplet edge embedding output size	64	64
Quadruplet edge embedding input size	32	32
Quadruplet edge embedding output size	32	32
Atom interaction embedding input size	64	64
Atom interaction embedding output size	64	64
Radial basis embedding size	16	16
Circular basis embedding size	16	16
Spherical basis embedding size	32	32
No. residual blocks before skip connection	2	2
No. residual blocks after skip connection	2	2
No. residual blocks after concatenation	1	1
No. residual blocks in atom embedding blocks	3	3
No. atom embedding output layers	3	3
Cutoff	12.0	12.0
Quadruplet cutoff	12.0	12.0
Atom edge interaction cutoff	12.0	12.0
Atom interaction cutoff	12.0	12.0
Max interaction neighbors	30	30
Max quadruplet interaction neighbors	8	8
Max atom edge interaction neighbors	20	20
Max atom interaction neighbors	1000	1000
Radial basis function	Gaussian	Gaussian
Circular basis function	Spherical harmonics	Spherical harmonics
Spherical basis function	Legendre Outer	Legendre Outer
Quadruplet interaction	True	True
Atom edge interaction	True	True
Edge atom interaction	True	True
Atom interaction	True	True
Direct forces	True	True
Activation	Silu	Silu
Optimizer	AdamW	AdamW
EMA decay	0.999	0.999
Gradient clip norm threshold	10	10
Learning rate	0.0005	0.0001
Scheduler	ReduceLROnPlateau	StepwiseLRDecay
LR Milestones	N/A	epochs 2-10, 0.5 after
Force loss function	AtomwiseL2	L2
Energy loss function	MAE	MAE
Force coefficient	1	100
Energy coefficient	1	1

Table C.7: Predicting total energy and force from a structure (*S2EF-Total*). Results are shared for the default, joint training, and fine-tuning training strategies. Experiments are evaluated on the validation set.

S2EF-Total Validation

Training	Model	Energy MAE [eV] ↓		Force MAE [eV/Å] ↓		Force Cosine ↑		EFwT [%] ↑	
		ID	OOD	ID	OOD	ID	OOD	ID	OOD
OC22-only	Median Baseline	169.733	164.316	0.076	0.074	0.002	0.002	0.00	0.00
	SchNet [232]	8.081	12.562	0.060	0.092	0.359	0.233	0.00	0.00
	DimeNet++ [133, 131]	2.354	2.838	0.043	0.061	0.599	0.459	0.00	0.00
	ForceNet [108]	-	-	0.057	0.066	0.343	0.302	0.00	0.00
	SpinConv [241]	1.268	2.359	0.048	0.087	0.507	0.379	0.00	0.00
	PaiNN [233]	1.125	2.948	0.046	0.062	0.478	0.364	0.00	0.00
	GemNet-dT [75]	1.108	1.844	0.032	0.041	0.657	0.560	0.01	0.00
	GemNet-OC [76]	0.543	1.011	0.030	0.040	0.683	0.580	0.03	0.00
OC20-2M + OC22	PaiNN[233]	0.572	1.576	0.048	0.069	0.460	0.337	0.02	0.00
	SpinConv[241]	1.050	2.138	0.035	0.063	0.626	0.462	0.00	0.00
	GemNet-OC [76]	0.602	1.092	0.030	0.038	0.685	0.589	0.03	0.01
OC20-20M + OC22	PaiNN[233]	0.542	1.321	0.047	0.064	0.472	0.359	0.02	0.00
	SpinConv[241]	1.097	2.189	0.036	0.060	0.602	0.468	0.00	0.00
	GemNet-OC [76]	0.485	1.109	0.028	0.036	0.713	0.615	0.08	0.02
OC20-All + OC22	SpinConv[241]	1.399	2.275	0.040	0.054	0.527	0.447	0.00	0.00
	GemNet-OC [76]	0.463	0.858	0.027	0.034	0.698	0.617	0.10	0.01
OC20→OC22	SpinConv[241]	1.173	2.518	0.035	0.056	0.604	0.468	0.00	0.00
	GemNet-dT [75]	0.878	1.300	0.032	0.042	0.660	0.567	0.02	0.00
	GemNet-OC [76]	0.394	1.042	0.030	0.040	0.671	0.569	0.11	0.00
	GemNet-OC-Large [76]	0.613	1.196	0.029	0.039	0.707	0.602	0.03	0.01

C.6 S2EF-Total, IS2RE-Total, IS2RS validation results

Full validation results are shown in Tables C.7, C.8, C.9 for *S2EF-Total*, *IS2RE-Total*, and *IS2RS*, respectively.

Table C.8: Predicting total relaxed energy from an initial structure (*IS2RE-Total*). Results are shared for the default, joint training, and fine-tuning training strategies. Experiments are evaluated on the validation set.

IS2RE-Total Validation

Approach	Training	Model	Energy MAE [eV] ↓		EwT [%] ↑	
			ID	OOD	ID	OOD
Direct	OC22-only	Median Baseline	183.987	177.349	0.00	0.00
		SchNet	2.019	5.287	1.14	0.47
		DimeNet++	1.992	4.336	0.91	0.50
		PaiNN	1.770	4.336	1.49	0.36
		GemNet-dT	1.690	4.522	1.37	0.47
	OC20+OC22	SchNet	3.030	5.076	0.65	0.43
		DimeNet++	1.989	4.450	0.91	0.61
		PaiNN	1.764	4.690	1.33	0.36
		GemNet-dT	2.519	5.150	0.61	0.40
	OC20→OC22	GemNet-OC*	1.227	2.360	4.08	1.08
Relaxation	OC22	SpinConv	2.014	3.422	1.07	0.50
		GemNet-dT	1.894	2.575	1.07	0.83
		GemNet-OC	1.328	1.883	2.06	1.29
	OC20+OC22	SpinConv	2.313	3.492	0.69	0.61
		GemNet-OC	1.247	2.059	3.05	1.12
	OC20→OC22	SpinConv	1.878	3.460	1.49	0.61
		GemNet-OC	1.173	1.901	5.18	1.83
		GemNet-OC-Large	1.270	2.040	1.30	1.22

*GemNet-OC pretrained on OC20+OC22 *S2EF-Total*

Table C.9: Predicting relaxed structures from an initial structure *IS2RS*. All models predicted relaxed structures through an iterative relaxation approach. The initial structure was used as a naive baseline (IS baseline). Experiments are evaluated on the validation set.

IS2RS Validation

Training	Model	ADwT [%] ↑	
		ID	OOD
OC22-only	IS baseline	44.77	42.59
	SpinConv	57.69	43.30
	GemNet-dT	59.68	51.25
	GemNet-OC	60.69	52.90
OC20+OC22	SpinConv	55.79	47.31
	GemNet-OC	60.99	53.85
OC20→OC22	SpinConv	56.69	45.78
	GemNet-OC	58.03	48.33
	GemNet-OC-Large	59.69	51.66

C.7 Additional DFT settings

All structure relaxations were performed using the Vienna ab initio simulation package (VASP) [140, 138, 139, 273, 141] with the projector augmented wave (PAW) approach. We modelled the exchange-correlation effects using the Perdew-Berke-Ernzerhof (PBE), generalized gradient approximation (GGA) [197] which is generally accepted for modeling surface reactions on oxides[83, 104, 271]. All calculations were performed with spin-polarization to account for the significant spin states in metal oxides. The external electrons were expanded in plane waves with kinetic energy cut-offs of 500 eV. The energies and atomic forces of all calculations were converged to within 1×10^{-4} eV and 0.05 eV Å^{-1}, respectively. We used Γ-centered k-point meshes of $\frac{50}{a} \times \frac{50}{b} \times \frac{50}{c}$ and $\frac{30}{a} \times \frac{30}{b} \times 1$ for bulk and slab calculations, respectively, with non-integer values rounded up to the nearest integer. We used a Gaussian smearing algorithm for setting the partial occupancies of each orbital. We defaulted to a mixture of the blocked Davidson iteration[118] and the RMM-DIIS[284, 204] scheme as the algorithm for electron minimization and withdrew to using only the blocked Davidson iteration for calculations containing Pb and In that failed to converge electronically. Ions were updated using the conjugated gradient algorithm.

In this study, we placed adsorbates on one of the two surfaces of our slab which results in uneven charges between the two surfaces. This results in an nonphysical dipole moments that can lead to diverging total DFT energies. To account for this dipole moment, we introduced an electrostatic potential to the local potential of our adsorbed slab.

C.8 Hubbard U corrections

Materials with certain combinations of transition metals and oxygen are known to have strongly correlated electrons, i.e. the movement of electrons significantly influences the properties of other electrons. It is well known that the GGA functional is unable to properly account for these strong electron correlations leading to inaccurate

Table C.10: Hubbard U values for transition metals available on the Materials Project.

	U (eV)
Co	3.32
Cr	3.7
Fe	5.3
Mn	3.9
Mo	4.38
Ni	6.2
V	3.25
W	6.2

calculations of thermodynamic and electronic properties. We account for this missing electron interaction by introducing the Hubbard U correction which uses a repulsive Coulombic force between the electrons. The strength of this repulsion stems from the "U" value which is empirically fitted to experimental quantities such as the band gap or formation enthalpy. To properly account for the effects of strong electron correlation on the thermodynamic properties of our dataset, we adapted the Hubbard U values from the Materials Project which were fitted to correctly calculate the experimental enthalpy of formation[113] (see Table C.10 for the list of Hubbard U values).

C.9 Chemical systems

The dataset of slabs in OC22 is constructed from a set of 51 elements shown in Figure C-2 resulting in $\binom{51}{1} = 51$ and $\binom{51}{2} = 1275$ possible unary and binary oxides respectively. We considered all transitions metals up to the 5d group with the exception of Tc due to its radioactivity (29 metals), all alkali and alkaline earth metals up to Fr (10), the lanthanides of Ce and Lu (2), and the p-block metals and metalloids: except Te and up to Po (11). We queried the materials project for materials with the top five lowest energy above hull and less than 150 atoms for all unary and binary oxides composed of these elements. This resulted in 4,728 bulk oxide structures considered. Figure C-3 provides a 2D grid heat map showing the frequency of chemical systems sampled in the OC22 dataset. Only 4,286 of the 4,728 bulk oxides were sampled in the dateset. Unary oxides are shown in the diagonal of the grid while all other blocks represent binary oxides. Not all chemical systems were sampled in the final dataset as some chemical systems did not exist in the Materials Project (red hatches) while other chemical systems had bulk oxide systems that were too large to create slabs of less than 150 atoms (grey blocks). We observe that for each chemical system considered, around 50 to 100 slabs and adsorbed slabs were included in the final dataset which demonstrates the even distribution of chemical space sampled in the dataset. Slabs and adsorbed slabs with Li-O, Sb-Cr-O and Ag-O were randomly over sampled with over 250 entries in the final dataset.

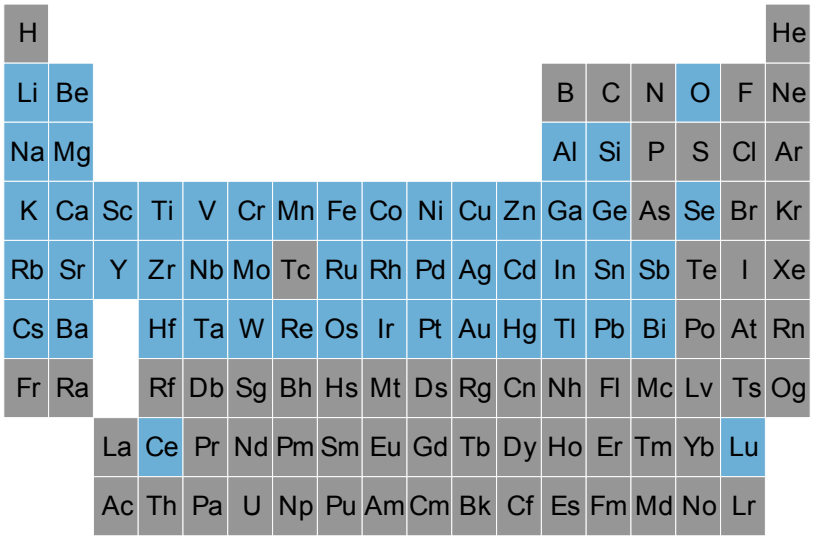


Figure C-2: Periodic table showing the 51 elements considered in the OC22 dataset in blue. Elements that were not considered are show in grey. All slabs were constructed from bulk oxides composed of one (unary) or two (binary) of these metals.

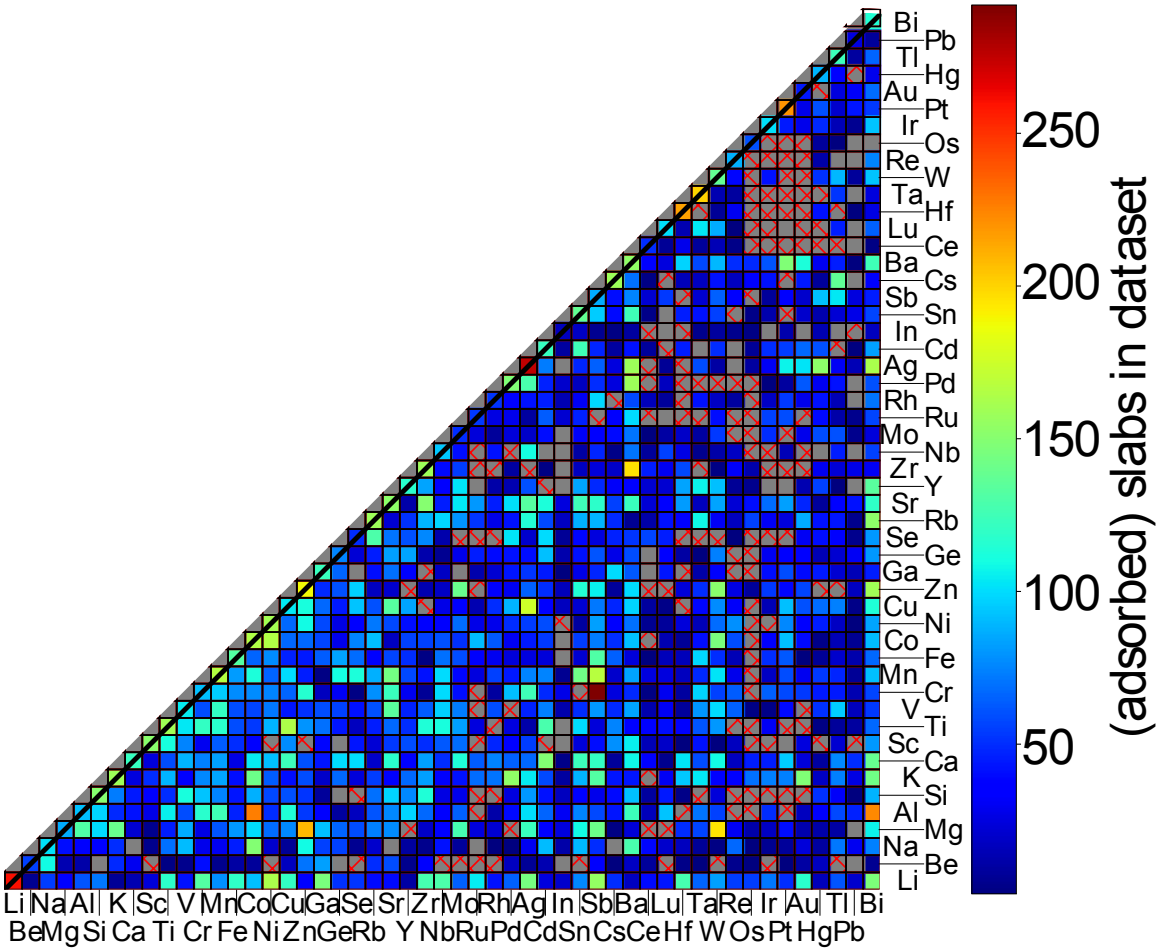


Figure C-3: A 2D grid heat map indicating the number of slabs and adsorbed slabs in the dataset containing specific pairs of metals of binary composition $A_xB_yO_z$. Grid points on the diagonal correspond to unary compositions of A_{xOy}. Grey grids containing red hatches correspond to compositions that were not available in the Materials Project. Grey grids without hatches indicate compositions that were in our possible sample set of materials, but were not randomly sampled during the construction of the dataset.