AQCat25: Unlocking spin-aware, high-fidelity machine learning potentials for heterogeneous catalysis

Omar Allam[†], Brook Wander[†], Aayush R. Singh[*]

SandboxAQ, Palo Alto, CA

Correspondence: aayush.singh@sandboxquantum.com

Large-scale datasets have enabled highly accurate machine learning interatomic potentials (MLIPs) for general-purpose heterogeneous catalysis modeling. There are, however, some limitations in what can be treated with these potentials because of gaps in the underlying training data. To extend these capabilities, we introduce AQCat25, a complementary dataset of 13.5 million density functional theory (DFT) single point calculations designed to improve the treatment of systems where spin polarization and/or higher fidelity are critical. We also investigate methodologies for integrating new datasets, such as AQCat25, with the broader Open Catalyst 2020 (OC20) dataset to create spin-aware models without sacrificing generalizability. We find that directly tuning a general model on AQCat25 leads to catastrophic forgetting of the original dataset's knowledge. Conversely, joint training strategies prove effective for improving accuracy on the new data without sacrificing general performance. This joint approach introduces a challenge, as the model must learn from a dataset containing both mixed-fidelity calculations and mixed-physics (spin-polarized vs. unpolarized). We show that explicitly conditioning the model on this system-specific metadata, for example by using Feature-wise Linear Modulation (FiLM), successfully addresses this challenge and further enhances model accuracy. Ultimately, our work establishes an effective protocol for bridging DFT fidelity domains to advance the predictive power of foundational models in catalysis.

SANDBOXAQ™

Introduction

Over the past three decades, computational approaches that couple first-principles density functional theory (DFT) with microkinetic modeling have become a cornerstone of modern heterogeneous catalysis research by providing a framework for rational catalyst design[1–5]. Numerous studies have linked atomic-scale surface chemistry to macroscopic kinetic observables, enabling the elucidation of complex reaction mechanisms for a variety of critical industrial heterogeneous catalytic processes including, but not limited to, ammonia synthesis[6,7], methanol synthesis[8,9], Fischer-Tropsch synthesis[10,11], selective hydrogenation[12,13], steam reforming of methane[14,15], the water-gas shift reaction[16,17], and ethylene epoxidation[18,19]. Many of these studies have leveraged unifying concepts such as d-band theory and the Brønsted-Evans-Polanyi relations[20–23], which correlate the binding and transition-state energies of elementary reactions, facilitating the construction of volcano plots that predict optimal catalyst performance, in some cases even leading to experimentally validated discovery of new catalysts[24,25]. Despite these successes, the prohibitive cost of DFT largely limits its application to relatively simple networks of reactions taking place over idealized low-index facets of unary and binary catalyst materials[2,23].

Machine learning interatomic potentials (MLIPs) have emerged as an attractive alternative to esti-

[†]Co-first authors; [*]Corresponding author

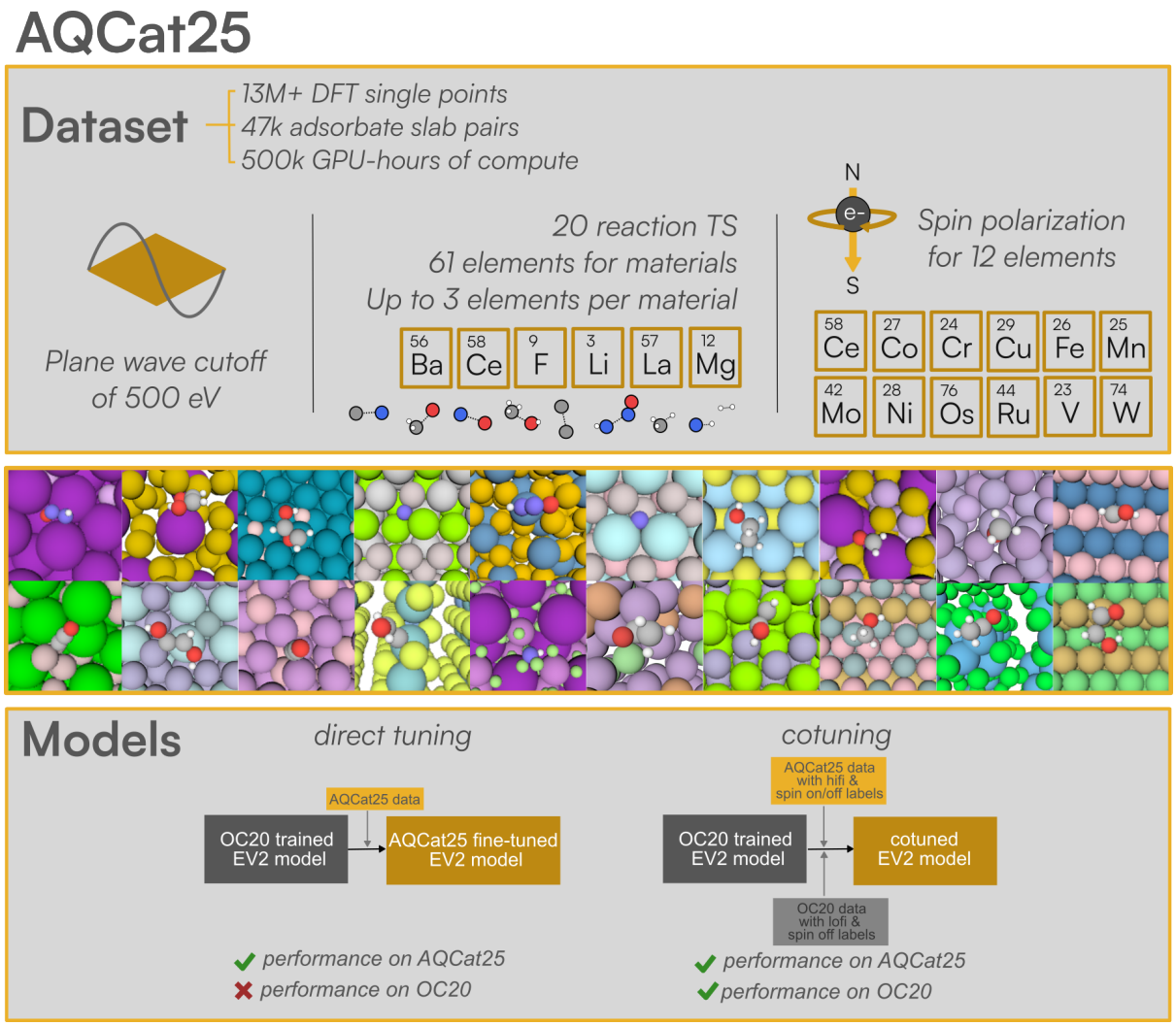

Figure 1: A summary of the AQCat25 dataset and models. Spin polarization has been included for 12 important elements. A plane wave cutoff of 500 eV is used. Six new elements are added, when compared to the Open Catalyst 2020 dataset (OC20 dataset)[26] as well as 20 transition state adsorbates. Models were jointly tuned/trained from scratch on OC20 and AQCat25 to achieve good performance on their validation and test splits, respectively.

mate electronic structure properties at near-quantum accuracy for a small fraction of the computational cost[27–32]. These models learn the interactions required to predict the potential energy landscape of atomistic systems from large-scale public databases of DFT calculations[33–36]. State-of-the-art models for heterogeneous catalysis are made possible by Meta FAIR's Open Catalyst 20, 22, and 25 datasets[26,37,38], which collectively consist of nearly 300 million single-point DFT calculations of adsorbate-surface interactions relevant for reactions of carbon, hydrogen, oxygen, and nitrogen over a diverse catalyst space spanning most of the periodic table. The introduction of machine learning methods into computational catalysis workflows has begun to enable studies of reaction network complexity[39–42] and catalyst structural dynamics[43–45] that were completely inaccessible just 10 years ago.

Although the sheer scale of these datasets has necessitated some compromises in the fidelity of the underlying training data, convergence of the resulting adsorption energies benefits from error cancellation and has been validated with respect to most DFT settings, with plane-wave cutoff and smearing width requiring some improvements to accurately capture total energies of non-metals[46]. One of the most significant gaps in existing large-scale heterogeneous catalysis datasets is the treatment of magnetism. Since spin-polarized DFT calculations are considerably more expensive than spin-unpolarized calculations, spin is often omitted in the interest of scale and throughput[47]. The consequence of this choice is that the resulting models are not suitable for many industrially relevant catalytic processes such as ammonia synthesis[6] and Fischer-Tropsch synthesis[10], which rely on on earth-abundant first-row transi-

tion metals (e.g., iron, cobalt, and nickel) that exhibit especially strong spin polarization effects on binding energies and activation barriers[48–51]. As the field moves towards discovering new, low-cost, and sustainable catalytic materials to replace precious metals, the importance of treating magnetic effects when training foundational MLIPs becomes increasingly paramount.

Alongside progress in data generation, developments in machine learning architectures, often based on equivariant graph neural networks, have improved performance on atomistic tasks. For heterogeneous catalysis, models like eSEN[52], EquiformerV2[53] and EScAIP[54] have achieved state-of-the-art results. A significant leap towards broader universality is the Universal Model for Atoms (UMA)[55], trained on ∼500 million structures across diverse chemical domains (molecules, materials, catalysts). UMA modifies the eSEN architecture with additive embeddings for global context (charge, spin, DFT task) and a Mixture of Linear Experts (MoLE) routed by this context plus element composition. UMA's core design goal is to accurately reproduce the original physics of each training task (e.g., spin-unpolarized OC20), operating as a multi-task surrogate rather than a model explicitly designed to perform cross-fidelity corrections between different levels of theory. A distinct advantage of total energy models like UMA is their ability to better capture, among other effects, restructuring of bare catalyst slabs[46].

Other approaches have focused on integrating low- and high-fidelity data for related tasks, such as by augmenting node features with a fidelity one-hot encoding and applying both common and fidelity-specific weights in modified linear layers[56] or utilizing a model's intrinsic global state feature to embed fidelity context during message passing[36]. For example, Ko and Ong demonstrated that a single multi-fidelity model trained with a small fraction of high-fidelity data could achieve similar accuracy to a single-fidelity model requiring eight times the amount of costly high-fidelity training data[36]. Alternatively, other methods utilize architectural separation, such as dynamically using separate prediction heads branching from a shared backbone for each fidelity level[57,58].

While recent work has produced model architectures that can incorporate spin[47,59–64], their predictive power is limited by the absence of high-quality, spin-polarized training data for heterogeneous catalysis. Given the success of methods for adapting models to new data[37,65–69], alongside advancements in training universal models from diverse datasets, we see a clear opportunity to develop improved foundational models specifically for spin-polarized, high-fidelity catalytic systems.

Here, we present the AQCat25 dataset and baseline AQCat25-EV2 models (Figure 1), which improve upon the performance of EquiformerV2-31M and EquiformerV2-153M adsorption energy MLIPs for heterogeneous catalysis in three key ways: increasing the fidelity of the reference DFT calculations, explicitly incorporating spin polarization for magnetic elements, and introducing new elements to the model domain that are underrepresented in existing datasets. Through this work, we demonstrate data-efficient methodologies for building multi-fidelity MLIPs that span distinct physical regimes (such as spin polarization) to new domains of chemistry while ensuring that the models maintain accuracy and generalizability across a wide range of catalysts and reactions. The dataset, models, and code are available publicly to support further developments by the academic community.

Methods

Density functional theory

DFT calculations were performed using the Vienna Ab Initio Simulation Package (VASP)[70–73]. A plane-wave cutoff energy of 500 eV was applied, and Gaussian smearing with a width of 0.1 eV was used. The revised Perdew-Burke-Ernzerhof (RPBE)[74,75] functional was chosen for its performance on heterogeneous catalyst systems, and the system's geometry was optimized using the conjugate gradient algorithm. For systems containing Ce, Co, Cr, Cu, Fe, Mn, Mo, Ni, Os, Ru, V, or W, spin polarization was enabled to account for magnetic effects. For a full list of VASP parameters, please see the Supplementary Information.

Although these settings represent a significant increase in fidelity over previous large-scale datasets and are considered nominal for catalysis research, we acknowledge that even higher-fidelity calculations are possible. However, any increase in per-calculation fidelity must be weighed against the loss of dataset diversity for a fixed computational budget. For foundational MLIPs that must generalize across a vast chemical space, this trade-off is critical.

Bulk selection

The AQCat25 bulk materials database was constructed by first updating and then expanding the OC20 dataset. Initially, the Materials Project (MP)[76,77] database was queried for all structures containing only

elements present in the OC20 dataset, subject to the constraints of a maximum of three unique elements per material and an energy above the convex hull (E_{hull}) of 0.1 eV/atom or less. Structures from the original OC20 dataset not found in this query were retained only if they were structurally unique. To expand the chemical space, a second query was performed using the same stability and size constraints but including six additional elements: Li, Ba, La, Ce, Mg, and F. The resulting set of new materials was then subsampled to ensure balanced representation. Up to 500 structures were randomly selected for each group containing a single new element, and up to 20 structures for each group containing a combination of new elements. The dataset was assembled by combining the updated OC20 dataset materials, the preserved unique structures, and the sampled new materials. Finally, the dataset was filtered to only contain bulks with up to 30 atoms per unit cell. Data splits were then assigned by attempting to preserve the original designations for all OC20 dataset materials and distributing new materials based on chemical composition to maintain consistency with the established OC20 dataset splitting methodology.

Adsorbate-slab selection

The number of single points and systems that make up the splits and data types included in the AQCat25 dataset is shown in Table 1. Here, a system is defined as a unique adsorbate-slab pair for that subsplit.

Dataset splits

Primary split	Secondary Split	N systems	N single points
In Domain	Relaxations	24,624	6,959 k
	Rattled	8,189	947 k
	Transition states	2,854	676 k
	Molecular Dynamics	2,098	249 k
	OC20 fidelity, spin on relaxations	4,831	863 k
Validation	OOD adsorbate relaxations	1,913	577 k
	OOD material relaxations	991	318 k
	OOD both relaxations	994	295 k
Test	OOD adsorbate relaxations	992	347 k
	OOD material relaxations	994	316 k
	OOD both relaxations	988	356 k
Slabs	ID	19,273	1,282 k
	ID OC20 fidelity, spin on	4,868	273 k
	OOD validation	497	29 k
	OOD test	498	36 k
Totals		47 k	13.5 M

Table 1: The number of systems and single points across data splits. The total system count reflects the number of unique adsorbate-slab combinations.

The dataset is structured into three primary splits: in-domain (ID), out-of-domain (OOD) validation, and OOD test. Each OOD split is further categorized by the type of novelty introduced, either in the adsorbate or in the material slab. This strategy is designed to evaluate the model's ability to generalize to novel systems it has not seen during training, in the same manner as the OC20 dataset[26]. The ID split contains configurations where both the adsorbate and the material slab are present in the training set. The test and validation ID splits serve as a baseline for the model's performance on familiar data and are sampled from the same distribution of the training set. Both OOD splits are designed to test the ability of machine learning models to generalize. The OOD validation set is used for hyperparameter tuning, while the OOD test set provides a final, unbiased evaluation of the model's performance on unseen data.

The following categories are included in both OOD splits: (1) OOD adsorbate, (2) OOD material, (3) OOD both. For OOD adsorbate, the material slab is ID, but the adsorbate is new and does not appear in the training data. The test OOD adsorbates also do not appear in the validation split and the validation OOD adsorbates also do not appear in the test split. For OOD material, the adsorbate is ID, but the bulk lattice structure (not necessarily its composition) used to construct the slab is new and does not appear in the training set. For OOD both, the adsorbate and the material slab are new and do not appear in the training set. The same segregation for validation and test also applies.

Sampling diverse states

To ensure models trained with this dataset have a robust understanding of different structural and energetic states, we employed several calculation types for data generation. The dataset samples both high-energy, off-equilibrium states and low-energy, near-equilibrium states. To sample low-energy states we performed adsorbate-slab structure relaxations. Relaxation calculations involve iteratively optimizing atomic positions to find a local energy minimum. A DFT call is made to determine forces, and atoms are moved along these force vectors. This process is repeated until the maximum force on any atom is less than 0.03 eV/Å or a maximum of 800 steps are reached. These trajectories sample a range of configurations from high to low forces. All OOD validation and test set calculations are relaxations.

To sample high-energy states we took three approaches: (1) running molecular dynamics (MD) calculations, (2) placing transition state (TS) systems, and (3) rattling atoms. To sample high-energy states accessible at elevated temperatures, we performed MD calculations. Starting from a relaxed structure, we ran 80 steps of MD at 900 K. To provide the model with examples of highly distorted configurations relevant to chemical reactions, we extracted transition state structures from the OC20NEB dataset[78]. These adsorbates were placed on new surfaces, followed by a short 5-step relaxation. This process generates data with high forces and energies, supporting the training of models that can handle reactive states. To further augment high-force data, we generated rattled configurations by randomly perturbing atomic positions. Two methods were used: (1) rattling all atoms and (2) rattling only adsorbate atoms, with displacements sampled from a normal distribution (σ = 0.05, 0.1, 0.15, or 0.2 Å). Some rattled systems underwent a single DFT calculation, while others had a short 5-step relaxation. Systems whose max absolute force or absolute adsorption energy exceeds 50 eV/Å and 10 eV were excluded from training and evaluation.

Additional data

To explore the opportunity to train models with less costly DFT data, we considered data that include spin polarization but with settings that otherwise match the OC20 dataset[26]. Notable differences between this data and the rest of the AQCat25 dataset are that we used a plane wave cutoff of 350 eV and Methfessel-Paxton smearing with a width of 0.2 eV. This data aids in understanding how the model handles the distinct physical regimes defined by fidelity and spin polarization. This dataset complements the existing high-fidelity spin-on/off (AQCat25) and spin-off OC20 data by filling a missing quadrant. Adsorption energies were computed using high-fidelity adsorbate references for all spin on systems. We found this to have little impact on the final target energies from preliminary tests.

We also wanted to form an understanding of model performance on the task of finding the minimum adsorption energy for an adsorbate-slab combination. To do this we constructed a small dense dataset, similar to the OC20dense dataset presented by Lan et al.[79]. For this dataset we selected 109 adsorbate-slab pairs. Adsorbates were selected to be disassociation reactants from the OC20NEB dataset[78]. Slabs were selected randomly, but we selected the materials they were cut from more strategically. We included five unary materials, five binary non-metal materials, 46 binary intermetallics, 30 ternary intermetallics, and 23 ternary non-metals. Within these categories, the bulks were also randomly selected from the bulk database. For each adsorbate-slab pair, we performed 50 placements using the random site with heuristic placement mode in `fairchem`[80]. These placements were relaxed with the same DFT settings as the broader AQCat25 dataset. The relaxed states were filtered using the same algorithms presented by Lan et al.[79] to find desorption, dissociation, intercalation, and significant surface change.

System enumeration

All systems were prepared using the publicly available `fairchem` package[80]. Slab enumeration was performed using the underlying `pymatgen`[81,82] algorithm. Adsorbate placement was performed heuristically at random sites. Rattled systems were perturbed after adsorbate placement using the rattle functionality in `ASE`[83,84]. For TS systems, they were placed as normal adsorbed intermediates would be by preparing a new adsorbate database with TS entries.

Machine Learning Experiments

A challenge in this work is training a single MLIP that can accurately predict energies and forces across a dataset containing multiple DFT settings. The combined training data spans four distinct physical regimes:

high-fidelity spin-on, high-fidelity spin-off, low-fidelity spin-on, and the original low-fidelity spin-off. We therefore explored methods to introduce this DFT context, namely spin treatment and calculation fidelity, directly to the model. Inspired by the effectiveness of Feature-wise Linear Modulation (FiLM)[85] and similar successful techniques[55] for multi-task learning, the baseline models presented in this paper focus on adapting the EquiformerV2 (EV2) architecture[53] using this approach. We did not focus on fine-tuning UMA[55] models because of licensing issues, but we expect those to have even better performance than the models presented here. The Feature-wise Linear Modulation (FiLM) technique[85] provides an expressive conditioning mechanism. Rather than adding context via feature concatenation, FiLM applies a learned, feature-wise affine transformation ($\gamma F + \beta$) that can scale, shift, or suppress activations. This strategy of deep, additive modulation is similar in principle to the additive embedding mechanism successfully employed by the UMA family of models[55]. To evaluate performance on AQCat25 while retaining knowledge from OC20, we used the EV2[53] model architecture with three variants and three training protocols. The variants were: (i) **EV2** (unmodified), (ii) **EV2-inFiLM**, which applies additive FiLM[85] shifts to the scalar ($l{=}0$) channels at the input, and (iii) **EV2-in+midFiLM**, which applies the same modulation at the input and after each equivariant block. The protocols were: direct fine-tuning of OC20-pretrained checkpoints, cotuning those checkpoints with OC20 replay, and cotraining from scratch on mixed data from both AQCat25 and OC20.

It is important to note how baseline performance was assessed in this context: evaluations of pretrained models on AQCat25 used the provided structures directly, without re-optimizing lattice constants using OC20 DFT settings. These metrics represent the performance inherited for subsequent tuning rather than the inherent capability of the OC20 model on these specific materials had geometries been fully relaxed with consistent settings. Similarly, all evaluations performed on the OC20 validation subset utilized structures with OC20-optimized lattice constants.

Architecture

EV2 is an E(3)-equivariant transformer over atomic graphs: atoms form nodes, edges use pairwise distances and spherical harmonics, attention layers are equivariant, and feed-forward layers use S^2 activations[53,86–88]. Energies are predicted by a scalar head, and forces are predicted directly via a vector head. Architectural hyperparameters are listed in Table 7.

FiLM conditioning supplies the network with compact context about the DFT settings of each structure. Two binary indicators encode spin treatment and fidelity (`spin_on` and `low_fi`). Each indicator is embedded; the embeddings are concatenated and passed through a small multilayer perceptron (MLP) to produce a modulation vector β. We apply β additively to the scalar channels broadcast across nodes (and per block for EV2-in+midFiLM). Preliminary tests with multiplicative factors γ did not improve validation metrics measurably, so we retained the additive shift only for simplicity. Figure 2 diagrams the module and its insertion points.

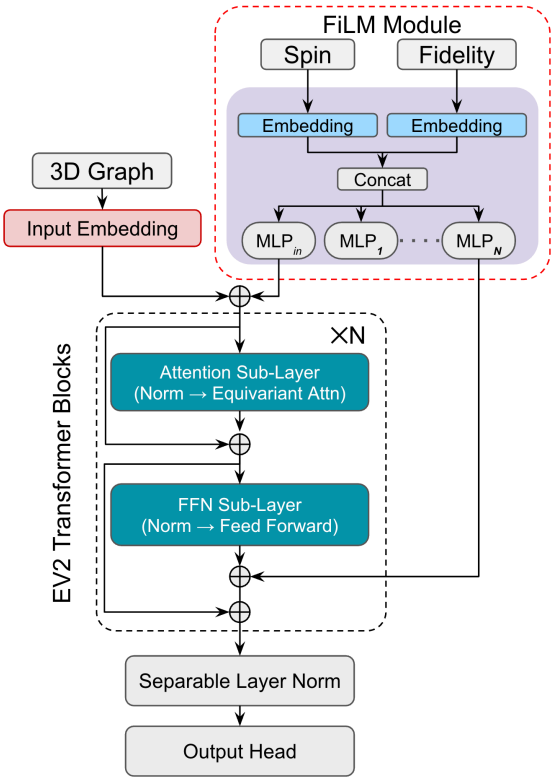

Figure 2: FiLM module: binary context (spin, fidelity) → embeddings → MLP → β; β additively modulates scalar channels at the input (inFiLM) and optionally mid-block (in+midFiLM).

Training and adaptation protocols

Direct fine-tuning Starting from public EV2 OC20 All+MD checkpoints (31M and 153M parameters), we fine-tuned on AQCat25 directly. These initial experiments tested the model's adaptation to the AQCat25 domain under distribution shift.

Cotuning with OC20 replay We fine-tuned the OC20 checkpoints on a composite stream consisting of AQCat25 high-fidelity and, when specified, a

small AQCat25 low-fidelity spin-on stream, mixed with OC20 spin-off data at 0M/2M/20M scales.

Cotraining from scratch We trained EV2, EV2-inFiLM, and EV2-in+midFiLM from random initialization on the same composite streams used for cotuning.

Hyperparameters, controls, and compute

Optimization and architectural settings are summarized in Tables 7 and 8. For each training run, 8×H100 NVIDIA GPUs were used. Unless noted, the force term dominated the objective; the default loss ratio was $\lambda_E : \lambda_F = 4 : 100$. To reduce the computational cost for the extensive model adaptation experiments, the AQCat25 dataset component was subsampled. Models were trained in single precision for a consistent comparison across the numerous experimental conditions and ablations. For tuning experiments, we tested stronger regularization by increasing weight decay and by lowering the learning rate; both choices produced early plateaus and higher validation errors on AQCat25 relative to fully thawed baselines. Except for models trained from scratch solely on AQCat25, energy and force targets were normalized using mean/standard deviation values from the OC20 distribution, as this yielded slightly improved performance in preliminary tests. We also tried incremental thawing schedules that kept the backbone frozen while adapting only input embeddings (to accommodate new elements), followed by gradual unfreezing. These schedules underperformed fully thawed tuning on AQCat25.

Additionally, we explored alternative conditioning mechanisms and found that a simpler baseline involving direct concatenation of context embeddings performed competitively with FiLM when cotuning with limited (2M) OC20 data replay. We further experimented with more complex architectural modifications aimed at adapting the pretrained weights, including adding separate prediction heads routed by the conditioning flags and incorporating lightweight adapter modules within the transformer blocks. However, these approaches did not yield significant performance enhancements over the FiLM-based conditioning and fully thawed training strategies presented here. Finally, we do not claim hyperparameter or schedule optimality. Alternative warmup/decay, replay curricula, batch-composition policies, weight decay, EMA, or gradient clipping may yield further gains. Our goal here is a consistent and reproducible setup that enables clear comparison across regimes and architectures for the baseline models being presented.

Results and Discussion

Dataset composition

Some summary statistics about the AQCat25 dataset and how it compares to the OC20 dataset are shown in Figure 3. As can be seen in Figure 3b, there is a significantly higher proportion of non-metal systems and lower proportion of metal systems in the AQCat25 dataset compared to the OC20 dataset. The proportion of the other two categories, however, (non-metal & metalloid and metalloid) are roughly equal. Because non-metal systems typically have poorer performance compared to intermetallics[89], we will look at key model performance metrics split over these material categories. The adsorption energy and maximum force distributions (Figure 3c-d) reveal that the AQCat25 dataset is biased towards higher force, higher energy systems when compared to OC20. This can be explained by the more aggressive approach taken when sampling high-force systems. Here, we used larger standard deviations to sample rattled configurations and also included the high energy transition state like systems.

Optimizing Data Generation Strategies for Fine-tuning

We wanted to explore opportunities to improve our data generation strategy to maximize model performance while minimizing the cost associated with dataset generation and model fine-tuning. To do this, we looked at the change in model performance as a function of two variables: number of DFT single points seen per system and number of slabs. For heterogeneous catalyst systems, there are two types of diversity the models must generalize across: (1) the adsorbates and (2) the material surfaces the adsorbates are adsorbed to. The latter is much more complex. We initially adopted a scheme of performing four adsorbate-slab relaxations per slab to reduce the number of slab relaxations that needed to be performed as this data is not directly used in model training for referenced energy models. This turned out to be a suboptimal choice, however, because material diversity is important to tackle. For the future, we

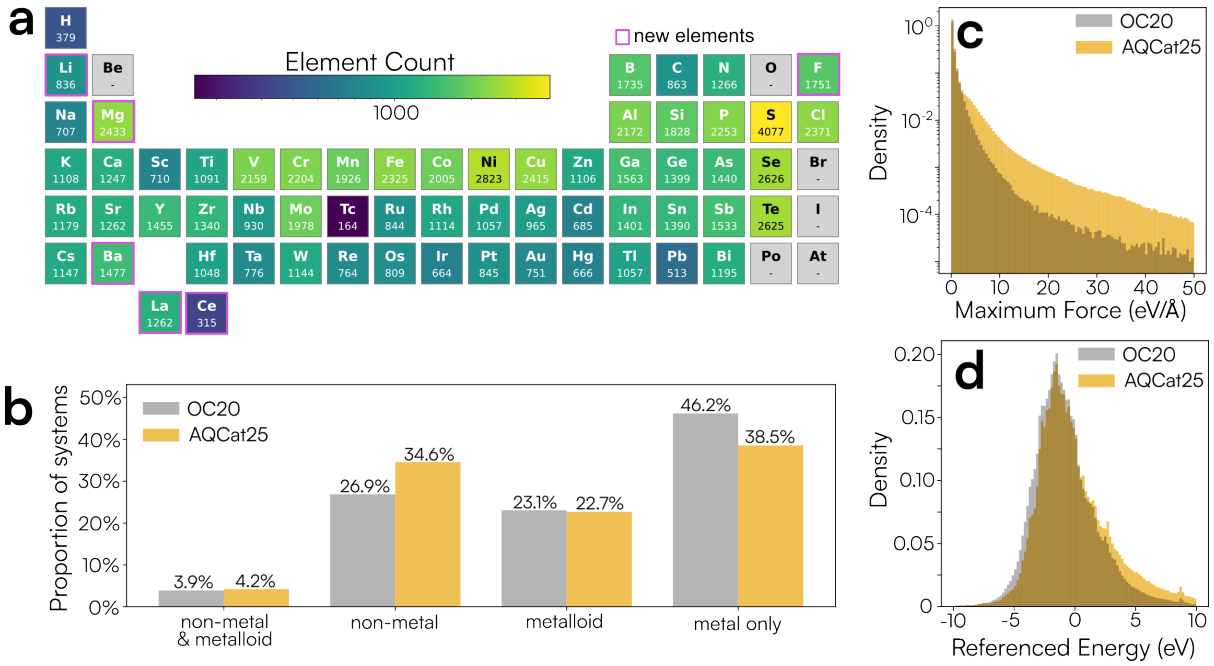

Figure 3: Summary statistics on the AQCat25 dataset and some comparisons to the OC20 dataset[26]. (a) Element counts showing the frequency with which each element appears in the dataset. (b) The proportion of systems that fit into four material-type categories for the entire training splits OC20 dataset and AQCat25. (c & d) The distribution of adsorption energies and maximum forces across the training split of AQCat25 and the 2M training subsplit of OC20.

would choose to perform one adsorbate-slab calculation per slab. In Figure 4, we show the relationship between model force performance, number of slabs seen, and number of DFT single points seen per system. Models were directly tuned starting from the publicly available 31M parameter EV2 model (All + MD)[53].

The amount of data used to train the models directly impacts the cost to train and to iterate between architectures and ablations. Gasteiger and colleagues have shown the OC20-2M subset to be representative of the full OC20 dataset, primarily because it preserved the underlying chemical diversity[90]. Other approaches use more complex stratified sampling (based on feature-space clustering) to ensure that diverse, high-energy, and uncommon configurations are explicitly captured to improve model robustness[91]. Given the precalculated training data, we explored the opportunity to reduce model training costs by sampling the frames along adsorbate-slab relaxation trajectories. Sampling was performed using force-stratified selections of the trajectory to obtain a representative distribution of systems. For consistency, models were trained for a nearly constant number of total gradient updates, approximately equivalent to the number of steps in one epoch of the largest split. Further, the data ablation for sampling frames from trajectories (Figure 4a) was designed to probe redundancy in highly autocorrelated data and thus only applied to the relaxations and MD data; the rattled and transition state data were included entirely for each model. To enable a cleaner evaluation of the data cost-benefit trade-off for tuning, without confounding the results with the model's ability to learn new, unseen elements, we restricted the subsampling experiments to AQCat25 systems with elements already seen by the 31M pretrained model. Figure 4a reports the validation MAE from the final training checkpoint, which highlights the risk of overfitting. Conversely, Figure 4b plots the MAE from the best-performing model during training for each slab count (averaged over all k values) against the data generation cost. As can be seen in Figure 4a, for small numbers of slabs, sampling a subset of frames rather than using the full trajectory actually leads to better force MAE values. This is likely due to a high propensity for overfitting, which is shown in Figure 10 and remedied by sampling. At larger slab numbers, the differences are minor but the cost to train will be lower if sampling is performed. For energy MAE there is not a substantial trend with changing sample size (see Figure 11). Therefore, sampling frames is a useful cost-saving strategy. We adopted a subsampling approach for the model adaptation experiments presented in subsequent sections. However, for those experiments, we employed random sampling per trajectory rather than the force-stratified approach. We found this yielded slightly improved performance, likely due to increasing the representation of low-force,

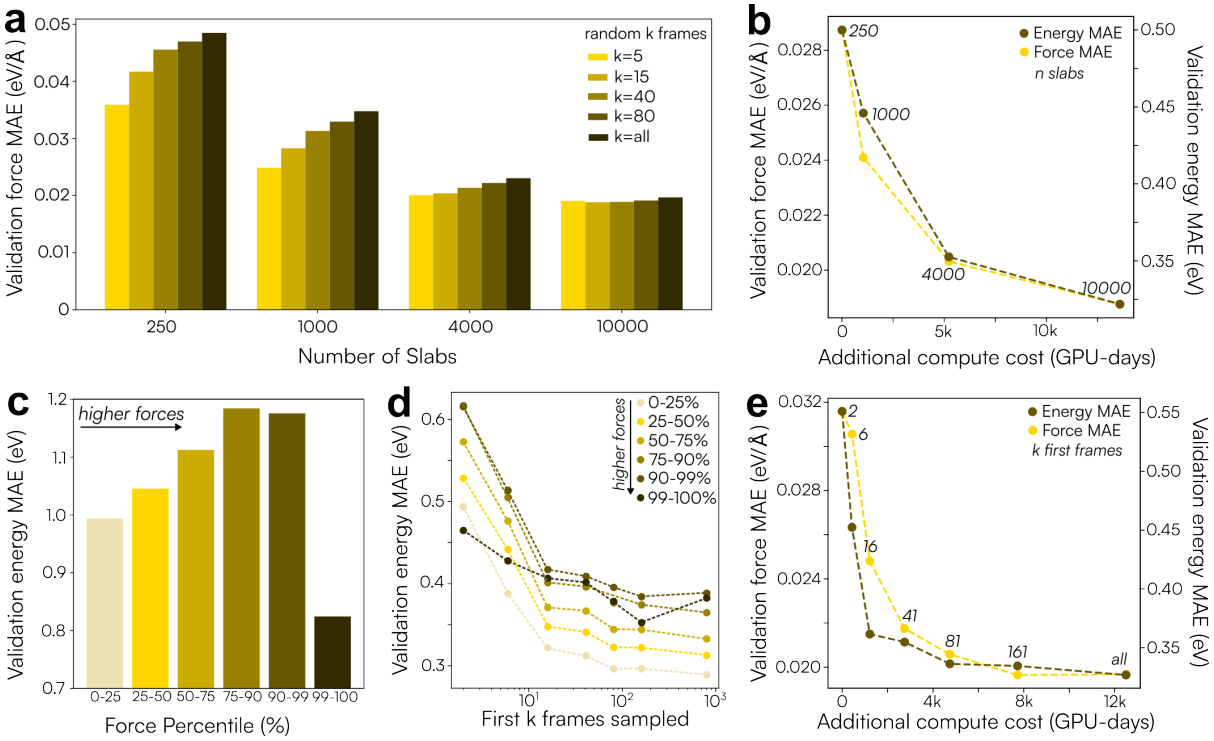

Figure 4: An exploration of opportunity for cost reduction in model training (a) and data generation (b-e) by directly fine-tuning 31M parameter Equiformer v2 models. (a) The validation force mean absolute error (MAE) obtained when training models on *random* subsamples of the data for four dataset sizes. (b) The incremental compute cost for increasing dataset size. (c) The pretrained 31M parameter Equiformer v2 model energy MAE on force percentile segregated data from the AQCat25 validation dataset. (d) Evolution in force percentile segregated energy MAE for fine-tuned models trained on variable *first k* frames from the dataset. (e) The trade-off between cost to generate training data and model performance when considering terminating relaxations after *first k* steps.

near-equilibrium frames that are critical for downstream adsorption energy tasks.

Unsurprisingly, having more unique slabs in the dataset improves performance. However, there is a cost trade-off to be made, which is explored in Figure 4b. Comparing 250 to 1,000 and 1,000 to 4,000 there is a large improvement in the metrics. Going from 4,000 to 10,000 slabs, however, we are beginning to enter the domain of having diminishing returns on our computational investment, indicating that the number of slabs we calculated in this dataset was a reasonable choice. Nonetheless, this experiment primarily assessed convergence with respect to the number of unique slabs, not the total number of unique adsorbate-catalyst combinations, which warrants further investigation. Here, the additional compute cost is referenced to the 250-slab dataset. This value serves as a proxy for the total computational investment, which is expected to correlate with the true data generation cost and illustrates the trend of diminishing returns.

The apparent redundancy revealed by randomly sampling offers a potential opportunity: what if instead of optimizing full relaxation trajectories we instead only calculate the first k points? This could greatly reduce the compute cost to generate the data, but it introduces a potential new problem. It biases the relaxation data towards higher force states which could cause models trained on the data to have poor performance on low force systems. To investigate this, we divided the validation set into force percentiles and examined changes in performance on the different percentiles. As a baseline, we first assessed the pretrained model performance in Figure 4c. Performance decreases with increasing force percentile with the exception of the highest force percentile considered, which has better performance than even the lowest force percentile. This is likely a reflection of the underlying OC20 dataset that contains, MD, rattled systems, and relaxations. MD and rattled data have high forces, while relaxations contain many frames in the low force regime. Figure 4d shows the evolution of model performance on the force percentile segmented validation split with an increasing number of frames sampled. Please note that here the data ablation also only applied to relaxations; all models were trained on the TS-like and rattled data, but none included the MD data. Performance overall improves with the number of frames sampled but it does not

occur in a way that disproportionally affects specific force segments from 0-99%. The one exception to this is the highest force percentile which modestly improves with increasing k. This is because all models used were trained using the very high force data (rattled and TS). Performance in this percentile is most influenced by high force data.

Using the first k relaxation frames presents an interesting trade-off between compute cost and model performance which is captured in Figure 4e. By only calculating between 40 and 80 frames instead of up to 800, we can achieve a Pareto optimum in model performance and compute cost for dataset generation. This would be our recommendation for future data generation campaigns. This exploration also revealed the advantage of training total energy models when designing a dataset to fine-tune models with cost in mind. If just 41 adsorbate-slab frames are computed, on average 75% of the compute would be spent relaxing the slab completely. At 81 frames, this cost decreases to 63%, but it is still substantial. For total energy models, this cost is not necessary because relaxed slab energies are not needed to train. We also recommend designing datasets to train total energy models for future campaigns.

Model Adaptation Strategies

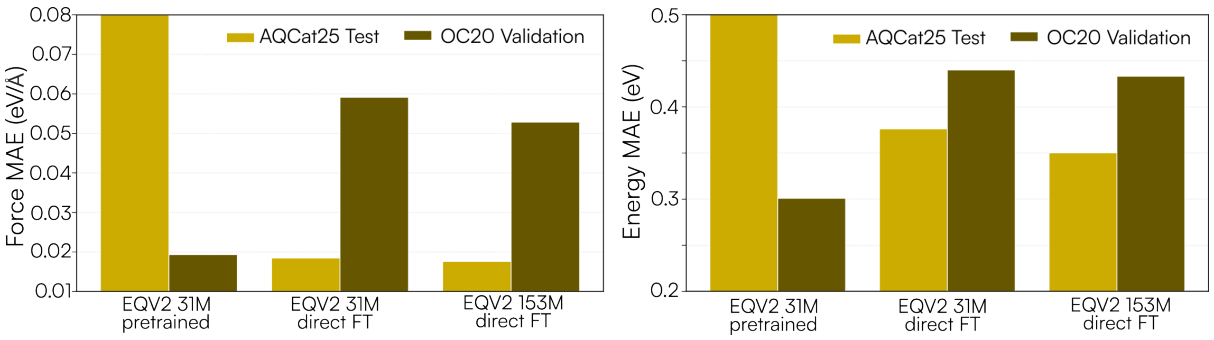

Figure 5: Test force and energy MAE on AQCat25 test and a system-stratified subsample of the OC20 validation OOD both split for three different models: a pretrained (on OC20 only) 31M parameter EV2 model, a 31M parameter EV2 directly fine-tuned (FT) on AQCat25, and a 153M parameter EV2 directly fine-tuned on AQCat25

Although total energy models offer a clear path forward for more efficient data generation, this study focused on the adsorption energy target. Though arguably a more challenging learning task, as the model must implicitly account for the bare slab reference energy and any restructuring, adsorption energy may offer a significant convenience in established catalysis workflows. It also provides a well-defined target that isolates the adsorbate-surface interaction, which is ideal for developing a mixed fidelity/mixed physics adaptation protocol. Moreover, our overall objective is to create an MLIP that is broadly applicable, so we also wanted to understand model performance on the OC20 validation set. To assess the performance drift on the original OC20 task during these and subsequent experiments, we utilized a system-stratified subsample of the OC20 Val OOD Both split. This subset, which was sized to be comparable to individual AQCat25 validation splits (∼300k frames), is a computationally efficient metric for relative comparisons between models. The resulting MAE values, however, may not reflect absolute performance on the full OC20 distributions. An evaluation of performance for a pretrained 31M parameter EV2 model and two directly fine-tuned (FT) EV2 models with 31M and 153M parameters on AQCat25 test and the subsampled OC20 validation split are shown in Figure 5. We find that direct fine-tuning delivers reasonable AQCat25 errors, but deviation from the OC20 baseline on its validation split is significant. As anticipated, increasing model capacity from 31M to 153M parameters generally improves energy metrics on the AQCat25 test set. This is also true for increasing the energy loss weight (λ_E) (see Table 6). The 153M model with $\lambda_E = 100$ yields the best energy MAE metrics in these direct fine-tuning experiments (Table 6). However, the gains achieved by the larger 153M model may not justify its increased computational cost for practical applications, and this larger model still suffers from a significant performance drift for the original OC20 task.

To mitigate this drift, we explored opportunities to cotune and cotrain models using both subsamples of the AQCat25 dataset and the OC20 dataset. The models evaluated in this context, including those jointly trained with no additional OC20 data, incorporate the low-fidelity, spin-on data alongside the high-fidelity AQCat25 data. As seen in Figure 6b and d, we observe that 0M models (models trained

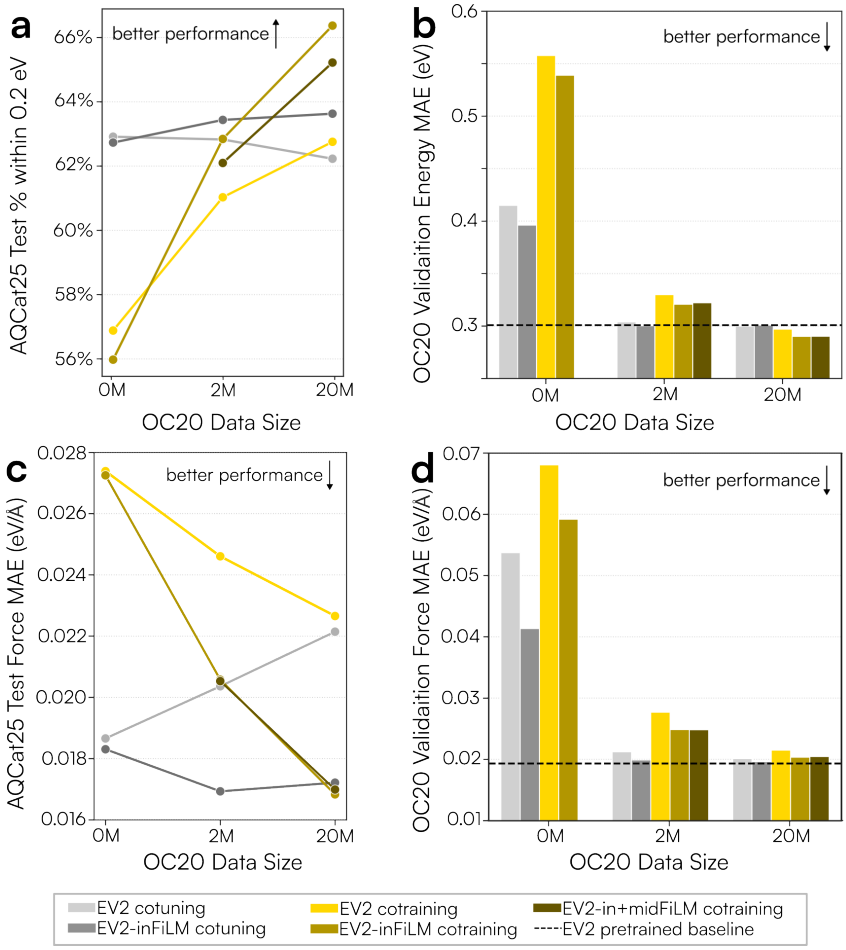

Figure 6: Model performance under cotuning and cotraining on AQCat25 and OC20. Part (a) shows the percentage of AQCat25 test set energies within 0.2 eV of the DFT value, while part (c) shows the force MAE on the AQCat25 test set, both as a function of the amount of OC20 data seen during training. Parts (b) and (d) explore the energy and force MAE trends on the OC20 validation set.

without OC20 data) exhibit a substantial deviation from the baseline OC20 performance (dashed black line). Unsurprisingly, the exclusion of the spin-off OC20 dataset leads to poor performance on OC20 validation relative to baseline metrics, even with the inclusion of the small lowfi spin-on set. Therefore, to produce a model that performs well across all domains (high/low fidelity, spin on/off) within a practical model size, we investigate cotuning and cotraining (from scratch) strategies that incorporate two amounts of the original OC20 data. Figure 6 summarizes the effect of including this OC20 data under two training regimes and three architecture variants.

Adding OC20 data consistently reduced deviation from the OC20 baseline on the examined validation split. For both cotuning and cotraining from scratch, the energy and force MAE trend toward the baseline as the amount of OC20 data increases for both energies and forces (Fig. 6b and d). We do not see this exact trend on the AQCat25 test split (Figure 6a and c). In this case for energies we are showing the percent of frames that have an absolute energy error less than or equal to 0.2 eV. For this metric, a perfect model would have 100%. This was done because we observed that the energy values had strong outliers. One group of systems contributing to this phenomenon are those where the slab is organic (entirely composed of non-metals). This approach as an alternative to MAE, is an unbiased way to ensure strong outliers do not skew the results. We have included some additional Figures in the Supplementary Information to explore this metric further with different cutoffs (0.1, 0.3, 0.4, 0.6, and 0.8 eV instead of 0.2 eV) and using the MAE for the energies instead on the test and validation splits. It seems as though the results are sensitive to this metric, so we will only make broad conclusions. The percentages of errors within the threshold for AQCat25 energies are largely unchanged when increasing data for cotuning, whereas with cotraining they increase (performance improves). For forces, there is a drastic increase in performance with more OC20 data for cotraining. For cotuning with FiLM there is a modest

improvement in forces when including OC20 data, but a substantial degradation when cotuning without FiLM. An economic approach when considering cost to train and performance on both AQCat25 test and OC20 validation is achieved when cotuning with 2M OC20 examples. Cotraining from scratch with FiLM improves performance for systems that have higher errors though, so a tradeoff exists.

These patterns follow from standard behavior under distribution shift and multi-domain supervision. OC20 is broader and uses different DFT settings than AQCat25. Fine-tuning only on AQCat25 moves the parameters toward that narrower distribution and forgets OC20-specific features. Adding replay during cotuning without FiLM counteracts forgetting but also pulls the solution toward OC20 conventions, which explains the performance degradation on AQCat25 forces in Figure 6c. Introducing FiLM provides a framework to distinguish these distributions, which rectifies the decrease in the force metric. Starting from scratch changes the optimization path. The performance trends are strongly dependent on the OC20 data size used. Unsuprisingly, at 0M, jointly tuning clearly outperforms jointly training from scratch, but as OC20 data is added, their performance becomes sensitive to the metric. While tuning holds a slight advantage at a very strict 0.1 eV low-error threshold, the models cotrained from scratch show an advantage at higher cutoffs (e.g., 0.6-0.8 eV), indicating they are more effective at capturing outliers (see Figures 16a-f. FiLM makes the domain information explicit. Conditioning on spin and fidelity yields feature rescaling that reduces gradient interference between magnetic vs. non-magnetic and high- vs. low-fidelity cases.

Exploring robustness and generalization

We also explored the robustness and generalizability of models to form a more complete assessment of model usability. To do this, we constructed an additional validation set aimed at assessing the ability of the models to identify the global minimum energy for a given adsorbate-slab combination in line with the approach presented by Lan et al[79]. We further explored differences in model performance when segmenting the data by interesting splits, namely the material type (metal-only, non-metal, metalloid, and metalloid+non-metal), whether spin was on or off, and whether the elements in a material were all included in OC20 or not.

Global minimum adsorption energy

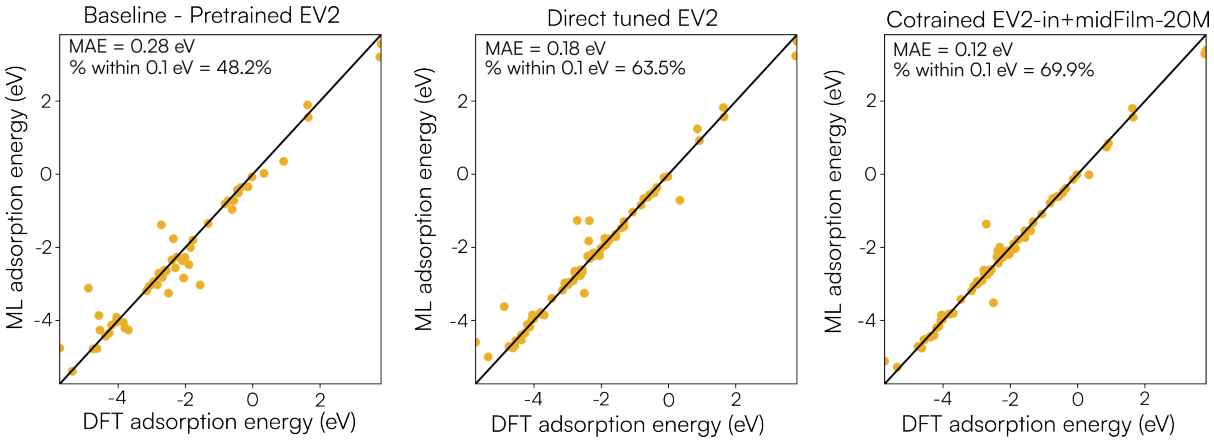

Figure 7: Parity plots between the DFT minimum adsorption energy and ML adsorption energy for a baseline pretrained 31M parameter EV2 model (left), the result of directly tuning that model on the AQCat25 dataset (center), and cotraining a 31M parameter EV2-in+midFilm model from scratch on both 20M examples from OC20 and the AQCat25 dataset (right).

The ultimate use of the MLIPs trained here will be for practical catalyst discovery where an important figure of merit is the global minimum adsorption energy. We explored this using the dense DFT validation set with 50 relaxations each for 109 adsorbate-slab combinations. We performed ML relaxation inference starting from the same initial configurations as DFT. The relaxed states were filtered using the same algorithms presented by Lan et al.[79] to find desorption, dissociation, intercalation, and significant surface change. Figure 7 compares the performance of three 31M parameter models on this task, using only the ML-predicted energies without DFT single-points on the ML-relaxed structures. On the left is the pretrained 31M EV2 model (trained on OC20 All+MD), taken from the publicly available fairchem checkpoint.

For this model, inference on systems containing new elements were omitted since performance would be poor. In the center is a directly fine-tuned EV2 model. On the right is the EV2-in+midFiLM model, which was cotrained using 20M OC20 examples and the AQCat25 dataset. As a point of comparison, when the OC20dense[79] dataset was released, the EV2 model was not available. The best performing model was eSCN-MD-Large and on this task it had a 56.5% success rate with an energy MAE of 0.17 eV[79]. Here, success rate is defined as the percent of systems where the minimum adsorption energy found by ML is within 0.1 eV of the DFT value. ML success metrics alone were included in a later release as 60.8% for an EV2 model of unspecified size, 68.4% for the UMA-S model, 71.1% for the UMA-M model, and 74.4% for the UMA-L[55]. The success metric and MAEs have been annotated on the plots. We see the trend we would expect to see between models, with increasing performance from left to right. This further validates the usefulness of the models, and also supports the fact that the loss function and training metrics are well designed and correlate with our downstream use case.

Evaluating material and magnetic subsplits

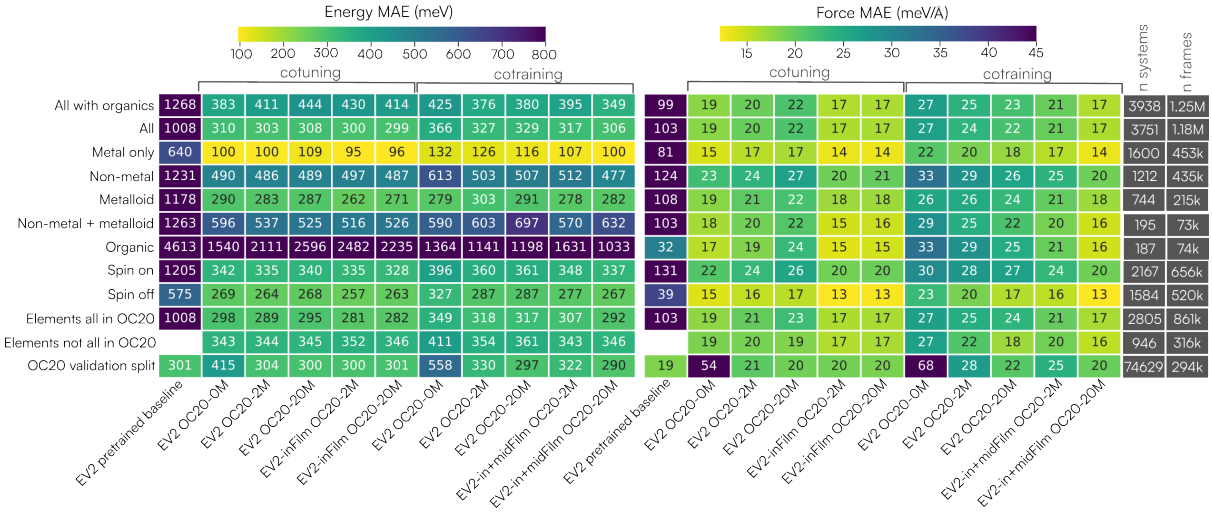

Figure 8: Comparison of Energy MAE (meV, left) and Force MAE (meV/Å, right) across different model training strategies and data subsets. Performance is evaluated for cotuning vs. cotraining, varying included OC20 data amounts (0M, 2M, 20M), and different architectures (vanilla EV2, EV2-inFiLM, EV2-in+midFiLM). Subsets include spin treatment, element novelty relative to OC20, and material type.

To further probe the robustness of the models presented and identify potential systematic biases related to specific chemical or physical properties, we next evaluate performance across distinct subsets of the test data. Specifically, we analyze error trends based on the material type, whether the elements contained were all included in OC20, and whether spin polarization was treated during the calculation. The results of this are shown in Figure 8 with model energy MAE on the left and force MAE on the right. The data presented here are evaluated on the AQCat25 test split for all rows except the OC20 validation split, which is the same subsample of the OC20 OOD both split discussed above.

This analysis exposed a segment of the dataset that has very poor performance for energies: organic materials. Materials that only contain H, O, N, C, S, P, F, Cl, Br, I, and/or Se have very poor metrics as seen in the "Organics" row of Figure 8. This poor performance, however, does not extend to forces. This is likely because of the referencing scheme used. These materials are more able to restructure and it is therefore far more likely that the relaxed slab state is very different than the adsorbate-slab along the relaxation trajectory. These materials are not necessarily of catalytic interest, so it could be beneficial to be more selective when including them in the dataset. Certainly total energy models would be better suited to handle these systems because they remove dependence on the referenced state. Because of the significance of the errors on these systems, we removed them from all other splits presented in the figure. An alternative version of this figure has been included in the Supplementary Information where the organics are not segmented out. Without removing organics, we observe trends that are opposite to those expected and presented here.

Here, we see that across the board model metrics for forces and energies are better for spin off than spin on. For most cases, when we compare the EV2 model to its corresponding EV2 + FiLM model, there

is an improvement in spin on. Performance on systems where all elements appear in OC20 is better than systems that contain at least one new element. Aligned with existing precedent, the models more accurately predict energies and forces on metals when compared to other material types. Metalloids are slightly better treated than non-metals. Interestingly, energies for non-metals + metalloids are worse than non-metals, but the forces are not.

The models jointly trained from scratch outperform corresponding jointly tuned models by ∼1 eV on the challenging organic material split, but achieve slightly worse performance on the other material splits excluding this category. The difference in optimization path and data exposure leads to these models marginally sacrificing performance on the broader set of materials to recover performance on this difficult class. Notably, as seen previously, the jointly trained from scratch model with 20M additional OC20 samples achieves the highest performance on the practical catalysis tasks described in Figures 6a and 7.

A summary of model performance for the models discussed here and some others is shown in Table 2. Model names denote the architecture (EV2 31M default, inFiLM, or in+midFiLM) and training protocol, where 'ft' signifies fine-tuning an OC20-pretrained model and its absence means cotraining from scratch. Dataset identifiers specify OC20 data added (+OC20-2M/20M). Direct tuning experiments excluded the low-fidelity spin-on subset.

Table 2: Model Performance Metrics (Energy in meV, Forces in meV/Å)

Category	Model	All E-MAE	All F-MAE	Spin On E-MAE	Spin On F-MAE	Spin Off E-MAE	Spin Off F-MAE	OC20 Val E-MAE	OC20 Val F-MAE
Pretrained	EV2-OC20	1268	98.63	1205	131.40	1378	37.06	301	19.32
Direct Tuning	EV2-OC20-ft-AQCat25-highfi only	376	18.46	337	21.63	419	14.80	440	59.13
	EV2-OC20-ft-AQCat25-highfi only (153M)	350	17.59	339	20.41	362	14.34	433	52.84
Cotuning	EV2-OC20-ft-AQCat25	383	18.65	342	21.81	428	15.00	415	53.73
	EV2-inFiLM-OC20-ft-AQCat25	379	18.30	346	21.23	415	14.91	396	41.34
	EV2-OC20-ft-AQCat25+OC20-2M	411	20.36	335	23.71	495	16.48	304	21.23
	EV2-inFiLM-OC20-ft-AQCat25+OC20-2M	430	16.93	335	19.90	536	13.49	300	19.90
	EV2-OC20-ft-AQCat25+OC20-20M	444	22.14	340	26.30	559	17.34	300	20.12
	EV2-inFiLM-OC20-ft-AQCat25+OC20-20M	414	17.21	328	20.28	510	13.66	301	19.62
	EV2-inFiLM-OC20-ft-AQCat25+OC20-20M ($\lambda_E = 100$)	412	21.03	325	24.32	508	17.22	289	21.79
Cotraining	EV2-AQCat25	425	27.38	396	29.86	457	24.53	558	68.06
	EV2-AQCat25+OC20-2M	376	24.60	360	28.01	394	20.68	330	27.69
	EV2-inFiLM-AQCat25+OC20-2M	392	20.57	345	23.79	442	16.86	321	24.86
	EV2-in+midFiLM-AQCat25+OC20-2M	395	20.53	348	23.75	447	16.80	322	24.83
	EV2-AQCat25+OC20-20M	380	22.65	361	26.85	402	17.81	297	21.51
	EV2-inFiLM-AQCat25+OC20-20M	367	**16.83**	334	19.90	403	13.28	290	20.35
	EV2-in+midFiLM-AQCat25+OC20-20M	**349**	16.98	337	20.05	363	13.44	290	20.46

Conclusion

This work tackled a significant gap limiting the application of large-scale MLIPs in heterogeneous catalysis: the proper treatment of magnetism and enhanced electronic fidelity to accurately model and discover novel catalysts containing earth-abundant, spin-polarized elements such as Fe, Co, and Ni. We demonstrated that while direct fine-tuning of a pretrained OC20 model on AQCat25 provides performance on the new data it leads to a significant degradation of performance on the original OC20 domain. We found that by combining the targeted high-fidelity physics captured in AQCat25 with the extensive chemical and structural diversity present in a large portion of the OC20 data, jointly training successfully enhances accuracy on the AQCat25 test set while mitigating degradation on the evaluated OC20 validation metrics. We further confirmed the applicability of our models for the practical catalysis task of identifying the global minimum adsorption energy on a diverse set of surfaces. This training methodology, utilizing multi-fidelity data and explicit conditioning, offers a promising path toward practical and broadly applicable MLIPs for heterogeneous catalysis.

Contributions

O.A.: Model implementation, experimental design, training, ablations, idea conceptualization, data processing, data analysis, data visualization, writing, editing. B.W.: Dataset generation, idea conceptualization, data analysis, data visualization, writing, editing. A.R.S.: Project leadership, idea conceptualization, writing, editing.

Acknowledgements

The authors acknowledge SandboxAQ leadership, especially Ang Xiao, Adam Lewis, Arman Zaribafiyan, Jeff Graf, Nadia Harhen, Andrew McLaughlin, and Jack Hidary, for their support of this research. The authors thank Joseph Gauthier, Kevin Ryczko, Tom Ludwig, Jia-Min Chu, Tyler Sours, Jiyoon Kim, and Jens Nørskov for valuable discussions, code reviews, and feedback. The authors recognize computational and engineering support from SungYeon Kim, Rudi Plesch, and the Nvidia DGX team.

References

[1] Jens K Nørskov, Thomas Bligaard, Jan Rossmeisl, and Claus H Christensen. Towards the computational design of solid catalysts. *Nature Chemistry*, 1(1):37—46, 2009.

[2] Jeffrey Greeley. Theoretical heterogeneous catalysis: scaling relationships and computational catalyst design. *Annual Review of Chemical and Biomolecular Engineering*, 7(1):605—635, 2016.

[3] Ali Hussain Motagamwala and James A Dumesic. Microkinetic modeling: a tool for rational catalyst design. *Chemical Reviews*, 121(2):1049—1076, 2020.

[4] Benjamin WJ Chen, Lang Xu, and Manos Mavrikakis. Computational methods in heterogeneous catalysis. *Chemical Reviews*, 121(2):1007—1048, 2020.

[5] Wenbo Xie, Jiayan Xu, Jianfu Chen, Haifeng Wang, and P Hu. Achieving theory—experiment parity for activity and selectivity in heterogeneous catalysis using microkinetic modeling. *Accounts of Chemical Research*, 55(9):1237—1248, 2022.

[6] K Honkala, Anders Hellman, IN Remediakis, Ashildur Logadottir, A Carlsson, Søren Dahl, Claus H Christensen, and Jens K Nørskov. Ammonia synthesis from first-principles calculations. *Science*, 307 (5709):555—558, 2005.

[7] Aayush R Singh, Joseph H Montoya, Brian A Rohr, Charlie Tsai, Aleksandra Vojvodic, and Jens K Nørskov. Computational design of active site structures with improved transition-state scaling for ammonia synthesis. *ACS Catalysis*, 8(5):4017—4024, 2018.

[8] LC Grabow and M Mavrikakis. Mechanism of methanol synthesis on Cu through CO_2 and CO hydrogenation. *ACS Catalysis*, 1(4):365—384, 2011.

[9] Felix Studt, Frank Abild-Pedersen, Qiongxiao Wu, Anker D Jensen, Burcin Temel, Jan-Dierk Grunwaldt, and Jens K Nørskov. CO hydrogenation to methanol on Cu—Ni catalysts: Theory and experiment. *Journal of Catalysis*, 293:51—60, 2012.

[10] RA Van Santen, AJ Markvoort, IAW Filot, MM Ghouri, and EJM24030478 Hensen. Mechanism and microkinetics of the Fischer—Tropsch reaction. *Physical Chemistry Chemical Physics*, 15(40):17038—17063, 2013.

[11] Zihao Yao, Chenxi Guo, Yu Mao, and P Hu. Quantitative determination of C—C coupling mechanisms and detailed analyses on the activity and selectivity for Fischer—Tropsch synthesis on Co (0001): Microkinetic modeling with coverage effects. *ACS Catalysis*, 9(7):5957—5973, 2019.

[12] Felix Studt, Frank Abild-Pedersen, Thomas Bligaard, Rasmus Z Sørensen, Claus H Christensen, and Jens K Nørskov. Identification of non-precious metal alloy catalysts for selective hydrogenation of acetylene. *Science*, 320(5881):1320—1322, 2008.

[13] Haoran He, Randall J Meyer, Robert M Rioux, and Michael J Janik. Catalyst design for selective hydrogenation of benzene to cyclohexene through density functional theory and microkinetic modeling. *ACS Catalysis*, 11(19):11831—11842, 2021.

[14] Glenn Jones, Jon Geest Jakobsen, Signe S Shim, Jesper Kleis, Martin P Andersson, Jan Rossmeisl, Frank Abild-Pedersen, Thomas Bligaard, Stig Helveg, Berit Hinnemann, et al. First principles calculations and experimental insight into methane steam reforming over transition metal catalysts. *Journal of Catalysis*, 259(1):147—160, 2008.

[15] Wen Zhu and Bo Yang. First-principles-based microkinetic modeling of methane steam reforming with improved description of product desorption. *The Journal of Physical Chemistry C*, 125(34): 18743—18751, 2021.

[16] N Schumacher, Astrid Boisen, Søren Dahl, Amit A Gokhale, Shampa Kandoi, Lars C Grabow, James A Dumesic, Manos Mavrikakis, and Ib Chorkendorff. Trends in low-temperature water—gas shift reactivity on transition metals. *Journal of Catalysis*, 229(2):265—275, 2005.

[17] Pushkar Ghanekar, Joseph Kubal, Yanran Cui, Garrett Mitchell, W Nicholas Delgass, Fabio Ribeiro, and Jeffrey Greeley. Catalysis at metal/oxide interfaces: density functional theory and microkinetic modeling of water gas shift at Pt/MgO boundaries. *Topics in Catalysis*, 63(7):673—687, 2020.

[18] Carsten Stegelmann, Niels Christian Schiødt, Charles T Campbell, and Per Stoltze. Microkinetic modeling of ethylene oxidation over silver. *Journal of Catalysis*, 221(2):630—649, 2004.

[19] Hao Li, Ang Cao, and Jens K Nørskov. Understanding trends in ethylene epoxidation on group IB metals. *ACS Catalysis*, 11(19):12052—12057, 2021.

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

[21] 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.

[22] Thomas Bligaard, Jens K Nørskov, Søren Dahl, J Matthiesen, Claus H Christensen, and JJJoC Sehested. The Brønsted—Evans—Polanyi relation and the volcano curve in heterogeneous catalysis. *Journal of Catalysis*, 224(1):206—217, 2004.

[23] Rutger A Van Santen, Matthew Neurock, and Sharan G Shetty. Reactivity theory of transition-metal surfaces: a brønsted- evans- polanyi linear activation energy- free-energy analysis. *Chemical Reviews*, 110(4):2005—2048, 2009.

[24] Flemming Besenbacher, Ib Chorkendorff, BS Clausen, Bjørk Hammer, AM Molenbroek, Jens K Nørskov, and Ivan Stensgaard. Design of a surface alloy catalyst for steam reforming. *Science*, 279(5358):1913—1915, 1998.

[25] Claus JH Jacobsen, Søren Dahl, Bjerne S Clausen, Sune Bahn, Ashildur Logadottir, and Jens K Nørskov. Catalyst design by interpolation in the periodic table: bimetallic ammonia synthesis catalysts. *Journal of the American Chemical Society*, 123(34):8404—8405, 2001.

[26] 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.

[27] Volker L Deringer, Miguel A Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019.

[28] Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Schutt, Alexandre Tkatchenko, and Klaus-Robert Muller. Machine learning force fields. *Chemical Reviews*, 121(16):10142—10186, 2021.

[29] Sergio Pablo-García, Santiago Morandi, Rodrigo A Vargas-Hernández, Kjell Jorner, Žarko Ivković, Núria López, and Alán Aspuru-Guzik. Fast evaluation of the adsorption energy of organic molecules on metals via graph neural networks. *Nature Computational Science*, 3(5):433—442, 2023.

[30] Deqi Tang, Rangsiman Ketkaew, and Sandra Luber. Machine learning interatomic potentials for heterogeneous catalysis. *Chemistry—A European Journal*, 30(60):e202401148, 2024.

[31] Carlota Bozal-Ginesta, Sergio Pablo-García, Changhyeok Choi, Albert Tarancón, and Alán Aspuru-Guzik. Developing machine learning for heterogeneous catalysis with experimental and computational data. *Nature Reviews Chemistry*, pages 1—16, 2025.

[32] Eric C-Y Yuan, Yunsheng Liu, Junmin Chen, Peichen Zhong, Sanjeev Raja, Tobias Kreiman, Santiago Vargas, Wenbin Xu, Martin Head-Gordon, Chao Yang, et al. Foundation models for atomistic simulation of chemistry and materials. *arXiv preprint arXiv:2503.10538*, 2025.

[33] Chi Chen, Yunxing Zuo, Weike Ye, Xiangguo Li, and Shyue Ping Ong. Learning properties of ordered and disordered materials from multi-fidelity data. *Nature Computational Science*, 1(1):46—53, 2021.

[34] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80—85, 2023.

[35] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, Matthew Avaylon, William J Baldwin, et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096*, 2023.

[36] Tsz Wai Ko and Shyue Ping Ong. Data-efficient construction of high-fidelity graph deep learning interatomic potentials. *npj Computational Materials*, 11(1):65, 2025.

[37] Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, et al. The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066—3084, 2023.

[38] Sushree Jagriti Sahoo, Mikael Maraschin, Daniel S Levine, Zachary Ulissi, C Lawrence Zitnick, Joel B Varley, Joseph A Gauthier, Nitish Govindarajan, and Muhammed Shuaibi. The Open Catalyst 2025 (OC25) Dataset and Models for Solid-Liquid Interfaces. *arXiv preprint arXiv:2509.17862*, 2025.

[39] 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):14621, 2017.

[40] Miao Zhong, Kevin Tran, Yimeng Min, Chuanhao Wang, Ziyun Wang, Cao-Thang Dinh, Phil De Luna, Zongqian Yu, Armin Sedighian Rasouli, Peter Brodersen, et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. *Nature*, 581(7807):178—183, 2020.

[41] Johannes T Margraf, Hyunwook Jung, Christoph Scheurer, and Karsten Reuter. Exploring catalytic reaction networks with machine learning. *Nature Catalysis*, 6(2):112—121, 2023.

[42] Santiago Morandi, Oliver Loveday, Tim Renningholtz, Sergio Pablo-García, Rodrigo A Vargas-Hernández, Ranga Rohit Seemakurthi, Pol Sanz Berman, Rodrigo García-Muelas, Alán Aspuru-Guzik, and Núria López. A foundational model for reaction networks on metal surfaces. *ChemRxiv preprint*, 2024. doi: 10.26434/chemrxiv-2024-bfv3d.

[43] Tianyou Mou, Hemanth Somarajan Pillai, Siwen Wang, Mingyu Wan, Xue Han, Neil M Schweitzer, Fanglin Che, and Hongliang Xin. Bridging the complexity gap in computational heterogeneous catalysis with machine learning. *Nature Catalysis*, 6(2):122—136, 2023.

[44] Cameron J Owen, Lorenzo Russotto, Christopher R O'Connor, Nicholas Marcella, Anders Johansson, Albert Musaelian, and Boris Kozinsky. Atomistic evolution of active sites in multi-component heterogeneous catalysts. *arXiv preprint arXiv:2407.13607*, 2024.

[45] Amir Omranpour, Jan Elsner, K Nikolas Lausch, and Jorg Behler. Machine learning potentials for heterogeneous catalysis. *ACS Catalysis*, 15(3):1616—1634, 2025.

[46] Kareem Abdelmaqsoud, Muhammed Shuaibi, Adeesh Kolluru, Raffaele Cheula, and John R Kitchin. Investigating the error imbalance of large-scale machine learning potentials in catalysis. *Catalysis Science & Technology*, 14(20):5899—5908, 2024.

[47] Wenbin Xu, Rohan Yuri Sanspeur, Adeesh Kolluru, Bowen Deng, Peter Harrington, Steven Farrell, Karsten Reuter, and John R Kitchin. Spin-informed universal graph neural networks for simulating magnetic ordering. *Proceedings of the National Academy of Sciences*, 122(27):e2422973122, 2025.

[48] X Sun, S Förster, QX Li, M Kurahashi, T Suzuki, JW Zhang, Y Yamauchi, Günter Baum, and Hans Steidl. Spin-polarization study of CO molecules adsorbed on Fe (110) using metastable-atom deexcitation spectroscopy and first-principles calculations. *Physical Review B—Condensed Matter and Materials Physics*, 75(3):035419, 2007.

[49] X Sun, Y Yamauchi, M Kurahashi, T Suzuki, ZP Wang, and S Entani. Spin polarization study of benzene molecule adsorbed on Fe (100) surface with metastable-atom deexcitation spectroscopy and density functional calculations. *The Journal of Physical Chemistry C*, 111(42):15289—15298, 2007.

[50] Ang Cao and Jens K Nørskov. Spin effects in chemisorption and catalysis. *ACS Catalysis*, 13(6): 3456—3462, 2023.

[51] Ke Zhang, Ang Cao, Lau Halkier Wandall, Jerome Vernieres, Jakob Kibsgaard, Jens K Nørskov, and Ib Chorkendorff. Spin-mediated promotion of Co catalysts for ammonia synthesis. *Science*, 383 (6689):1357—1363, 2024.

[52] Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S Levine, Meng Gao, Misko Dzamba, and C Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. *arXiv preprint arXiv:2502.12147*, 2025.

[53] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023.

[54] Eric Qu and Aditi Krishnapriyan. The importance of being scalable: Improving the speed and accuracy of neural network interatomic potentials across chemical domains. *Advances in Neural Information Processing Systems*, 37:139030—139053, 2024.

[55] Brandon M Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R Kitchin, Daniel S Levine, et al. UMA: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*, 2025.

[56] Jaesun Kim, Jisu Kim, Jaehoon Kim, Jiho Lee, Yutack Park, Youngho Kang, and Seungwu Han. Data-efficient multi-fidelity training for high-fidelity machine learning interatomic potentials. *arXiv preprint arXiv:2409.07947*, 2409(07947):1—17, 2024. arXiv:2409.07947.

[57] Mitchell Messerly, Sakib Matin, Alice E A Allen, Benjamin Nebgen, Kipton Barros, Justin S Smith, Nicholas Lubbers, and Richard Messerly. Multi-fidelity learning for interatomic potentials: Low-level forces and high-level energies are all you need. *arXiv preprint arXiv:2505.01590*, 2505(01590):1—16, 2025. arXiv:2505.01590.

[58] John L.A. Gardner, Hannes Schulz, Jean Helie, Lixin Sun, and Gregor N.C. Simm. Understanding multi-fidelity training of machine-learned force-fields. *arXiv preprint arXiv:2506.14963v1*, 2025.

[59] Jacob BJ Chapman and Pui-Wai Ma. A machine-learned spin-lattice potential for dynamic simulations of defective magnetic iron. *Scientific Reports*, 12(1):22451, 2022.

[60] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031—1041, 2023.

[61] Dylan M Anstine and Olexandr Isayev. Machine learning interatomic potentials and long-range physics. *The Journal of Physical Chemistry A*, 127(11):2417—2431, 2023.

[62] Taiping Hu, Teng Yang, Jianchuan Liu, Bin Deng, Zhengtao Huang, Xiaoxu Wang, Fuzhi Dai, Guobing Zhou, Fangjia Fu, Ping Tuo, et al. A spin-dependent machine learning framework for transition metal oxide battery cathode materials. *arXiv preprint arXiv:2309.01146*, 2023.

[63] Teng Yang, Zefeng Cai, Zhengtao Huang, Wenlong Tang, Ruosong Shi, Andy Godfrey, Hanxing Liu, Yuanhua Lin, Ce-Wen Nan, Meng Ye, et al. Deep learning illuminates spin and lattice interaction in magnetic materials. *Physical Review B*, 110(6):064427, 2024.

[64] Hongyu Yu, Yang Zhong, Liangliang Hong, Changsong Xu, Wei Ren, Xingao Gong, and Hongjun Xiang. Spin-dependent graph neural network potential for magnetic materials. *Physical Review B*, 109(14):144426, 2024.

[65] Joseph Musielewicz, Xiaoxiao Wang, Tian Tian, and Zachary Ulissi. FINETUNA: fine-tuning accelerated molecular simulations. *Machine Learning: Science and Technology*, 3(3):03LT01, 2022.

[66] Bowen Deng, Yunyeong Choi, Peichen Zhong, Janosh Riebesell, Shashwat Anand, Zhuohan Li, KyuJung Jun, Kristin A Persson, and Gerbrand Ceder. Systematic softening in universal machine learning interatomic potentials. *npj Computational Materials*, 11(1):9, 2025.

[67] Xiaoqing Liu, Kehan Zeng, Zedong Luo, Yangshuai Wang, Teng Zhao, and Zhenli Xu. Fine-tuning universal machine-learned interatomic potentials: A tutorial on methods and applications. *arXiv preprint arXiv:2506.21935*, 2025.

[68] Emma King-Smith. Transfer learning for a foundational chemistry model. *Chemical Science*, 15(14): 5143—5151, 2024.

[69] Mariia Radova, Wojciech G Stark, Connor S Allen, Reinhard J Maurer, and Albert P Bartók. Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning. *npj Computational Materials*, 11(1):237, 2025.

[70] 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, 1994.

[71] 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.

[72] 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, 1996.

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

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

[75] Yingkai Zhang and Weitao Yang. Comment on "Generalized gradient approximation made simple". *Physical Review Letters*, 80(4):890, 1998.

[76] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[77] Matthew K Horton, Patrick Huck, Ruo Xi Yang, Jason M Munro, Shyam Dwaraknath, Alex M Ganose, Ryan S Kingsbury, Mingjian Wen, Jimmy X Shen, Tyler S Mathis, et al. Accelerated data-driven materials science with the Materials Project. *Nature Materials*, pages 1—11, 2025.

[78] Brook Wander, Muhammed Shuaibi, John R Kitchin, Zachary W Ulissi, and C Lawrence Zitnick. Cattsunami: Accelerating transition state energy calculations with pretrained graph neural networks. *ACS Catalysis*, 15(7):5283—5294, 2025.

[79] Janice Lan, Aini Palizhati, Muhammed Shuaibi, Brandon M Wood, Brook Wander, Abhishek Das, Matt Uyttendaele, C Lawrence Zitnick, and Zachary W Ulissi. AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials. *npj Computational Materials*, 9(1):172, 2023.

[80] Muhammed Shuaibi, Abhishek Das, Anuroop Sriram, Misko, Luis Barroso-Luque, Ray Gao, Siddharth Goyal, zulissimeta, Brandon Wood, Tian Xie, Junwoong Yoon, Brook Wander, Adeesh Kolluru, Richard Barnes, Ethan Sunshine, Kevin Tran, Xiang, Daniel Levine, Nima Shoghi, Ilias Chair, Janice Lan, Kaylee Tian, Joseph Musielewicz, clz55, Weihua Hu, Kyle Michel, Facebook Community Bot, willis, and vbttchr. facebookresearch/fairchem: fairchem_data_oc-1.0.1, 2025. URL `https://doi.org/10.5281/zenodo.15594818`.

[81] Wenhao Sun and Gerbrand Ceder. Efficient creation and convergence of surface slabs. *Surface Science*, 617:53—59, 2013.

[82] Richard Tran, Zihan Xu, Balachandran Radhakrishnan, Donald Winston, Wenhao Sun, Kristin A Persson, and Shyue Ping Ong. Surface energies of elemental crystals. *Scientific Data*, 3(1):1—13, 2016.

[83] S. R. Bahn and K. W. Jacobsen. An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.*, 4(3):56—66, May 2002.

[84] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, et al. The atomic simulation environment — a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

[85] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[86] Taco S. Cohen, Mario Geiger, Jonas Koehler, and Max Welling. Spherical CNNs, 2018. Published at ICLR 2018.

[87] C. Lawrence Zitnick, Abhishek Das, Adeesh Kolluru, Janice Lan, Muhammed Shuaibi, Anuroop Sriram, Zachary Ulissi, and Brandon Wood. Spherical Channels for Modeling Atomic Interactions, 2022. Published at NeurIPS 2022.

[88] Saro Passaro and C. Lawrence Zitnick. Reducing SO(3) Convolutions to SO(2) for Efficient Equivariant GNNs, 2023. Published at ICML 2023.

[89] 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. *ACS Catalysis*, 12(14):8572—8581, 2022.

[90] Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ulissi, C Lawrence Zitnick, and Abhishek Das. Gemnet-oc: Developing graph neural networks for large and diverse molecular simulation datasets. *Transactions on Machine Learning Research*, 2022(9), 2022. URL `https://openreview.net/forum?id=u8tvSxm4Bs`. Published 09/2022; arXiv:2204.02782v3 [cs.LG] 30 Sep 2022.

[91] Ji Qi, Tsz Wai Ko, Brandon C Wood, Tuan Anh Pham, and Shyue Ping Ong. Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling. *arXiv preprint arXiv:2307.13710*, 2307(13710), 2023.

Supplementary Information

VASP parameters

The VASP parameters are summarized in Tables 3 and 4. The Bloch vectors (kpoints) were set using the lattice vectors using the same technique implemented in the fairchem repository for slabs and adsorbate-slab systems. The z-direction is set to 1, while x and y are set using Equation 1. For bulks, this calculation was also applied to the z-direction. For systems containing Ce, Co, Cr, Cu, Fe, Mn, Mo, Ni, Os, Ru, V, or W, spin polarization was enabled to account for magnetic effects.

$$k = max\left[\lfloor\frac{40}{c}\rceil, 1\right] \tag{1}$$

Variable	Setting Slabs Systems	Setting Bulks
IBRION	2	1
NSW	800	250
ISIF	0	7
ISPIN	1 or 2	1 or 2
ISYM	0	0
ALGO	Normal	Normal
ISMEAR	0	0
SIGMA	0.1	0.1
EDIFFG	-0.03	1E-5
ENCUT	500	500
PREC	Accurate	Accurate
POTIM	0.5	0.5
NELM	250	250
EDIFF	1E-4	1E-4
SYMPREC	1E-10	1E-5
LREAL	Auto	False

Table 3: VASP parameters.

Variable	Setting
TEBEG	900
TEEND	900
MDALGO	1
ANDERSEN PROB	0.0
NSW	80
POTIM	2
IBRION	0
NELMIN	4

Table 4: MD specific VASP parameters.

Adsorbate referencing

The adsorbate gas phase references were constructed using the energies of CO, H_2, H_2O, and N_2. Calculations were performed with the molecules separately in vacuum cubes of 10, 20, and 30 Å. There was not a significant energy difference between 20 and 30Å, so 30Å was taken to be converged. The resulting per atom/element energies are summarized in Table 5.

Atom	Energy [eV]
H	-3.4944
O	-7.1590
C	-7.2654
N	-8.1351

Table 5: Adsorbate per atom energy corrections.

Element counts OC20 versus AQCat25

Figure 9 shows a comparison between the frequency with which elements occur in the AQCat25 and OC20 datasets. There are some notable differences like the presence of the six additional elements in AQCat25, the higher relative presense of boron in AQCat25, and the lower presense of Tc in AQCat25.

Figure 9: Element counts for all of the train splits of OC20 and AQCat25.

Sampling

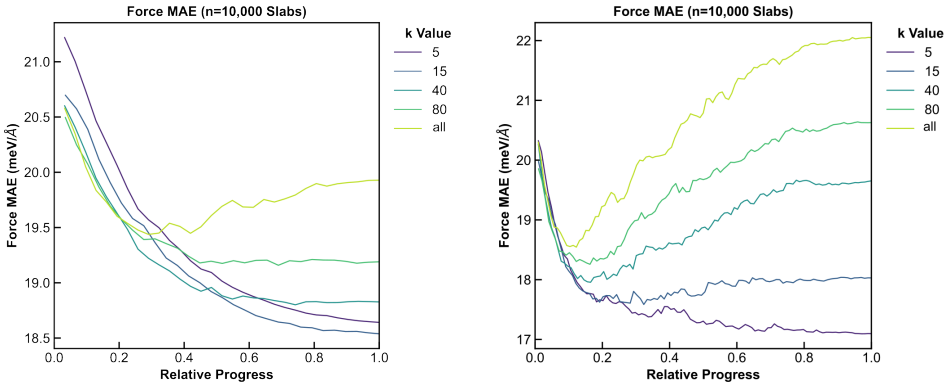

Figure 10: Model overfitting for direct tuning using the 31M parameter (left) and 153M parameter Equiformer v2 model. Sampling frames (as indicated by the k-values) reduces overfitting.

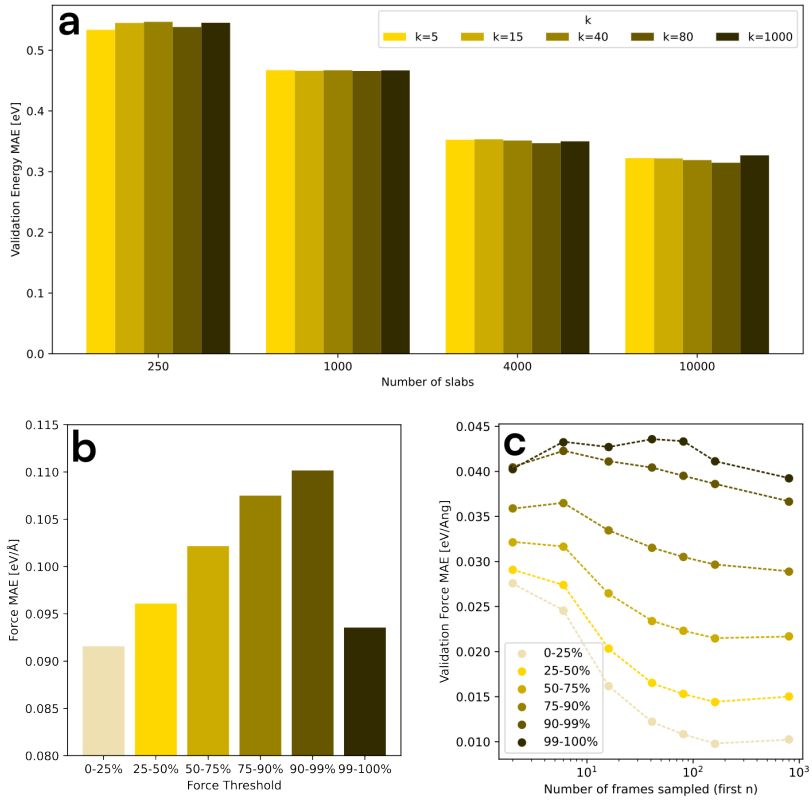

Figure 11: Complementary parts to the figure describing subsampling in the main text.

Figure 11 shows: (a) Energy MAE which does not show a substantial trend with changing k for random subsampling, but is improved by increasing the number of slabs. (b) The AQCat25 validation force MAE for the pretrained 31M parameter Equiformer v2 model across stratified force bins, which shows roughly the same trend as energy: performance decreases on higher force systems with the exception of very high force frames which have better performance. (c) The AQCat25 validation force MAE for naively fintuned models using different values of first k samples of the AQCat25 dataset to fine-tune.

Additional direct tuning metrics

Table 6: Model performance metrics for direct tuning (energy in meV, forces in meV/Å)

Category	Model	All E-MAE	All F-MAE	Spin On E-MAE	Spin On F-MAE	Spin Off E-MAE	Spin Off F-MAE	OC20 Val E-MAE	OC20 Val F-MAE
31M	$\lambda_E = 4$	376	18.46	337	21.63	419	14.80	440	59.13
	$\lambda_E = 100$	372	20.48	349	23.90	398	16.52	458	57.67
	$\lambda_E = 100$, with lowfi spin-on	383	18.65	342	21.81	428	15.00	415	53.73
153M	$\lambda_E = 4$	350	17.59	339	20.41	362	14.34	433	52.84
	$\lambda_E = 100$	343	19.71	335	22.77	352	16.16	420	47.07

Effect of toggling fidelity and spin flags

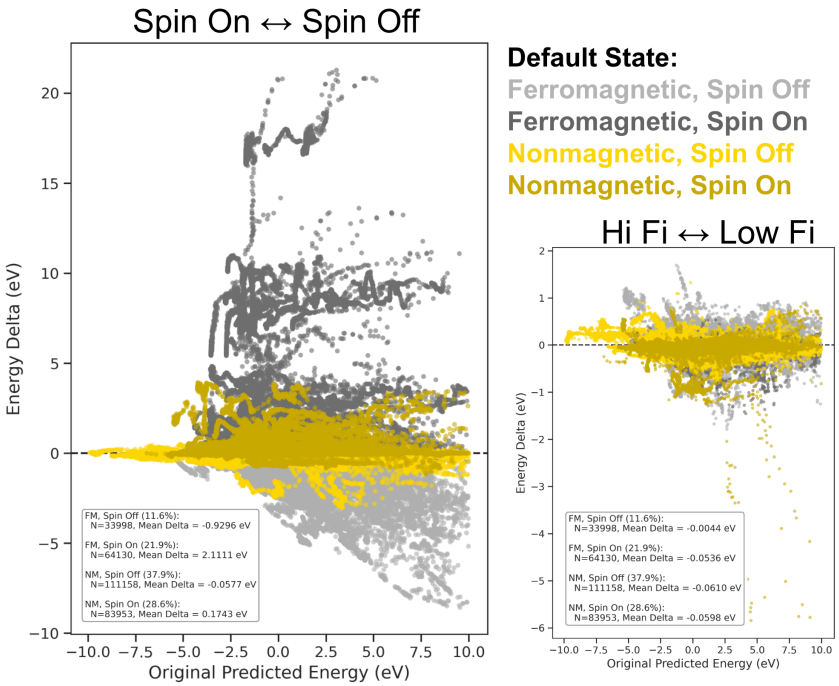

Figure 12: Effect of toggling conditioning flags during inference on predicted energy delta (ΔE). The left panel shows the impact of switching between `spin_on` and `spin_off` flags, while the right panel shows switching between high and low fidelity flags. Colors distinguish between ferromagnetic (FM, grey) and nonmagnetic (NM, yellow) systems based on their ground truth magnetic state (from the MP) and spin treatment in the dataset. Toggling the spin flag has a much larger effect on FM systems, including FM systems labeled as spin-off in the training data.

We also wanted to ablate the impact of the spin and fidelity flags on the resultant energies. Toggling the spin flag induces large energy shifts in opposite directions for systems categorized as ferromagnetic by the MP, depending on their original spin treatment: destabilizing correctly labeled spin-on systems (energy increases) and stabilizing incorrectly labeled spin-off systems (energy decreases). In contrast, NM systems show minimal energy changes when the spin flag is toggled, indicating the model correctly associates strong spin effects primarily with the FM materials (even those that excluded the elements that we categorized as necessitating spin treatment). Further analysis is needed to fully validate that the model has learned the correct underlying physics across domains. For instance, the observed asymmetry could simply reflect that evaluating FM systems with spin turned off represents a significant deviation from the training data distribution. The model may underperform in this regime because it has primarily learned patterns associated with spin-polarized FM states and lacks sufficient training examples or capacity to accurately model the less common or physically distinct spin-unpolarized state for these materials.

Model Training Parameters

Table 7: Architectural hyperparameters for the EquiformerV2 models, including FiLM specifics. We refer to the EquiformerV2 paper[53] for a complete description of all architectural components, including normalization and activation functions.

Hyperparameter	Value
Core EquiformerV2 Architecture	
Number of Transformer blocks	8 (31M), 20 (153M)
Embedding dimension d_{embed}	128
$f_{ij}^{(L)}$ dimension d_{attn_hidden}	64
Hidden dimension in feed forward networks d_{ffn}	128
Number of attention heads	8
Maximum spherical harmonic degree (L_{max})	4 (31M), 6 (153M)
Maximum spherical harmonic order (M_{max})	2 (31M), 3 (153M)
Dropout rate	0.1
Stochastic depth	0.1
Cutoff radius (Å)	12.0
Maximum number of neighbors	20
FiLM Architecture Addendum (EV2-FiLM)	
Auxiliary feature embedding dimension	16
MLP hidden dimension for modulation	128
MLP dropout	0.1
FiLM modulation strategy	**Cotuning**: Input layer only **Training from Scratch**: - Input layer only - Input layer & all Transformer blocks

Table 8: Training and optimization hyperparameters for each experimental strategy.

Parameter	Direct Finetuning	Cotuning	Training from Scratch
Pre-trained Checkpoint	OC20 All+MD	OC20 All+MD	None
Optimizer			
Optimizer	`AdamW`	`AdamW`	`AdamW`
Weight decay	1×10^{-3}	1×10^{-3}	1×10^{-3}
Learning rate (31M)	7×10^{-5}	7×10^{-5}	4×10^{-4}
Learning rate (153M)	8×10^{-5}	8×10^{-5}	4×10^{-4}
LR scheduling	Cosine annealing with linear warmup		
Warmup epochs	0.01	0.01	0.1
Model EMA decay	0.999	0.999	0.999
Batch Size & Epochs			
Batch size per GPU (31M)	20	20	20
Batch size per GPU (153M)	6	6	6
Gradient accumulation (153M)	3 steps	3 steps	3 steps
Effective batch size (31M)	160	160	160
Effective batch size (153M)	144	144	144
Max epochs	30	30	30
Loss & Regularization			
Energy coefficient (λ_E)	4, 100	4	4
Force coefficient (λ_F)	100	100	100
Gradient clipping norm threshold	5	5	100

Probing the impact of spin and fidelity

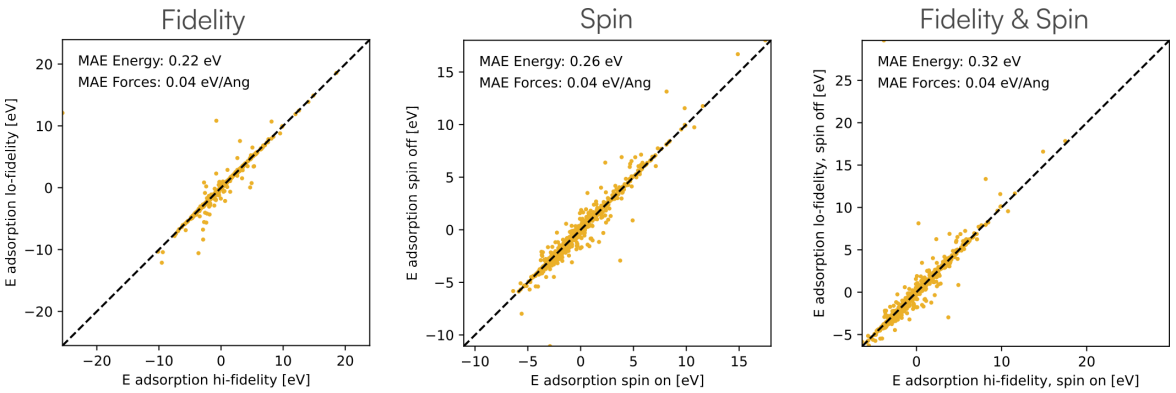

Figure 13: An investigation of the impact of fidelity being the same as OC20 (left), spin being off, rather than on (center), and both spin and fidelity being ablated simultaneously.

We also wanted to investigate the impact of spin and fidelity on the resultant energies. It is difficult to do this in a well posed way because the underlying bulk structure can be impacted by these DFT settings, so making a direct comparison is difficult. To attempt to do so, here we performed DFT single points on the DFT relaxed (with AQCat25 settings elections) adsorbate-slab configuration and the DFT relaxed slab (again with AQCat25 settings elections). The energies presented here are the difference between these two energies to exploit a cancellation of error from any differences in the true lattice constant. The single points were performed specifically ablating the settings highlighted. For fidelity (Fig. 13 - left), 500 spin-on systems and 500 spin-off systems were selected and single points were performed with ENCUT = 350 eV, and Methfessel-Paxton smearing with a width of 0.2 eV. For spin (Fig. 13 - center) 1000 systems with spin on were selected and single points were performed with spin off. For both spin and fidelity, 1000 systems with spin on were selected and single points were performed with the alternative fidelity and spin off. This can give some idea of the independent and combinatorial impact of these two factors on the DFT result.

Minimum adsorption energy task segmented by element category and spin category

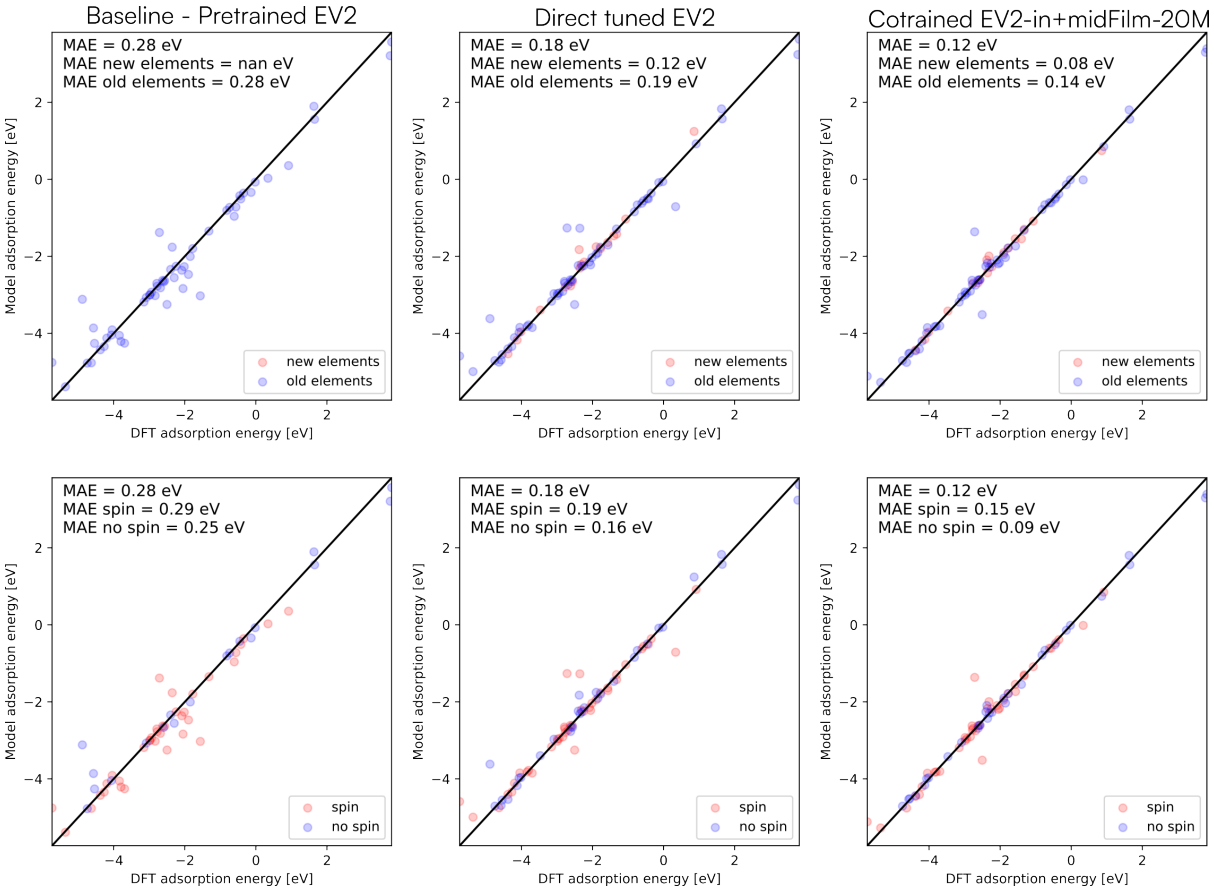

Figure 14: Model performance for finding the global minimum adsorption energy segmented by whether elements appear in OC20 (old elements) or not (new elements) - top and by whether the system was run with spin on or spin off - bottom.

The results looking at the dense dataset but split over whether the system was spin off (Figure 14 or on and whether the system contains new elements reveals that there are not any strong discrepancies between these groups. This is in alignment with Figure 8.

Material and spin splits when including organics

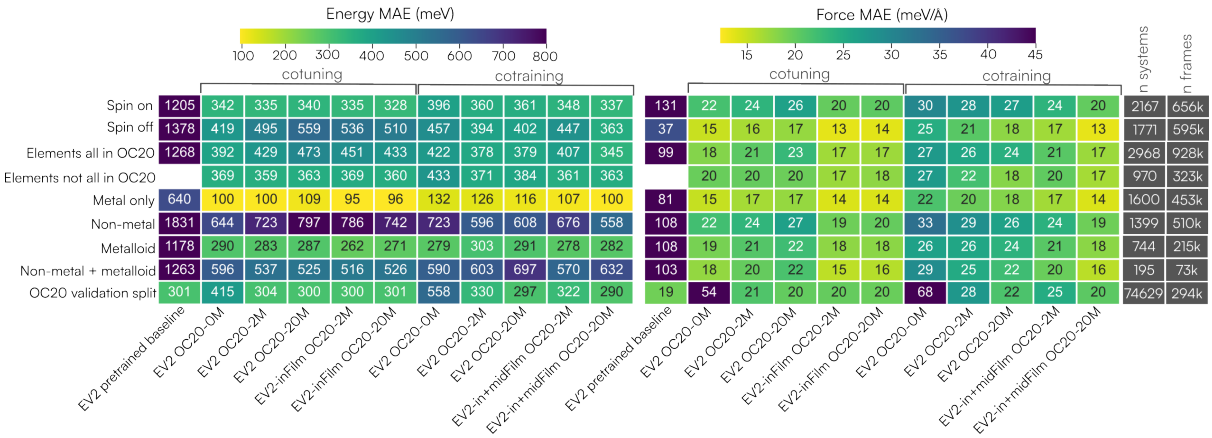

Figure 15: The same results presented in Figure 8, but without segregating the organic (fully non-metal) materials.

When looking at the results including the organic materials, we see that the trends we would expect to see disappear. We would expect that performance on spin off systems should be better in most cases since that is the majority of data seen by the model, but because organic materials were all treated as spin off, the performance on spin off is dragged down. Metrics on non-metals are also pulled down. The same opposing trend is observed for new and in-OC20 elements. These organic materials will always be classed as in-OC20 element materials, and we see that performance is actually better on new elements because they drag down results for in-OC20 elements.

Additional looks at cotuning and cotraining energy metrics

Figure 16 shows the evolution of performance with changing energy cutoff. The cutoff is used to determine the proportion of systems with absolute energies errors less than the value. For the most strict cutoff, cotuning with FiLM has an advantage. For looser thresholds, cotraining has an advantage. Figure 17 shows the energy and force MAE metrics on val and test for the cotuned and cotrained models. For forces, the trends are the same between the two. This is not true, however, for energies. This inspired us to investigate the cause which is that some very high energy errors are skewing the result. This is captured in Figure 8, which shows that performance is poor for organic materials. Cotraining models perform better on these materials at the expense of a slight reduction for other material classes. This shows that the trends for energy performance are sensitive to the metric selected.

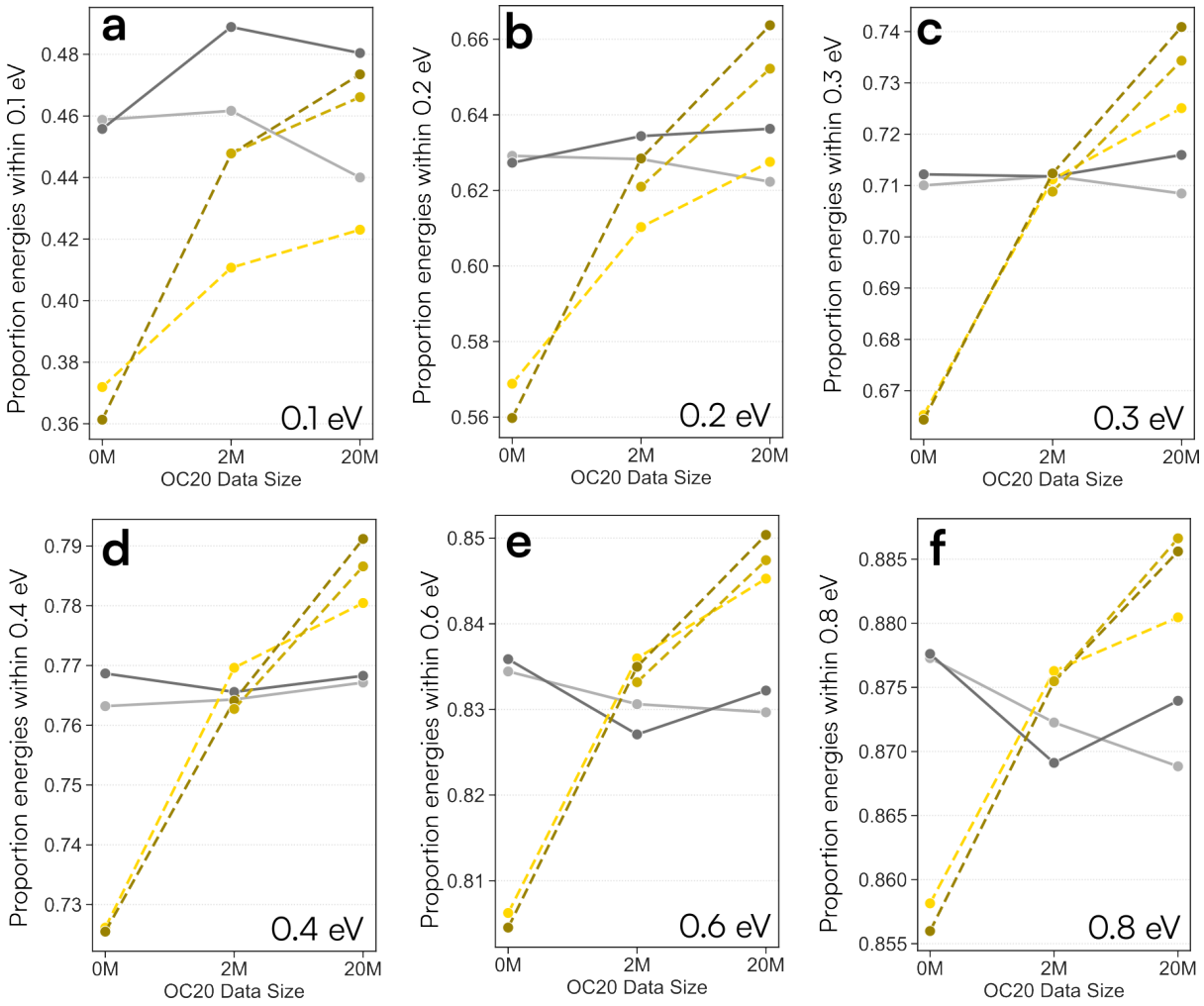

Figure 16: The evolution of trends with changing energy cutoff thresholds.

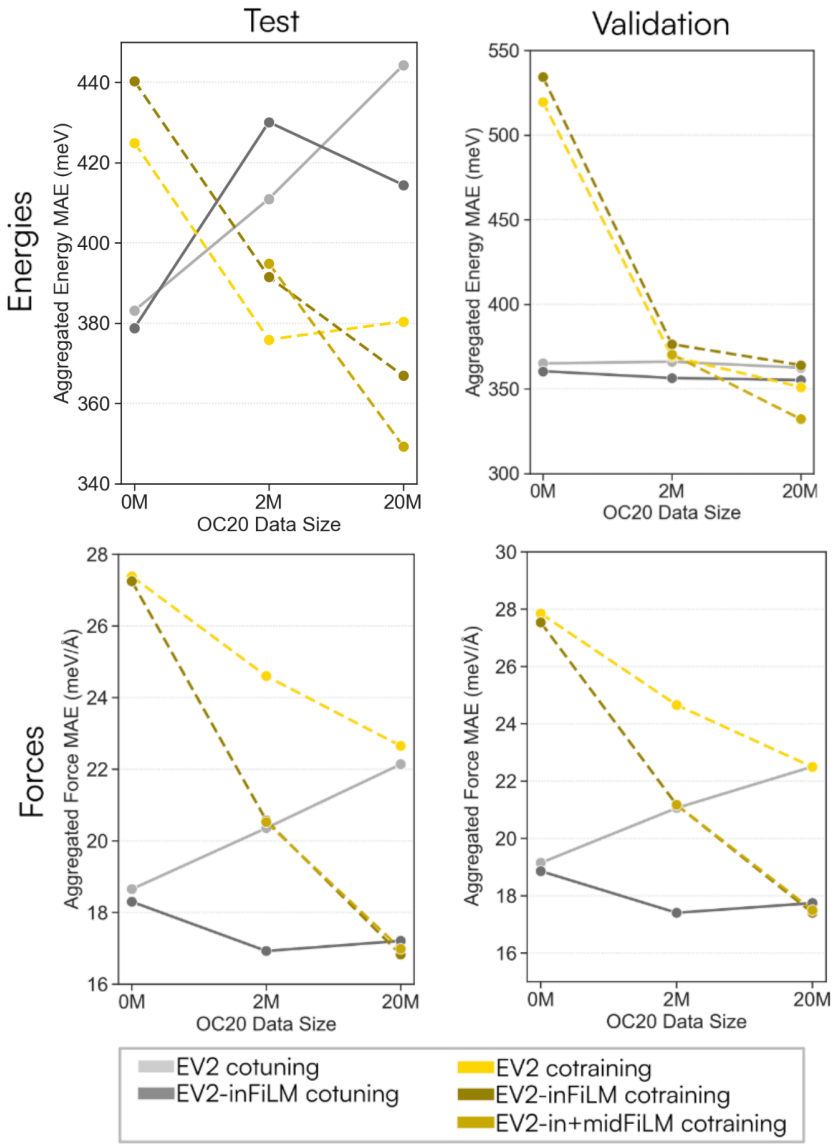

Figure 17: Performance of contuned and cotrained models on test and val for energies (top) and forces (bottom).