

# Cross Learning between Electronic Structure Theories for Unifying Molecular, Surface, and Inorganic Crystal Foundation Force Fields

Ilyes Batatia,<sup>1,\*</sup> Chen Lin,<sup>2,\*</sup> Joseph Hart,<sup>1,3</sup> Elliott Kesoar,<sup>4,1</sup> Alin M. Elena,<sup>4</sup> Sam Walton Norwood,<sup>5</sup> Thomas Wolf,<sup>6</sup> and Gábor Csányi<sup>1</sup>

<sup>1</sup>*Engineering Laboratory, University of Cambridge, Trumpington St, Cambridge, UK*

<sup>2</sup>*University of Oxford, UK*

<sup>3</sup>*Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge, UK*

<sup>4</sup>*Scientific Computing Department, Science and Technology Facilities Council, Daresbury Laboratory, Keckwick Lane, Daresbury WA4 4AD, UK*

<sup>5</sup>*Mirror Physics, 31 Hudson Yards, Floor 11, New York, NY 10001*

<sup>6</sup>*Hugging Face, 20 Jay Street Suite 620, Brooklyn, NY 11201*

(Dated: October 30, 2025)

Creating a single unified interatomic potential capable of attaining *ab initio* accuracy across all chemistry remains a long-standing challenge in computational chemistry and materials science. This work introduces a training protocol for foundation machine-learning interatomic potentials (MLIPs) that bridge molecular, surface, and materials chemistry through cross-domain learning. First, we introduce enhancements to the MACE architecture that improve its performance on chemically diverse databases by increasing weight sharing across chemical elements and introducing non-linear factors into the tensor decomposition of the product basis. Second, we develop a multi-head replay post-training methodology that enables efficient knowledge transfer across diverse chemical domains. By fine-tuning on datasets at different levels of electronic structure theory—including inorganic crystals, molecular systems, surface chemistry, and reactive organic chemistry—we demonstrate that a single unified model achieves state-of-the-art performance across several chemical domains. Comprehensive benchmarking reveals superior cross-domain transferability compared with existing specialised and multi-task models, with notable improvements in molecular and surface properties while maintaining state-of-the-art performance in materials-property prediction.

## I. Introduction

The development of accurate and transferable machine learning interatomic potentials (MLIPs) remains one of the most challenging problems in computational chemistry and materials science [1–11]. Traditional approaches have created a fragmented landscape where distinct models are required for molecular systems [5, 12–14], surface chemistry [15–17], and bulk materials [18–20], creating substantial barriers when studying phenomena that naturally span multiple chemical domains, such as heterogeneous catalysis, crystal growth, or interfacial processes. Recent advances in foundation MLIPs have demonstrated remarkable capabilities through large pre-training on diverse datasets [18–24]. However, most existing foundation models suffer from limited chemical scope, often excelling in one domain while showing inadequate performance in others. This limitation stems from both dataset constraints and architectural choices that favour domain-specific optimisation over cross-domain transferability.

The challenge of unifying multiple chemical domains within a single MLIP framework involves several key considerations: (i) training on datasets with different levels of electronic structure theory, (ii) efficient knowledge transfer between chemical domains without catastrophic forgetting (if training involves multiple stages),

and (iii) maintaining computational efficiency while expanding chemical coverage. One strategy, used recently by UMA [25], DPA-3 [26] and SevenNet [27], is to make the model output explicitly depend on the task by embedding the task as part of the input along with the atomic coordinates. This makes most model layers task-dependent, and allows for significant flexibility while benefiting from some cross-learning. An alternative is meta-learning, where during a pre-training stage multiple tasks are sequentially attempted, resulting in a model that is easy to specialise to a specific task by performing fine-tuning training [28]. The downside of both these approaches is that at inference time the task needs to be specified by the user; there is no single model applicable to all tasks out of the box. DPA-2 [24] and JMP [29] used pre-training with multiple readout heads (where only the last computational layer is specific to each task) and downstream fine-tuning for a given specific task. In this work, we create a single foundation MLIP using multiple diverse datasets, also using a multi-head approach, and evaluate the performance of its “main” head (in this case corresponding to DFT with the PBE functional) on all tasks. The ultimate goal is to have a single continuous potential energy function applicable to all chemical contexts. We therefore design a training protocol that enhances cross-learning and knowledge sharing from all heads to the main head.

Our first contribution is to introduce a set of changes to the state-of-the-art MACE architecture that improves its performance for large databases containing a large

\* These authors contributed equally.

number of chemical species. These changes, explicitly highlighted below, include the use of more weight-sharing between the different species to enable the model to learn more powerful compressions of the chemical domains. We also introduce non-linear factors in the tensor decomposition that are demonstrated to provide better accuracy than purely polynomial features.

Second, we introduce a multi-head cross-learning fine-tuning protocol that aims to produce a single model bridging materials, molecular, and surface chemistry through two key innovations: (1) a multi-head architecture that enables simultaneous learning across potentially inconsistent levels of electronic structure theory while maintaining shared chemical and geometrical representations; (2) pre-training followed by a replay fine-tuning methodology that facilitates knowledge transfer across domains while preventing catastrophic forgetting from the base model. We perform comprehensive validation across molecular, surface, and materials benchmarks, establishing new standards for evaluating unified foundation force field accuracy.

Our results show that cross-domain learning attains competitive in-domain accuracy and yields measurable cross-domain generalisation via knowledge transfer, improving performance on molecular, surface, and crystal benchmarks without degrading materials accuracy. This work establishes a clear path toward MLIPs that can seamlessly handle physical and chemical phenomena across all areas of chemistry and materials science.

## II. Methods

### A. Theoretical Foundation

Our approach builds upon the MACE architecture [10], which employs many-body equivariant message passing to build fast and accurate machine learning interatomic potentials. MACE parameterises a mapping from atomic coordinates and atomic numbers (element indices) to the potential energy by decomposing it into atomic energies. Each atomic energy term is a function of invariant descriptors that embed the chemical and geometrical many-body information of the neighbourhood of the atom. The total energy  $E$  of a system is expressed as:

$$E = \sum_i E_i(\{\mathbf{r}_{ij}, z_j\}_{j \in \mathcal{N}_i}) \quad (1)$$

where  $E_i$  represents the atomic energy contribution of atom  $i$ ,  $\mathcal{N}_i$  denotes the set of its neighbours within a cutoff radius  $r_{\text{cut}}$ ,  $\mathbf{r}_{ij}$  are relative vectors from atom  $i$  to its neighbour  $j$ , and  $z_j$  are atomic numbers.

The multi-head architecture enables simultaneous learning across multiple potentially inconsistent levels of electronic structure theory by employing distinct shallow readout functions that map shared latent feature repre-

sentations to each desired theoretical framework:

$$E_i^{(\text{head})} = \sum_s \mathcal{R}^{(\text{head}, s)}(\mathbf{h}_i^{(s)}) + E_{0, z_i}^{(\text{head})} \quad (2)$$

where (head) enumerates readout heads corresponding to different levels of theory,  $\mathbf{h}_i^{(s)}$  represents the node features at layer  $s$ , and  $E_0^{(\text{head})}$  are head-specific atomic reference energies. For all the fine-tuning heads, we use the method outlined in Section XV A to estimate their  $E_0$ , and for the superior cross-domain transferability head we use DFT computed atomic reference energies. This multi-head approach parallels recent developments in machine learning potentials, including the DPA-2 model [24]. In our case, for each head, we use a simple linear readout layer for the first layer and a single-hidden-layer fully-connected feed-forward network (multilayer perceptron, MLP) for the second layer.

### B. MACE with non-linear tensor decomposition

Equations (3)-(19) represent the complete functional form of MACE as they were introduced in Ref. [10, 30] together with the blue parts corresponding to the modifications we introduce in this paper. We arrived at this particular set of changes by performing a grid-search over proposed changes using large-dataset training runs to maximise the accuracy of the model with the fewest changes. For ease of reference, we keep the formulas together and provide a detailed explanation of them below. Note that in order to help readability by limiting the number of symbols and symbol modifiers, we reuse the symbol  $W$  to refer to the free parameters of the model; each occurrence below corresponds to independent weights with appropriate dimensions and indices. There are additional free parameters in the MLPs which are not explicitly shown.

The finite neighbourhood of each atom creates a graph topology on the atomic structures. Each node  $i$  carries a feature vector  $\mathbf{h}_i$ , expanded in a spherical-harmonic basis so that components are indexed by  $(l, m)$ ; we write  $\mathbf{h}_i^{(s)}$  for the features after iteration  $s$  (the  $s$ -th message-passing layer) and denote the total number of layers by  $S$ .

We start by initialising the node features  $\mathbf{h}_i^{(0)}$  as a learnable embedding of the chemical elements with atomic number  $z_i$  into  $K_{\text{node}}$  learnable channels indexed by  $k$ , cf. Eq. (3). This kind of mapping has been used extensively for graph neural networks [31–34] and elsewhere [35, 36] and has been shown to lead to some transferability between molecules with different elements [37]. The zeros for the  $lm$  indices correspond to these initial features being scalars, i.e. rotationally invariant. The higher-order elements of  $\mathbf{h}_i^{(0)}$  with nonzero  $lm$  indices are initialised to zero. At the beginning of each subsequent iteration, the node features (both scalars and

higher order) are linearly mixed together resulting in  $\bar{h}_j$ , cf. Eq. (4).

Next, we combine the features of each of the neighbouring atoms  $j$  with the interatomic displacement vectors pointing to them from the central atom  $i$  (corresponding to the  $i$ - $j$  edge in the graph) expressed using radial and spherical harmonic basis. This is analogous to the construction of the one-particle basis of neighbour density representations, such as SOAP [38] and ACE [7, 37], and we construct it in a similar way to Cormorant [39] and NequIP [33]. The relationships between these different approaches are discussed in Ref. [37]. We construct the radial basis set using the first spherical Bessel function  $j_0^n$  for different wavenumbers,  $n$ , up to some small maximum (typically 8), in Eq. (5) as proposed in Ref. [34]. For each channel  $k$ , the radial information is then passed through a separate MLP, Eq. (7), whose inputs are the Bessel functions with different frequencies, and which has many outputs, indexed by  $(\eta_1, l_1, l_2, l_3)$ . In a modification to the original MACE architecture, we compute separate embeddings of the source and target elements following [25], and concatenate them with the Bessel features before passing them into the radial MLP. The next modification compared to the original MACE design is that we apply the radial cutoff function  $f_{\text{cut}}$  outside the MLP, and not directly to the Bessel function as in the original MACE. This forces a smoother decay near the cutoff. When combining positional information (itself an equivariant that transforms under rotation like a vector) with equivariant node features, we use the spherical tensor product formalism of angular momentum addition [40]. All possible combinations of equivariants are constructed using the appropriate Clebsch-Gordan coefficients, Eq. (8).

The one-particle basis  $\phi$  is summed over the atoms in the neighbourhood in Eq. (10). This is where permutation invariance of the MACE descriptors over the atoms in the neighbourhood is achieved - note that the identity of chemical elements has already been embedded, and hence this sum is over all atoms, regardless of their atomic number. A linear mixing of  $k$  channels with learnable weights yields the initial atomic basis,  $\tilde{A}_i$ . As the tensor product operation in Eq. (8), which happens on the edges, is the computational bottleneck of MACE, we let the freedom to use fewer channels for this operation and use  $K_{\text{edge}}^{(s)} \leq K$  instead. The dependence  $s$  is used to specify more edge channels for the inexpensive first layer  $s = 0$  compared to other layers.

To ensure internal normalization of the features and smooth extrapolation to systems with different densities, we divide the atomic basis in each layer by a learnable quantity called density normalization  $n_i$ . We extend our density normalization [18] to be a (learnable) weighted sum of the density term and a constant term. This weighted sum acts like a mixture of a body-ordered feature (with the constant term) and a mean-field feature with the density term.

$$h_{i,k00}^{(0)} = \sum_z W_{kz} \delta_{zz_i} \quad (3)$$

$$\bar{h}_{i,kl_2m_2}^{(s)} = \sum_{\bar{k}}^{K_{\text{node}}} W_{k\bar{k}l_2}^{(s)} h_{i,\bar{k}l_2m_2}^{(s)} \quad (4)$$

$$j_0^n(r_{ij}) = \sqrt{\frac{2}{r_{\text{cut}}}} \frac{\sin\left(n\pi \frac{r_{ij}}{r_{\text{cut}}}\right)}{r_{ij}} f_{\text{cut}}(r_{ij}) \quad (5)$$

$$e_{i,k}^{(s),\text{src}} = \sum_z W_{kz} \delta_{zz_i}, \quad e_{i,k}^{(s),\text{trg}} = \sum_z W_{kz} \delta_{zz_i} \quad (6)$$

$$R_{k\eta_1l_1l_2l_3}^{(s)}(r_{ij}) = \text{MLP}\left(\{j_0^n(r_{ij})\}_n, e_i^{(s),\text{src}}, e_j^{(s),\text{trg}}\right) f_{\text{cut}}(r_{ij}) \quad (7)$$

$$\phi_{ij,k\eta_1l_3m_3}^{(s)} = \sum_{l_1l_2m_1m_2} C_{\eta_1l_1m_1l_2m_2}^{l_3m_3} R_{k\eta_1l_1l_2l_3}^{(s)}(r_{ij}) \times Y_{l_1}^{m_1}(\hat{\mathbf{r}}_{ij}) \bar{h}_{j,\bar{k}l_2m_2}^{(s)} \quad (8)$$

$$n_i^{(s)} = \sum_{j \in \mathcal{N}(i)} \tanh\left(\text{MLP}\left(\{j_0^n(r_{ij})\}_n, e_i^{(s),\text{src}}, e_j^{(s),\text{trg}}\right)^2\right) f_{\text{cut}}(r_{ij}) \quad (9)$$

$$A_{i,kl_3m_3}^{(s)} = \frac{1}{\alpha^{(s)} + \beta^{(s)} n_i^{(s)}} \sum_{\eta_1} \sum_{\bar{k}}^{K_{\text{edge}}^{(s)}} W_{k\bar{k}\eta_1l_3}^{(s)} \sum_{j \in \mathcal{N}(i)} \phi_{ij,\bar{k}\eta_1l_3m_3}^{(s)} \quad (10)$$

$$\tilde{A}_{i,kl_3m_3}^{(s)} = A_{i,kl_3m_3}^{(s)} + \sum_{\bar{k}}^{K_{\text{node}}} W_{k\bar{k}l_3}^{(s)} \bar{h}_{i,\bar{k}l_3m_3}^{(s)} \quad (11)$$

$$A_{i,kl_3}^{(s),\text{scalars}} = \frac{1}{\alpha^{(s)} + \beta^{(s)} n_i^{(s)}} \sum_{\eta_1} \sum_{\bar{k}}^{K_{\text{edge}}^{(s)}} W_{k\bar{k}\eta_1l_3}^{(s)} \sum_{j \in \mathcal{N}(i)} \phi_{ij,\bar{k}\eta_100}^{(s)} \quad (12)$$

$$\Omega_{i,kl_3}^{(s)} = A_{i,kl_3}^{(s),\text{scalars}} + \sum_{\bar{k}}^{K_{\text{node}}} W_{k\bar{k}l_3}^{(s)} \bar{h}_{i,\bar{k}00}^{(s)} \quad (13)$$

$$g(x_{i,k00}, y_{i,klm}) = \begin{cases} x_{i,k00} \sigma(x_{i,k00}) y_{i,k00}, & l = 0, \\ \sigma(x_{i,k00}) y_{i,klm}, & l > 0, \end{cases} \quad (14)$$

$$A_{i,kl_3m_3}^{(s),\text{gated}} = \sum_{\bar{k}}^{K_{\text{node}}} W_{k\bar{k}l_3}^{(s)} g(\Omega_{i,kl_3}, \tilde{A}_{i,\bar{k}l_3m_3}^{(s)}) \quad (15)$$

$$A_{i,klm}^{(s),\nu} = \prod_{\xi=1}^{\nu} A_{i,kl_{\xi}m_{\xi}}^{(s),\text{gated}} \quad (16)$$

$$B_{i,\eta_{\nu}kLM}^{(s),\nu} = \sum_{lm} C_{\eta_{\nu}lm}^{LM} A_{i,klm}^{(s),\nu} \quad (17)$$

$$m_{i,kLM}^{(s)} = \sum_{\nu} \sum_{\eta_{\nu}} W_{\eta_{\nu}kL}^{(s),\nu} B_{i,\eta_{\nu}kLM}^{(s),\nu} \quad (18)$$

$$h_{i,kLM}^{(s+1)} = \sum_{\bar{k}}^{K_{\text{node}}} W_{kL,\bar{k}}^{(s)} m_{i,\bar{k}LM}^{(s)} + \sum_{\bar{k}}^{K_{\text{node}}} W_{kL,\bar{k}}^{(s)} h_{i,\bar{k}LM}^{(s)} \quad (19)$$

We then construct an updated atomic basis,  $A_i^{\text{gated}}$  in Eq. (11), using a learnable residual connection from the initial node features, that contains both information about the neighbourhood of the atom and also information about the atom  $i$  itself. This new modification enables the model to construct a polynomial that contains not only factors from the neighbours of  $i$ , but also factors that depend on the features of  $i$  directly, allowing for richer many-body features.

We also construct a set of scalar features  $\Omega$  on each atom, from a learnable combination of extra scalars computed from the atomic basis  $A_i^{(s),\text{scalars}}$  in Eq. (12) and the node features as a residual connection in Eq. (13). The scalar features are used to compute a gated non-linearity [41] in equations (14) and (15). We found that a sigmoid ( $\sigma$ ) gate for the equivariant channels and a SiLU for the invariant channels works best.

We use these gated atomic basis to construct many-body messages in the same way as in our original MACE design using tensor-decomposed symmetric contractions. For more details on these operations, see the original MACE paper [10]. Additionally, for more explanation on the role of the tensor decomposition, refer to the TRACE paper [42]. One important modification compared to the original tensor decomposition is that, due to the non-linearity in Eq. 15, the model can learn non-linear rank-1 factors in the tensor decomposition. We also use element-agnostic weights for the message construction (Eq. (18)) and the update (Eq. (19)). The rest of the architecture is identical to that of our previous MACE models, with the computation of the product basis in Eq. 16, the symmetrization with the generalised Clebsch–Gordan coefficients in Eq. 17, the message construction in Eq. 18 and the update in Eq. 19. After  $S$  layers of message passing (usually we use  $S=2$ ), the node features are used to predict site energies per head in Eq. 2.

### III. Multi-Head Replay Post-Training

Figure 1 provides an overview of our cross-learning framework, illustrating the progression from foundation pre-training through Multi-Head Replay Post-Training. Our multi-head replay post-training strategy [18] facilitates efficient knowledge transfer from a broad foundation dataset to multiple specialised datasets while preventing catastrophic forgetting through strategic replay sampling. The workflow comprises two distinct stages:

- **Stage 1: Pre-training at base level theory.** We first train a unified backbone model on a large, diverse dataset of inorganic crystals, producing shared feature representations that capture fundamental chemical and geometrical patterns.
- **Stage 2: Multi-head Fine-tuning with Replay.** We simultaneously fine-tune multiple shallow readout heads, starting from the pre-trained weights on domain-specific datasets. We do not

freeze any weights, i.e. we keep fine-tuning the backbone weights as well. Each head targets a different chemical domain (molecular systems, surfaces, inorganic materials) or specific level of theory. To mitigate catastrophic forgetting, we construct a replay buffer by sampling representative configurations from the original pre-training data. During fine-tuning, each minibatch combines randomly new domain-specific samples with replay samples, ensuring retention of foundational chemical knowledge.

## IV. Datasets

We employ a combination of large-scale foundation data and targeted fine-tuning datasets to comprehensively cover inorganic crystals, surfaces, and molecular chemistry.

### A. Pre-training Dataset

**OMAT** [43]: A large inorganic crystal dataset containing 100 million configurations spanning 89 elements. All calculations are performed at the PBE and PBE+ $U$  level of theory, providing broad coverage of diverse inorganic crystals and serving as our pre-training dataset.

### B. Fine-tuning Datasets

- **OMAT Replay** (10% subset): We randomly selected 10 million configurations from OMAT as a replay buffer to prevent catastrophic forgetting of foundational inorganic structures during fine-tuning. We will keep this head as our PBE head for the model.
- **RGD1** [44]: Contains 300K configurations of small organic reaction intermediates and transition states computed at the B3LYP/6-31G\* level, enabling accurate modelling of organic reaction pathways.
- **MPTraj**: This dataset comprises 1.5 million configurations from the Materials Project (MP) [45] including static calculations and structural optimisation trajectories. The dataset emphasises dynamic lattice distortions in small periodic unit cells (90% under 70 atoms) describing inorganic crystals with some molecular components. DFT calculations employ the PBE exchange-correlation functional with Hubbard  $U$  terms for selected transition metal oxides [46]. This dataset was originally compiled for CHGNet [21].
- **SPICE-1** [47]: Encompasses  $\sim 500$ K geometries of small to medium-sized organic molecules calculated

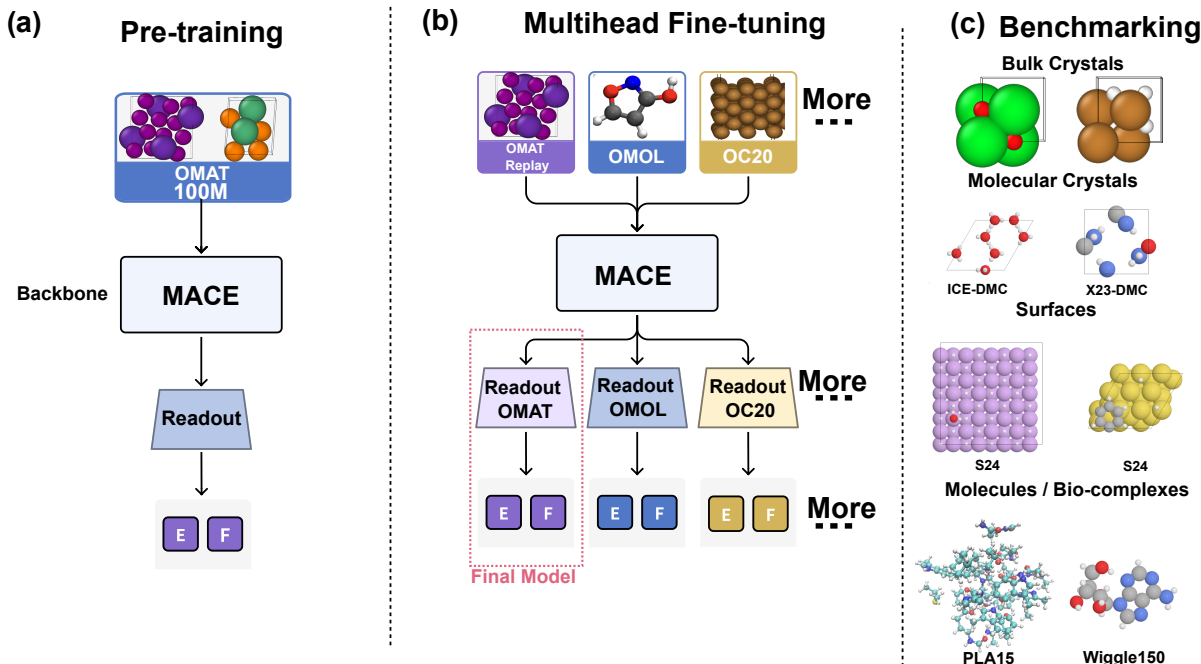


FIG. 1. **Workflow for cross-domain machine learning interatomic potential development.** (a) Stage 1 establishes foundation chemical knowledge through pre-training on large-scale inorganic materials data (OMAT-24). (b) Stage 2 implements multi-head fine-tuning with strategic replay across diverse chemical domains, enabling knowledge transfer while preventing catastrophic forgetting. (c) The resulting unified model is benchmarked across molecular, materials, and surface chemistry tests and achieves state-of-the-art performance.

at the  $\omega$ B97M-D3(BJ)/def2-TZVP level, providing comprehensive coverage of conformational landscapes and intramolecular interactions.

- **OC20** [15]: We randomly subsample 2 million metal surface slabs and adsorbate complexes computed at the PBE level, specifically targeting catalytic surface processes and gas-surface interactions.
- **OMOL-1%** [48]: Subsampled 1.2 million neutral, closed-shell and diverse organic, organometallic, and transition-metal configurations, ensuring comprehensive coverage of coordination chemistries, calculated using hybrid DFT with the  $\omega$ B97M-VV10 functional.
- **MATPES R2SCAN** [49]: Comprises 400K inorganic crystal snapshots sampled via molecular dynamics using machine learning force fields at the  $r^2$ SCAN level (without Hubbard  $U$  corrections).

## V. Model Descriptions

The hyperparameters for all MACE models can be found in the supplementary Table XVII. We benchmark several models to evaluate the effectiveness of our multi-head replay approach:

- **mace-omat-1**: Initial pre-training MACE model on the full OMAT dataset. It uses the new block proposed in IIB.
- **mace-mh-1**: Our proposed MACE model, fine-tuned with multi-head replay fine-tuning on the mace-omat-1 backbone. All of the results in the main text use the single “OMAT head” of the model referred to as **mace-mh-1-omat**. We also benchmark the “OMOL” head fine-tuned on the OMOL dataset referred to as **mace-mh-1-omol-1%**. The 1% reflects that it was trained on just 1% of the full OMOL dataset. We also present results for R2SCAN heads of our models in Table XVIII. For the hyper-parameters of the MACE models, please see Table XVII.
- **mace-omat-0**: MACE model trained on the full OMAT dataset, using the original blocks of MACE, with slightly smaller model size (corresponding to a medium sized model, see the supplementary Table XVII).
- **mace-omol**: MACE model trained on the full 100M OMOL dataset, with total charge and total spin global embedding (denoted in the text as **mace-omol-100%**). As this model is only trained on molecular systems, we only benchmark it when appropriate.

- **uma** [25]: eSEN [50] model trained on 100M inorganic crystals of OMAT, 230M surfaces/small molecules of OC20/OC22, 100M molecular configurations of OMOL, 25M molecular crystals configurations of OMC and 29M metal organic frameworks of ODAC. The dataset types are embedded as one-hot encoding in the model as a global input. We benchmark the OMAT and OMOL variant throughout this paper, with both the S-1.1 and the larger M-1.1 variant referred to as **uma-s-1p1-omat/omol-100%** and **uma-m-1p1-omat/omol-100%** respectively in the text. We use omol-100% to highlight the fact that it is trained on the full dataset.
- **mace-mp-0a** [18]: MACE trained on the MPTraj dataset, representing a strong baseline for inorganic materials modelling.
- **mattersim-5M** [22]: A recent foundation model trained on diverse chemical systems, providing state-of-the-art comparison.
- **orb-v3** [51]: Orb model trained on the OMAT AIMD subset; we use the most accurate orb-v3-conservative-inf-omat in all the tests, referred to as **orb-v3-consv-inf-omat**.

## VI. Benchmark Overview

Figure 2 summarises cross-domain performance using five normalised domain scores—**Materials**, **Molecular Crystals**, **Surfaces**, **Molecules**, and **Physicality**—and a global score defined as a weighted sum (Materials 0.25, Molecules 0.25, Surfaces 0.20, Molecular Crystals 0.20, Physicality 0.10). We evaluate model physicality through several benchmarks assessing size extensivity, additivity, and smoothness of dimer interactions, see the Section XII for details. The breakdown of the weighting of each benchmark towards the scores can be found in Table XX. We acknowledge that aggregating metrics is a difficult task, and our proposed weighting is necessarily subjective. To normalise the scores, we defined two sets of bounds, one for a model that would be deemed inaccurate and one bound for the accuracy at which the benchmark would be saturated, either because it would reach the intrinsic accuracy of the DFT in the case of comparison with wavefunction methods, or because it is a mathematical upper bound. We believe that going forward, these weights need to be improved by a joint discussion of the community. Panel (a) shows a radar plot comparison of domain scores for the leading models. The multi-head model, **mace-mh-1-omat**, attains the most balanced profile, with strong Molecules and Molecular Crystals while remaining competitive on Materials, Surfaces, and Physicality. Panel (b) ranks models by the global score and highlights the same model as the top overall performer, indicating that gains in molecular

chemistry do not come at the expense of bulk materials or surface physics. Panel (c) isolates the architecture trajectory within the MACE family—**mace-omat-0** (linear block)  $\rightarrow$  **mace-omat-1** (non-linear block)  $\rightarrow$  **mace-mh-1-omat**—and shows monotonic improvements in the Molecules and Surfaces categories together with stable Materials performance, which together drive the increase in the global score. The improvements from **mace-omat-0** to **mace-omat-1** are due to two factors: a choice of a slightly larger model size, going from  $L = 1$  messages to  $L = 2$ , and to the changes to the mace architecture outlined in the previous section. Panel (d) plots per-category ranks (lower is better; axis inverted), revealing that the best models are not specialists: they maintain good ranks across all five categories, with especially consistent behaviour for **mace-mh-1-omat**.

In the following sections, we give a detailed performance breakdown of the different tested models on each benchmark, with materials benchmarks in section VIII, molecular crystals in section IX, surfaces in section X, molecular systems in section XI, physicality benchmark in section XII and computational performance in section XIII. In each table, we **bold** the best model(s) and underline the second best model(s). When we benchmark both OMAT and OMOL heads or tasks for different benchmarks, we bold and underline the best and second best models within each task.

## VII. Dispersion corrections (D3)

We incorporate D3(BJ) dispersion corrections [52, 53] when evaluating systems with significant van der Waals interactions with PBE-trained models, using the torchDFTD3 Python package [54] with the PBE parametrization. All D3(BJ) evaluations use the same parameterisation across models to ensure fair comparison. The models trained on OMOL at the  $\omega$ B97M-VV10 level of theory were run without additional dispersion correction.

## VIII. Materials Benchmarks

We evaluate model performance across comprehensive benchmarks covering inorganic materials properties, including elastic moduli, thermal conductivity, and phonon spectra. These benchmarks assess the models’ ability to predict basic physical properties for inorganic materials. These metrics go beyond just energy and force errors by probing the MLIPs’ understanding of the curvature of the PES. Note that most reference data for these benchmarks are computed at the PBE+ $U$  level, matching the MPtraj and OMAT dataset specifications.

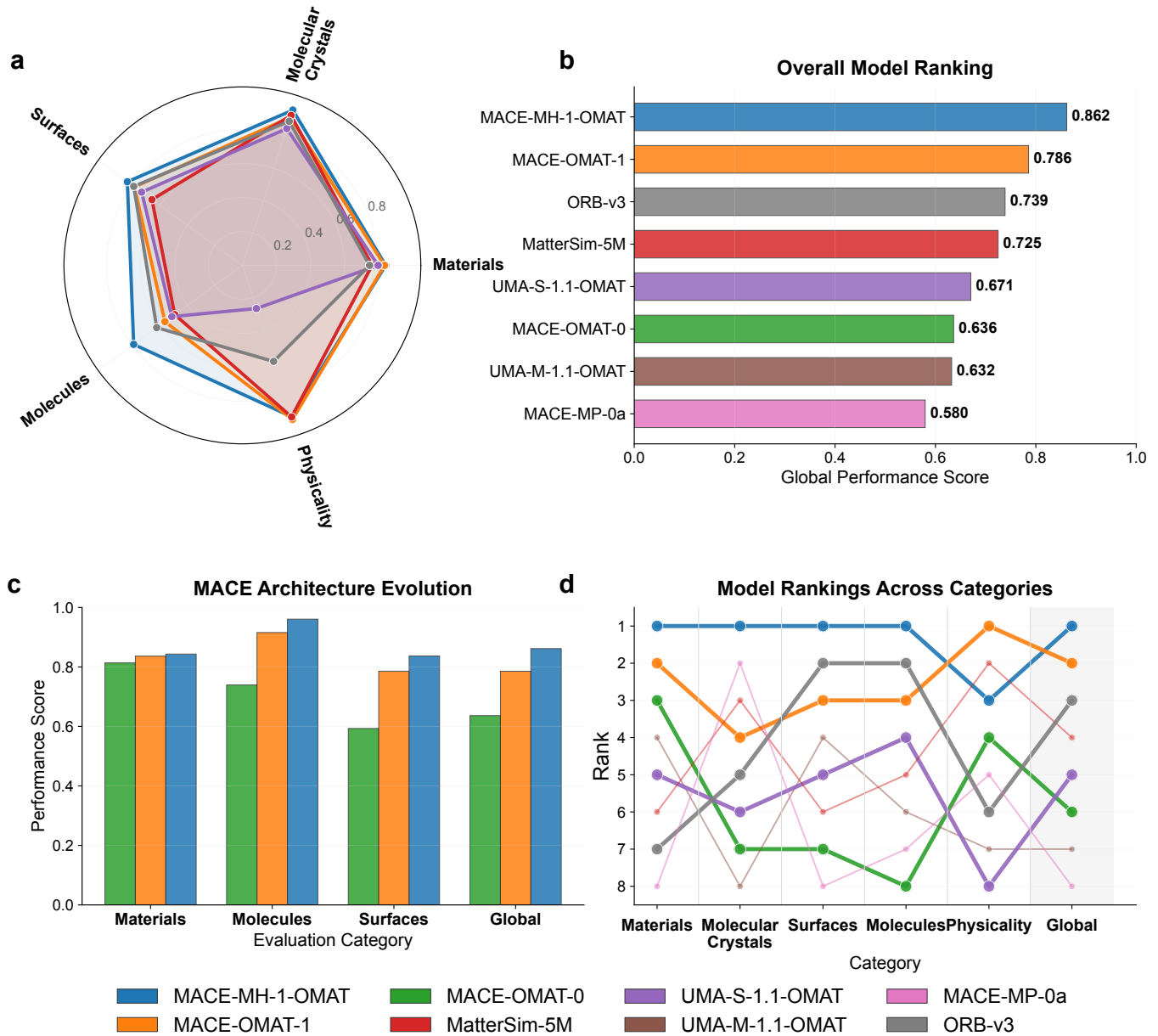


FIG. 2. **Cross-domain performance summary of foundation interatomic potentials.** (a) Polar chart of domain scores for five evaluation groups—Materials, Molecular Crystals, Surfaces, Molecules, and Physicality—where values are normalised to  $[0, 1]$  (1 = best; higher is better) and computed as per-metric means within each group. (b) Overall ranking by a global score, defined as a weighted sum of domain scores (Materials 0.25, Molecules 0.25, Surfaces 0.20, Molecular Crystals 0.20, Physicality 0.10). (c) Ablation within the MACE family comparing the baseline linear block (mace-omat-0), the non-linear block (mace-omat-1), and the multi-head model (mace-mh-1-omat), shown across Materials, Molecules, Surfaces, and the resulting global score. (d) Per-category model ranks (smaller is better; axis inverted so the top is rank 1) illustrating consistency across domains; the shaded region marks the global rank. The scoring procedure is described in Sec. VI and normalisation bounds and metric definitions are given in Table XX.

### A. Elastic Moduli

We benchmark bulk ( $B$ ) and shear ( $G$ ) moduli against DFT PBE and PBE+ $U$  references to evaluate the models’ ability to capture potential energy surface curvature and mechanical response under small strains.

We use MatCalc’s ElasticityCalc [55] to deform the structures with normal (diagonal) strain magnitudes of  $\pm 0.01$  and  $\pm 0.005$  for  $\epsilon_{11}$ ,  $\epsilon_{22}$ ,  $\epsilon_{33}$ , and off-diagonal strain magnitudes of  $\pm 0.06$  and  $\pm 0.03$  for  $\epsilon_{23}$ ,  $\epsilon_{13}$ ,  $\epsilon_{12}$ . The stress tensor  $\sigma$  is calculated for each applied strain  $\epsilon_{ij}$  and the resulting sets of stress-strain pairs are used in linear regression to obtain the elastic tensor [56]. Then,

the bulk and shear moduli are obtained from the elastic and stress tensors via the Voigt-Reuss-Hill (VRH) average, which combines the upper (Voigt) and lower (Reuss) bounds of the moduli for a more accurate approximation [57–59].

TABLE I. Bulk and shear moduli benchmark results. Materials Project elasticity dataset containing 12,122 materials (excluding  $B$  or  $G < -50$  GPa and  $B$  or  $G > 600$  GPa). Both the initial and deformed structures were relaxed.

| Model                 | B MAE (GPa)   | G MAE (GPa)   |
|-----------------------|---------------|---------------|
| mace-mh-1-omat        | 12.49         | <b>7.95</b>   |
| mace-omat-1           | 11.50         | 8.09          |
| mace-mp-0a            | 11.02         | 20.86         |
| mace-omat-0           | 12.47         | 8.95          |
| mattersim-5M          | <u>10.47</u>  | 9.69          |
| orb-v3-consv-inf-omat | <b>7.18</b>   | <u>8.03</u>   |
| uma-m-1p1-omat        | 13.60         | 9.65          |
| uma-s-1p1-omat        | 14.33         | 8.18          |
| mace-mh-1-omol-1%     | 18.00         | 10.77         |
| uma-m-1p1-omol-100%   | Not converged | Not converged |
| uma-s-1p1-omol-100%   | Not converged | Not converged |

Table I reports mean absolute errors for predicted bulk and shear moduli against Materials Project elasticity dataset references [56]. The **orb-v3-consv-inf-omat** model achieves the lowest bulk modulus error, while **mace-mh-1-omat** achieves the lowest shear modulus error. Overall, the **mace-mh-1-omat** model demonstrates competitive performance, with accuracy comparable to the backbone OMAT model, confirming that multi-head replay fine-tuning effectively retains bulk property accuracy.

## B. Thermal Conductivity

Thermal conductivity represents a critical property for electronics, thermoelectrics, and energy storage applications. We employ a comprehensive thermal conductivity benchmark [60, 61] evaluating both microscopic anharmonic phonon properties and resulting lattice thermal conductivity, capturing particle-like (Boltzmann transport equation) and wave-like (Wigner) heat-transport mechanisms across 103 chemically and structurally diverse solids at near first-principles accuracy.

The benchmark encompasses 103 binary crystals—rock salt, zinc-blende, and wurtzite phases spanning 34 chemical elements—with accompanying first-principles harmonic and anharmonic force-constant data enabling reference Wigner-transport conductivities and mode-resolved metrics for rigorous quantitative comparison.

Table II presents root-mean-square errors for predicted lattice thermal conductivity on the 103-compound Wigner-transport benchmark. The **uma-m-1p1-omat** model achieves the lowest  $\kappa_{\text{RMSE}}$ , closely followed by **mace-omat-1** and **uma-s-1p1-omat**. Our **mace-mh-1-omat** model maintains strong performance compared

TABLE II. Thermal conductivity benchmark performance

| Model                 | $\kappa_{\text{RMSE}}$ (W/mK) |
|-----------------------|-------------------------------|
| mace-mh-1-omat        | 0.24                          |
| mace-omat-1           | <u>0.20</u>                   |
| mace-mp-0a            | 0.62                          |
| mace-omat-0           | 0.24                          |
| mattersim-5M          | 0.57                          |
| orb-v3-consv-inf-omat | 0.21                          |
| uma-m-1p1-omat        | <b>0.17</b>                   |
| uma-s-1p1-omat        | <u>0.20</u>                   |

to the baseline, confirming robust bulk property prediction.

## C. Phonons

We evaluate phonon frequencies and derived thermodynamic properties using the Materials Data Repository (MDR) phonon benchmark [62], testing approximately 10,000 materials against reference PBE+ $U$  calculations. We assess maximum, average, and minimum phonon frequencies, plus mean absolute error (MAE) across the Brillouin zone. Additional thermodynamic properties include entropy ( $S$ ), Helmholtz free energy ( $F$ ), and heat capacity at constant volume ( $C_V$ ) at 300 K, all derived from phonon frequencies. Structures are relaxed with fixed symmetries matching DFT references, and phonon frequencies are computed using finite difference methods with identical displacement parameters. For the “no-sym” pipeline, we deliberately do not apply symmetry fixing because we observed numerical instabilities in finite-displacement workflows when enforcing symmetry constraints for these models; all “no-sym” results are therefore obtained without symmetry restoration. The full set of results showing all models with and without symmetry constraints applied is shown in Table XIX.

TABLE III. MDR Phonon benchmark on the phonon frequencies and thermodynamic properties (300 K) of roughly ten thousand materials. Each column corresponds to the MAE of the named quantity. BZ refers to the MAE across the whole Brillouin Zone.

| Model                          | $\omega_{\text{max}}$<br>(K) | $\omega_{\text{avg}}$<br>(K) | $\omega_{\text{min}}$<br>(K) | BZ<br>(K) | $S$<br>(J/mol-K) | $F$<br>(kJ/mol) | $C_V$<br>(J/mol-K) |
|--------------------------------|------------------------------|------------------------------|------------------------------|-----------|------------------|-----------------|--------------------|
| mace-mh-1-omat                 | 12                           | <b>3</b>                     | <b>11</b>                    | <b>5</b>  | <u>8</u>         | <b>2</b>        | <u>3</u>           |
| mace-omat-1                    | 13                           | <b>3</b>                     | <u>12</u>                    | 8         | <u>8</u>         | <b>2</b>        | <u>3</u>           |
| mace-mp-0a                     | 65                           | 32                           | 19                           | 33        | 60               | 23              | 14                 |
| mace-omat-0                    | 16                           | 4                            | 13                           | 7         | 10               | 3               | 3                  |
| mattersim-5M                   | 19                           | 5                            | 16                           | 10        | 14               | 4               | 4                  |
| orb-v3-consv-inf-omat (no-sym) | 12                           | 5                            | 29                           | 15        | 13               | 3               | 4                  |
| uma-m-1p1-omat (no-sym)        | <b>9</b>                     | <b>3</b>                     | 18                           | 8         | <u>8</u>         | <b>2</b>        | <b>2</b>           |
| uma-s-1p1-omat (no-sym)        | <u>11</u>                    | 4                            | 21                           | 9         | <b>7</b>         | <b>2</b>        | <u>3</u>           |
| mace-mh-1-omol-1%              | <b>49</b>                    | <b>12</b>                    | <b>18</b>                    | <b>16</b> | <b>16</b>        | <b>7</b>        | <b>7</b>           |
| uma-s-1p1-omol-100% (no-sym)   | 155                          | 50                           | 64                           | 64        | 82               | 32              | 19                 |

Table III shows mean absolute errors for phonon frequencies and thermodynamic properties. The **mace-mh-1-omat** model achieves the lowest errors for most

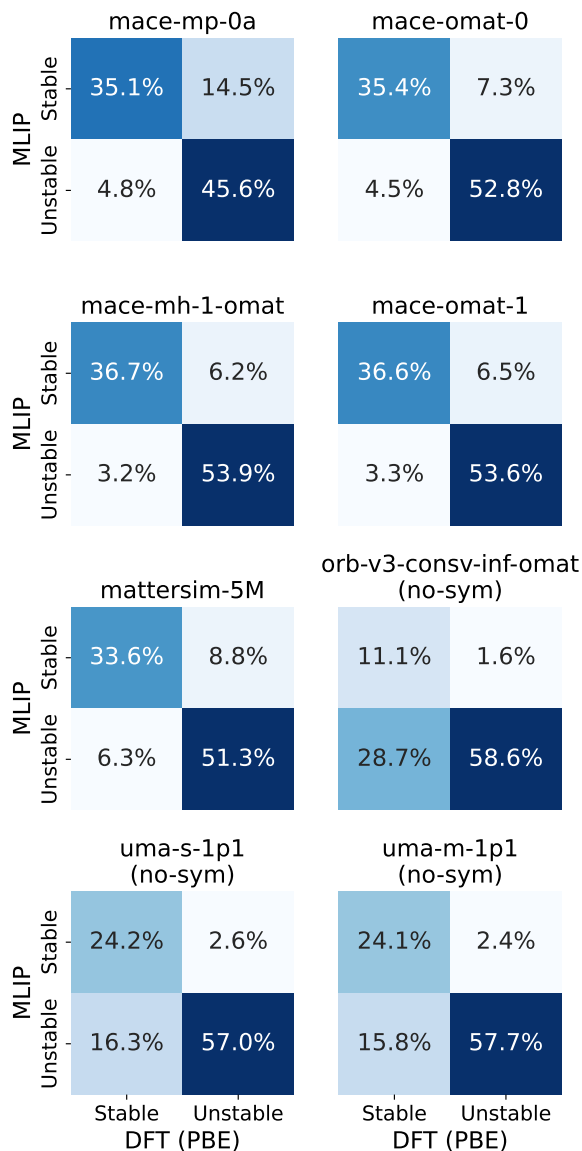


FIG. 3. **Phonon dynamical stability classification confusion matrices.** Materials are classified as unstable if  $|\omega_{\text{imag}}| > 0.05$  THz ( $\approx 2.4$  K).

properties, only ranked second best behind **uma-m-1p1-omat** for  $C_V$ , third behind **uma-s-1p1-omat** and **uma-m-1p1-omat** for  $\omega_{\text{max}}$  and second behind **uma-s-1p1-omat** for  $S$ , demonstrating excellent phonon frequency and thermodynamic property prediction. Notably, symmetry fixing significantly impacted **uma** and **orb-v3** models’ performance while leaving other models unaffected. Our **mace-mh-1-omat** model maintains accuracy or slightly outperforms the backbone model **mace-omat-1**, confirming the effectiveness of retaining high bulk materials accuracy through multi-head replay fine-tuning. In Figure 3, we report dynamical stability classification performance, showing that our **mace-mh-1-omat** achieves the best overall classification perfor-

mance. We also compare the OMOL head of **mace-mh-1** and of **uma-s-1p1** to further assess cross-learning in multidomain models. We observe a clearly superior degree of cross-learning for **mace-mh-1** than **uma-s-1p1**, with the mace model achieving a phonon performance superior in each domain, even outperforming the specialised MPTraj model **mace-mp-0a**. The **uma-m-1p1-omol** was unable to geometry optimise a large portion of the structures, so we did not include it in the table.

## IX. Molecular Crystal Benchmarks

Molecular crystal formation energies test the models’ accuracy in describing intermolecular interactions and the cohesive assembly of molecular solids.

### A. X23 Molecular Crystals

The X23 benchmark [63] comprises 23 experimentally characterised organic molecular crystals selected to span diverse noncovalent binding motifs, including hydrogen bonding, dispersion-dominated packing, and mixed electrostatic-van der Waals interactions. Reference formation enthalpies are computed via periodic calculations using Diffusion Monte Carlo (DMC) [64]. Comparing per-molecule cohesive energies against these high-level references evaluates the ability to capture subtle intermolecular forces governing crystal stability.

TABLE IV. X23 benchmark formation energy mean absolute errors.

| Model                 | MAE (kJ/mol) |
|-----------------------|--------------|
| mace-mh-1-omat-D3     | <b>15.82</b> |
| mace-omat-1-D3        | 19.60        |
| mace-mp-0a-D3         | <u>17.42</u> |
| mace-omat-0-D3        | 53.23        |
| mattersim-5M-D3       | 20.11        |
| orb-v3-consv-inf-omat | 28.76        |
| uma-m-1p1-omat-D3     | 78.89        |
| uma-s-1p1-omat-D3     | 27.99        |
| mace-mh-1-omol-1%     | <u>7.41</u>  |
| uma-m-1p1-omol-100%   | 9.33         |
| uma-s-1p1-omol-100%   | <b>6.81</b>  |

Table IV shows that **mace-mh-1-omat-D3** achieves the lowest MAE among the tested models for the X23 benchmark, indicating improved molecular crystal formation energy prediction. Notably, it outperforms **uma-s-1p1-omat-D3**, despite the latter being trained on larger datasets including 20M molecular crystals. The substantial improvement over **mace-omat-1-D3** demonstrates clear knowledge transfer of molecular chemistry from OMOL and SPICE heads to the PBE head, transferring molecular chemical knowledge. We observe that the OMOL-trained models outperform the PBE models,

which is likely due to a much better description of intermolecular interaction by the OMOL range-separated hybrid DFT  $\omega$ B97M-VV10.

### B. DMC Water Ice Polymorphs

We test model performance on formation energies of common ice phases (Ih, II, III, etc.), benchmarking hydrogen-bonded networks under different pressures [65].

TABLE V. Ice phase mean absolute errors

| Model                 | MAE (meV)     |
|-----------------------|---------------|
| mace-mh-1-omat-D3     | <b>11.23</b>  |
| mace-omat-1-D3        | 144.57        |
| mace-mp-0a-D3         | <u>59.79</u>  |
| mace-omat-0-D3        | 447.08        |
| mattersim-5M-D3       | 96.95         |
| orb-v3-consv-inf-omat | 138.44        |
| uma-m1-p1-D3          | 545.72        |
| uma-s1-p1-D3          | 310.82        |
| mace-mh-1-omol-1%     | <b>79.60</b>  |
| uma-m-1p1-omol-100%   | 120.62        |
| uma-s-1p1-omol-100%   | <u>109.20</u> |

Table V demonstrates that **mace-mh-1-omat-D3** achieves the lowest MAE for ice polymorph relative energies among the tested models, indicating improved modelling of hydrogen bonding interactions in this benchmark. The significant improvement over **mace-omat-1-D3** again demonstrates effective fine-tuning enhancement for molecular crystal systems and better description of subtle intermolecular interactions. Surprisingly, the **mace-mh-1-omat-D3** even outperforms the omol-trained models.

### X. Surface Benchmarks

Surface interaction benchmarks probe adsorption energetics and site-specific binding, critical for catalysis and gas-surface processes.

#### A. General Surface Adsorption: S24 Dataset

Accurately describing molecule-surface interactions at the first-principles level is essential for designing advanced catalysts, gas-separation membranes, and sensing materials. We employ the S24 benchmark set [18], covering 24 prototypical adsorption systems across diverse surface types:

- **Covalent surfaces:** graphene, silicene, and other 2D covalent networks
- **Ionic surfaces:** alkali halide (e.g., NaCl) and metal oxide (e.g., MgO) slabs

- **Metallic facets:** close-packed (111) and stepped surfaces of Pt, Au, and Cu
- **Porous materials:** representative metal-organic frameworks (MOFs) and zeolites

Reference adsorption energies are computed at PBE+D3(BJ) level using standardised VASP/MPRelaxSet protocols. We report mean absolute deviations for predicted versus reference energies (in eV) both overall and per surface category.

TABLE VI. S24 benchmark mean absolute errors for the adsorption energies of small molecules on surfaces and porous materials.

| Model                    | MAE (eV)     |
|--------------------------|--------------|
| mace-mh-1-omat-D3        | <b>0.095</b> |
| mace-omat-1-D3           | <u>0.140</u> |
| mace-mp-0a-D3            | 0.152        |
| mace-omat-0-D3           | 0.550        |
| mattersim-5M-D3          | 0.141        |
| orb-v3-consv-inf-omat-D3 | 0.174        |
| uma-m-1p1-omat-D3        | 0.523        |
| uma-s-1p1-omat-D3        | 0.329        |
| mace-mh-1-omol-1%        | <b>0.288</b> |
| uma-m-1p1-omol-100%      | 10.774       |
| uma-s-1p1-omol-100%      | <u>4.636</u> |

Table VI shows that **mace-mh-1-omat-D3** achieves the lowest MAE for surface adsorption energies, demonstrating a good description of interactions of molecules with surfaces and porous materials. In particular, we observe a 68% improvement compared to the backbone model **mace-omat-1-D3** on the adsorption energies. When testing the OMOL tasks of the models, we observe much better performance for the **mace-mh-1-omol-1%** compared to the **uma** models confirming superior cross-learning on this benchmark.

#### B. OC20 Metal Surface Benchmark

We benchmark small-molecule adsorption on metal surfaces using structures generated during the Open Catalyst Challenge 2023 [16, 66] to evaluate catalytic reaction intermediate modelling. We use the set of configurations from the Open Catalyst Challenge 2023 recomputed at the MPtraj DFT level introduced in Ref. [18].

We observe that **uma-m-1p1-omat-D3** achieves the best performance with an MAE of 0.097 eV followed by **mace-mh-1-omat-D3** with an MAE of 0.138 eV. Note that **uma-m-1p1-omat-D3** and **uma-s-1p1-omat-D3** are trained on the full OC20 dataset representing 100 times more surface configurations than our **mace-mh-1-omat-D3** model. When benchmarking the OMOL tasks, we confirm the better cross-learning of the **mace-mh** model compared to the **uma** models.

TABLE VII. Open Catalyst Challenge 2023 generated set of adsorption energies of small molecules on metallic surfaces.

| Model                    | MAE (eV)     | Pearson's r |
|--------------------------|--------------|-------------|
| mace-mh-1-omat-D3        | <u>0.138</u> | <u>0.98</u> |
| mace-omat-1-D3           | 0.171        | 0.97        |
| mace-mp-0a-D3            | 0.412        | 0.86        |
| mace-omat-0-D3           | 0.243        | 0.94        |
| mattersim-5M-D3          | 0.285        | 0.92        |
| orb-v3-consv-inf-omat-D3 | 0.159        | 0.974       |
| uma-m-1p1-omat-D3        | <b>0.097</b> | <b>0.99</b> |
| uma-s-1p1-omat-D3        | 0.172        | 0.97        |
| mace-mh-1-omol-1%        | <b>0.214</b> | <b>0.96</b> |
| uma-m-1p1-omol-100%      | 3.394        | 0.28        |
| uma-s-1p1-omol-100%      | <u>1.120</u> | <u>0.50</u> |

## XI. Molecular Benchmarks

We evaluate model performance across comprehensive molecular property benchmarks, including conformer energies, reaction energies, and noncovalent interactions, assessing capability to predict key chemical properties governing molecular behaviour.

### A. Wiggle 150

Wiggle150 is a benchmark comprising 150 highly strained conformations of adenosine, benzylpenicillin, and efavirenz. Table VIII reports mean absolute errors for strained conformer relative energies against Wiggle150 benchmark references [67] computed at the DLPNO-CCSD(T)/CBS level of theory.

TABLE VIII. Wiggle150 strained conformer relative energy benchmark with D3(BJ) dispersion correction

| Model                    | MAE (kcal mol <sup>-1</sup> ) |
|--------------------------|-------------------------------|
| $\omega$ B97M-D3         | 1.18                          |
| PBE-D3                   | 4.91                          |
| mace-mh-1-omat-D3        | <b>4.80</b>                   |
| mace-omat-1-D3           | 10.39                         |
| mace-mp-0a-D3            | 25.90                         |
| mace-omat-0-D3           | 9.62                          |
| mattersim-5M-D3          | 12.12                         |
| orb-v3-consv-inf-omat-D3 | 7.65                          |
| uma-m-1p1-omat-D3        | <u>5.04</u>                   |
| uma-s-1p1-omat-D3        | 6.60                          |
| mace-mh-1-omol-1%        | 1.30                          |
| mace-omol-100%           | <b>0.83</b>                   |
| uma-m-1p1-omol-100%      | 0.93                          |
| uma-s-1p1-omol-100%      | <u>0.91</u>                   |

Our **mace-mh-1-omat-D3** model achieves 4.80 kcal mol<sup>-1</sup> MAE, comparable to the underlying PBE-D3 functional performance. The multi-head fine-tuning im-

proves by a factor of 2 over the pre-trained model's baseline (**mace-omat-1-D3**).

### B. GMTKN55 Main Group Chemistry

We evaluate gas-phase chemical accuracy using the GMTKN55 benchmark suite [68], which represents the most comprehensive test of electronic structure methods for main group thermochemistry, kinetics and noncovalent interactions. The complete suite comprises 55 individual test sets totalling 1,505 relative energies, systematically categorized into five chemical domains:

- 1. Basic Properties (8 sets):** Atomic energies, ionization potentials, electron affinities
- 2. Reaction Energies (15 sets):** Thermochemistry including G2/97, G3/99 test sets
- 3. Barrier Heights (12 sets):** Transition state energetics for fundamental organic reactions
- 4. Intramolecular Noncovalent (7 sets):** Conformational energetics, hydrogen bonding
- 5. Intermolecular Noncovalent (13 sets):** Dimers, clusters, host-guest complexes

All reference energies employ high-level coupled cluster methods, primarily CCSD(T) extrapolated to the complete basis set limit with core correlation corrections where appropriate. The weighted total mean absolute deviation (WTMAD) provides a single metric combining performance across all chemical domains, with typical values of 2-3 kcal mol<sup>-1</sup> representing chemical accuracy for electronic structure methods. We keep only the neutral, singlet subset as most of the models do not handle charged systems yet.

TABLE IX. GMTKN55 subset mean absolute errors across benchmark categories (neutral singlet subset only). All values represent weighted mean absolute errors in kcal mol<sup>-1</sup> (lower is better) with weights taken from WMAD-2.

| Model                    | Basic Small  | Large Systems | Barrier Heights | Intra-Noncov | Inter-Noncov | All          |
|--------------------------|--------------|---------------|-----------------|--------------|--------------|--------------|
| $\omega$ B97M-D3BJ       | 2.86         | 5.77          | 2.34            | 4.54         | 3.63         | 4.04         |
| PBE-D3(BJ)               | 11.24        | 10.61         | 13.16           | 9.95         | 9.89         | 10.58        |
| mace-mh-1-omat-D3        | <b>12.88</b> | <b>13.74</b>  | <b>9.58</b>     | <b>11.45</b> | <b>9.54</b>  | <b>11.23</b> |
| mace-omat-1-D3           | 25.05        | 33.65         | 18.95           | <u>26.47</u> | <u>20.59</u> | 24.90        |
| mace-mp-0a-D3            | 42.62        | 59.75         | 31.41           | 59.60        | 23.27        | 45.04        |
| mace-omat-0-D3           | 40.61        | 44.51         | 30.49           | 57.73        | 111.69       | 64.15        |
| mattersim-5M-D3          | 29.27        | 41.36         | 20.34           | 37.40        | 23.50        | 31.46        |
| orb-v3-consv-inf-omat-D3 | 26.30        | <u>14.72</u>  | <u>13.88</u>    | 28.09        | 23.08        | <u>22.30</u> |
| uma-m-1-p1-omat-D3       | <u>25.02</u> | 20.77         | 28.34           | 23.33        | 138.21       | 52.89        |
| uma-s-1-p1-omat-D3       | 29.35        | 28.31         | 15.24           | 33.64        | 36.75        | 30.83        |
| AIMNet2                  | 11.91        | 18.67         | 10.80           | 12.19        | 20.76        | 15.15        |
| MACE-OFF23(L)            | <u>8.74</u>  | 10.40         | 22.52           | 6.79         | 37.84        | 16.46        |
| mace-mh-1-omol-1%        | 10.62        | 7.64          | 7.10            | 7.47         | 10.11        | 8.44         |
| mace-omol-100%           | 9.41         | 4.33          | 3.06            | <b>4.96</b>  | 3.59         | 4.65         |
| uma-m-1p1-omol-100%      | 12.18        | <u>3.93</u>   | <u>2.91</u>     | <b>4.96</b>  | <b>2.34</b>  | <u>4.49</u>  |
| uma-s-1p1-omol-100%      | <b>8.23</b>  | <b>3.50</b>   | <b>2.76</b>     | 5.15         | <u>2.86</u>  | <b>4.22</b>  |

Table IX shows that **mace-mh-1-omat-D3** significantly outperforms other PBE-trained foundational models, including **uma-m-1-p1-omat-D3** achieving a mean error on the GMTKN55 of 11.23 kcal mol<sup>-1</sup>, comparable to the accuracy of PBE-D3(BJ). The substantial improvement over **mace-omat-1-D3** confirms the effective knowledge transfer from the molecular heads (SPICE and OMOL) to the material head. Notably, our model achieves performance comparable to specialised molecular models (AIMNet2, MACE-OFF23(L)) that were trained on hybrid meta-GGA functional  $\omega$ B97M-D3BJ, which is much more accurate for molecular system.

### C. Protein fragments: PLF547

Accurately capturing intramolecular hydrogen bonding and side-chain/backbone contacts in polypeptide fragments is a stringent test for any MLIP intended to model biomolecular chemistry. We therefore evaluate on the PLF547 benchmark suite [69] of 547 shorter peptide-like fragments (PLF547) with interaction energies referenced to high-level DLPNO-CCSD(T)/CBS calculations. Following prior work, we compute single-point energies on the published geometries and report the mean absolute error between the predicted and reference interaction energies, which provides a good proxy to the quality of intermolecular interactions. We used only the subset of molecules that are neutral singlet.

TABLE X. PLF547 neutral subset interaction energy MAE (kcal mol<sup>-1</sup>) for the different tested models.

| Model                    | PLF547 MAE (kcal mol <sup>-1</sup> ) |
|--------------------------|--------------------------------------|
| mace-mh-1-omat-D3        | <b>0.626</b>                         |
| mace-omat-1-D3           | <u>0.839</u>                         |
| mace-mp-0a-D3            | 1.040                                |
| mace-omat-0-D3           | 4.926                                |
| mattersim-5M-D3          | 1.017                                |
| orb-v3-consv-inf-omat-D3 | 1.829                                |
| uma-m-1p1-omat-D3        | 12.057                               |
| uma-s-1p1-omat-D3        | 2.935                                |
| mace-omat-100%           | <b>0.334</b>                         |
| mace-mh-1-omat-1%        | <u>0.394</u>                         |
| uma-m-1p1-omat-100%      | 0.839                                |
| uma-s-1p1-omat-100%      | 0.655                                |

Table X summarizes the results. The **mace-mh-1-omat-D3** model achieves the lowest MAE, with 0.626 kcal mol<sup>-1</sup>, substantially improving over the OMAT-only backbone. Overall, these results demonstrate that cross-domain replay enhances the model’s ability to describe biomolecular fragment energetics.

### D. S30L Molecular complexes

Large host-guest and  $\pi - \pi$  stacked complexes probe the long-range dispersion and subtle many-body polarization effects that drive supramolecular binding. We assess these regimes with the S30L benchmark [70], a set of 30 noncovalent complexes ranging from crown-ether inclusion compounds to charged receptor-ligand pairs. Reference binding energies are empirical binding energies obtained by back-correcting the experimental association free energies, making S30L a challenging benchmark for dispersion-dominated interactions.

TABLE XI. S30L [70] benchmarking of host-guest binding energies in large molecular complexes. Binding energies mean absolute errors in kcal mol<sup>-1</sup> compared to experimental references.

| Model                    | MAE (kcal mol <sup>-1</sup> ) |
|--------------------------|-------------------------------|
| mace-mh-1-omat-D3        | <b>10.13</b>                  |
| mace-omat-1-D3           | 15.35                         |
| mace-mp-0a-D3            | 14.22                         |
| mace-omat-0-D3           | 25.87                         |
| mattersim-5M-D3          | <u>11.92</u>                  |
| orb-v3-consv-inf-omat-D3 | 13.64                         |
| uma-m-1p1-omat-D3        | 13.09                         |
| uma-s-1p1-omat-D3        | 15.14                         |
| mace-mh-1-omat-1%        | <b>6.66</b>                   |
| uma-m-1p1-omat-100%      | <u>7.66</u>                   |
| uma-s-1p1-omat-100%      | 14.59                         |

We evaluate single-point interaction energies on the published geometries and report mean absolute errors (MAE) in kcal mol<sup>-1</sup> (Table XI). The **mace-mh-1-omat-D3** model attains the best performance, with an MAE of 10.13 kcal mol<sup>-1</sup>, substantially outperforming pre-trained model **mace-omat-1-D3** at 15.35 kcal mol<sup>-1</sup>. In contrast, UMA models—despite training on vastly more molecular data—achieve larger errors, reinforcing that architectural choices on how to merge different datasets are critical for enhancing cross-learning behaviour. Note also that the total charge information was not given to the UMA OMAT task models as it was causing very large errors, which highlights that naive inclusion of total charge via global embedding is not transferable between the models’ tasks.

## XII. Physicality Benchmarks

While extensive accuracy benchmarks are essential, ensuring physically realistic predictions is equally important, given the models’ broad applicability across chemical domains that makes exhaustive validation challenging. We evaluate model physicality through several benchmarks assessing size extensivity, additivity, and smoothness of pair interactions.

### A. Size Extensivity and Locality

The size extensivity and locality of the potential energy surface represent fundamental quantum mechanical properties, ensuring that the energy scales correctly with particle number and that sufficiently separated system energies equal their component sums. Large violations of these properties can lead to unphysical simulations.

TABLE XII. Slab test for size extensivity evaluation.  $\Delta$  refers to the difference between the isolated slab energies and the combined system energy ( $\Delta = E_{1,2} - (E_1 + E_2)$ ). Results to 1 d.p.

| Model                 | $E_1$<br>(eV) | $E_2$<br>(eV) | $E_{12}$<br>(eV) | $\Delta$<br>(meV) |
|-----------------------|---------------|---------------|------------------|-------------------|
| mace-mh-1-omat        | -468.2        | -615.0        | -1083.3          | <b>0.0</b>        |
| mace-omat-1           | -467.3        | -617.0        | -1084.3          | <b>0.0</b>        |
| mace-mp-0a            | -468.1        | -649.5        | -1117.6          | <b>0.0</b>        |
| mace-omat-0           | -468.8        | -616.0        | -1084.7          | <b>0.0</b>        |
| mattersim-5M          | -467.8        | -646.7        | -1114.5          | 0.2               |
| orb-v3-consv-inf-omat | -466.3        | -614.1        | -1081.1          | -709.7            |
| uma-m-1p1-omat        | -467.2        | -616.4        | -1081.2          | 1436.9            |
| uma-s-1p1-omat        | -467.4        | -616.6        | -1084.5          | -453.8            |
| mace-mh-1-omol-1%     | -844543.4     | -5253593.4    | -6098136.8       | <b>0.0</b>        |
| uma-m-1p1-omol-100%   | -843663.9     | -5253001.4    | -6097051.2       | -385824.5         |
| uma-s-1p1-omol-100%   | -843500.5     | -5253244.0    | -6096655.8       | 88670.6           |

For the slab test, we compare the sum of the energies of isolated FCC(111) aluminium ( $E_1$ ) and nickel ( $E_2$ ) slabs, each 8 layers thick in a 4x4 in-plane supercell, against the energy of the combined configuration ( $E_{12}$ ), with a 100 Å gap. The expected non-interacting behaviour is confirmed if  $E_{12} = E_1 + E_2$ .

Table XII shows all models except UMA models, and ORB-V3-ConsV-Inf-omat maintain proper extensivity, while UMA models and ORB models exhibit large energy deviations, indicating the presence of unphysical interactions. For the UMA models, this non-local interaction is likely arising from global chemical element embeddings creating non-local interactions. For the ORB model, it is due to its non-local readout in which it passes a summed or averaged energy over the entire structure into a non-linear function, creating potential unphysical non-local interactions.

We further test size additivity by placing a hydrogen atom 50 Å from an aluminium slab and computing forces at various distances. Table XIII shows all models except **uma-s-1p1-omat**, **uma-m-1p1-omat** and **orb-v3-consv-inf-omat** produce zero forces as expected for non-interacting systems. The **uma** models and orb models exhibit large spurious forces, with for example **uma-s-1p1-omat** showing a spurious force of max 969.2 meV/Å, confirming large unphysical interactions.

TABLE XIII. Addition of a hydrogen atom to an aluminium slab at 50 Å distance to test for size additivity.

| Model                 | max $ \Delta F $<br>(meV/Å) | mean $ \Delta F $<br>(meV/Å) | std $ \Delta F $<br>(meV/Å) |
|-----------------------|-----------------------------|------------------------------|-----------------------------|
| mace-mh-1-omat        | <b>0.0000</b>               | <b>0.0000</b>                | <b>0.0000</b>               |
| mace-omat-1           | <b>0.0000</b>               | <b>0.0000</b>                | <b>0.0000</b>               |
| mace-mp-0a            | <b>0.0000</b>               | <b>0.0000</b>                | <b>0.0000</b>               |
| mace-omat-0           | <b>0.0000</b>               | <b>0.0000</b>                | <b>0.0000</b>               |
| mattersim-5M          | <b>0.0000</b>               | <b>0.0000</b>                | <b>0.0000</b>               |
| orb-v3-consv-inf-omat | 61.65                       | 19.21                        | <b>0.0000</b>               |
| uma-m-1p1-omat        | 1520.0                      | 11.73                        | 0.0006                      |
| uma-s-1p1-omat        | 969.20                      | 16.48                        | 0.0005                      |
| mace-mh-1-omol-1%     | <b>0.0000</b>               | <b>0.0000</b>                | <b>0.0000</b>               |
| uma-m-1p1-omol-100%   | 38.17                       | 0.227                        | 0.0003                      |
| uma-s-1p1-omol-100%   | 719.2                       | 1.018                        | 0.0002                      |

### B. Homonuclear and Heteronuclear Diatomics

Diatomic potential curve analysis provides fundamental tests of model smoothness and physicality for low-order body interactions. As diatomic molecules have

TABLE XIV. Homonuclear and heteronuclear diatomic physicality assessment across key smoothness and correlation metrics. Force flips count the number of sign changes in the force, with the ideal being a single flip from repulsive to attractive. Energy minima denote the number of distinct local minima in the potential energy curve, with the physical expectation of only one bound state. Energy inflections capture the number of inflections in the potential energy curve, where the ideal curve would have one inflection point. The Spearman's rank correlation coefficients for the repulsive and attractive regimes have ideal values of  $\rho_{E_{rep}} = -1$  and  $\rho_{E_{at}} = 1$  respectively for perfectly monotonic relationships.

| Model                 | Mean No.<br>Force Flips | Mean No.<br>Energy Minima | Mean No.<br>Energy Inflections | Mean $\rho_{E_{rep}}$ | Mean $\rho_{E_{at}}$ |
|-----------------------|-------------------------|---------------------------|--------------------------------|-----------------------|----------------------|
| mace-mh-1-omat        | 2.09                    | <u>1.42</u>               | 2.72                           | <u>-0.99</u>          | <u>0.88</u>          |
| mace-omat-1           | <b>1.52</b>             | <b>1.18</b>               | <b>2.15</b>                    | <b>-1.00</b>          | 0.82                 |
| mace-mp-0a            | 3.75                    | 2.15                      | 3.32                           | -0.96                 | 0.15                 |
| mace-omat-0           | <u>1.97</u>             | 1.45                      | <u>2.24</u>                    | <b>-1.00</b>          | 0.78                 |
| mattersim-5M          | 2.17                    | 1.61                      | 2.45                           | <b>-1.00</b>          | <b>0.89</b>          |
| orb-v3-consv-inf-omat | 2.91                    | 1.62                      | 2.56                           | -0.98                 | 0.63                 |
| uma-m-1p1-omat        | 8.83                    | 4.16                      | 12.85                          | -0.57                 | 0.56                 |
| uma-s-1p1-omat        | 10.73                   | 4.82                      | 15.55                          | -0.70                 | 0.42                 |
| mace-mh-1-omol-1%     | <b>2.03</b>             | <b>3.39</b>               | <b>3.70</b>                    | <b>-0.99</b>          | <b>0.87</b>          |
| uma-m-1p1-omol-100%   | 12.09                   | 5.56                      | 16.11                          | -0.83                 | <u>0.77</u>          |
| uma-s-1p1-omol-100%   | <u>11.40</u>            | <u>5.27</u>               | <u>16.06</u>                   | <u>-0.93</u>          | 0.76                 |

shown convergence issues for plane-wave DFT reference data, one can instead define a set of general physical requirements for the ideal dimer curve, where the choice of these metrics was inspired by MLIP Arena [71]. Such a curve should exhibit a single energy minimum, a single energy inflection, and one change in force sign (force flip). To quantify smoothness and monotonicity, Spearman's rank correlation coefficients are used. The ideal dimer potential energy curve can be decomposed into two monotonic functions of energy and atomic separation either side of the energy minimum, with a repulsive regime at shorter separations and an attractive regime at

longer separations. Table XIV evaluates diatomic curve quality using mean values of each metric over all compatible homonuclear and heteronuclear diatomics. The **mace-omat-1** and **mace-mh-1-omat** models demonstrate good performance with minimal force flips and energy minima. However, **uma-s-1p1-omat(omol)** and **uma-m-1p1-omat(omol)** exhibit poor diatomic behaviour with numerous force sign flips, energy minima, and inflection points, indicating problematic two-body interaction smoothness. This is confirmed upon visual inspection of the **uma** models’ diatomic curves. We also attach in the supplementary material compressed files containing all the homonuclear diatomic curves for the models tested in the paper.

### XIII. Computational efficiency

The computational efficiency is assessed by comparing the time required to calculate the energy and forces of 1,000 atoms in Carbon FCC structures with a lattice constant  $a = 3.8 \text{ \AA}$  on a single NVIDIA H100 80GB GPU using FP32 (TF32-high precision). Numbers for UMA models and ORB are taken from [25], from which we reproduce the timing protocol. For the new MACE timings, we use both NVIDIA’s cuEquivariance kernels [72] and torch.compile with ‘reduce-overhead’ settings.

TABLE XV. Single GPU (NVIDIA H100 80GB GPU) speed comparison between models to compute energy and forces of a 1000 atoms diamond structure, using torch.compile and for MACE models cuEquivariance [72] kernels, excluding graph construction.

| Model                 | Steps per second |
|-----------------------|------------------|
| mace-mh-1             | 43               |
| mace-omat-1           | 43               |
| mace-mp-0a            | <b>83</b>        |
| mace-omat-0           | <b>83</b>        |
| orb-v3-consv-inf-omat | 30               |
| uma-m-1p1             | 3                |
| uma-s-1p1             | 16               |

We observe that our models achieve competitive speed compared to the other state-of-the-art models in the tested setup, achieving a speed of 43 steps per second (around 3.8 Megasteps/day) in a best-case scenario (neglecting any molecular dynamics (MD) overheads). The smaller inference speed of the **mace-omat-1** and **mace-mh-1-omat** compared to **mace-mp-0a** and **mace-omat-0** is mainly due to a choice of large hyperparameters ( $L=2$  messages and 512 channels for the node features) and not to the architecture changes outlined in Section II B. We note that a fair assessment of the speed of MLIPs is a hard task, that depends on many factors, such as MD drivers, floating-point precisions, compilations and kernels, system density and size, and type of GPU.

TABLE XVI. Single GPU (Nvidia H100 80GB GPU) speed comparison between models to run molecular dynamics NVT simulation of a 1000 atoms diamond structure, using MACE models cuEquivariance [72] kernels in LAMMPS MLIAP (no torch.compile).

| Model       | Mega-steps per day |
|-------------|--------------------|
| mace-mh-1   | 1.4                |
| mace-omat-1 | 1.4                |
| mace-mp-0a  | 2.2                |
| mace-omat-0 | 2.2                |

We also benchmark the MACE models during real molecular dynamics in the LAMMPS MLIAP KOKKOS interface of the same carbon structure, using cuEquivariance in float32 but not torch.compile. We observe around 1.4 to 2.2 mega-steps per day of simulation. Compared to the idealised timings of Table XV, this represents a slowdown of around a factor of 2.5, which is mainly due to not using the torch.compile and to a smaller extent, to the various overheads of graph construction and LAMMPS updates. We are working on further engineering optimisations that will enable us to get closer to the ideal models’ speed in real simulations.

### XIV. Discussion and Outlook

Our results provide strong evidence for effective knowledge transfer across chemical domains through shared representational learning. The substantial improvement observed in molecular systems for the OMAT head when including the SPICE and OMOL datasets as part of the multi-head model demonstrates that molecular knowledge can be transferred to the material head through an improved description of local atomic environments and coordination patterns. The multi-head architecture enables knowledge sharing while maintaining consistency across different levels of electronic structure theory in its loss functions. We observe that the more flexible global embedding of the level of theory of UMA does not exhibit a similar level of transfer of knowledge, suggesting that architecture choices and reduced flexibility may enhance efficient transfer..

This work establishes the foundation for next-generation simulation capabilities where single models seamlessly handle complex multiscale phenomena spanning molecular, surface, and materials chemistry. As the field moves toward foundation MLIPs that are out of the box as accurate as a GGA DFTs for most of chemistry with a single model, the principles demonstrated here provide valuable guidance for achieving both breadth and accuracy in chemical modelling applications.

Several promising extensions emerge from this work: (i) incorporation of additional chemical domains including solid/liquid interfaces [17], molecular crystals [73], or amorphous systems, (ii) development of uncertainty

quantification methods for reliable out-of-domain predictions, (iii) integration with experimental data through hybrid learning approaches, and (iv) extension to charged and magnetic systems with the inclusion of electrostatic interactions and spin states (available in some datasets) that will further enhance both accuracy and transferability.

### Supplementary Material

The supplementary material contains details of the hyperparameters of the models, weighting schemes for the benchmark scores, and additional R2SCAN model results. We also attach in the supplementary material compressed files containing all the homonuclear diatomic curves for the models tested in the paper.

### Conflict of Interest

GC is a partner in Symmetric Group LLP that licenses force fields commercially and also has equity interest in Ångström AI. SWN has financial interest and equity stake in Mirror Physics, a company working on AI and atomistic modelling.

### Data and Models Availability

The MACE code is available on <https://github.com/ACEsuit/mace> and example input scripts and pre-trained models to reproduce the results are provided on the MACE foundation GitHub: <https://github.com/ACEsuit/mace-foundations>. The datasets used for training are all public and referenced in the text. Processed data and analysis scripts to reproduce the benchmarks will be made available in an upcoming publication on ML Potential Usability and Performance Guide (ML-PEG) <https://github.com/ddmms/ml-peg> and live <http://ml-peg.stfc.ac.uk>. The cuEquivariance kernels for MACE are available here: <https://github.com/NVIDIA/cuEquivariance>.

### Acknowledgments

We would like to thank Domantas Kuryla for providing the RGD1 dataset that he recomputed at the B3LYP/6-31G\* level of theory. J. H. would like to thank Balázs Póta for discussions on phonons. We acknowledge the Jean Zay cluster access to compute as part of the Grand Challenge: GC010815458 (Grand Challenge Jean Zay H100). We would like to thank the Jean Zay cluster team and administration, as well as GENCI, for the continual help in using the Jean Zay cluster. We would like to thank the Max Planck Computing and Data Facility for providing access to the

Raven HPC system, which enabled the computation of many benchmarks. We would like to thank Sovereign AI and Isambard-AI for providing additional compute to run experiments. We are grateful for computational support from the UK national high-performance computing service, ARCHER2, for which access was obtained via the UKCP consortium and funded by EPSRC grant reference EP/P022065/1 and EP/X035891/1. I.B. was supported by the Harding Distinguished Postgraduate Scholarship. J. H. was supported by The Lennard-Jones Centre Ruth Lynden-Bell Scholarship in Scientific Computing. E.K and A.M.E were supported by Ada Lovelace centre at Science and Technology Facilities Council (<https://adalovelacecentre.ac.uk/>), Physical Sciences Databases Infrastructure (<https://psdi.ac.uk>, jointly STFC and University of Southampton) under grants EP/X032663/1 and EP/X032701/1, and EPSRC under grants EP/W026775/1 and EP/V028537/1.

- [1] J. Behler and M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, *Phys. Rev. Lett.* **98**, 146401 (2007).
- [2] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, *Phys. Rev. Lett.* **104**, 136403 (2010).
- [3] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, *Journal of Computational Physics* **285**, 316 (2015).
- [4] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, The Journal of Chemical Physics **148**, 241722 (2018).
- [5] J. S. Smith, O. Isayev, and A. E. Roitberg, Ani-1: an extensible neural network potential with dft accuracy at force field computational cost, *Chemical Science* **8**, 3192–3203 (2017).
- [6] V. L. Deringer, M. A. Caro, and G. Csányi, Machine Learning Interatomic Potentials as Emerging Tools for Materials Science, *Advanced Materials* **31**, 1902765 (2019).
- [7] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, *Phys. Rev. B* **99**, 014104 (2019).
- [8] O. A. von Lilienfeld and K. Burke, Retrospective on a decade of machine learning for chemical discovery, *Nature Communications* **11**, 4895 (2020).
- [9] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, *Nature communications* **13**, 2453 (2022).
- [10] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csányi, Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, *Advances in Neural Information Processing Systems* **35**, 11423 (2022).
- [11] T. W. Ko and S. P. Ong, Recent advances and outstanding challenges for machine learning interatomic potentials, *Nature Computational Science*, 1 (2023).
- [12] C. Devereux, J. S. Smith, K. K. Huddleston, K. Barros, R. Zubatyuk, O. Isayev, and A. E. Roitberg, Extending the applicability of the ani deep learning molecular potential to sulfur and halogens, *Journal of Chemical Theory and Computation* **16**, 4192–4202 (2020).
- [13] D. M. Anstine, R. Zubatyuk, and O. Isayev, Aimnet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs, *Chemical Science* **16**, 10228–10244 (2025).
- [14] D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, Y. Pu, V. Kapil, W. C. Witt, I.-B. Magdău, D. J. Cole, and G. Csányi, Mace-off: Short-range transferable machine learning force fields for organic molecules, *Journal of the American Chemical Society* **147**, 17598–17611 (2025).
- [15] L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C. L. Zitnick, and Z. Ulissi, Open catalyst 2020 (oc20) dataset and community challenges, *ACS Catalysis* **11**, 6059–6072 (2021).
- [16] R. Tran, J. Lan, M. Shuaibi, B. M. Wood, S. Goyal, A. Das, J. Heras-Domingo, A. Kolluru, A. Rizvi, N. Shoghi, et al., The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts, *ACS Catalysis* **13**, 3066 (2023).
- [17] S. J. Sahoo, M. Maraschin, D. S. Levine, Z. Ulissi, C. L. Zitnick, J. B. Varley, J. A. Gauthier, N. Govindarajan, and M. Shuaibi, The open catalyst 2025 (oc25) dataset and models for solid-liquid interfaces (2025), arXiv:2509.17862 [cond-mat.mtrl-sci].
- [18] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. D. Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, P. Grigorev, S. Han, W. Handley, H. H. Heenen, K. Hermanson, C. Holm, J. Jaafar, S. Hofmann, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimtarzi, J. R. Kermode, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O’Neill, C. Ortner, K. A. Persson, K. Reuter, A. S. Rosen, L. L. Schaaf, C. Schran, B. X. Shi, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, T. D. Swinburne, J. Tilly, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, F. Zills, and G. Csányi, A foundation model for atomistic materials chemistry (2024), arXiv:2401.00096 [physics.chem-ph].
- [19] C. Chen and S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, *Nat. Comput. Sci.* **2**, 718 (2022).
- [20] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, Scaling deep learning for materials discovery, *Nature*, 1 (2023), publisher: Nature Publishing Group.
- [21] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling, *Nat. Mach. Intell.* **5**, 1031 (2023).
- [22] H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S. Chen, C. Zeni, M. Horton, R. Pinsler, A. Fowler, D. Zügner, T. Xie, J. Smith, L. Sun, Q. Wang, L. Kong, C. Liu, H. Hao, and Z. Lu, Mattersim: A deep learning atomistic model across elements, temperatures and pressures, arXiv preprint arXiv:2405.04967 (2024), arXiv:2405.04967 [cond-mat.mtrl-sci].
- [23] A. Mazitov, F. Bigi, M. Kellner, P. Pegolo, D. Tisi, G. Fraux, S. Pozdnyakov, P. Loche, and M. Ceriotti, Pet-mad, a lightweight universal interatomic potential for advanced materials modeling (2025), arXiv:2503.14118 [cond-mat.mtrl-sci].
- [24] D. Zhang, X. Liu, X. Zhang, C. Zhang, C. Cai, H. Bi, Y. Du, X. Qin, A. Peng, J. Huang, B. Li, Y. Shan, J. Zeng, Y. Zhang, S. Liu, Y. Li, J. Chang, X. Wang, S. Zhou, J. Liu, X. Luo, Z. Wang, W. Jiang, J. Wu, Y. Yang, J. Yang, M. Yang, F.-Q. Gong, L. Zhang,

- M. Shi, F.-Z. Dai, D. M. York, S. Liu, T. Zhu, Z. Zhong, J. Lv, J. Cheng, W. Jia, M. Chen, G. Ke, W. E. L. Zhang, and H. Wang, Dpa-2: a large atomic model as a multi-task learner, *npj Computational Materials* **10**, 10.1038/s41524-024-01493-2 (2024).
- [25] B. M. Wood, M. Dzamba, X. Fu, M. Gao, M. Shuaibi, L. Barroso-Luque, K. Abdelmaqsoud, V. Gharakhanyan, J. R. Kitchin, D. S. Levine, K. Michel, A. Sriram, T. Cohen, A. Das, A. Rizvi, S. J. Sahoo, Z. W. Ulissi, and C. L. Zitnick, Uma: A family of universal models for atoms (2025), arXiv:2506.23971 [cs.LG].
- [26] D. Zhang, A. Peng, C. Cai, W. Li, Y. Zhou, J. Zeng, M. Guo, C. Zhang, B. Li, H. Jiang, T. Zhu, W. Jia, L. Zhang, and H. Wang, A graph neural network for the era of large atomistic models (2025), arXiv:2506.01686 [physics.comp-ph].
- [27] J. Kim, J. Kim, J. Kim, J. Lee, Y. Park, Y. Kang, and S. Han, Data-efficient multifidelity training for high-fidelity machine learning interatomic potentials, *Journal of the American Chemical Society* **147**, 1042–1054 (2024).
- [28] A. E. A. Allen, N. Lubbers, S. Martin, J. Smith, R. Messerly, S. Tretiak, and K. Barros, Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning, *npj Computational Materials* **10**, 10.1038/s41524-024-01339-x (2024).
- [29] N. Shoghi, A. Kolluru, J. R. Kitchin, Z. W. Ulissi, C. L. Zitnick, and B. M. Wood, From molecules to materials: Pre-training large generalizable models for atomic property prediction, in *The Twelfth International Conference on Learning Representations* (2024).
- [30] D. P. Kovács, I. Batatia, E. S. Arany, and G. Csányi, Evaluation of the mace force field architecture: From medicinal chemistry to materials science, *The Journal of Chemical Physics* **159**, 10.1063/5.0155322 (2023).
- [31] K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, Schnet: A continuous-filter convolutional neural network for modeling quantum interactions, *Advances in neural information processing systems* **30** (2017).
- [32] K. Schütt, O. Unke, and M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra, in *International conference on machine learning* (PMLR, 2021) pp. 9377–9388.
- [33] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, *Nature communications* **13**, 2453 (2022).
- [34] J. Gastegger, S. Giri, J. T. Margraf, and S. Günnemann, Fast and uncertainty-aware directional message passing for non-equilibrium molecules, in *Machine Learning for Molecules Workshop, NeurIPS* (2020).
- [35] M. J. Willatt, F. Musil, and M. Ceriotti, Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements, *Phys. Chem. Chem. Phys.* **20**, 29661 (2018).
- [36] K. Gubaev, E. V. Podryabinkin, G. L. Hart, and A. V. Shapeev, Accelerating high-throughput searches for new alloys with active learning of interatomic potentials, *Computational Materials Science* **156**, 148 (2019).
- [37] I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. C. Simm, R. Drautz, C. Ortner, B. Kozinsky, and G. Csányi, The design space of e(3)-equivariant atom-centred interatomic potentials, *Nature Machine Intelligence* **7**, 56–67 (2025).
- [38] A. P. Bartók, R. Kondor, and G. Csányi, On representing chemical environments, *Physical Review B* **87**, 184115 (2013).
- [39] B. Anderson, T. S. Hy, and R. Kondor, Cormorant: Covariant molecular neural networks, *Advances in neural information processing systems* **32** (2019).
- [40] E. Wigner, *Group theory*, Vol. 5 (Elsevier, 2012).
- [41] M. Geiger and T. Smidt, e3nn: Euclidean neural networks (2022), arXiv:2207.09453 [cs.LG].
- [42] J. P. Darby, D. P. Kovács, I. Batatia, M. A. Caro, G. L. W. Hart, C. Ortner, and G. Csányi, Tensor-reduced atomic density representations, *Phys. Rev. Lett.* **131**, 028001 (2023).
- [43] L. Barroso-Luque, M. Shuaibi, X. Fu, B. M. Wood, M. Dzamba, M. Gao, A. Rizvi, C. L. Zitnick, and Z. W. Ulissi, Open materials 2024 (omat24) inorganic materials dataset and models (2024), arXiv:2410.12771 [cond-mat.mtrl-sci].
- [44] Q. Zhao, S. M. Vaddadi, M. Woulfe, L. A. Ogunfowora, S. S. Garimella, O. Isayev, and B. M. Savoie, Comprehensive exploration of graphically defined reaction spaces, *Scientific Data* **10**, 10.1038/s41597-023-02043-z (2023).
- [45] M. K. Horton, P. Huck, R. X. Yang, J. M. Munro, S. Dwarknath, A. M. Ganose, R. S. Kingsbury, M. Wen, J. X. Shen, T. S. Mathis, A. D. Kaplan, K. Berket, J. Riebesell, J. George, A. S. Rosen, E. W. C. Spotte-Smith, M. J. McDermott, O. A. Cohen, A. Dunn, M. C. Kuner, G.-M. Rignanese, G. Petretto, D. Waroquier, S. M. Griffin, J. B. Neaton, D. C. Chrzan, M. Asta, G. Hautier, S. Cholia, G. Ceder, S. P. Ong, A. Jain, and K. A. Persson, Accelerated data-driven materials science with the materials project, *Nature Materials* **10.1038/s41563-025-02272-0** (2025).
- [46] Materials project calculation details, <https://docs.materialsproject.org/methodology/materials-methodology/calculation-details>, accessed: 2023-12-18.
- [47] P. Eastman, P. K. Behara, D. L. Dotson, R. Galvelis, J. E. Herr, J. T. Horton, Y. Mao, J. D. Chodera, B. P. Pritchard, Y. Wang, G. De Fabritiis, and T. E. Markland, Spice, a dataset of drug-like molecules and peptides for training machine learning potentials, *Scientific Data* **10**, 10.1038/s41597-022-01882-6 (2023).
- [48] D. S. Levine, M. Shuaibi, E. W. C. Spotte-Smith, M. G. Taylor, M. R. Hasyim, K. Michel, I. Batatia, G. Csányi, M. Dzamba, P. Eastman, N. C. Frey, X. Fu, V. Gharakhanyan, A. S. Krishnapriyan, J. A. Rackers, S. Raja, A. Rizvi, A. S. Rosen, Z. Ulissi, S. Vargas, C. L. Zitnick, S. M. Blau, and B. M. Wood, The open molecules 2025 (omol25) dataset, evaluations, and models (2025), arXiv:2505.08762 [physics.chem-ph].
- [49] A. D. Kaplan, R. Liu, J. Qi, T. W. Ko, B. Deng, J. Riebesell, G. Ceder, K. A. Persson, and S. P. Ong, A foundational potential energy surface dataset for materials (2025), arXiv:2503.04070 [cond-mat.mtrl-sci].
- [50] X. Fu, B. M. Wood, L. Barroso-Luque, D. S. Levine, M. Gao, M. Dzamba, and C. L. Zitnick, Learning smooth and expressive interatomic potentials for physical property prediction (2025), arXiv:2502.12147 [physics.comp-]

- ph].
- [51] B. Rhodes, S. Vandenhaute, V. Šimkus, J. Gin, J. Godwin, T. Duignan, and M. Neumann, Orb-v3: atomistic simulation at scale (2025), arXiv:2504.06231 [cond-mat.mtrl-sci].
  - [52] S. Grimme, J. Antony, S. Ehrlich, and H. 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**, 10.1063/1.3382344 (2010).
  - [53] S. Grimme, S. Ehrlich, and L. Goerigk, Effect of the damping function in dispersion corrected density functional theory, *Journal of Computational Chemistry* **32**, 1456–1465 (2011).
  - [54] S. Takamoto, C. Shinagawa, D. Motoki, K. Nakago, W. Li, I. Kurata, T. Watanabe, Y. Yayama, H. Iriguchi, Y. Asano, T. Onodera, T. Ishii, T. Kudo, H. Ono, R. Sawada, R. Ishitani, M. Ong, T. Yamaguchi, T. Kataoka, A. Hayashi, and T. Ibuka, Pfp: Universal neural network potential for material discovery (2021), arXiv:2106.14583 [cond-mat.mtrl-sci].
  - [55] R. Liu, E. Liu, J. Riebesell, J. Qi, S. P. Ong, and T. W. Ko, MatCalc (2024).
  - [56] M. De Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. Van Der Zwaag, J. J. Plata, et al., Charting the complete elastic properties of inorganic crystalline compounds, *Scientific data* **2**, 1 (2015).
  - [57] D. Chung and W. Buessem, The voigt-reuss-hill approximation and elastic moduli of polycrystalline mgo, caf<sub>2</sub>,  $\beta$ -zns, znse, and cdte, *Journal of applied physics* **38**, 2535 (1967).
  - [58] R. Hill, The elastic behaviour of a crystalline aggregate, *Proceedings of the Physical Society. Section A* **65**, 349 (1952).
  - [59] R. Golesorkhtabar, P. Pavone, J. Spitaler, P. Puschnig, and C. Draxl, Elastic: A tool for calculating second-order elastic constants from first principles, *Computer Physics Communications* **184**, 1861 (2013).
  - [60] B. Póta, P. Ahlawat, G. Csányi, and M. Simoncelli, Thermal conductivity predictions with foundation atomistic models (2025), arXiv:2408.00755 [cond-mat.mtrl-sci].
  - [61] J. Riebesell, R. E. A. Goodall, P. Benner, Y. Chiang, B. Deng, G. Ceder, M. Asta, A. A. Lee, A. Jain, and K. A. Persson, A framework to evaluate machine learning crystal stability predictions, *Nature Machine Intelligence* **7**, 836–847 (2025).
  - [62] A. Loew, D. Sun, H.-C. Wang, S. Botti, and M. A. L. Marques, Universal machine learning interatomic potentials are ready for phonons, *npj Computational Materials* **11**, 10.1038/s41524-025-01650-1 (2025).
  - [63] A. M. Reilly and A. Tkatchenko, Understanding the role of vibrations, exact exchange, and many-body van der waals interactions in the cohesive properties of molecular crystals, *The Journal of chemical physics* **139** (2013).
  - [64] F. Della Pia, A. Zen, D. Alfè, and A. Michaelides, How accurate are simulations and experiments for the lattice energies of molecular crystals?, *Physical Review Letters* **133**, 10.1103/physrevlett.133.046401 (2024).
  - [65] F. Della Pia, A. Zen, D. Alfè, and A. Michaelides, Dmcc-ice13: Ambient and high pressure polymorphs of ice from diffusion monte carlo and density functional theory, *The Journal of Chemical Physics* **157** (2022).
  - [66] L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, et al., Open catalyst 2020 (oc20) dataset and community challenges, *Acs Catalysis* **11**, 6059 (2021).
  - [67] R. R. Brew, I. A. Nelson, M. Binayeva, A. S. Nayak, W. J. Simmons, J. J. Gair, and C. C. Wagen, Wiggle150: Benchmarking density functionals and neural network potentials on highly strained conformers, *Journal of Chemical Theory and Computation* **21**, 3922 (2025).
  - [68] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, A look at the density functional theory zoo with the advanced gmtkn55 database for general main group thermochemistry, kinetics and noncovalent interactions, *Physical Chemistry Chemical Physics* **19**, 32184 (2017).
  - [69] K. Kriz and J. Rezac, Benchmarking of semiempirical quantum-mechanical methods on systems relevant to computer-aided drug design, *Journal of Chemical Information and Modeling* **60**, 1453 (2020).
  - [70] R. Sure and S. Grimme, Comprehensive benchmark of association (free) energies of realistic host–guest complexes, *Journal of Chemical Theory and Computation* **11**, 3785 (2015).
  - [71] Y. Chiang, T. Kreiman, E. Weaver, I. Amin, M. Kuner, C. Zhang, A. Kaplan, D. Chrzan, S. M. Blau, A. S. Krishnapriyan, and M. Asta, MLIP arena: Advancing fairness and transparency in machine learning interatomic potentials through an open and accessible benchmark platform, in *AI for Accelerated Materials Design - ICLR 2025* (2025).
  - [72] J. Firoz, F. Pellegrini, M. Geiger, D. Hsu, J. A. Bilibrey, H.-Y. Chou, M. Stadler, M. Hoehnerbach, T. Wang, D. Lin, E. Kucukbenli, H. W. Sprueill, I. Batatia, S. S. Xantheas, M. Lee, C. Mundy, G. Csanyi, J. S. Smith, P. Sadayappan, and S. Choudhury, Optimizing data distribution and kernel performance for efficient training of chemistry foundation models: A case study with mace (2025), arXiv:2504.10700 [cs.DC].
  - [73] V. Gharakhanyan, L. Barroso-Luque, Y. Yang, M. Shuaibi, K. Michel, D. S. Levine, M. Dzamba, X. Fu, M. Gao, X. Liu, H. Ni, K. Noori, B. M. Wood, M. Uyttendaele, A. Boromand, C. L. Zitnick, N. Marom, Z. W. Ulissi, and A. Sriram, Open molecular crystals 2025 (omc25) dataset and models (2025), arXiv:2508.02651 [physics.chem-ph].
  - [74] G. Strang, *Linear algebra and its applications*, Chapter 3 (Thomson, Brooks/Cole, 2000).
  - [75] S. Weisberg, *Applied linear regression*, Chapter 2, Vol. 528 (John Wiley & Sons, 2005).
  - [76] Å. Björck, *Numerical methods for least squares problems* (SIAM, 2024).
  - [77] S. J. Reddi, S. Kale, and S. Kumar, On the convergence of adam and beyond, in *International Conference on Learning Representations* (2018).
  - [78] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

## XV. Appendix

### A. Re-estimation of atomic reference energies (E0s)

During the fine-tuning phase, one needs to adapt the reference energies, to ensure proper normalization of the target energies as the MACE architecture learns to predict relative energies instead of the total energy, as do many other MLIPs. This relative energy is the atomization energy:

$$E^{\text{atm}} = E^{\text{tot}} - \sum_i^N E^0. \quad (20)$$

There are currently two choices for E0s: computed isolated atom energies using the same reference method as the training set, or using MACE’s “average” argument which re-estimates the E0s using averaging. A linear system can be formulated to provide a more robust approach for E0 re-estimation for fine-tuning. The energy prediction error  $\epsilon_i$  for a configuration  $i$  is defined as:

$$\epsilon_i = E_i^{\text{true}} - E_i^{\text{predicted}}. \quad (21)$$

We assume this error can be systematically corrected for each element  $j$  by adjusting its value of  $E_0$ :

$$\epsilon_i = \sum_j N_{ij} \times c_j, \quad (22)$$

where  $N_{ij}$  is the number of atoms of element  $j$  in configuration  $i$  and  $c_j$  is the correction for element  $j$ . In matrix notation, we can write this as  $N\mathbf{c} = \boldsymbol{\epsilon}$  [74]:

$$\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1n} \\ n_{21} & n_{22} & \cdots & n_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ n_{m1} & n_{m2} & \cdots & n_{mn} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{bmatrix} \quad (23)$$

Since we are adjusting the E0 values of different elements, which are present in many configurations with different energies, we therefore have an overdetermined system with more equations than unknowns. Since an exact solution may not exist, we can instead minimize the sum of squared residuals,  $\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ , and efficiently solve this via the least squares method [75]. The least-squares solution follows from the normal equations [76]:

$$\mathbf{c} = (N^T N)^{-1} N^T \boldsymbol{\epsilon}. \quad (24)$$

Now the E0 values can be re-estimated:

$$E0_j^{\text{new}} = E0_j^{\text{old}} + c_j. \quad (25)$$

Note that the E0s of the replay head are kept fixed to the original DFT values from the pre-training stage.

## XVI. Hyperparameters

**a. Training loss** The models were trained using a weighted sum of Huber losses of energy, forces, and stress:

$$\begin{aligned} \mathcal{L} = & \frac{\lambda_E}{N_b} \sum_{b=1}^{N_b} \mathcal{L}_{\text{Huber}} \left( \frac{\hat{E}_b}{N_a}, \frac{E_b}{N_a}, \delta_E \right) \\ & + \frac{\lambda_F}{3 \sum_{b=1}^{N_b} N_a} \sum_{b=1}^{N_b} \sum_{a=1}^{N_a} \sum_{i=1}^3 \mathcal{L}_{\text{Huber}}^* \left( -\frac{\partial \hat{E}_b}{\partial r_{b,a,i}}, F_{b,a,i}, \delta_F \right) \\ & + \frac{\lambda_\sigma}{9N_b} \sum_{b=1}^{N_b} \sum_{i=1}^3 \sum_{j=1}^3 \mathcal{L}_{\text{Huber}} \left( \frac{1}{V_b} \frac{\partial \hat{E}_b}{\partial \varepsilon_{b,ij}}, \sigma_{b,ij}, \delta_\sigma \right), \end{aligned} \quad (26)$$

where  $\lambda_E, \lambda_F, \lambda_\sigma$  are predetermined weights of energy ( $E$ ), forces ( $F$ ), and stress ( $\sigma$ ) losses, the symbols under a hat correspond to predicted values, and  $N_b$  and  $N_a$  are the batch size and the number of atoms in each structure. In the last term involving the stress,  $\varepsilon_b$  and  $\sigma_b$  correspond to the strain and stress tensors, respectively. We used  $(\lambda_E, \lambda_F, \lambda_\sigma) = (1, 10, 10)$  and Huber deltas of  $\delta_E = 0.01, \delta_F = 0.01, \delta_\sigma = 0.01$ . We use a conditional Huber loss  $\mathcal{L}_{\text{Huber}}^*$  for forces, where the Huber delta  $\delta_F$  is adaptive to the force magnitude on each atom, as used in [18]. The Huber delta  $\delta_F$  decreases step-wise by a factor from 1.0 to 0.1 as the atomic force increases from 0 to 300 eV/Å.

**b. Optimisation** The models are trained with the AMSGrad [77] variant of Adam [78] with default parameters  $\beta_1 = 0.9, \beta_2 = 0.999$ , and  $\epsilon = 10^{-8}$ .

**c. Pre-training** For the pre-training phase, we use a learning rate of 0.001 and an exponential moving average (EMA) learning scheduler with a decaying factor of 0.999. We employ gradient clipping of 100. The model is trained for 7 epochs on 48 NVIDIA H100 GPUs across 12 nodes.

**d. Fine-tuning** For the fine-tuning phase, we use a learning rate of 0.001 and an exponential moving average (EMA) learning scheduler with a decaying factor of 0.999. We employ gradient clipping of 100. The model is trained for 20 epochs on 48 NVIDIA H100 GPUs across 12 nodes.

TABLE XVII. Hyper-parameter settings for the three MACE variants used in this work.

| Hyperparameter     | Models      |            |             |
|--------------------|-------------|------------|-------------|
|                    | mace-omat-1 | mace-mh-1  | mace-omat-0 |
| max_ell            | 3           | 3          | 3           |
| correlation        | 3           | 3          | 3           |
| max_L              | 2           | 2          | 1           |
| num_channels_edge  | 128         | 128        | 128         |
| num_channels_node  | 512         | 512        | 128         |
| num_interactions   | 2           | 2          | 2           |
| num_radial_basis   | 8           | 8          | 8           |
| r_max              | 6           | 6          | 6           |
| interactions_class | non-linear  | non-linear | linear      |
| irreps             | 16x0e       | 16x0e      | 16x0e       |
| batch size         | 256         | 256        | 256         |
| energy coefficient | 1           | 1          | 1           |
| force coefficient  | 10          | 10         | 10          |
| stress coefficient | 10          | 10         | 10          |

TABLE XVIII. Summary of all reported benchmark results for **mace-mh-0-r2scan** and **mace-mh-1-r2scan**. When only a “-D3” value was reported in the manuscript, it is used directly here (per the instruction to treat D3 and non-D3 as the same).

| Domain              | Benchmark / Metric                         | Unit                             | mace-mh-0-r2scan | mace-mh-1-r2scan |
|---------------------|--------------------------------------------|----------------------------------|------------------|------------------|
| Materials           | Elastic moduli (no relaxation) – Bulk MAE  | GPa                              | 24.07            | 33.93            |
|                     | Elastic moduli (no relaxation) – Shear MAE | GPa                              | 15.59            | 20.80            |
|                     | Elastic moduli (relaxed) – Bulk MAE        | GPa                              | 12.10            | 20.19            |
|                     | Elastic moduli (relaxed) – Shear MAE       | GPa                              | 9.03             | 10.60            |
|                     | Thermal conductivity RMSE                  | $\text{W m}^{-1}\text{K}^{-1}$   | 0.33             | 0.32             |
|                     | Lattice constants MAE                      | $\text{\AA}$                     | 0.048            | 0.027            |
| Materials (Phonons) | $\omega_{\text{max}}$ MAE                  | K                                | 21               | 36               |
|                     | $\omega_{\text{avg}}$ MAE                  | K                                | 7                | 10               |
|                     | $\omega_{\text{min}}$ MAE                  | K                                | 14               | 14               |
|                     | Brillouin zone RMSE                        | K                                | 13               | 19               |
|                     | Entropy $S$ MAE (300 K)                    | $\text{J mol}^{-1}\text{K}^{-1}$ | 16               | 14               |
|                     | Helmholtz free energy $F$ MAE (300 K)      | $\text{kJ mol}^{-1}$             | 5                | 6                |
|                     | Heat capacity $C_V$ MAE (300 K)            | $\text{J mol}^{-1}\text{K}^{-1}$ | 5                | 6                |
| Molecular Crystals  | X23 formation energy MAE                   | $\text{kJ mol}^{-1}$             | 10.05            | 28.13            |
|                     | Ice polymorphs relative energy MAE         | meV                              | 7.85             | 25.69            |
| Surfaces            | S24 adsorption energy MAE                  | eV                               | 0.150            | 0.249            |
|                     | OC20 adsorption RMSD                       | eV                               | 0.242            | 0.185            |
|                     | OC20 Pearson $r$                           | –                                | 0.97             | 0.98             |
| Molecules           | Wiggle150 strained conformers MAE          | $\text{kcal mol}^{-1}$           | 5.28             | 1.59             |
|                     | GMTKN55 – Basic (small) MAE                | $\text{kcal mol}^{-1}$           | 16.24            | 12.92            |
|                     | GMTKN55 – Large systems MAE                | $\text{kcal mol}^{-1}$           | 21.15            | 7.42             |
|                     | GMTKN55 – Barrier heights MAE              | $\text{kcal mol}^{-1}$           | 11.47            | 8.06             |
|                     | GMTKN55 – Intramolecular NC MAE            | $\text{kcal mol}^{-1}$           | 14.76            | 10.84            |
|                     | GMTKN55 – Intermolecular NC MAE            | $\text{kcal mol}^{-1}$           | 13.56            | 12.94            |
|                     | GMTKN55 – Overall (WTMAD-like)             | $\text{kcal mol}^{-1}$           | 15.17            | 10.72            |
|                     | S30L MAE                                   | $\text{kcal mol}^{-1}$           | 10.72            | 10.68            |

TABLE XIX. MDR Phonon benchmark on the phonon frequencies and thermodynamic properties (300 K) of roughly ten thousand materials. BZ refers to the RMSE across the whole Brillouin Zone.

| Material                       | $\omega_{\max}$<br>(K) | $\omega_{\text{avg}}$<br>(K) | $\omega_{\min}$<br>(K) | BZ<br>(K) | $S$<br>(J/mol-K) | $F$<br>(kJ/mol) | $C_V$<br>(J/mol-K) |
|--------------------------------|------------------------|------------------------------|------------------------|-----------|------------------|-----------------|--------------------|
| uma-s-1p1-omat                 | 25                     | <b>3</b>                     | 138                    | 16        | 13               | 3               | 6                  |
| uma-m-1p1-omat                 | 17                     | <b>3</b>                     | 101                    | 13        | 11               | <b>2</b>        | 5                  |
| orb-v3-consv-inf-omat          | <b>10</b>              | <b>3</b>                     | 38                     | 10        | 9                | <b>2</b>        | 4                  |
| mace-omat-0                    | 16                     | 4                            | 13                     | 10        | 10               | 3               | 3                  |
| mace-mp-0a                     | 65                     | 32                           | 19                     | 41        | 60               | 23              | 14                 |
| mattersim-5M                   | 19                     | 5                            | 16                     | 13        | 14               | 4               | 4                  |
| mace-mh-1-omat                 | 12                     | <b>3</b>                     | <b>11</b>              | <b>7</b>  | 8                | <b>2</b>        | 3                  |
| mace-omat-1                    | 13                     | <b>3</b>                     | 12                     | 8         | 8                | <b>2</b>        | 3                  |
| uma-s-1p1-omat (no-sym)        | 11                     | 4                            | 21                     | 14        | <b>7</b>         | <b>2</b>        | 3                  |
| orb-v3-consv-inf-omat (no-sym) | 12                     | 5                            | 29                     | 24        | 13               | 3               | 4                  |
| uma-m-1p1-omat (no-sym)        | <b>9</b>               | <b>3</b>                     | 18                     | 13        | 8                | <b>2</b>        | <b>2</b>           |
| mattersim-5M (no-sym)          | 20                     | 6                            | 18                     | 16        | 13               | 4               | 3                  |
| mace-omat-0 (no-sym)           | 18                     | 5                            | 16                     | 13        | 10               | 3               | 3                  |
| mace-mp-0a (no-sym)            | 67                     | 33                           | 21                     | 43        | 60               | 24              | 13                 |
| mace-mh-1-omat (no-sym)        | 13                     | 4                            | 13                     | 10        | 8                | 3               | <b>2</b>           |
| mace-omat-1 (no-sym)           | 15                     | 4                            | 13                     | 11        | 8                | 3               | <b>2</b>           |

TABLE XX. Normalization bounds for all benchmarks. Direction: Lower-is-better (L) or Higher-is-better (H).

| Group              | Benchmark / Metric               | Unit                             | Dir. | Best (b) | Worst (w) | Rationale                                        |
|--------------------|----------------------------------|----------------------------------|------|----------|-----------|--------------------------------------------------|
| Materials          | Elastic moduli (bulk) MAE        | GPa                              | L    | 0.0      | 50.0      | Mathematical (zero error)                        |
| Materials          | Elastic moduli (shear) MAE       | GPa                              | L    | 0.0      | 50.0      | Mathematical (zero error)                        |
| Materials          | Thermal conductivity RMSE        | $\text{W m}^{-1}\text{K}^{-1}$   | L    | 0.0      | 2.0       | Mathematical (zero error)                        |
| Materials          | Phonon $\omega_{\max}$ MAE       | K                                | L    | 0.0      | 50.0      | Mathematical (zero error)                        |
| Materials          | Phonon $\omega_{\text{avg}}$ MAE | K                                | L    | 0.0      | 50.0      | Mathematical (zero error)                        |
| Materials          | Phonon $\omega_{\min}$ MAE       | K                                | L    | 0.0      | 50.0      | Mathematical (zero error)                        |
| Materials          | Phonon BZ MAE                    | K                                | L    | 0.0      | 50.0      | Mathematical (zero error)                        |
| Materials          | Entropy MAE                      | $\text{J mol}^{-1}\text{K}^{-1}$ | L    | 0.0      | 50.0      | Mathematical (zero error)                        |
| Materials          | Helmholtz free energy MAE        | $\text{kJ mol}^{-1}$             | L    | 0.0      | 50.0      | Mathematical (zero error)                        |
| Materials          | Heat capacity $C_V$ MAE          | $\text{J mol}^{-1}\text{K}^{-1}$ | L    | 0.0      | 50.0      | Mathematical (zero error)                        |
| Materials          | Phonon stability                 | count                            | L    | 0.0      | 50.0      | Mathematical (zero instabilities)                |
| Materials          | Phonon stability $n$             | count                            | L    | 0.0      | 50.0      | Mathematical (zero instabilities)                |
| Molecular Crystals | X23 formation energy MAE         | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Molecular Crystals | Ice polymorphs energy MAE        | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Surfaces           | S24 adsorption energy MAE        | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Surfaces           | OC20 adsorption MAE              | eV                               | L    | 0.021    | 0.50      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Molecules          | Wiggle150 MAE                    | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Molecules          | GMTKN55 Basic small MAE          | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Molecules          | GMTKN55 Large systems MAE        | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Molecules          | GMTKN55 Barrier heights MAE      | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Molecules          | GMTKN55 Intra-NC MAE             | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Molecules          | GMTKN55 Inter-NC MAE             | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Molecules          | GMTKN55 overall MAE              | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Molecules          | PLF547 MAE                       | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Molecules          | S30L MAE                         | $\text{kcal mol}^{-1}$           | L    | 0.5      | 50.0      | Reference accuracy (0.5 $\text{kcal mol}^{-1}$ ) |
| Physicality        | Slab extensivity $\Delta$        | meV                              | L    | 0.0      | 50.0      | Mathematical (extensivity; zero deviation)       |
| Physicality        | H-additivity max $ \Delta F $    | $\text{meV } \text{\AA}^{-1}$    | L    | 0.0      | 50.0      | Mathematical (additivity; zero deviation)        |
| Physicality        | Diatomic force flips             | count                            | L    | 1.0      | 10.0      | Mathematical/physical (one sign change)          |
| Physicality        | Diatomic energy minima           | count                            | L    | 1.0      | 10.0      | Mathematical/physical (single minimum)           |
| Physicality        | Diatomic inflections             | count                            | L    | 2.0      | 10.0      | Mathematical/physical (typical Morse shape)      |
| Physicality        | Spearman $\rho(E_{\text{rep}})$  | –                                | H    | –1.0     | –0.5      | Mathematical (perfect anticorrelation)           |
| Physicality        | Spearman $\rho(E_{\text{at}})$   | –                                | H    | 1.0      | 0.0       | Mathematical (perfect correlation)               |