

Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C. Lawrence Zitnick, Zachary W. Ulissi

Fundamental AI Research (FAIR) at Meta

The ability to discover new materials with desirable properties is critical for numerous applications from helping mitigate climate change to advances in next generation computing hardware. AI has the potential to accelerate materials discovery and design by more effectively exploring the chemical space compared to other computational methods or by trial-and-error. While substantial progress has been made on AI for materials data, benchmarks, and models, a barrier that has emerged is the lack of publicly available training data and open pre-trained models. To address this, we present a Meta FAIR release of the Open Materials 2024 (OMat24) large-scale open dataset and an accompanying set of pre-trained models. OMat24 contains over 110 million density functional theory (DFT) calculations focused on structural and compositional diversity. Our EquiformerV2 models achieve state-of-the-art performance on the Matbench Discovery leaderboard and are capable of predicting ground-state stability and formation energies to an F1 score above 0.9 and an accuracy of 20 meV/atom, respectively. We explore the impact of model size, auxiliary denoising objectives, and fine-tuning on performance across a range of datasets including OMat24, MPtraj, and Alexandria. The open release of the OMat24 dataset and models enables the research community to build upon our efforts and drive further advancements in AI-assisted materials science.

Date: October 18, 2024

Correspondence: L. Barroso-Luque (lbluque@meta.com), C.L. Zitnick (zitnick@meta.com), Z. Ulissi (zulissi@meta.com)

Code: <https://github.com/FAIR-Chem/fairchem>, Permissive open source license

Dataset: <https://huggingface.co/datasets/fairchem/OMAT24>, Creative Commons 4.0 License

Checkpoints: <https://huggingface.co/fairchem/OMAT24>, Permissive open source license

Blogpost: <https://ai.meta.com/blog/fair-news-segment-anyting-2-1-meta-spirit-lm-layer-skip-salsa-sona>



1 Introduction

The discovery of new materials lies at the foundation of many pressing global problems. This includes finding new catalysts materials for renewable energy storage or for producing carbon neutral fuels^{1–3} and the design of direct air capture sorbents⁴, among many others^{5–11}. The search space of possible materials is enormous, and remains a significant challenge for both computational and experimental approaches to material science. Identifying promising candidates through computational screening with machine learning models offers the potential to dramatically increase the search space and the rate of experimental discovery.

To aid in the discovery of new materials, computational approaches are typically used as filters for identifying promising materials for synthesis in the lab. This is done by computing the formation energy of a candidate material and any materials in its local neighborhood of composition space. An indication that a candidate material may be stable experimentally is whether its formation energy is on or below the convex hull of the energies of its neighbors. The challenge computationally is that the formation energy calculations are typically performed using Density Functional Theory (DFT), which is computationally very expensive and limits its utility in exploring the combinatorial search space of new materials.

Recently, significant advancements have been made in the training of machine learning interatomic potentials to replace costly DFT calculations. Most of these approaches use graph neural network architectures (among

many others^{12–23}) with large (>1M configurations) training datasets for molecules^{24,25}, catalysts^{1,26}, and metal organic frameworks⁴. Open datasets and models have led to rapid progress in each of these areas.

For inorganic materials, Matbench²⁷ was one of the first community benchmarks and had datasets with up to ~130k examples. As model development quickly advanced to architectures that could leverage more data, the training data was expanded to most of the relaxation data (~1.6M examples) from the Materials Project²⁸ to create the MPtrj dataset²⁹. At the same time, the Matbench Discovery Benchmark³⁰ was introduced and remains the primary method for evaluating new models and approaches. Recently, further advances and state-of-the-art models on Matbench Discovery have been realized using even larger datasets (>10M)^{22,23,31}. However, most of these datasets and the procedures used to train these models have remained proprietary (as of 10/2024) making it challenging for the research community to build upon these advances.

In this paper, we introduce the Open Materials 2024 (OMat24) dataset and models (Figure 1) from Meta FAIR to further the rapid advancement of AI and material science. The OMat24 dataset contains over 100 million single point DFT calculations sampled from a diverse set of non-equilibrium atomic configurations and elemental compositions for inorganic bulk materials. The dataset builds upon other public datasets such as MPtrj²⁹, the Materials Project²⁸ and Alexandria³² which contain equilibrium or near-equilibrium configurations³³. We pre-train several variants of the EquiformerV2 model^{20,34} on the OMat24 dataset, and demonstrate state-of-the-art results on the MatBench Discovery leaderboard³⁰ after fine-tuning on the MPtrj dataset²⁹ and a subset of the Alexandria dataset³². Further, we show that the EquiformerV2 model trained with DeNS³⁴ on just the MPtrj dataset is also competitive with some of the largest proprietary models. The dataset, models and code are all open sourced to ensure the reproducibility of the results and to allow the community to build upon and further improve our results.

2 OMat24 Dataset

The OMat24 dataset consists of a combination of DFT single-point calculations, structural relaxations, and molecular dynamic trajectories over a diverse set of inorganic bulk materials. In total, ~118 million structures labeled with total energy, forces and cell stress were calculated with 400M+ core hours of compute. OMat24 includes physically-important non-equilibrium structures with wide energy, forces and stress distributions, as well as significant compositional diversity. Given the focus on non-equilibrium structures, we expect that models trained on OMat24 will be better at far-from-equilibrium and dynamic properties than models trained solely on relaxations. We also demonstrate that pre-training on the diverse OMat24 data substantially improves fine-tuning performance on MPtrj.

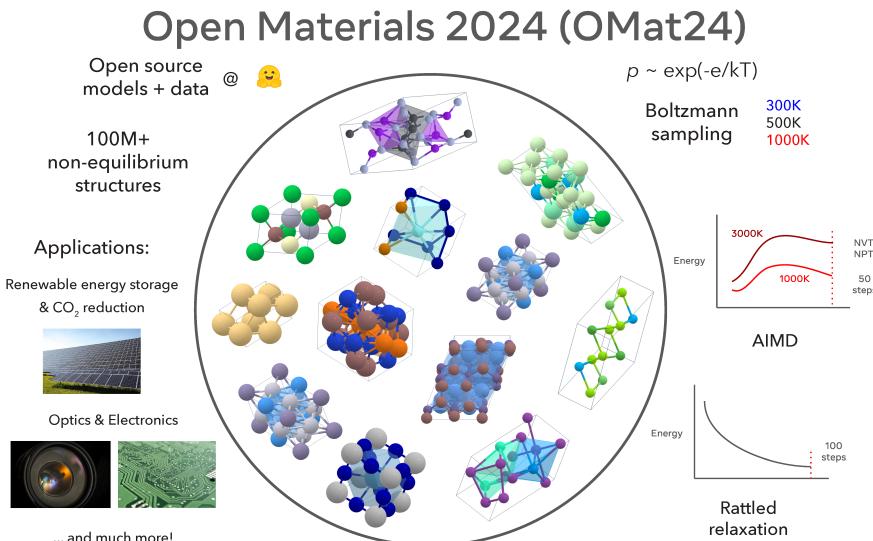


Figure 1 Overview of the OMat24 dataset generation, application areas, and sampling strategies. Inset images are a random sample across the different sampling strategies.

2.1 OMat24 summary and statistics

The OMat24 dataset includes a total of 118 million structures labeled with energy, forces and cell stress. The number of atoms per structure ranges from 1 to 100 atoms per structure, with the majority of structures having 20 atoms or less. A histogram of the number of atoms per structure is shown in Figure 2a. The majority of OMat24 structures have less than 20 atoms per structure as a direct result of starting from structures in the Alexandria dataset, which are predominantly structures with 16 or less atoms per structure. The OMat24 structures with more than 50 atoms per structure are the larger input structures that were used for AIMD.

OMat24 was designed to enable property predictions from equilibrium (see sampling strategies below), and this is captured in the label distributions. The distributions of energy, forces and stress labels along with those of the MPtrj and Alexandria datasets are shown in Figure 2(a). We observe that the energy distribution is slightly higher than the Alexandria dataset, which was used for input structures, and significantly higher than the MPtrj dataset. As expected, the distribution of forces of the OMat24 dataset is notably wider than that of both MPtrj and Alexandria. The distribution of maximum cell stress values of OMat24 is also much wider than that of the MPtrj and Alexandria datasets.

The elemental distribution of OMat24 largely covers the periodic table as illustrated in Figure 2(b). As expected, the elemental distribution is similar to that of the Alexandria and MPtrj datasets in Figures 6 and 7. The dataset covers most of the elements relevant to inorganic materials discovery. Oxides are somewhat over-represented compared to other elements due to their abundance in most open datasets.

2.2 Dataset generation

2.2.1 Crystal structure generation

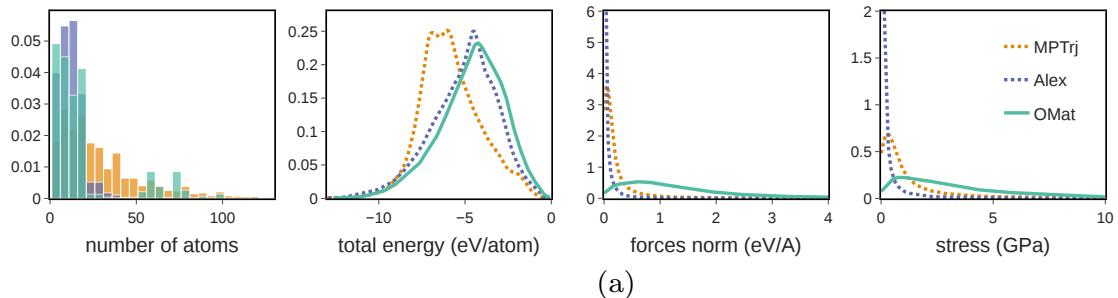
Input structure generation for the OMat24 dataset consists of three different processes intended to obtain diverse non-equilibrium structures: Boltzmann sampling of rattled (random Gaussian perturbations of atomic positions) structures, ab initio molecular dynamics (AIMD), and relaxations of rattled structures. These methods were used to increase the diversity of sampled configurations, similar to prior large dataset efforts ^{22,31}.

In all three approaches, initial structures were obtained by randomly sampling the relaxed structures in the Alexandria PBE bulk materials dataset (3D compounds)³². The Alexandria dataset was chosen as a starting point because it was the largest openly available DFT dataset of equilibrium and near equilibrium structures (~ 4.5 million materials). By randomly sampling relaxed structures from the Alexandria dataset we are able to cover a wide elemental composition diversity. Additionally, using Alexandria relaxed structures as a starting point prevents generating structures too far from equilibrium that could result in DFT convergence errors, or unphysical starting configurations. The details for the three processes are as follows:

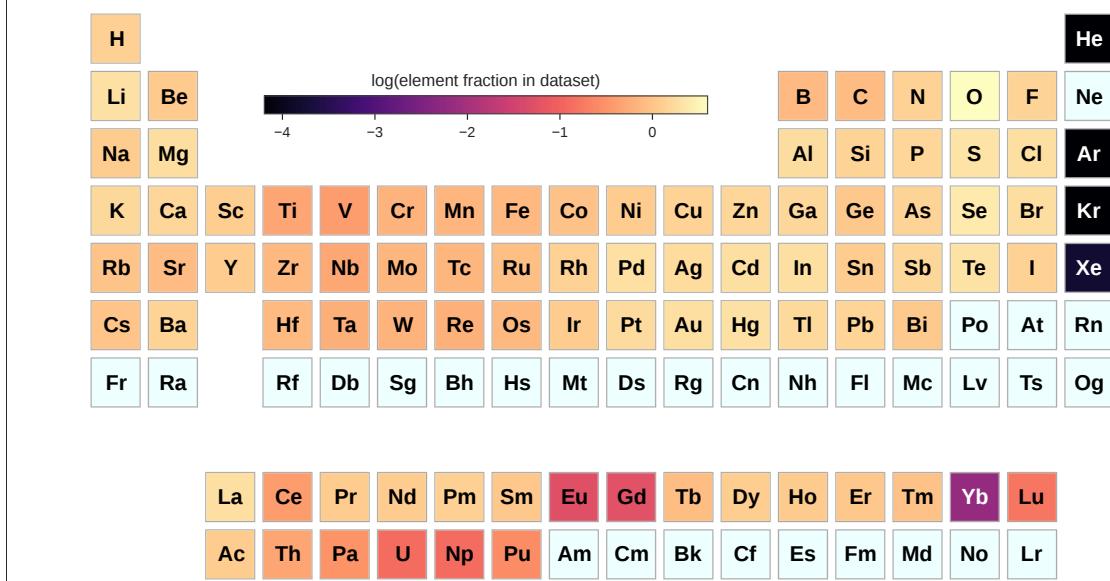
Rattled Boltzmann sampling: For each randomly sampled Alexandria structure we generated 500 candidate non-equilibrium structures by scaling the unit cell to contain at least 10 atoms. Atomic positions were then rattled with displacements sampled from a Gaussian distribution of $\mu = 0\text{\AA}$ and $\sigma = 0.5\text{\AA}$. Unit cells were also deformed isotropically and anisotropically from a Gaussian distribution of $\mu = 1\%$ and $\sigma = 5\%$. For each set of 500 candidate structures we selected 5 of them from a Boltzmann-like distribution based on total energies predicted by an EquiformerV2 model trained on the MPtrj dataset. The sampling procedure was done using three different sampling temperatures: 300K, 500K and 1000K.

Ab-Initio Molecular Dynamics (AIMD): Short-length (50 ionic steps) ab-initio molecular dynamics were carried out starting from randomly sampled relaxed structures in the Alexandria dataset. Structures were recorded from constant temperature and volume (NVT), and constant temperature and pressure (NPT) AIMD trajectories at temperatures of 1000K and 3000K. Unit cells were scaled to contain at least 50 atoms for structures sampled at 3000K.

Rattled relaxation: Relaxed structures from Alexandria were selected at random, rattled (both atomic positions and unit cell), and re-relaxed. Atomic displacements were sampled from a Gaussian distribution of $\mu = 0\text{\AA}$ and $\sigma = 0.2\text{\AA}$. Similarly, isotropic and anisotropic cell deformations were sampled with a $\mu = 1\%$ and $\sigma = 4\%$. All structures along the relaxation trajectory were included in the dataset.



(a)



(b)

Figure 2 (a) Energy per atom, forces norm and max absolute stress element distributions for MPTrj, Alexandria and OMat24 datasets. (b) Distribution of elements in the OMat24 dataset.

We note that these strategies were chosen to increase diversity, but there are many possible sampling strategies that could be used to maximize information content^{35,36}, and we expect these strategies to be useful for future sampling efforts. Active learning^{22,31} sampling strategies have the potential to further enhance these approaches but it remains unclear how they compare to random baselines when considering large scale dataset sizes.

2.2.2 DFT calculation settings and details

DFT calculations generally followed Material Project default settings³³ with some important exceptions. 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 Universität Wien^{37,38} with periodic boundary conditions and the projector augmented wave (PAW) pseudopotentials. Exchange and correlation effects were calculated using the generalized gradient approximation and the Perdew-Burke-Ernzerhof (PBE) with Hubbard U corrections for oxide and fluoride materials containing Co, Cr, Fe, Mn, Mo, Ni, V, or W, following Materials Project defaults³³.

VASP input sets were generated using the MPRELAXSET class defined in the PYMATGEN³⁹ library with the following modifications to account for recent updates and changes in the underlying algorithms and pseudopotentials:

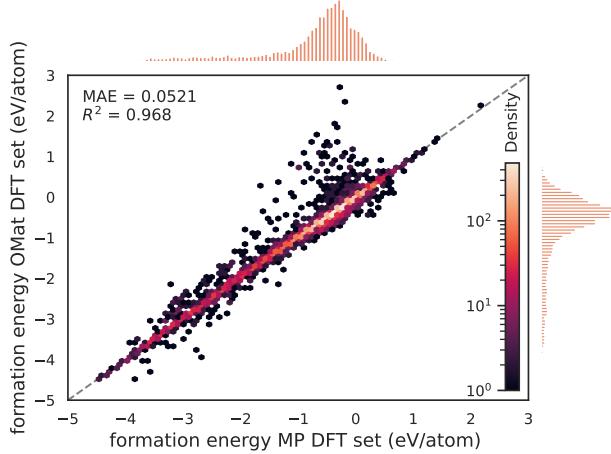


Figure 3 Formation energy taken directly from the WBM dataset⁴⁶ and formation energy calculated from DFT calculations with OMat DFT settings. Outliers are primarily elements with updated pseudopotentials.

1. Version 54 of pseudopotentials provided by VASP were used, rather than the legacy PBE MPRelaxSet defaults³³. The Yb_3 and W_sv pseudopotentials were used for Yb and W to account for changes between version 52 and 54 of VASP PBE pseudopotentials^{40,41}.
2. All calculations were done with the ALGO flag set to “Normal”.

Relaxations were conducted with the MPRELAX set defaults for IBRION/etc. All AIMD calculations were carried out for 50 steps at a time-step of 2 femtoseconds. We note that 2 fs is a large time-step for typical AIMD simulations, especially for hydrogen-containing materials, but was chosen as the goal was to sample diverse configurations rather than perfectly integrate the trajectories. Finally, for static calculations resulting from inputs generated using rattling and Boltzmann sampling, only the NSW and IBRION flags were updated as appropriate for single point calculations.

2.3 Dataset limitations

The OMat24 dataset is the largest open dataset of its kind for training DFT surrogate models for materials. However, the dataset has limitations similar to many high-throughput datasets that impact the predictions of models trained using the dataset. OMat24 is calculated with PBE and PBE+U levels of DFT, which includes inherent errors in their approximation and resulting calculations⁴² that are addressed to some extent in other functionals such as PBEsol⁴³, SCAN⁴⁴, R2SCAN⁴⁵ or hybrid functionals. The OMat24 dataset includes only periodic bulk structures and excludes important effects from point defects, surfaces, non-stoichiometry, and lower dimensional structures. Finally, the OMat24 dataset includes a small fraction of structural relaxations (45,000 total relaxations) starting from distorted relaxed structures in the Alexandria dataset, and does not provide additional or novel information about stable structures.

Note the calculations in OMat24 differ from those found in the Materials Project PBE and PBE+U calculations. Care must be taken when mixing calculations for analysis or training models. Although the difference in settings is small (the pseudopotential in version 5.4 and the choice of pseudopotential for Yb and W), predictions of total and formation energies differ. To illustrate this, we compare calculated energies and formation energies for MP settings and OMat24 settings using calculations in the WBM dataset⁴⁶. To understand the impact of these changes, we used the original WBM calculations, which were done with parameters fully compatible with MP calculations⁴⁶. They are compared to single-point calculations of the same relaxed structures with the OMat24 DFT settings. In order to compute formation energies, we computed elemental references and fit anion and PBE/PBE + U corrections following the methodology used in the MATERIALS2020COMPATIBILITY class in PYMATGEN. The mean absolute error between WBM calculations and those we computed with OMat24 DFT settings is 52.25 meV/atom. Figure 3 shows parity plots for DFT calculated total energy and predicted formation energies for 240,000 compounds from the WBM dataset.

2.4 OMat24 train, validation, and test splits

OMat24 is divided into several splits to ensure consistent training and evaluation by the community. Training and validation splits are released to allow for model development and iteration. The test set is divided into four different splits. The first split (WBM Test) is to ensure that the training dataset does not overlap with the Matbench Discovery leaderboard³⁰ created from the WBM dataset⁴⁶. The other three splits measure the accuracy of the models on in-domain training data (ID) and the ability of the models to generalize to out-of-distribution compositions (OOD-Composition) and elemental compositions (OOD-Element).

The WBM Test split was created using the AFLOW structure prototype labels⁴⁷ in the AVIARY package⁴⁸. The prototype label of a structure is a standardized way to classify crystal structures by elemental stoichiometry, space group, Pearson symbol, and Wyckoff positions⁴⁷. The split includes all OMat24 structures that were generated starting from an Alexandria relaxed structure with a prototype label matching any of prototype labels from the initial or relaxed structures included in the WBM dataset. Additionally, all OMat24 structures with a prototype label matching an initial or relaxed WBM structure are also included in the WBM Test split. The WBM Test split includes a total of 5.3 million structures. Note, filtering the dataset and creating this test split is important to ensure that there was no inadvertant data leakage from the training data to the final Matbench Discovery results, as there is overlap in materials between the Alexandria and WBM datasets.

The OOD-Composition split is constructed by picking approximately $\sim 5,000$ unique elemental compositions and adding all structures with matching compositions to the split. This OOD-Composition test split includes 573,000 total structures. The OOD-Elemental split is made by picking $\sim 3,000$ unique element combinations and retrieving all structures matching these element combinations. This resulted in 619,000 total structures included in the OOD-Elemental split.

The training, validation and ID test splits includes all remaining OMat24 structures after creating the 3 test splits described above, and includes a total of 111 million structures. We randomly split this dataset into a training, validation and ID test split containing 100 million, 5 million and 5 million structures respectively for the model training in this work, included in Table 1.

Table 1 Size of the OMat24 train, validation and test dataset splits.

Split	Size	Fraction %
Train	100,824,585	85.3
Validation	5,320,549	4.5
WBM Test	5,373,339	4.5
ID Test	5,453,320	4.6
OOD Composition Test	573,301	0.5
OOD Element Test	619,021	0.5

3 OMat24 models and training strategies

Progress in artificial intelligence and deep learning has led to the development of models that can efficiently and accurately predict and simulate materials properties^{12,22,29,48–51}. Recently, Graph Neural Network (GNN) machine learning potentials have surpassed the accuracy of other ML models when predicting and simulating mechanical properties^{30,52,53}. All of the top models on the OC20 leaderboard⁵⁴, a dataset with similar size and diversity to OMat24, are GNNs. Similarly, for the task of predicting ground-state formation energy and energy above the convex hull—as a proxy for *materials discovery*—GNN interatomic potentials have surpassed all other methodologies³⁰ and have set a new standard in terms of scale (number of different materials) and accuracy (predicted energy errors)^{21–23,50,51}.

We leverage the OMat24 dataset along with the MPtrj²⁹ and Alexandria³² datasets to train GNNs. Since similar structures exist in the Alexandria and the WBM dataset used for testing, we subsampled the Alexandria dataset for training to ensure there was no leakage between the training and testing datasets. The new subset of Alexandria (sAlexandria) was created by removing all trajectories in which any structure matched a

structure in the WBM initial and relaxed structures using structure prototype labels⁴⁷. Next, we reduced the size of the dataset by removing all structures with energies > 0 eV, forces norm > 50 eV/Å, and stress > 80 GPa. Finally, we only sampled structures in the remaining trajectories that had a energy difference greater than 10 meV/atom. The resulting datasets used for training and validation had 10 million and 500 thousand structures respectively.

For model architectures, we focused solely on EquiformerV2²⁰ since it is currently the top performing model on the OC20¹, OC22²⁶ and ODAC23⁴ leaderboards among the many contributions from the broader scientific community. The models are trained to predict energy, forces and stress given an input structure. The models are optimized for relaxed energy prediction for aiding materials discovery by directly predicting forces and stress, instead of relying on energy derivatives. For model training, we explored three strategies:

1. EquiformerV2 models trained solely on the OMat24 dataset, with and without denoising augmentation objectives. These are the models with the most physical meaning as they are fit solely on datasets containing important updates to the underlying pseudopotentials relative to the legacy Materials Project settings.
2. EquiformerV2 models trained solely on the MPtrj dataset, with and without denoising augmentation objectives, useful for direct comparison on the Matbench Discovery leaderboard (denoted “compliant” models).
3. EquiformerV2 models from (1) or OC20 checkpoints further fine-tuned on either the MPtrj or sAlexandria datasets, leading to the highest performing models for the Matbench Discovery leaderboard (denoted “non-compliant”).

In each case, several model sizes were chosen. Table 2 lists the total number of parameters for the models trained. Additional model specifications, training information and parameters are given in Section B, Table 7 and Table 8. Similar to many of the leading OC20 models and the recently released ORB models²³, we use a non-conservative model here with separate energy, force, and stress heads for GPU-efficient training.

Table 2 Total number of parameters and their inference throughput in the EquiformerV2 models in this work. Throughput evaluated on Nvidia A100 GPUs with batch size 1 and no inference-time optimization with samples from the MPtrj dataset.

Model	# of Parameters	Throughput
		(Samples / GPU sec. (MPtrj))
eqV2-S (small)	31,207,434	9.4
eqV2-M (medium)	86,589,068	7.4
eqV2-L (large)	153,7698,68	4.9

We also explored the role of including auxiliary denoising objectives for training models with different dataset types and sizes. We use the Denoising Non-equilibrium Structures (DeNS) protocol⁵⁵ to train EquiformerV2 models on the OMat24 dataset and the MPtrj relaxation dataset. In addition to predicting energy, forces, and stress given input structures, DeNS introduces data augmentation by letting the model predict the energy and added noise from perturbed input structures and unperturbed forces.

4 Results

We include results for each of the three training strategies described above (training models from scratch on MPtrj, training models from scratch on OMat24, and fine-tuning OMat24 or OC20-trained models on MPtrj and Alexandria). OMat24 pre-training is performed using only the 100M training split to avoid dataset contamination with WBM.

We provide all of the OMat24 training data⁵⁶ with a Creative Commons 4.0 license, and the necessary source code⁵⁷ and model weights⁵⁸ with a permissive open source license (with some geographic and acceptable use restrictions) necessary to reproduce our results.

We evaluate the models on the popular Matbench-Discovery benchmark³⁰. The benchmark provides a useful set of metrics to measure the impact and progress that machine learning potentials have in accelerating materials discovery. Specifically it evaluates the task of predicting ground-state (0 K) thermodynamic stability, which examines a material’s decomposition energy (stability) with respect to all possible competing phases⁵⁹. Predicting the energy above the convex hull with DFT level accuracy is a reliable proxy for thermodynamic stability predictions^{52,59,60}.

When evaluated on Matbench-Discovery, we observe the EquiformerV2 model achieves state-of-the-art performance on the leaderboard (with F1 being the primary metric) for both compliant (trained using MPtrj only) and non-compliant (trained with additional data) benchmarks. When training on MPtrj only, our best model achieves substantial performance improvement from DeNS³⁴. Because MPtrj is a relatively small dataset with relaxation trajectories only, we interpret the improved performance as a result of the effective data augmentation through the noisy input structures that DeNS provides. When training on both the OMat24 and MPtrj datasets, we observe that pre-training on OMat24 (a dataset with high composition diversity and predominantly non-equilibrium structures) and fine-tuning on MPtrj and sAlexandria substantially improves model performance to an F1 score of 0.916 and an energy MAE of 20 meV/atom - the top scores for non-compliant models.

4.1 Models trained solely on OMat24

We trained three models of different sizes listed in Table 2 using the OMat24 training data for a total of two epochs. Tables 3 and 4 list validation and test mean absolute error (MAE) for predicted energy, forces and stress. The ID and OOD test splits result in similar results demonstrating the ability of the models to generalize to new compositions and elemental combinations. Interestingly, the WBM split performs worse, and is likely due to the materials in the split having greater material diversity than the OOD splits. When training eqV2-S with DeNS, the results are slightly worse with a energy MAE of 11.3 meV/atom and force MAE of 52.0 meV/Å, suggesting that due to the scale and diversity in the OMat24 dataset, denoising does not provide additional training improvements.

Table 3 Validation mean absolute error metrics of equiformerV2 models trained on the OMat24 dataset. Energy errors are in units meV/atom, forces errors are in meV/Å and stress errors are in meV/Å³.

Model	Energy ↓	Forces ↓	Stress ↓	Forces cos ↑
eqV2-S	11	49.2	2.4	0.985
eqV2-M	10	44.8	2.3	0.986
eqV2-L	9.6	43.1	2.3	0.987

Table 4 Test mean absolute error metrics of equiformerV2 models trained on OMat24 dataset. Energy errors are in units meV/atom, forces errors are in meV/Å and stress errors are in meV/Å³.

Model	WBM test			ID test			OOD composition test			OOD element test		
	energy	forces	stress	energy	forces	stress	energy	forces	stress	energy	forces	stress
eqV2-S	15.96	50.68	3.69	12.09	51.14	2.73	11.05	48.85	2.53	9.34	49.14	2.16
eqV2-M	14.87	46.26	3.61	11.09	46.62	2.64	9.98	44.51	2.45	8.83	44.71	2.06
eqV2-L	14.57	44.72	3.57	10.70	44.88	2.60	9.70	43.00	2.42	8.38	43.08	2.04

4.2 Models trained from scratch on MPtrj as “compliant” Matbench-Discovery models

We use models trained on the MPtrj dataset as a baseline for performance. By training only on the smaller MPtrj dataset of relaxation trajectories, we can determine how much pre-training with the OMat24 and OC20 datasets can improve model performance. These baseline models also fall into the *compliant* model category in the Matbench-Discovery benchmark, which allows a disciplined comparison of our EquiformerV2 models to other architectures.

We trained models using the original MPTrj dataset²⁹, removing all structures in which all the atoms are farther than 12 Å apart. We used the uncorrected total energy, forces and stress as labels. For model size, we trained a small variant (eqV2-S in Table 2). We also trained small, medium and large (eqV2-S, eqV2-M, and eqV2-L) models with DeNS³⁴. More details on model and training hyper-parameters are summarized in Table 7 and Table 8.

Table 5 shows the results of the compliant models on the Matbench-Discovery benchmark. The EquiformerV2-S model outperforms prior models with an F1 score of 0.77. With the addition of DeNS, the F1 scores significantly improve to 0.815 with the smaller model. For the larger model, the regularization provided by DeNS enables effective training even with the smaller MPTrj dataset to achieve the highest F1 score of 0.823 for compliant models. The energy MAE does not vary significantly (35–36 meV/atom) for different model sizes, and indicates that the smaller model may be the most useful in practice.

Table 5 Matbench-Discovery benchmark results of compliant models trained only on MPTraj with results on the unique prototype split. Mean absolute error (MAE) and Root mean squared error (RMSE) are in units of eV/atom.

Model	eqV2-L-DeNS	eqV2-M-DeNS	eqV2-S-DeNS	eqV2-S	ORB MPTrj	SevenNet	MAE
F1 ↑	0.823	0.818	0.815	0.77	0.765	0.724	0.669
DAF ↑	5.184	5.109	5.042	4.64	4.702	4.252	3.777
Precision ↑	0.792	0.781	0.771	0.709	0.719	0.65	0.577
Recall ↑	0.856	0.858	0.864	0.841	0.817	0.818	0.796
Accuracy ↑	0.944	0.942	0.941	0.926	0.922	0.904	0.878
TPR ↑	0.856	0.858	0.864	0.841	0.817	0.818	0.796
FPR ↓	0.041	0.044	0.047	0.063	0.059	0.081	0.107
TNR ↑	0.959	0.956	0.953	0.937	0.941	0.919	0.893
FNR ↓	0.144	0.142	0.136	0.159	0.183	0.182	0.204
MAE ↓	35	35	36	42	45	48	57
RMSE ↓	82	82	85	87	91	92	101
R2 ↑	0.802	0.803	0.788	0.778	0.756	0.75	0.697

4.3 Models pre-trained on OMat24 or OC20, and fine-tuned for Matbench-Discovery

Table 6 shows the resulting metrics for the models pre-trained with the OMat24 and OC20 datasets and fine-tuned on MPTrj or sAlex and MPTrj jointly, and other non-compliant models on the Matbench-Discovery leaderboard. Additional results for the total and 10K most stable predictions splits are listed in Supplementary section D. Several trends can be observed. Pre-training on the OMat24 leads to significantly better results on almost all metrics, and significantly outperforms previous approaches on the F1 (0.916) and energy MAE (20 meV / atom) metrics. Fine-tuning on both sAlexandria and MPTrj outperforms just fine-tuning on MPTrj. Pre-training on OC20, a dataset not intended for training models for materials, provides surprisingly strong results once fine-tuned with MPTrj. Larger models (eqV2-M) do outperform the smaller models (eqV2-S), but depending on the applications the accuracy improvements may not be worth the additional compute cost. Finally, we did train models using DeNS for both pre-training and fine-tuning, but did not see an improvement in accuracy. We hypothesize this was due to the diversity of the larger datasets not requiring the additional regularization provided by DeNS.

5 Conclusions and future directions

In summary, the EquiformerV2 model trained from scratch on MPTrj is state-of-the-art for “compliant” models on MatBench Discovery with a remarkable 35 meV/atom MAE. When models are pre-trained on OMat24, the results improve substantially with an energy above hull MAE of 20 meV/atom on MatBench Discovery and are the first models of any kind to reach the F1=0.9 threshold. In practice, the models trained from scratch on OMat24 are likely to be even more useful as they have energy MAE of only 10 meV/atom on the OMat24 test splits despite most of the structures being at elevated temperatures.

We are encouraged by the results of applying denoising objectives for improved model accuracy and data

Table 6 Matbench-Discovery benchmark results of non-compliant models on the unique prototype split. Mean absolute error (MAE) and Root mean squared error (RMSE) are in units of eV/atom.

Model Pre-train Dataset Fine-tune Dataset	eqV2-M OMat MPtrj-sAlex	eqV2-M OMat MPtrj	eqV2-S OMat MPtrj-sAlex	eqV2-S OMat MPtrj	eqV2-L OC20 MPtrj	eqV2-S OC20 MPtrj	ORB	MatterSim	GNoME
F1 ↑	0.916	0.909	0.901	0.89	0.86	0.837	0.88	0.859	0.829
DAF ↑	6.040	5.948	5.902	5.752	5.639	5.392	6.041	5.646	5.523
Precision ↑	0.923	0.909	0.902	0.879	0.862	0.824	0.924	0.863	0.844
Recall ↑	0.91	0.909	0.9	0.901	0.858	0.849	0.841	0.856	0.814
Accuracy ↑	0.974	0.973	0.97	0.966	0.957	0.951	0.965	0.957	0.955
TPR ↑	0.91	0.909	0.9	0.901	0.858	0.849	0.841	0.856	0.814
FPR ↓	0.014	0.017	0.018	0.023	0.025	0.033	0.013	0.025	0.028
TNR ↑	0.986	0.983	0.982	0.977	0.975	0.967	0.987	0.975	0.972
FNR ↓	0.09	0.091	0.1	0.099	0.142	0.151	0.159	0.144	0.186
MAE ↓	20	21	24	26	29	33	28	26	35
RMSE ↓	72	72	80	81	78	80	77	80	85
R2 ↑	0.848	0.849	0.811	0.807	0.823	0.81	0.824	0.812	0.785

efficiency. Applying strategies developed and honed on the OC20 dataset⁵⁵, we obtain strong results when trained solely on the MPtrj dataset. Perhaps unsurprisingly these denoising objectives become less important in the limit of very large well sampled datasets, with no improvements when training on OMat24. We believe that the observed transferability of model performance from OC20 to OMat24 is yet another strong indication that model training strategies developed for very large simulation datasets are likely transferable to other large dataset challenges (similar to correlations seen between OC20, OC22, and ODAC23).

Perhaps most importantly, the release of the OMat24 dataset, one of the largest open DFT datasets, and the models trained in this work, will provide a valuable resource for future model and dataset development. We hope the community builds upon our efforts, driving further advancements in the capabilities of ML models in material science. This work builds on a number of open science datasets such as the Materials Project³³ and the Alexandria datasets³².

As the MatBench Discovery benchmark approaches saturation (the best models are now F1>0.9), it is essential to develop newer and improved benchmarks that address its inherent limitations. In particular, discrepancies between the PBE functional and experimental results limit the effective value of predictions with respect to experiments. Developing training sets and benchmarks using more accurate DFT functionals such as SCAN and R2SCAN could help overcome these challenges⁶¹. As of release v2022.10.28⁶², the Materials Project now includes R2SCAN calculations for a subset of its data. Large scale datasets similar to OMat24 but with SCAN, R2SCAN, or other functionals are an obvious next step. We hope that the release of OMat24 will enable these efforts to proceed in an even more data-efficient manner.

While our work has focused on materials discovery, there are still many opportunities for exploration beyond this application that are unlocked by OMat24 but have not yet been thoroughly explored. For instance, molecular dynamics (MD) and Monte Carlo (MC) simulations are critical methodologies needed to predict and study properties beyond formation energy at zero Kelvin. Substantial progress and work has been published on the reliability and shortcomings of MD with GNN interatomic potential^{63–65}. Future research could investigate the applicability of models trained using the OMat24 dataset to these areas, where non-equilibrium training data may be even more important.

6 Acknowledgements

We acknowledge helpful discussions with Miguel Marques (Ruhr University Bochum), Janosh Riebesall (Radical AI), Shyue Ping Ong (UCSD), Yi-Lun Liao (MIT), Tess Smidt (MIT) and Matt Uyttendaele. The authors acknowledge valuable assistance from Kyle Michel and Lowik Chanussot (FAIR at Meta) in enabling and scaling the simulation datasets on Meta compute resources. We generated many of our figures using CRYSTAL-TOOLKIT⁶⁶ and PYMATVIZ⁶⁷.

References

- [1] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021.
- [2] C Lawrence Zitnick, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Thibaut Lavril, Aini Palizhati, Morgane Riviere, et al. An introduction to electrocatalyst design using machine learning for renewable energy storage. *arXiv preprint arXiv:2010.09435*, 2020.
- [3] Arunima K Singh, Joseph H Montoya, John M Gregoire, and Kristin A Persson. Robust and synthesizable photocatalysts for co₂ reduction: a data-driven materials discovery. *Nature Communications*, 10(1):443, 2019.
- [4] Anuroop Sriram, Sihoon Choi, Xiaohan Yu, Logan M. Brabson, Abhishek Das, Zachary Ulissi, Matt Uyttendaele, Andrew J. Medford, and David S. Sholl. The open dac 2023 dataset and challenges for sorbent discovery in direct air capture. *ACS Central Science*, 10(5):923–941, 2024.
- [5] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon JL Billinge, et al. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1):59, 2022.
- [6] G. Ceder, Y. M. Chiang, D. R. Sadoway, M. K. Aydinol, Y. I. Jang, and B. Huang. Identification of cathode materials for lithium batteries guided by first-principles calculations. *Nature*, 392(6677):694–696, 1998.
- [7] Mark W. Tibbitt, Christopher B. Rodell, Jason A. Burdick, and Kristi S. Anseth. Progress in material design for biomedical applications. *Proceedings of the National Academy of Sciences*, 112(47):14444–14451, 2015.
- [8] George Crabtree, Sharon Glotzer, Bill McCurdy, and Jim Roberto. Computational materials science and chemistry: Accelerating discovery and innovation through simulation-based engineering and science. Technical report, United States, 2010.
- [9] Timothy Erps, Michael Foshey, Mina Konaković Luković, Wan Shou, Hanns Hagen Goetzke, Herve Dietsch, Klaus Stoll, Bernhard von Vacano, and Wojciech Matusik. Accelerated discovery of 3d printing materials using data-driven multiobjective optimization. *Science advances*, 7(42):eabf7435, 2021.
- [10] Shulin Luo, Tianshu Li, Xinjiang Wang, Muhammad Faizan, and Lijun Zhang. High-throughput computational materials screening and discovery of optoelectronic semiconductors. *WIREs Computational Molecular Science*, 11(1):e1489, 2021.
- [11] Elizabeth A. Pogue, Alexander New, Kyle McElroy, Nam Q. Le, Michael J. Pekala, Ian McCue, Eddie Gienger, Janna Domenico, Elizabeth Hedrick, Tyrel M. McQueen, Brandon Wilfong, Christine D. Piatko, Christopher R. Ratto, Andrew Lennon, Christine Chung, Timothy Montalbano, Gregory Bassan, and Christopher D. Stiles. Closed-loop superconducting materials discovery. *npj Computational Materials*, 9(1):181, 2023.
- [12] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.
- [13] Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.
- [14] Ilyes Batatia, David Peter Kovacs, Gregor N. C. Simm, Christoph Ortner, and Gabor Csanyi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

- [15] Ilyes Batatia, Simon Batzner, Dávid Péter Kovács, Albert Musaelian, Gregor NC Simm, Ralf Drautz, Christoph Ortner, Boris Kozinsky, and Gábor Csányi. The design space of e (3)-equivariant atom-centered interatomic potentials. *arXiv preprint arXiv:2205.06643*, 2022.
- [16] 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):2453, 2022.
- [17] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.
- [18] Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ward 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.
- [19] Saro Passaro and C Lawrence Zitnick. Reducing so (3) convolutions to so (2) for efficient equivariant gnns. In *International Conference on Machine Learning*, pages 27420–27438. Proceedings of Machine Learning Research, 2023.
- [20] 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.
- [21] Yutack Park, Jaesun Kim, Seungwoo Hwang, and Seungwu Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 2024.
- [22] Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.
- [23] Orbital Materials. orb-models github repository. <https://github.com/orbital-materials/orb-models>, 2024.
- [24] Peter Eastman, Pavan Kumar Behara, David L Dotson, Raimondas Galvelis, John E Herr, Josh T Horton, Yuezhi Mao, John D Chodera, Benjamin P Pritchard, Yuanqing Wang, Gianni De Fabritiis, and Thomas E. Markland. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023.
- [25] 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):134, 2020.
- [26] 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.
- [27] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.
- [28] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.
- [29] 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.
- [30] Janosh Riebesell, Rhys EA Goodall, Anubhav Jain, Philipp Benner, Kristin A Persson, and Alpha A Lee. Matbench discovery—an evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920*, 2023.

- [31] 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.
- [32] Jonathan Schmidt, Tiago FT Cerqueira, Aldo H Romero, Antoine Loew, Fabian Jäger, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, page 101560, 2024.
- [33] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), 2013.
- [34] Yi-Lun Liao, Tess Smidt, and Abhishek Das. Generalizing denoising to non-equilibrium structures improves equivariant force fields. *arXiv preprint arXiv:2403.09549*, 2024.
- [35] Daniel Schwalbe-Koda, Aik Rui Tan, and Rafael Gómez-Bombarelli. Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nature Communications*, 12(1):5104, 2021.
- [36] Kangming Li, Daniel Persaud, Kamal Choudhary, Brian DeCost, Michael Greenwood, and Jason Hattrick-Simpers. Exploiting redundancy in large materials datasets for efficient machine learning with less data. *Nature Communications*, 14(1):7283, 2023.
- [37] 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.
- [38] 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.
- [39] 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*, 68:314–319, 2013.
- [40] Materials Project. Issue #2968: Add support for new crystal structure prediction method. <https://github.com/materialsproject/pymatgen/issues/2968>, 2023.
- [41] Materials Project. Issue #3016: Add support for new crystal structure prediction method. <https://github.com/materialsproject/pymatgen/issues/3016>, 2023.
- [42] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865, 1996.
- [43] John P Perdew, Adrienn Ruzsinszky, Gábor I Csonka, Oleg A Vydrov, Gustavo E Scuseria, Lucian A Constantin, Xiaolan Zhou, and Kieron Burke. Restoring the density-gradient expansion for exchange in solids and surfaces. *Physical review letters*, 100(13):136406, 2008.
- [44] Jianwei Sun, Adrienn Ruzsinszky, and John P Perdew. Strongly constrained and appropriately normed semilocal density functional. *Physical Review Letters*, 115(3):036402, 2015.
- [45] James W Furness, Aaron D Kaplan, Jinliang Ning, John P Perdew, and Jianwei Sun. Accurate and numerically efficient r2scan meta-generalized gradient approximation. *The Journal of Physical Chemistry Letters*, 11(19):8208–8215, 2020.
- [46] Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Predicting stable crystalline compounds using chemical similarity. *npj Computational Materials*, 7(1):12, 2021.
- [47] Michael J Mehl, David Hicks, Cormac Toher, Ohad Levy, Robert M Hanson, Gus Hart, and Stefano Curtarolo. The aflow library of crystallographic prototypes: part 1. *Computational Materials Science*, 136:S1–S828, 2017.
- [48] Rhys EA Goodall, Abhijith S Parackal, Felix A Faber, Rickard Armiento, and Alpha A Lee. Rapid discovery of stable materials by coordinate-free coarse graining. *Science Advances*, 8(30):eabn4117, 2022.
- [49] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

- [50] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, William J Baldwin, Noam Bernstein, et al. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096*, 2023.
- [51] 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.
- [52] Christopher J Bartel, Amalie Trewartha, Qi Wang, Alexander Dunn, Anubhav Jain, and Gerbrand Ceder. A critical examination of compound stability predictions from machine-learned formation energies. *njp Computational Materials*, 6(1):97, 2020.
- [53] Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Lio, Yoshua Bengio, and Michael Bronstein. A hitchhiker’s guide to geometric gnns for 3d atomic systems. *arXiv preprint arXiv:2312.07511*, 2023.
- [54] Meta Fundamental AI Research and collaborators. The oc20 leaderboard. <https://opencatalystproject.org/leaderboard.html>, 2024.
- [55] Yi-Lun Liao, Tess Smidt, and Abhishek Das. Generalizing denoising to non-equilibrium structures improves equivariant force fields. *arXiv preprint arXiv:2403.09549*, 2024.
- [56] Meta Fundamental AI Research. The omat24 dataset. <https://huggingface.co/datasets/fairchem/OMAT24>, 2024.
- [57] Meta Fundamental AI Research and Collaborators. The fair chemistry (fairchem) model repository. <https://github.com/FAIR-Chem/fairchem>, 2024.
- [58] Meta Fundamental AI Research. The omat24 trained model checkpoints. <https://huggingface.co/fairchem/OMAT24>, 2024.
- [59] Christopher J Bartel. Review of computational approaches to predict the thermodynamic stability of inorganic solids. *Journal of Materials Science*, 57(23):10475–10498, 2022.
- [60] Wenhao Sun, Stephen T Dacek, Shyue Ping Ong, Geoffroy Hautier, Anubhav Jain, William D Richards, Anthony C Gamst, Kristin A Persson, and Gerbrand Ceder. The thermodynamic scale of inorganic crystalline metastability. *Science Advances*, 2(11):e1600225, 2016.
- [61] Ryan Kingsbury, Ayush S Gupta, Christopher J Bartel, Jason M Munro, Shyam Dwaraknath, Matthew Horton, and Kristin A Persson. Performance comparison of r2-scan and scan metagga density functionals for solid materials via an automated, high-throughput computational workflow. *Physical Review Materials*, 6(1):013801, 2022.
- [62] The Materials Project. Materials project database versions. <https://docs.materialsproject.org/changes/database-versions>, 2024.
- [63] Sina Stocker, Johannes Gasteiger, Florian Becker, Stephan Günnemann, and Johannes T Margraf. How robust are modern graph neural network potentials in long and hot molecular dynamics simulations? *Machine Learning: Science and Technology*, 3(4):045010, 2022.
- [64] Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *arXiv preprint arXiv:2210.07237*, 2022.
- [65] S. Raja, I. Amin, F. Pedregosa, and A. S. Krishnapriyan. Stability-aware training of neural network interatomic potentials with differentiable boltzmann estimators. *arXiv preprint arXiv:2402.13984*, 2024.
- [66] Matthew Horton, Jimmy-Xuan Shen, Jordan Burns, Orion Cohen, François Chabbey, Alex M Ganose, Rishabh Guha, Patrick Huck, Hamming Howard Li, Matthew McDermott, et al. Crystal toolkit: A web app framework to improve usability and accessibility of materials science research algorithms. *arXiv preprint arXiv:2302.06147*, 2023.
- [67] Janosh Riebesell, Haoyu Yang, Rhys Goodall, and Sterling G. Baird. Pymatviz: visualization toolkit for materials informatics, 2022. 10.5281/zenodo.7486816 - <https://github.com/janosh/pymatviz>.

Appendix

A Dataset statistics

Figures 4 shows histograms for the number of atoms per structure in the sub-datasets that make up the OMat24 dataset. Similarly figure 5 shows the distributions of energy, forces and stress in the different sub-datasets. The number of atoms per system is shifted to larger sizes, since these structures were generated by tiling the unit cell of relaxed Alexandria structures. The rest of the distributions are concentrated at smaller sizes, as a direct result of the distribution in the Alexandria dataset shown in Figure 2.

From Figure 5 we can see that the AIMD portion of the dataset has a much narrower force and stress distribution compared to that of structures generated from rattled relaxations and rattled Boltzmann sampling. We also observe a shift to lower energies for structures generated using AIMD.

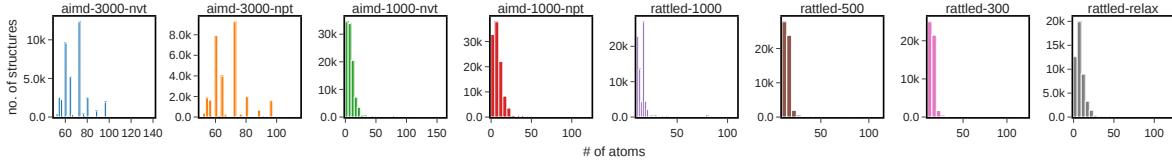


Figure 4 Histogram of number of atoms per structure per sub-dataset in OMat-24 dataset.

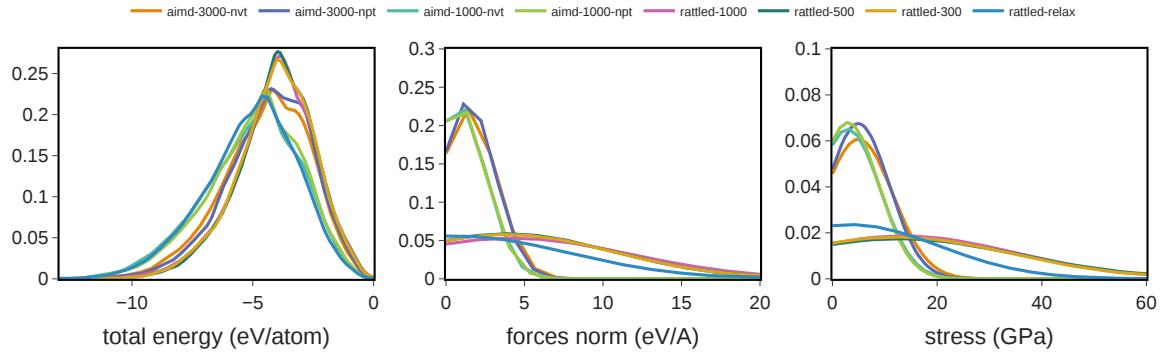


Figure 5 Energy, forces norm and maximum absolute stress densities for all sub-datasets in OMat-24.

Figures 6 and 7 show per element histograms of the number of occurrences in each dataset. The histograms are created by randomly sampling 1% and 10% of each dataset respectively, as a result elements that appear very rarely may have been missed. By comparing Figures 6 and 7 with the corresponding histogram for the OMat24 dataset in Figure 2b we observe that the distributions between OMat24 and Alex are very similar and more uniformly distributed than that of MPtrj.

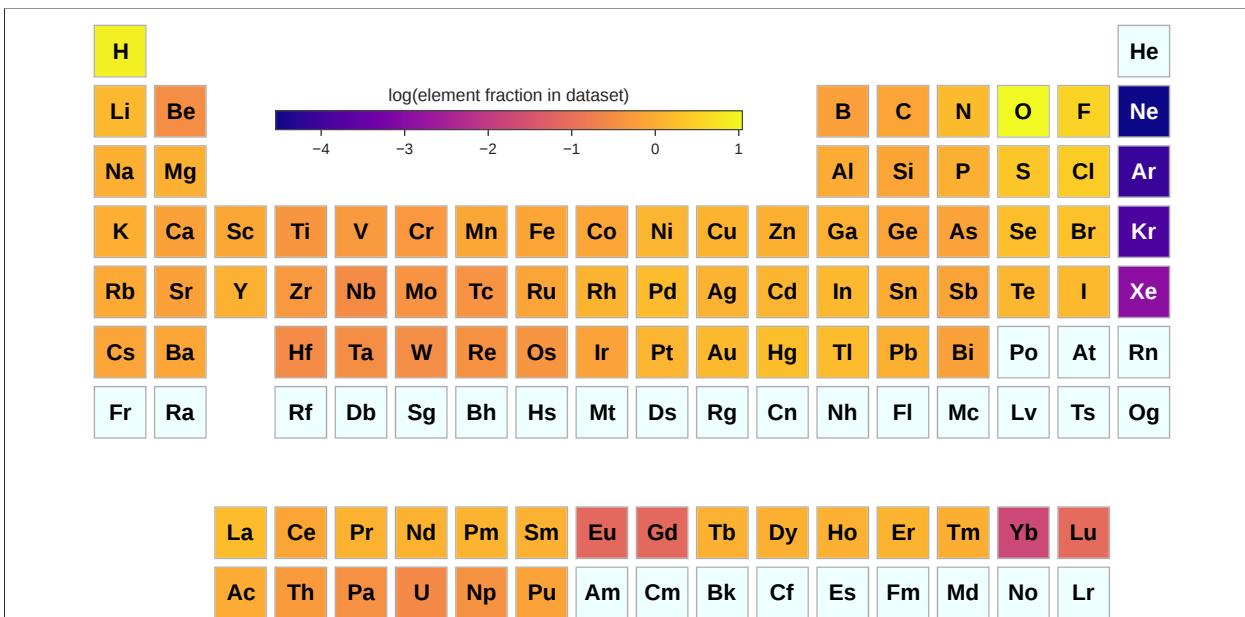


Figure 6 Element distribution in Alexandria bulk PBE dataset³².

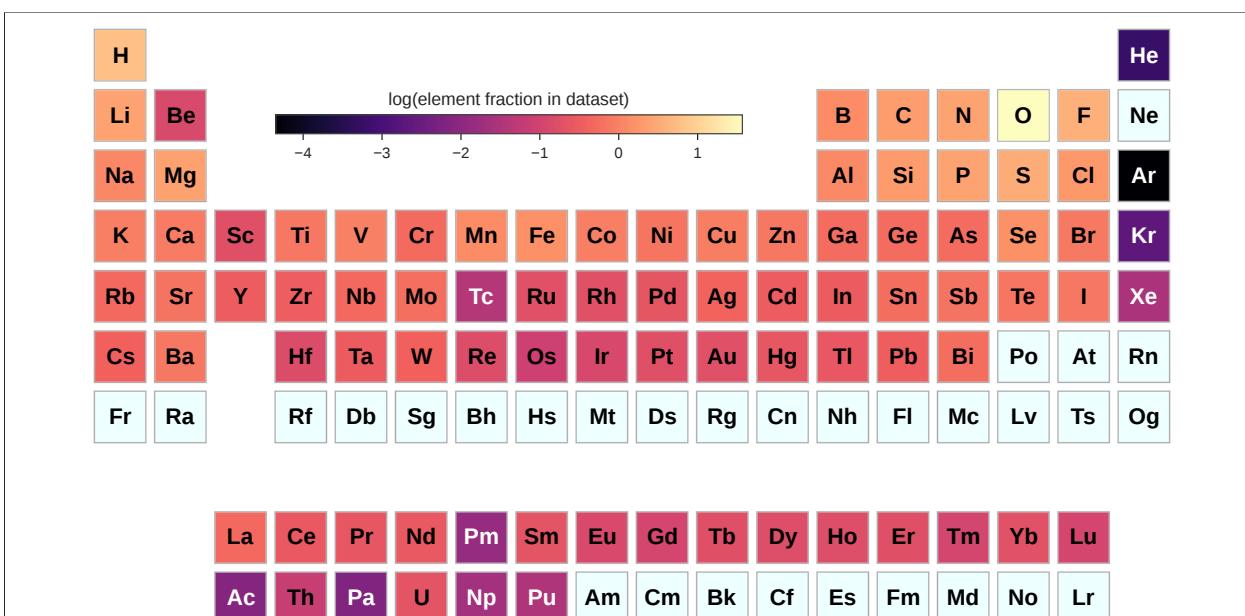


Figure 7 Element distribution in MPtrj bulk dataset²⁹.

B Model training hyper-parameters and configuration

All models trained in this work use the EquiformerV2 architecture²⁰ with separate heads for energy, forces and stress prediction. We use a per-atom MAE loss for energy and an l_2 norm loss for forces. For stress prediction, we decompose the 3×3 symmetric stress tensor to a 1-dimensional isotropic component ($L = 0$) and a 5-dimensional anisotropic component ($L = 2$). We use a prediction head that outputs a scalar and irreps of $L = 2$, then use an MAE loss for the isotropic and the anisotropic component separately. At test time, we recover the stress tensor by combining the isotropic and anisotropic components. For models trained with DeNS, an additional head is added to predict the noise added to a perturbed structure. For a DeNS forward pass, we input the noisy structure and predict the unperturbed energy as well as the noise vectors, then compute a per-atom MAE loss for energy and an MSE loss for the noise vectors. We refer to the readers to the original EquiformerV2 and DeNS papers for more details on model architectures^{20,34}. The hyper-parameters for our models are summarized in Table [table 7](#). The hyper-parameters for training are summarized in Table [table 8](#). We use the same mixed precision strategy as the original EquiformerV paper²⁰ to speed up the training process. Models were trained on 64 NVIDIA A100 GPUs for pre-training and 32 NVIDIA A100 GPUs for fine-tuning with distributed data parallel.

Table 7 Hyper-parameters for the EquiformerV2 models of different sizes. All hyper-parameters for a given model size is used for all dataset settings. We denote the dimension of irreps features as (L_{max}, C) where L_{max} is the maximum degree and C is the number of channels.

Hyper-parameters	eqV2-S	eqV2-M	eqV2-L
Maximum degree L_{max}	4	6	6
Maximum order M_{max}	2	4	3
Number of Transformer blocks	8	10	20
Cutoff radius (Å)	12	12	12
Maximum number of neighbors	20	20	20
Number of radial bases	600	600	600
Dimension of hidden scalar features in radial functions d_{edge}	(0, 128)	(0, 128)	(0, 128)
Embedding dimension d_{embed}	(4, 128)	(6, 128)	(6, 128)
$f_{ij}^{(L)}$ dimension d_{attn_hidden}	(4, 64)	(6, 64)	(6, 64)
Number of attention heads h	8	8	8
$f_{ij}^{(0)}$ dimension d_{attn_alpha}	(0, 64)	(0, 64)	(0, 64)
Value dimension d_{attn_value}	(4, 16)	(6, 16)	(6, 16)
Hidden dimension in feed forward networks d_{ffn}	(4, 128)	(6, 128)	(6, 128)
Resolution of point samples R	18	18	18

Table 8 Hyper-parameters for EquiformerV2 model training for different dataset setting. All model sizes use the same set of hyper-parameters for a given dataset setting.

Hyper-parameters	MPTraj training	OMat training	MPTraj Fine-tuning	MPTraj+sAlex Fine-tuning
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate scheduling	Cosine	Cosine	Cosine	Cosine
Warmup epochs	0.1	0.01	0.1	0.1
Warmup factor	0.2	0.2	0.2	0.2
Maximum learning rate	2×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}
Minimum learning rate factor	0.01	0.01	0.01	0.01
Batch size	512	512	256	256
Number of epochs	150	2	32	8
Gradient clipping norm threshold	100	100	100	100
Model EMA decay	0.999	0.999	0.999	0.999
Weight decay	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}
Dropout rate	0.1	0.1	0.1	0.1
Stochastic depth	0.1	0.1	0.1	0.1
Energy loss coefficient	20	20	20	20
Force loss coefficient	20	20	10	10
Stress loss coefficient	5	5	1	1
DeNS settings				
Probability of optimizing DeNS	0.5	0.25	-	-
Standard deviation of Gaussian noise	0.1	0.1	-	-
DeNS loss coefficient	10	10	-	-

C Validation metrics for models trained and fine-tuned with MPTrj

Tables 9 and 10 list the validation metrics for the models trained or fine-tuned solely on MPTrj.

Table 9 Validation mean absolute error metrics for models trained on the MPTrj dataset. Energy errors are in units meV/atom, forces errors are in meV/Å and stress errors are in meV/Å³.

model	energy ↓	forces ↓	stress ↓	forces cos ↑
eqV2-S	12.4	32.22	1.55	0.72
eqV2-S-DeNS	11.43	31.67	1.44	0.72
eqV2-M-DeNS	11.17	31.46	1.48	0.728
eqV2-L-DeNS	10.58	30.48	1.47	0.738

Table 10 Validation metrics for finetuning OMat pretrained models on MPTrj

model	energy (meV/atom) ↓	forces (meV/Å) ↓	stress (meV/Å ³) ↓	forces cos ↑
eqV2-S-OMat-MP	8.52	23.86	1.3	0.764
eqV2-L-OMat-MP	7.99	22.63	1.28	0.777

D Matbench-Discovery metrics for full and 10k most stable splits

We report the performance of our compliant EquiformerV2 models on the full WBM test set and the 10k materials predicted by each model to be most stable in Table 11 and Table 12. We report the performance of our non-compliant EquiformerV2 models on the full and 10k most stable splits in Table 13 and Table 13.

Notably the best models—those fine-tuned jointly with MPTrj and sAlex—perform worse on the 10,000 materials predicted to be most stable. We hypothesize that this is a result of undertraining and/or the sAlex data overwhelming that in MPTrj.

Table 11 Matbench-Discovery benchmark results of compliant models trained only on MPTraj with results on all of the WBM test set. Mean absolute error (MAE) and Root mean squared error (RMSE) are in units of eV/atom.

Model	eqV2-L-DeNS	eqV2-M-DeNS	eqV2-S-DeNS	eqV2-S
F1 ↑	0.806	0.8	0.798	0.758
DAF ↑	4.497	4.414	4.362	4.053
Precision ↑	0.772	0.757	0.748	0.696
Recall ↑	0.844	0.847	0.855	0.833
Accuracy ↑	0.931	0.929	0.927	0.912
TPR ↑	0.844	0.847	0.855	0.833
FPR ↓	0.052	0.056	0.059	0.076
TNR ↑	0.948	0.944	0.941	0.924
FNR ↓	0.156	0.153	0.145	0.167
MAE ↓	34	34	35	41
RMSE ↓	81	81	84	85
R2 ↑	0.798	0.8	0.785	0.777

Table 12 Matbench-Discovery benchmark results of compliant models trained only on MPTraj with results on the 10K materials predicted to be most stable. Mean absolute error (MAE) and Root mean squared error (RMSE) are in units of eV/atom.

Model	eqV2-L-DeNS	eqV2-M-DeNS	eqV2-S-DeNS	eqV2-S
F1 ↑	0.985	0.984	0.983	0.974
DAF ↑	6.347	6.33	6.326	6.21
Precision ↑	0.97	0.968	0.967	0.949
Accuracy ↑	0.97	0.968	0.967	0.949
MAE ↓	30	28	31	37
RMSE ↓	91	79	91	94
R2 ↑	0.821	0.865	0.823	0.812

Table 13 Matbench-Discovery benchmark results of non-compliant models on all of the WBM test set. Mean absolute error (MAE) and Root mean squared error (RMSE) are in units of eV/atom.

Model	eqV2-M	eqV2-M	eqV2-S	eqV2-S	eqV2-L	eqV2-S
Pre-train Dataset	OMat	OMat	OMat	OMat	OC20	OC20
Fine-tune Dataset	MPtrj-sAlex	MPtrj	MPtrj-sAlex	MPtrj	MPtrj	MPtrj
F1 ↑	0.895	0.887	0.88	0.868	0.84	0.817
DAF ↑	5.24	5.143	5.106	4.942	4.874	4.661
Precision ↑	0.899	0.882	0.876	0.848	0.836	0.8
Recall ↑	0.892	0.892	0.884	0.888	0.843	0.835
Accuracy ↑	0.964	0.962	0.959	0.954	0.946	0.938
TPR ↑	0.892	0.892	0.884	0.888	0.843	0.835
FPR ↓	0.021	0.025	0.026	0.033	0.034	0.043
TNR ↑	0.979	0.975	0.974	0.967	0.966	0.957
FNR ↓	0.108	0.108	0.116	0.112	0.157	0.165
MAE ↓	0.02	0.021	0.024	0.025	0.028	0.032
RMSE ↓	0.071	0.071	0.079	0.08	0.076	0.079
R2 ↑	0.843	0.843	0.809	0.804	0.819	0.807

Table 14 Matbench-Discovery benchmark results of non-compliant models on the 10K materials predicted to be most stable. Mean absolute error (MAE) and Root mean squared error (RMSE) are in units of eV/atom.

Model	eqV2-M	eqV2-M	eqV2-S	eqV2-S	eqV2-L	eqV2-S
Pre-train Dataset	OMat	OMat	OMat	OMat	OC20	OC20
Fine-tune Dataset	MPtrj-sAlex	MPtrj	MPtrj-sAlex	MPtrj	MPtrj	MPtrj
F1 ↑	0.987	0.987	0.99	0.991	0.989	0.987
DAF ↑	6.368	6.37	6.413	6.424	6.399	6.368
Precision ↑	0.974	0.974	0.98	0.982	0.978	0.974
Accuracy ↑	0.974	0.974	0.98	0.982	0.978	0.974
MAE ↓	17	17	16	17	26	28
RMSE ↓	72	71	61	63	95	91
R2 ↑	0.887	0.889	0.917	0.91	0.806	0.824