

Appendix Table of Contents

A	Training Details and Hyperparameters	2
A.1	Two-stage training	2
A.2	Parallelisms	2
A.3	Model	2
A.4	MPtrj and sAlex Fine-tuning	3
A.5	Training Compute Resources	3
A.6	Referencing and Normalization	3
B	Training Data	4
B.1	Materials	5
B.2	Molecules	5
B.3	Catalysis	5
B.4	Molecular Crystals	5
B.5	Metal-organic frameworks (MOFs)	5
C	Scaling Laws Methods	6
C.1	FLOP counting	6
C.2	Compute optimal fits	6
C.3	Fitting loss vs. parameters	6
D	Inference	7
E	Evaluations	8
E.1	Additional UMA Results	9
E.2	Materials	9
E.3	Catalysis	10
E.4	Molecules	10
E.5	Molecular Crystals	12
E.6	Metal–Organic Frameworks	12
F	Single-task vs Multi-task	13
G	Additional MoLE Analysis	13
G.1	Expert Analysis	13
G.2	Generalization Across Architectures	13

A Training Details and Hyperparameters

A.1 Two-stage training

While conservative models have been found to provide reliable performance in diverse physical property prediction tasks [14], the backward pass required for force/stress prediction significantly increases inference costs. Models with direct force prediction are much more efficient, and have been found to be effective as a pre-training strategy to save compute when subsequently fine-tuned as a conservative model [14, 5]. The UMA-S and UMA-M both follow this procedure. In addition, we introduce low-precision training, max-atom batching, and max neighbor switching to further enhance the scalability and efficiency of our training process. These novel strategies are discussed in turn below. Detailed hyper-parameters are summarized in Table 1.

Precision. For pretraining, we used BF16 numerical format, commonly used in training LLMs but uncommon in MLIP models due high numerical precision requirements. In our experiments, we found that BF16 is significantly more stable than AMP-FP16 (automatic mixed precision) especially in our multi-modal setting where data distributions can vary dramatically, frequently causing gradient and loss spikes that would destabilize AMP training. However, it suffers an accuracy drop compared to AMP-FP16 and FP32. We found the degradation can be nearly completely recovered after a very small number of finetuning steps in FP32 (<1% of data).

Max-atom Batching. Due to the large differences in system topology and number of atoms/edges per system, using a fixed number of systems as batch size is infeasible. Instead we chose to use a max-atom batching scheme where we randomly pack batches that contain as close to an upper bound (max atoms) as possible without going over to guarantee an upper bound on memory usage.

Max Neighbors. For training efficiency purposes, we use a significantly smaller number of neighbors per atom during pretraining and found that it has no effect on the final performance, energy conservation, and smoothness properties of the model after finetuning with effectively “infinite” max neighbors.

A.2 Parallelisms

Although our models are designed with inference efficiency in mind, training models with a large number of MoLE experts is memory intensive. In particular, for the finetuning stages, a combination of infinite neighbors, FP32 precision, and autograd forces puts significant constraints on memory and training speed. We used a combination three parallelism training techniques summarized as follows:

- Graph parallelism (GP) [42]: Partitioning graphs across GPUs during message passing layers is used when scaling up to a large number of atoms at large model sizes. Graph partitioning is only used within a node with a fixed graph parallel rank size (2 or 4) during conservative fine-tuning stages.
- Fully-sharded data parallel (FSDP): We use the Pytorch FSDPv1 implementation on MoLE expert layers only for models with a total parameter count exceeding 1B during conservative fine-tuning when memory is scarce. Parameters are sharded within a node and replicated across nodes.
- Distributed Data Parallel (DDP): We use the standard PyTorch DDP implementation, with modifications for per-atom loss averaging and compatibility with graph parallelism.

Furthermore, we leveraged Pytorch’s Distributed Checkpointing framework to ensure saving and loading extremely large checkpoints is efficient and stable across different node configurations. Exponential moving average (EMA) is used for stable validation performance.

A.3 Model

One model for all tasks. UMA is designed for multi-task learning under diverse DFT settings. For two inputs with exactly the same atomic numbers and positions, the DFT labels will be different when different DFT settings are used. Such DFT settings include the level of theory and system total charge/spin. These task specifications are global information of the atomic system, and we process them through initial embedding layers. In this paper, five levels of theories are involved –

Table 1: Summary of main training-related hyper-parameters for the pre-training and fine-tuning stages. These hyper-parameters are shared among model sizes.

Hyper-parameter	Pre-training	Finetuning
Precision	BF16	FP32
Radius cutoff Å	6	6
Max neighbors	30	300
Force prediction	Direct	Autograd
Stress prediction	None	Autograd
Optimizer	AdamW	AdamW
Learning rate scheduling	Cosine	Cosine
Maximum learning rate	8×10^{-4}	4×10^{-4}
Warmup epochs	0.01	0.01
Warmup factor	0.2	0.2
Gradient clipping norm threshold	100	100
Model EMA decay	0.999	0.999
Weight decay	1×10^{-3}	1×10^{-3}
Energy loss coefficient	10	20
OMol energy loss coefficient	30	-
Force loss coefficient	30	2
Stress loss coefficient	-	1

Table 2: Hyper-parameters for UMA models of different sizes.

Hyper-parameters	UMA-S	UMA-M	UMA-L
Number of MoLE experts	32	32	Dense
Number of layer blocks	4	10	16
Maximum degree L_{\max}	2	4	6
Maximum order M_{\max}	2	2	2
Number of channels N_{channel}	128	128	256
Number of radial basis functions	64	128	256
Global batch size (atoms)	88k	44k	44k
Number of pre-training steps	1.68M	2.08M	2.58M
Number of fine-tuning steps	1M	545k	350k

OMat24, OC20, OMol25, OMC25, and ODAC25 all use different DFT levels of theory. OMol25 further contains systems with non-neutral charge/spin.

For this iteration of the UMA models, these global information are embedded as follows:

Furthermore, to use a single model for all tasks, it is crucial to normalize the labels such that targets from different datasets fall into similar numerical ranges. We specifically design a referencing scheme that brings the diverse datasets to a similar level, detailed in Appendix B. The model hyperparameters are shown in Table 3.

A.4 MPTrj and sAlex Fine-tuning

For materials evaluations, UMA models were fine-tuned on the MPTrj [13] and sAlex [41, 3] datasets to ensure consistent DFT settings. The fine-tuning procedure is the same as eSEN-30M-OAM as documented in [14].

A.5 Training Compute Resources

The resources used to train UMA models are described in Table 4.

A.6 Referencing and Normalization

While each dataset used in this work comes with its own specific set of DFT settings, we wanted a referencing scheme that provides a way to make energy magnitudes comparable across datasets. We

Table 3: Hyper-parameters for UMA models of different sizes.

Hyper-parameters	UMA-S	UMA-M	UMA-L
Number of MoLE experts	32	32	Dense
Number of layer blocks	4	10	16
Maximum degree L_{\max}	2	4	6
Maximum order M_{\max}	2	2	2
Number of channels N_{channel}	128	128	256
Number of radial basis functions	64	128	256
Global batch size (atoms)	88k	44k	44k
Number of pre-training steps	1.68M	2.08M	2.58M
Number of fine-tuning steps	1M	545k	350k

Table 4: Training Times for UMA models.

Model	Stage	GPUs in Parallel	Training Days	GPU-Type
UMA-S	Direct Pre-train	128	5	H200 140GB
UMA-S	Conserve Fine-tune	256	5	H200 140GB
UMA-M	Direct Pre-train	128	14	H200 140GB
UMA-M	Conserve Fine-tune	256	14	H200 140GB
UMA-L	Direct Pre-train	128	25	H100 80GB
UMA-L	Stress Fine-tune	128	4	H100 80GB
UMA-L	FP32 Fine-tune	128	2	H100 80GB

do this through a "heat of formation" (HOF) reference that is applied to the energies:

$$E_{ref} = E_{DFT} - \sum_i^N [E_{i,DFT} - \Delta H_{f,i}]$$

Where E_{DFT} corresponds to the total DFT energy of the system, i is the atom number, N is the total number of atoms in the system, $E_{i,DFT}$ is the DFT energy of an isolated atom i in a box, and $\Delta H_{f,i}$ is the heat of formation of atomic number i as taken directly from Mendelev [33]. $E_{i,DFT}$ was calculated using the DFT settings for each of the unique datasets in this work. Additionally, we apply a linear reference to the above energies to help with the convergence and training stability of our models. We follow the same protocol as described in the OC22 Appendix [47].

We use a custom normalization $x' = \frac{x-\mu}{\sigma}$ for all targets (energy, forces, stress), where the shift term $\mu = 0$ and the scale term $\sigma = \text{Force root mean square (RMS)}$. For the combined dataset the Force RMS is computed as the weighted average (based on the number of systems) of all individual dataset Force RMS values.

B Training Data

A summary of the five datasets used to train UMA is shown in Table 5. In total, the dataset has 459 million training examples, containing up to 350 atoms. The average number of atoms varies based on the dataset, from 19 for OMat24 to 178 for ODAC25.

Table 5: Overview of the five datasets used to train UMA. For each dataset various statistics are provided alongside the sampling ratio used for training.

Dataset	Domain	Training Size	Labels	# Elements	Avg Size	Force RMS	Sampling ratio
OMat24	Materials	100,824,585	E,F,S	89	19	2.83	4
OMol25	Molecules	75,889,983	E,F	83	52	0.985	4
OC20++	Catalysis	229,054,043	E,F	56	77	0.624	1
OMC25	Molecular Crystals	24,870,226	E,F,S	12	130	0.103	2
ODAC25	MOFs	28,517,826	E,F	70	178	0.046	1
Total		459,156,663					

B.1 Materials

The field of inorganic bulk materials is moving at an incredibly fast pace. Here we train on the Open Materials (OMat24) dataset (100M) [3], one of the largest and most diverse datasets in the community. All DFT calculations from this domain were run with VASP [26, 25, 27, 28] and used the PBE [37] functional. Due to the differences in pseudopotential version and different pseudopotentials for certain elements in OMat24 and Materials Project [3] calculation settings used for the data in most third party benchmarks, finetuning was also performed on MPtrj [13] and subsampled Alexandria (sAlex) [41] to ensure consistent evaluation on the materials benchmarks.

B.2 Molecules

The community has seen dozens of molecular datasets spanning different scales for a variety of applications []. However, the varying levels of DFT theory and quality makes it challenging to unify under a single model. The release of the Open Molecules 2025 (OMol25) dataset [30] helps address this by providing the largest single dataset (100M+) spanning 80+ elements covering metal-complexes, biomolecules, electrolytes, and several existing datasets under a single, high-quality level of theory. All DFT calculations were performed using Orca [34] (ω B97M-V/def2-TZVPD). At the time of training, only 75M samples from OMol25 were available for use, and we refer to this as OMol-preview. Splits were constructed to ensure that this snapshot of the dataset is consistent with the full dataset release. All ablations and results were trained with this OMol-preview, unless stated otherwise. Released models, however, will be retrained with the full OMol25 dataset to ensure the best models are accessible by the community.

B.3 Catalysis

The Open Catalyst (OC20) dataset [9] provides the largest adsorbate+surface dataset in the community. OC20 enumerates 1M+ unique surface + adsorbate combinations, spanning 55 elements, and runs local geometry optimizations. Here, we train on the OC20 All (133M), MD (38M), and Rattled (17M) datasets. Unlike prior work, we also leverage OC20’s clean surface data (14M) since models here are trained on total energies. One limitation of OC20 is that it only contains single adsorbates that interact with a surface. To address this, we introduce the OC20-Multi-Adsorbate (mAds) dataset (22M) to better capture coverage effects and adsorbate-adsorbate interactions. All DFT calculations were performed using VASP [26, 25, 27, 28] with the RPBE exchange-correlation functional [18].

B.4 Molecular Crystals

The most recent 7th Crystal Structure Prediction (CSP) Blind Test organized by Cambridge Crystallographic Data Center (CCDC) demonstrated the effectiveness of tailored machine learning interatomic potentials (MLIPs) in predicting, filtering, and ranking molecular crystal structures [21, 20]. However, despite the widespread applications of molecular crystals, there has been limited focus on developing universal MLIPs for molecular crystals, mostly because of the scarcity of publicly available datasets. Currently, publicly available datasets of molecular crystals are scarce, with at most 60,000 materials represented [1, 7]. To address this data gap, we use the Open Molecular Crystals (OMC25) dataset, which comprises 25 million molecular crystal structures. The dataset includes multiple relaxation trajectories of various molecular crystal packings generated with Genarris [46] starting from organic molecules from the OE62 dataset [45]. The dataset includes 12 elements (all elements from OE62, excluding Li, As, Se, and Te) and the maximum number of atoms is capped at 300. The dataset is computed using VASP [26, 25, 27, 28] with the PBE exchange-correlation functional [37] and D3 dispersion correction [17]. The OMC25 dataset, along with in-depth details on data and polymorph structure generation, will be released in an upcoming publication [16].

B.5 Metal-organic frameworks (MOFs)

The Open Direct Air Capture 2025 (ODAC25) dataset represents the largest MOF dataset (78M) for DAC applications [44]. ODAC25 extends the earlier ODAC23 [43] dataset, and is derived from the CoREMOF [10, 11] dataset. ODAC25 focuses on the adsorption and co-adsorption of CO₂ and H₂O in MOFs among other adsorbates, which is critical for DAC performance evaluation. This work uses

the subset of ODAC25 that overlaps with ODAC23, where all DFT-computed adsorption energies have been upgraded to a higher k-points sampling grid density, providing improved accuracy over ODAC23 (29M calculations). All DFT calculations were performed using VASP [26, 25, 27, 28] with the PBE exchange-correlation functional [37] and D3 dispersion correction [17].

C Scaling Laws Methods

The scaling law experiments were only performed on pretraining; due to compute constraints, we did not study the effect on finetuning with energy conservation. We used an 8-expert MoLE version of the model with the UMA-M settings ($l_{max} = 4$, $m_{max} = 2$, and $n_{neighbors} = 30$) for consistency.

C.1 FLOP counting

We approximate our training FLOPs by

$$C(N, D) \approx \kappa ND, \quad (1)$$

where N is the total number of parameters in the network, and D is the number of inputs (tokens for LLMs and atoms or edges for MLIPs) computed on. This relationship holds for any network that is dominated by linear layers where κ indicates how many times a parameters is re-used on an input. For a single forward pass of a linear network, $\kappa = 2$. A full training cycle for such a network with a single backward pass brings $\kappa = 6$ as we need to compute the gradient with respect to both the parameters and inputs. Thus, for most LLMs $\kappa = 6$ FLOPs/parameter/token [23] for a single forward and backwards pass where D is in units of tokens.

For our edge-based SO2 equivariant networks [36], κ is a function of l_{max} and m_{max} spherical harmonic orders and the number of edges per atom $n_{neighbors}$. For the scaling law experiments, we used $l_{max} = 4$, $m_{max} = 2$, and $n_{neighbors} = 30$ which corresponds to the settings of UMA-M model, resulting in $\kappa \approx 270$ FLOPs/parameter/atom (or 9 FLOPs/parameter/edge) per training step, which can be computed or experimentally determined. As long as the number of input edges and parameters are sufficiently large (which holds for UMA models), this flop approximation holds as contributions from all other operators are marginal. We also verified this assumption with direct FLOP counting in our model code.

Overall, parameter reuse is significantly higher in Equivariant-GNNs compared to LLMs, and hence the flop count is 2-3 orders of magnitude higher for a similar parameter-sized LLM network.

C.2 Compute optimal fits

The compute optimal model and dataset sizes can then be fitted to power laws [23, 19]:

$$\begin{aligned} \log(N^*(C)) &= \alpha \log(C) + A \\ \log(D^*(C)) &= \beta \log(C) + B \end{aligned} \quad (2)$$

Where C is the compute in FLOPs described in Appendix C.1. $N^*(C)$ represents the optimal model size (in parameters) as a function of compute, and $D^*(C)$ represents the optimal dataset size (in units of atoms). $N^*(C)$ is determined by finding the minima of fitted parabolas for each Isoflop curve. The 10% and 90% percentile bootstrap errors are shown in Figure 3(e) in the main text. α and β are the scaling coefficients w.r.t. model size and dataset size respectively. A and B are offset constants of the fit. Fit coefficients and bootstrap errors are shown in Table 6.

C.3 Fitting loss vs. parameters

To understand the minimum achievable loss for dense vs MoLE models we can fit the more general parameterized ansatz of $L(N, D)$ proposed by [23].

$$\tilde{L}(N, D) = \hat{E} + \frac{\hat{A}}{N^{\hat{\alpha}}} + \frac{\hat{B}}{D^{\hat{\beta}}} \quad (3)$$

This maps the power law coefficients from Equation 3 to those in Equation 2 by using $\alpha = \frac{\hat{\beta}}{\hat{\alpha} + \hat{\beta}}$ and $\beta = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}}$. Here we can either minimize the $\tilde{L}(N, D)$ directly by fitting the 5 parameters $\hat{E}, \hat{A}, \hat{B}, \hat{\alpha}, \hat{\beta}$ with a iterative minimization procedure such as LBFGS [19, 8] or by examining the loss as a power relationship of N^* using

$$\log \tilde{L}(N^*) = \hat{\alpha} \log(N^*) + \gamma \quad (4)$$

where $\gamma = \log([1 + \frac{\hat{\alpha}}{\hat{\beta}}]\hat{A})$ with $\hat{E} \approx 0$. We found both methods yielded similar results but the minimization of \tilde{L} was more sensitive to the choice of hyperparameters.

Table 6: Power Law Coefficients determined from fitting Equations 2 and 4. Error bounds are determined by bootstrap sampling 1000 times and taking the 10th and 90th percentile values, quoted in brackets.

Parameter	Dense	MoLE
α	0.61 (0.57, 0.65)	0.56 (0.49, 0.59)
β	0.39 (0.35, 0.43)	0.44 (0.39, 0.43)
A	-4.5 (-3.8, -5.3)	-3.8 (-2.56, -4.65)
B	3.6 (2.9, 4.4)	2.9 (1.6, 3.7)
$\hat{\alpha}$	-0.29 (-0.27, -0.31)	-0.25 (-0.2, -0.3)
γ	2.16 (2.02, 2.34)	1.82 (1.61, 2.12)

D Inference

For inference benchmarking we use a periodic fcc carbon system with lattice constant $a = 3.8\text{\AA}$. This results in a fixed density of approximately 50 edges per atom within 6\AA . For UMA models, we use a combination of torch.compile, cuda graphs and pre-merged MoLE experts for inference speed. For large number of atoms (> 1000), we use edge-based activation checkpointing to trade off memory for some speed, allowing us to fit 100k+ atoms for the UMA-S into memory. We checked all our optimizations chosen does not degrade simulation accuracy, equivariance or energy conservation properties. While our benchmarks do not include graph generation, our internal CUDA based graph generation algorithm is very fast and decreases throughput by no more than 10% even for the largest systems tested. In the case of non-MoLE merging, we found the inference speed was comparable but the parameters require more GPU memory to store.

Table 7: Single-GPU simulation speeds for energy-conservative models in *steps per second*: comparing conservative UMA models to the top two models (eSEN and OrbV3) on the Matbench Discovery leaderboard [40] and the MACE materials and molecules models. Benchmarks are run, excluding graph generation, on a single Nvidia H100 80GB GPU using FP32 (TF32-high precision) and torch compile when possible. Test systems are a standard periodic atomic system that have ≈ 50 neighbors per atom with a 6\AA cutoff. OOM indicates the model ran out of memory. Refer to Sec.D for more details.

Atoms	UMA-S (6.6M)	UMA-M (50M)	eSEN-30M- OAM (30M)	Orb-v3 conservative- inf-omat (25M)	MACE- MPA-0 (9M)	MACE- OFF23-L (4.7M)
100	44	21	8	77	38	89
1,000	16	3	1.7	30	24	20
10,000	1.6	0.2	OOM	3.7	2.9	OOM
50,000	0.2	OOM	OOM	OOM	OOM	OOM
100,000	0.1	OOM	OOM	OOM	OOM	OOM

For fair comparisons against other models, we used pytorch2.6.0, cuda12.4, python3.12 and TF-32 precision universally on a H100 80GB GPU. We use standard torch.compile settings whenever possible (only MACE-MPA-0 failed to compile). Different models have different radius cutoffs and

max neighbors settings. We made sure that all models was receiving roughly 50 neighbors per atom for the same number of atoms.

E Evaluations

Section E.1 provides results for all versions of UMA.

Sections E.2-E.6 provides additional results for the original version of UMA, which was trained on a preview subset of the OMol25 dataset.

Table 8: Test MAE results on held out test splits for materials [40], catalysis [9], molecules [30], molecular crystals [16] and MOFs [43]. All energies are in meV, forces are in meV/Å and stresses are in meV/Å³. Results for UMA are compared against the SOTA literature results. Target accuracies for practical utility are provided as an approximate guide for reference.

Model	Materials						Catalysis				Molecules				Molecular crystals			MOFs	
	WBM Energy/Atom	Forces	Stress	HEA Energy/Atom	Forces	Stress	ID Ads. Energy	Forces	OOD-Both Ads. Energy	Forces	OOD-Comp Energy/Atom	Forces	PDB-TM Energy/Atom	Forces	OMC-Test Energy/Atom	Forces	Stress	OOD-LT Ads. Energy	Forces
UMA																			
UMA-S	20.0	60.8	4.4	22.0	72.8	3.1	52.1	24.3	70.2	30.9	3.64	10.80	0.88	16.12	0.91	4.77	0.97	292.4	16.0
UMA-M	18.1	51.4	4.3	19.0	62.2	3.2	33.4	16.0	46.5	21.0	3.26	9.09	0.69	10.37	0.82	3.00	0.98	290.2	10.7
UMA-L	17.6	45.5	3.8	24.8	48.3	2.8	32.4	12.2	43.5	15.9	2.33	5.19	0.81	8.76	0.59	2.28	0.10	291.1	6.5
UMA-S-1	19.4	62.2	4.5	21.9	73.5	3.1	53.1	24.5	70.4	31.2	0.96	8.25	0.93	15.56	0.93	5.15	1.01	287.1	13.6
UMA-S-1.1	20.2	62.8	4.4	24.9	83.7	3.5	51.5	24.1	68.8	30.7	0.95	8.64	0.84	15.47	1.03	5.04	0.93	289.9	13.3
UMA-M-1.1	18.2	50.7	4.2	21.9	69.0	3.5	31.8	15.5	45.5	20.2	0.74	5.44	0.50	10.14	0.84	2.83	0.90	294.1	10.3
Literature																			
eSEN-OMat [14]	16.2	49.6	4.1	20.0	59.5	3.2	-	-	-	-	-	-	-	-	-	-	-	-	-
eqV2-OMat [3]	14.9	46.3	3.6	20.3	47.0	2.7	-	-	-	-	-	-	-	-	-	-	-	-	-
eqV2-OC20 [31]	-	-	-	-	-	-	149.1	11.6	306.5	15.7	-	-	-	-	-	-	-	-	-
GemNet-OC20 [15]	-	-	-	-	-	-	163.5	16.3	343.3	23.1	-	-	-	-	-	-	-	-	-
eSEN-sm-cons. [30]	-	-	-	-	-	-	-	-	-	-	1.35	7.39	0.83	12.72	-	-	-	-	-
eSEN-S-OMC [16]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.05	5.39	0.94	-	-
eqv2-ODAC [43]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	316.0	7.2
Target																			
Practical Utility	10-20	-	-	10-20	-	-	100	-	100	-	1-3	-	1-3	-	1-3	-	-	100	-

Table 9: Evaluation results on Matbench-Discovery [40], MDR phonon [32], elastic tensor [12, 22], and AdsorbML benchmarks [29]. Results are also provided for a diverse set of molecule [30] and molecular crystal [20, 16] benchmarks. NVE MD [14] tests whether energy conservation is observed when running the model for molecular dynamics. SOTA results from literature are reported where available. For the materials evaluations, UMA models were fine-tuned on MPtrj [13] and sAlex [41, 3] to be consistent with the benchmark’s DFT settings.

	Materials									Catalysis	Molecules					Molecular Crystals		
Model	Matbench [40] <i>F1</i>	<i>RMSD</i>	<i>MAE [eV/atom]</i>	κ_{symc} [38]	Phonons [32] ω_{max} [K]	Free Energy [kJ/mol]	Elasticity [12, 22] G_{vib} [GPa]	K_{vib} [GPa]	NVE MD [14] Conserve	AdsorbML [29] Success Rate	OMol25 [30] Ligand-strain [meV]	PDB-pocket [meV]	Dist-SR [meV]	Dist-LR [meV]	NVE MD [14] Conserve	CSP Targets [20] Lattice Energy [kJ/mol]	Kendall Rank	RMSD [Å]
UMA																		
UMA-S	0.916	0.064	0.020	0.203	17.59	5.00	8.54	4.96	✓	68.35%	4.39	150.3	67.6	432.1	✓	2.695	0.82	0.12
UMA-M	0.930	0.061	0.018	0.195	13.91	3.39	8.40	4.76	✓	71.12%	2.45	89.7	41.6	588.7	✓	2.664	0.81	0.13
UMA-L	0.928	0.065	0.018	0.671	78.50	18.20	20.56	14.48	✗	74.41%	3.37	71.7	16.6	246.1	✗	2.488	0.84	0.12
UMA-S	0.916	0.064	0.020	0.203	17.59	5.0	8.54	4.96	✓	68.35%	4.39	150.3	67.6	432.1	✓	2.70	0.82	0.12
UMA-S-1	0.914	0.064	0.020	0.231	18.65	5.51	13.72	5.19	✓	64.85%	5.19	138.3	23.8	-	✓	2.57	0.81	0.13
UMA-S-1.1	0.913	0.064	0.020	0.204	18.82	5.48	9.47	5.16	✓	66.80%	5.2	127.7	26.7	256.7	✓	2.13	0.86	0.13
UMA-M-1.1	0.929	0.061	0.018	0.176	14.81	3.87	8.57	4.78	✓	72.25%	2.3	76.8	21.5	214.8	✓	3.24	0.82	0.14
Literature																		
eSEN-30M-OAM [14]	0.925	0.061	0.018	0.170	15.00	4.00	9.13	5.73	✓	-	-	-	-	-	-	-	-	-
ORB v3 [39]	0.905	0.075	0.024	0.210	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SevenNet-MF-ompa [24]	0.901	0.064	0.021	0.317	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GRACE-2L-OAM [6]	0.880	0.067	0.023	0.294	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MACE-MPA-0 [4]	0.852	0.073	0.028	0.412	-	-	-	-	-	-	-	-	-	-	-	-	-	-
eqv2-OC20 [29]	-	-	-	-	-	-	-	-	-	60.80%	-	-	-	-	-	-	-	-
GemNet-OC20 [29]	-	-	-	-	-	-	-	-	-	54.88%	-	-	-	-	-	-	-	-
eSEN-sm-cons. [30]	-	-	-	-	-	-	-	-	-	-	4.52	147.3	28.6	268.6	✓	-	-	-
eSEN-S-OMC [16]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6.18	0.74	0.18

E.1 Additional UMA Results

In this section, we provide results for the UMA, UMA-1, and UMA-1.1 models. The original UMA models (S, M, L) were trained on a preview subset of the OMol25 dataset ($\approx 70\%$), since portions of the dataset were still being calculated when UMA model training began. Both UMA-1 and UMA-1.1 were trained on the full OMol25 as released 05/14/2025. The other datasets used for training were unchanged, with the exception of the sampling ratios (Table 5) where the OMat24 coefficient was changed from 4 to 2 for both UMA-1 and UMA-1.1. UMA-1.1 resolved a size extensive bug later discovered. These models were fine-tuned on our 1.0 models with the fix in place. Tables 8 and 9 correspond to Tables 2 and 3 in the main text. Additional UMA 1 series models will be included in the arXiv version of the paper.

E.2 Materials

Table 10: Materials validation and test evaluations from OMat24 [3] and HEA. All energies are in meV, forces are in meV/Å and stresses are in meV/Å³.

Model	Val [3]				Test											
	Energy/Atom	Forces	Stress	Force Cosine	WBM [3]	OOD Composition [3]	OOD Element [3]	HEA	Energy/Atom	Forces	Stress	Energy/Atom	Forces	Stress	Energy/Atom	Forces
UMA																
UMA-S	11.3	57.1	2.9	0.98	20.0	60.8	4.4	11.5	57.0	3.0	9.9	56.9	2.6	22.0	72.8	3.1
UMA-M	10.0	47.3	2.7	0.99	18.1	51.4	4.3	10.2	47.6	2.9	8.5	47.1	2.4	19.0	62.2	3.2
UMA-L	9.7	43.5	2.5	0.99	17.6	45.5	3.8	9.8	43.6	2.6	8.1	43.6	2.3	24.8	48.3	2.8
Literature																
eSEN-30M-OMat [14]	10.7	47.3	2.6	0.99	16.2	49.6	4.1	10.7	47.3	2.8	9.0	47.2	2.3	20.0	59.5	3.2
eqV2-86M-OMat [31]	10.0	44.9	2.4	0.99	14.9	46.3	3.6	10.0	44.5	2.5	8.8	44.7	2.1	20.3	47.0	2.7

Table 11: **Materials evals results.** We note that UMA models are fine-tuned on the MPTrj [13] and sAlex [41, 3] datasets to be consistent with the DFT settings of the benchmarks.

	Matbench [40]					Kappa 103 [38]				MDR Phonons [32]				Elasticity [12, 22]		Binary Elasticity		NVE MD [14]	
Model	F1	DAF	Precision	Accuracy	MAE [eV/atom]	RMSE [eV/atom]	r ²	κ _{src}	κ _{srne}	ω _{max} , MAE [K]	ω _{avg} , MAE [K]	Entropy, MAE [kJ/K/mol]	C _v , MAE [kJ/K/mol]	Free Energy, MAE [kJ/mol]	G _{vib} , MAE [GPa]	K _{vib} , MAE [GPa]	G _{vib} , MAE [GPa]	K _{vib} , MAE [GPa]	Conservative
UMA																			
UMA-S	0.92	6.00	0.92	0.97	0.02	0.07	0.87	0.09	0.20	17.59	7.41	13.59	3.49	5.00	8.54	4.96	8.57	5.25	✓
UMA-M	0.93	6.08	0.93	0.98	0.02	0.07	0.87	0.09	0.20	13.91	5.11	9.63	2.66	3.39	8.40	4.76	7.07	4.75	✓
UMA-L	0.93	6.09	0.93	0.98	0.02	0.07	0.86	0.45	0.67	78.50	27.68	43.04	15.85	18.20	20.56	14.48	21.95	17.01	✗
Literature																			
eSEN-30M-OAM [14]	0.93	6.07	0.93	0.98	0.02	0.07	0.87	-	0.17	15.00	10.21	10.00	3.00	4.00	9.13	5.73	9.02	5.73	✓
eqV2-86M-OAM [3]	0.92	6.05	0.92	0.98	0.02	0.07	0.85	1.82	1.94	840.33	377.96	426.79	102.72	251.14	19.60	26.25	22.02	26.50	✗

Full results on materials' benchmarks are in Tables 10 and 11. Table 10 shows both validation and test results following OMat24 [3] along with the new high entropy alloy HEA test introduced in this paper. The HEA dataset contains relaxation trajectories for over 5000 alloys with atomic configuration disorder. Input structures were generated by sampling metallic element combinations of up to 6 different unique elements and using the special quasirandom structure method [48, 2] to decorate face-centered cubic, body-centered cubic and hexagonal close packed structures. DFT relaxations were carried out following the settings used in the OMat24 dataset [3].

In Table 11 we show full results for Matbench discovery [40], MDR phonon, elastic tensors, high entropy alloy IS2RE and NVE MD conservation benchmarks. The Matbench-Discovery benchmark evaluates a model's ability to predict ground-state thermodynamic stability by optimizing geometry and predicting energy. The thermal conductivity prediction task demands accurate modeling of harmonic and anharmonic phonons, which are crucial for precise predictions of thermal transport. The MDR Phonon benchmark assesses a model's performance in predicting phonon and vibrational thermodynamic properties. The MP elastic constant benchmark tests a model's accuracy in predicting

bulk and shear moduli, requiring precise calculations of stress tensors and their derivatives with respect to cell deformations.

E.3 Catalysis

For catalysis, we show full validation and test results for OC20 [9] in Table 12. The structures in the dataset contain molecules, called adsorbates, interacting with surfaces. The goal is to estimate the adsorption energy, which is the change in energy as the adsorbates come into contact with the surface, and the forces on the atoms. Force MAEs are comparable across UMA-M and UMA-L to other state-of-the-art models. Adsorption energies for UMA are calculated by subtracting two total energy calculations as described in the main text, which improves results over prior models.

In addition to energy and forces MAEs, we use AdsorbML to evaluate the performance on the practically relevant task of finding the global minima adsorption energy [29]. One limitation of the originally proposed AdsorbML pipeline is that it requires a DFT evaluation of the ML-identified global minima structure. To make this benchmark more accessible to those without access to DFT, we propose a slightly altered version of this benchmark that only requires ML. Here, we follow the same AdsorbML pipeline but allow the use of the ML-predicted global minima energy, but success is now only considered if the energy is within 0.1 eV of the DFT minima. Previously, success did not enforce a lower bound so long a DFT evaluation confirmed that energy prediction is realized. Without DFT, we also set a lower bound of 0.1 eV to provide some flexibility to finding a better global minima than DFT. We show that metrics are still highly correlated when you perform the original evaluation to the one proposed here.

Table 12: Catalysis validation and test results on OC20 [9] metrics. All energies are in meV and forces are in meV/Å.

Model	Val (Total Energy)						Test (Ads. Energy)					
	ID			OOD-Both			ID			OOD-Both		
	Energy	Forces	Force Cosine	Energy	Forces	Force Cosine	Energy	Force		Energy	Force	
UMA												
UMA-S	63.6	24.1	0.63	107.0	29.2	0.65	52.1	24.3		70.2	30.9	
UMA-M	43.1	15.8	0.73	70.0	19.2	0.75	33.4	16.0		46.5	21.0	
UMA-L	32.6	12.0	0.77	49.8	14.5	0.79	32.4	12.2		43.5	15.9	
Literature												
eqV2-OC20 [31]	-	-	-	-	-	-	149.1	11.63		306.5	15.74	
GemNet-OC20 [15]	-	-	-	-	-	-	163.5	16.33		343.3	23.11	

E.4 Molecules

Table 13: Open Molecule 2025 [30] validation evaluations across biomolecules, electrolytes, metal complexes, neutral organics and OOD-comp. All energies are in meV and forces are in meV/Å. All models are trained with preview OMol25.

Model	Biomolecules		Electrolytes		Metal Complexes		Neutral Organics		OOD-Comp	
	Energy/Atom	Force	Energy/Atom	Force	Energy/Atom	Force	Energy/Atom	Force	Energy/Atom	Force
UMA										
UMA-S	0.53	5.69	2.69	11.65	4.63	37.85	1.00	13.15	3.62	12.02
UMA-M	0.44	3.95	2.28	10.21	3.60	28.81	0.68	7.00	3.21	9.90
UMA-L	0.33	2.90	1.13	4.52	3.35	24.85	0.65	5.02	2.39	5.83
Baseline										
eSEN-S-OMol	0.54	6.06	2.52	12.63	4.27	37.30	0.84	13.00	3.69	12.78

We report results on molecules following the Open Molecules 2025 [30]. These include energy and force estimates for validation and test splits, as well as, numerous other tasks. The validation and test results are shown in Tables 13 and 14, respectively. Note that the eSEN-S-OMol model is only trained on the preview OMol25 dataset, which is $\approx 70\%$ of the full OMol25 dataset for a fair

Table 14: Open Molecule 2025 [30] test evaluations across biomolecules, electrolytes, metal complexes, neutral organics and OOD-comp, metal ligand, PDB-TM, reactivity, COD and anions. All energies are in meV and forces are in meV/Å. All models are trained with preview OMol25.

Model	Biomolecules		Electrolytes		Metal Complexes		Neutral Organics		OOD-Comp		Metal Ligand		PDB-TM		Reactivity		COD		Anions	
	Energy/Atom	Force	Energy/Atom	Force	Energy/Atom	Force	Energy/Atom	Force	Energy/Atom	Force	Energy/Atom	Force	Energy/Atom	Force	Energy/Atom	Force	Energy/Atom	Force	Energy/Atom	Force
UMA																				
UMA-S	0.51	5.70	3.80	13.95	3.07	33.56	1.49	20.33	3.64	10.80	1.23	17.71	0.88	16.12	4.82	61.80	2.92	29.82	0.66	10.85
UMA-M	0.42	3.89	3.29	13.19	2.40	24.85	0.98	10.83	3.26	9.09	0.99	12.04	0.69	10.37	3.88	47.75	2.19	20.11	0.50	8.03
UMA-L	0.34	2.92	1.41	5.42	2.47	21.67	1.03	6.91	2.33	5.19	1.05	10.00	0.81	8.76	3.93	41.97	2.57	15.44	0.62	5.90
Baseline																				
eSEN-S-OMol	0.52	6.06	3.52	15.23	2.86	33.30	1.35	20.06	3.67	11.56	1.18	17.49	0.79	14.11	4.89	61.16	2.93	24.12	0.47	10.38

comparison with the UMA models that were trained on the same subset. In general, the UMA-S and eSEN-S-OMol models provide comparable results.

Table 15: Open Molecule 2025 [30] single point evaluations for protein-ligand, IE/EA, spin gap and distance scaling. All energies are in meV and forces are in meV/Å. All models are trained with preview OMol25.

Model	Protein-ligand		IE/EA		Spin gap		Distance scaling	
	Exn Energy MAE	Exn Forces MAE	Δ Energy MAE	Δ Forces MAE	Δ Forces coorite sim.	Δ Energy MAE	Δ Forces MAE	Δ Forces coorite sim.
UMA								
UMA-S	150.25	5.09	336.16	80.60	0.77	665.75	112.09	0.69
UMA-M	89.69	4.06	236.76	66.28	0.81	547.73	98.15	0.74
UMA-L	71.68	2.27	244.23	57.18	0.82	568.36	93.09	0.73
Baselines								
eSEN-S-OMol	154.48	5.59	310.19	77.48	0.77	634.02	110.28	0.70

Table 16: Open Molecule 2025 [30] optimization evaluations including ligand-strain, conformer prediction, and protonation states. Results are reported across a variety of energy and structure based metrics for each task. All energies are in meV. All models are trained with preview OMol25.

Model	Ligand strain		Conformers		Protonation	
	Strain energy MAE [meV] ↓	RMSD min. [Å] ↓	RMSD ensemble [Å] ↓	RMSD boltz. [Å] ↓	Δ Energy MAE [meV] ↓	RMSD reprot. [Å] ↓
UMA						
UMA-S	4.39	0.28	0.06	0.05	5.76	0.02
UMA-M	2.45	0.19	0.04	0.03	3.19	0.01
UMA-L	3.37	0.25	0.05	0.05	4.97	0.01
Baselines						
eSEN-S-OMol	5.15	0.31	0.06	0.05	5.25	0.03

OMol25 [30] provides numerous other tasks for evaluation. These include evaluations which only require the estimation of a single point DFT calculation and evaluations that require optimizations. The comparison for single point DFT calculations is shown in Table 15. Similar to the test results, the UMA-S and eSEN-S-OMol models perform similarly with the UMA-M and UMA-L performing significantly better. Table 16 shows the results for tasks that require optimizations, which require repeated calls to the model to update atoms positions until a local energy minima is found. The small models report similar performance, and the UMA-M demonstrates the highest accuracies. It is likely that UMA-M outperforms UMA-L due to UMA-M being energy conserving and better behaved during optimization tasks.

Table 17: Open Molecular Crystals 2025 [16] validation and test table. All energies are in meV, forces are in meV/Å and stresses are in meV/Å³.

Model	Val				Test			
	Energy/Atom	Forces	Stress	Force Cosine	Energy/Atom	Forces	Stress	Force Cosine
UMA								
UMA-S	0.9	4.9	1.0	0.92	0.9	4.8	1.0	0.93
UMA-M	0.8	3.1	1.0	0.95	0.8	3.0	1.0	0.95
UMA-L	0.6	2.3	0.1	0.96	0.6	2.3	0.1	0.96
Baselines								
eSEN-S-OMC	1.06	5.58	0.96	0.92	1.05	5.39	0.94	0.92

Table 18: Open Molecular Crystals 2025 [16] evaluation for polymorphs from the Structure Ranking Phase of CCDC 7th CSP Blind Test [20]. All lattice energies are per molecule basis in kJ/mol.

Model	CCDC 7th CSP Blind Test Polymorphs							
	Lattice Energy MAE [kJ/mol]	RMSE [kJ/mol]	r^2	Rank correlation Kendall	Spearman	Structures RMSD [Å]	RMSD sd [Å]	Match rate
UMA								
UMA-S	2.69	3.67	0.73	0.82	0.93	0.12	0.07	0.99
UMA-M	2.66	3.71	0.60	0.81	0.91	0.13	0.07	0.99
UMA-L	2.49	3.70	0.81	0.84	0.95	0.12	0.07	1.00
Baselines								
eSEN-S-OMC	6.18	7.38	0.07	0.74	0.87	0.18	0.08	0.91

E.5 Molecular Crystals

Open Molecular Crystals (OMC25) [16] evaluates whether a model can predict the packing of molecule into crystal structures. This task requires the accurate estimation of inter-molecular forces. Results on validation and test splits are shown in Table 17. It is notable that all sizes of UMA models outperform the eSEN-S-OMC model trained on only OMC25. This indicates that the other datasets, such as OMol25, provide useful complementary information for the task. One important and real-world task for molecular crystals is to predict the lowest energy packing, called a polymorph, for a molecule. Results for this task for a subset of molecular crystal polymorphs from the most recent 7th Crystal Structure Prediction (CSP) Blind Test [20] are shown in Table 18. The pymatgen’s [35] StructureMatcher class with default settings is used to match DFT and UMA-relaxed polymorphs, and root mean square deviation (RMSD) is computed for matches. Similar to the test metrics, the UMA models outperform the eSEN-S-OMC trained on only OMC25.

E.6 Metal–Organic Frameworks

Table 19: OpenDAC [43] val and test table. All energies are in meV and forces are in meV/Å.

Model	Val (Total Energy)			Test (Ads. Energy)					
				ID			OOD-LT		
	Energy	Forces	Force Cosine	Energy	Forces	Force Cosine	Energy	Forces	Force Cosine
UMA									
UMA-S	60.4	5.9	0.82	169.5	16.7	0.63	292.4	16.0	0.57
UMA-M	59.3	3.8	0.91	167.3	14.8	0.62	290.2	10.7	0.76
UMA-L	38.7	3.3	0.91	177.1	7.8	0.82	291.1	6.5	0.91
Literature									
eqV2-ODAC [43]	-	-	-	145.0	8.2	0.69	316.0	7.2	0.72

The results on OpenDAC [43] are shown in Table 19. The OpenDAC dataset contains Metal-Organic Frameworks (MOFs) with CO₂ and water molecules. The goal is to estimate the change in energy in the presence with and without the CO₂ and water molecules. These adsorption energies are computed in the same manner as for catalysts. The use of total energies leads to significantly better adsorption

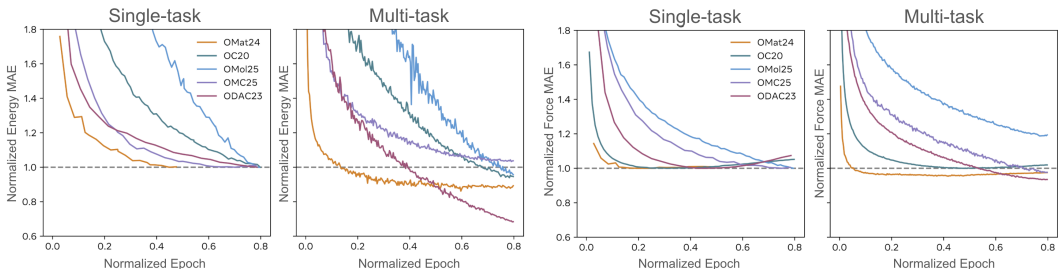


Figure 1: Pre-training curves of UMA-L for both single-task and multi-task models. Errors are normalized based on single-task performance. Note single-task models can overfit (forces on right), and the multi-task model generally converges to lower errors.

energy estimates, similar to catalysis. The forces of UMA-M and UMA-L are similar to the SOTA eqV2-ODAC [43] model.

F Single-task vs Multi-task

For large models, multi-task training offers benefits even without MoLE. In Figure 1, we plot the direct-force pre-training curves of UMA-lg and single-task models with the same model architecture and size. All metrics are normalized to those achieved with the models trained on single tasks to easily compare their relative performance with multi-task models. We observe that single task models frequently overfit to forces (OMat overfits upon further training), while the multi-task UMA model does not. Furthermore, UMA achieves lower losses in most cases. The one exception is OMol forces, for which errors are already small (< 10 meV/Å) for both models.

G Additional MoLE Analysis

G.1 Expert Analysis

Using a limited validation set consisting of 10,000 OMat24, 5,000 OC20, 20,000 OMol25, 10,000 OMC25, and 5,000 ODAC23 samples, we calculate the mean expert coefficient for each element-expert pair across all systems where the pair appears. We visualize these results using 32 periodic tables, each representing one of the 32 experts in UMA-S (Figure 2).

Additionally we visualize the expert coefficient mean and variance across datasets (tasks). Many experts utilized in OC20 are also used in OMat24. In contrast, the experts associated with OMol25 show minimal overlap with other datasets, aside from a small subset shared with OM25C. Lastly ODAC23 and OMC25 utilize the fewest experts, and these two share a single expert between them.3).

G.2 Generalization Across Architectures

Table 20: Testing the generalization capability of MoLE layers. We applied 8-expert MoLE layers to both eSEN[14] and EquiformerV2[31] model architectures and measured their relative performance on direct pretraining against the versions without no MoLE. For each row, we show the relative improvement over the baseline.

OMol25 Val subset	eSEN 6M	EquiformerV2 6M
Metal Complexes	+10%	+10%
Spice	+5%	+5%
Neutral organics	0%	+3%
Biomolecules	+25%	+21%
Electrolytes	+12%	+10%
Mean across subsets	+10%	+10%

We performed additional analysis to ablate the benefit of the MoLE layers. Since our MoLE implement is very general and can be applied any linear layers inside a model, we ran experiments adding MoLE

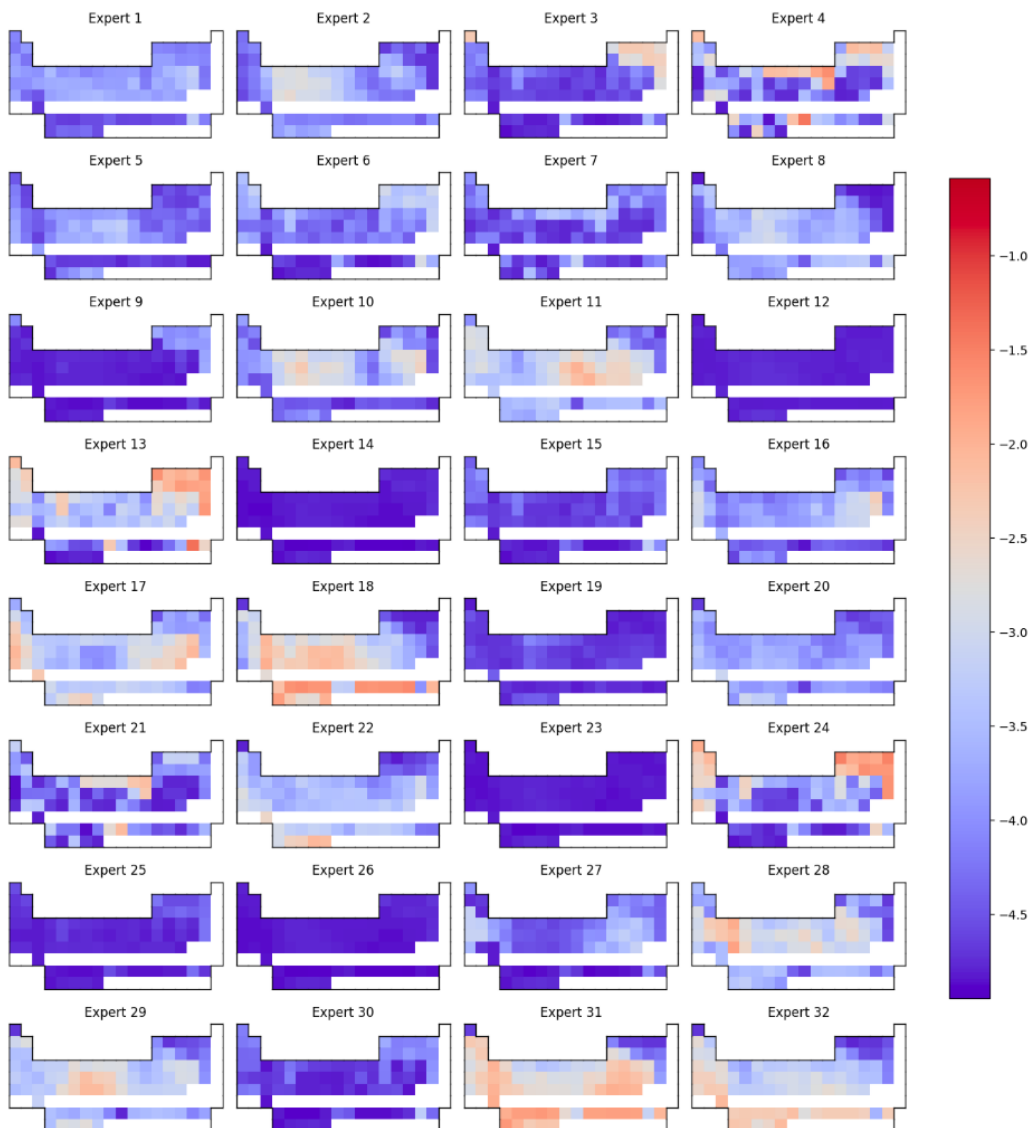


Figure 2: Log mean expert coefficient across element-expert pairs.

to both eSEN and EquiformersV2 architectures, controlling for model size (6M base parameters) and 8 training epochs on Omol dataset, direct pretraining only.

20 shows that despite EquiformersV2 and eSEN being very different architectures, the relative benefit of adding MoLE is fairly consistent in both cases. This suggests that MoLE is a simple and general approach to boost the capacity of MLIP models without incurring inference speed and memory overhead.

References

- [1] Qianxiang Ai, Vinayak Bhat, Sean M. Ryno, Karol Jarolimek, Parker Sornberger, Andrew Smith, Michael M. Haley, John E. Anthony, and Chad Risko. OCELOT: An infrastructure for data-driven research to discover and design crystalline organic semiconductors. *The Journal of Chemical Physics*, 154(17):174705, May 2021. ISSN 0021-9606. doi: 10.1063/5.0048714.
- [2] Luis Barroso-Luque, Julia H. Yang, Fengyu Xie, Tina Chen, Ronald L. Kam, Zinab Jadidi, Peichen Zhong, and Gerbrand Ceder. SMOL: A Python package for cluster expansions and

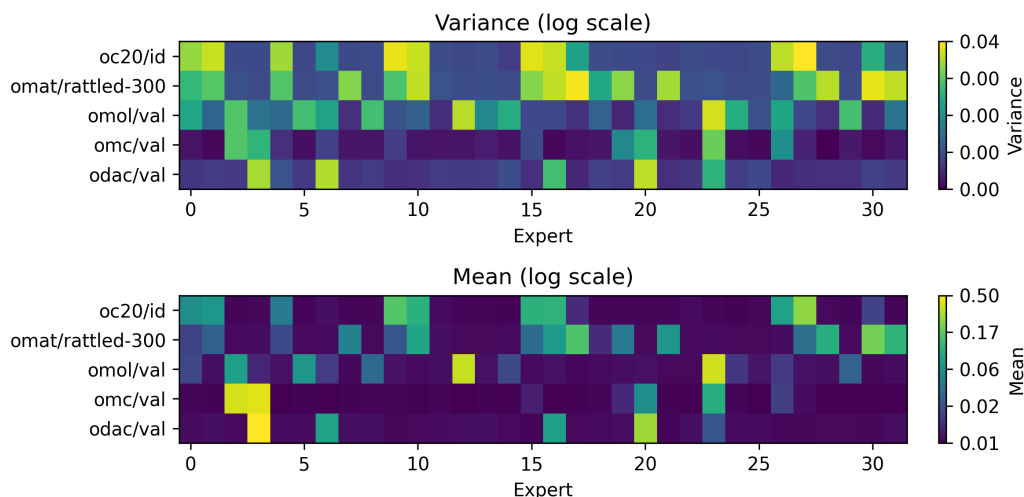


Figure 3: Log mean expert coefficient across element-expert pairs.

- beyond. *Journal of Open Source Software*, 7(77):4504, 2022. ISSN 2475-9066. doi: 10.21105/joss.04504.
- [3] Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C Lawrence Zitnick, and Zachary W Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint arXiv:2410.12771*, 2024.
 - [4] 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.
 - [5] Filippo Bigi, Marcel Langer, and Michele Ceriotti. The dark side of the forces: assessing non-conservative force models for atomistic machine learning. *arXiv preprint arXiv:2412.11569*, 2024.
 - [6] Anton Bochkarev, Yury Lysogorskiy, and Ralf Drautz. Graph atomic cluster expansion for semilocal interactions beyond equivariant message passing. *Physical Review X*, 14(2):021036, 2024.
 - [7] Stanislav S. Borysov, R. Matthias Geilhufe, and Alexander V. Balatsky. Organic materials database: An open-access online database for data mining. *PLOS ONE*, 12(2):e0171501, February 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0171501.
 - [8] Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale? *arXiv preprint arXiv:2410.23179*, 2024.
 - [9] 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.
 - [10] Yongchul G. Chung, Jeffrey Camp, Maciej Haranczyk, Benjamin J. Sikora, Wojciech Bury, Vaiva Krungleviciute, Taner Yildirim, Omar K. Farha, David S. Sholl, and Randall Q. Snurr. Computation-ready, experimental metal–organic frameworks: A tool to enable high-throughput screening of nanoporous crystals. *Chem. Mater.*, 26(21):6185–6192, 2014. doi: 10.1021/cm502594j.
 - [11] Yongchul G. Chung, Emmanuel Haldoupis, Benjamin J. Bucior, Maciej Haranczyk, Seulchan Lee, Hongda Zhang, Konstantinos D. Vogiatzis, Marija Milisavljevic, Sanliang Ling, Jeffrey S. Camp, Ben Slater, J. Ilja Siepmann, David S. Sholl, and Randall Q. Snurr. Advances, updates, and analytics for the computation-ready, experimental metal–organic framework database: Core mof 2019. *J. Chem. Eng. Data*, 64(12):5985–5998, 2019. doi: 10.1021/acs.jced.9b00835.

- [12] Maarten de Jong, Wei Chen, Thomas Angsten, Anubhav Jain, Randy Notestine, Anthony Gamst, Marcel Sluiter, Chaitanya Krishna Ande, Sybrand van der Zwaag, Jose J. Plata, Cormac Toher, Stefano Curtarolo, Gerbrand Ceder, Kristin A. Persson, and Mark Asta. Charting the complete elastic properties of inorganic crystalline compounds. *Scientific Data*, 2(1):150009, March 2015. ISSN 2052-4463. doi: 10.1038/sdata.2015.9.
- [13] 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.
- [14] 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.
- [15] 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. *arXiv preprint arXiv:2204.02782*, 2022.
- [16] Vahe Gharakhanyan, Luis Barroso-Luque, Yi Yang, Muhammed Shuaibi, Kyle Michel, Daniel S. Levine, Misko Dzamba, Xiang Fu, Meng Gao, Xingyu Liu, Haoran Ni, Keian Noori, Brandon M. Wood, Matt Uyttendaele, Arman Boromand, C. Lawrence Zitnick, Noa Marom, Zachary W. Ulissi, and Anuroop Sriram. Open molecular crystals 2025 (omc25) dataset and models, 2025. URL <https://arxiv.org/abs/2508.02651>.
- [17] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics*, 132(15):154104, April 2010. ISSN 0021-9606. doi: 10.1063/1.3382344.
- [18] Bjørk Hammer, Lars Bruno Hansen, and Jens Kehlet Nørskov. Improved adsorption energetics within density-functional theory using revised perdew-burke-ernzerhof functionals. *Physical Review B*, 59(11):7413, 1999.
- [19] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [20] Lily M Hunnisett, Nicholas Francia, Jonas Nyman, Nathan S Abraham, Srinivasulu Aitipamula, Tamador Alkhidir, Mubarak Almehairbi, Andrea Anelli, Dylan M Anstine, John E Anthony, et al. The seventh blind test of crystal structure prediction: Structure ranking methods. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 80(6): 548–574, December 2024. ISSN 2052-5206. doi: 10.1107/S2052520624008679.
- [21] Lily M Hunnisett, Jonas Nyman, Nicholas Francia, Nathan S Abraham, Claire S Adjiman, Srinivasulu Aitipamula, Tamador Alkhidir, Mubarak Almehairbi, Andrea Anelli, Dylan M Anstine, et al. The seventh blind test of crystal structure prediction: Structure generation methods. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 80(6): 517–547, December 2024. ISSN 2052-5206. doi: 10.1107/S2052520624007492.
- [22] Aaron D. Kaplan, Runze Liu, Ji Qi, Tsz Wai Ko, Bowen Deng, Janosh Riebesell, Gerbrand Ceder, Kristin A. Persson, and Shyue Ping Ong. A Foundational Potential Energy Surface Dataset for Materials, March 2025.
- [23] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [24] Jaesun Kim, Jisu Kim, Jaehoon Kim, Jiho Lee, Yutack Park, Youngho Kang, and Seungwu Han. Data-efficient multifidelity training for high-fidelity machine learning interatomic potentials. *Journal of the American Chemical Society*, 147(1):1042–1054, 2024.
- [25] 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.

- [26] Georg Kresse and Jürgen Hafner. Ab initio molecular dynamics for liquid metals. *Physical review B*, 47(1):558, 1993.
- [27] Georg Kresse and Jürgen Hafner. Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements. *Journal of Physics: Condensed Matter*, 6(40):8245, 1994.
- [28] Georg Kresse and Daniel Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical review b*, 59(3):1758, 1999.
- [29] 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.
- [30] Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The open molecules 2025 (omol25) dataset, evaluations, and models, 2025. URL <https://arxiv.org/abs/2505.08762>.
- [31] 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.
- [32] Antoine Loew, Dewen Sun, Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Universal Machine Learning Interatomic Potentials are Ready for Phonons, December 2024.
- [33] Łukasz Mentel. mendeleeev - A Python package with properties of chemical elements, ions, isotopes and methods to manipulate and visualize periodic table., March 2021. URL <https://github.com/lmentel/mendeleeev>.
- [34] Frank Neese, Frank Wennmohs, Ute Becker, and Christoph Riplinger. The orca quantum chemistry program package. *The Journal of chemical physics*, 152(22), 2020.
- [35] 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.
- [36] Saro Passaro and C. Lawrence Zitnick. Reducing SO(3) convolutions to SO(2) for efficient equivariant GNNs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202. PMLR, 2023.
- [37] John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical review letters*, 77(18):3865, 1996.
- [38] Balázs Póta, Paramvir Ahlawat, Gábor Csányi, and Michele Simoncelli. Thermal Conductivity Predictions with Foundation Atomistic Models, September 2024.
- [39] Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale. *arXiv preprint arXiv:2504.06231*, 2025.
- [40] Janosh Riebesell, Rhys EA Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Alpha A Lee, Anubhav Jain, and Kristin A Persson. Matbench discovery—a framework to evaluate machine learning crystal stability predictions. *arXiv preprint arXiv:2308.14920*, 2023.
- [41] 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*, 48:101560, 2024.

- [42] Anuroop Sriram, Abhishek Das, Brandon M Wood, and C. Lawrence Zitnick. Towards training billion parameter graph neural networks for atomic simulations. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=OjP2n0YFmKG>.
- [43] 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, 2024.
- [44] Anuroop Sriram, Logan M. Brabson, Xiaohan Yu, Sihoon Choi, Kareem Abdelmaqsoud, Elias Moubarak, Pim de Haan, Sindy Löwe, Johann Brehmer, John R. Kitchin, Max Welling, C. Lawrence Zitnick, Zachary Ulissi, Andrew J. Medford, and David S. Sholl. The open dac 2025 dataset for sorbent discovery in direct air capture, 2025. URL <https://arxiv.org/abs/2508.03162>.
- [45] Annika Stuke, Christian Kunkel, Dorothea Golze, Milica Todorović, Johannes T. Margraf, Karsten Reuter, Patrick Rinke, and Harald Oberhofer. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Scientific Data*, 7(1):58, February 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0385-y.
- [46] Rithwik Tom, Timothy Rose, Imanuel Bier, Harriet O’Brien, Álvaro Vázquez-Mayagoitia, and Noa Marom. Genarris 2.0: A random structure generator for molecular crystals. *Computer Physics Communications*, 250:107170, 2020.
- [47] 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.
- [48] Alex Zunger, S.-H. Wei, L. G. Ferreira, and James E. Bernard. Special quasirandom structures. *Physical Review Letters*, 65(3):353–356, July 1990. doi: 10.1103/PhysRevLett.65.353.