npj Computational Materials

https://doi.org/10.1038/s41524-026-01979-1

Article in Press

Graph atomic cluster expansion for foundational machine learning interatomic potentials

Received: 28 August 2025

Accepted: 24 January 2026

Cite this article as: Lysogorskiy, Y., Bochkarev, A., Drautz, R. Graph atomic cluster expansion for foundational machine learning interatomic potentials. *npj Comput Mater* (2026). https://doi.org/10.1038/s41524-026-01979-1

Yury Lysogorskiy, Anton Bochkarev & Ralf Drautz

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Graph atomic cluster expansion for foundational machine learning interatomic potentials

Yury Lysogorskiy*, Anton Bochkarev, and Ralf Drautz

Interdisciplinary Centre for Advanced Materials Simulation (ICAMS),
Ruhr-University Bochum, 44780 Bochum, Germany
*Corresponding author: yury.lysogorskiy@rub.de

(Dated: Originally submitted: August 26 2025; Revised: January 3 2026)

ABSTRACT

Foundational machine learning interatomic potentials that can accurately and efficiently model a vast range of materials are critical for accelerating atomistic discovery. We introduce universal potentials based on the graph atomic cluster expansion (GRACE) framework, trained on several of the largest available materials datasets. Through comprehensive benchmarks, we demonstrate that the GRACE models establish a new Pareto front for accuracy versus efficiency among foundational interatomic potentials. We further showcase their exceptional versatility by adapting them to specialized tasks and simpler architectures via fine-tuning and knowledge distillation, achieving high accuracy while preventing catastrophic forgetting. This work establishes GRACE as a robust and adaptable foundation for the next generation of atomistic modeling, enabling high-fidelity simulations across the periodic table.

I. INTRODUCTION

The ability to predict materials properties from atomistic simulations is essential for modern materials design. Machine learning interatomic potentials (MLIPs), trained on data from electronic structure methods like density functional theory, have recently emerged as a powerful tool, achieving excellent accuracy for diverse systems[1–11]. However, the applicability of conventional MLIPs is constrained by their limited elemental scope. Extending an existing MLIP to include a new chemical element requires generating thousands of new reference calculations and retraining the entire model - a process that represents a significant computational bottleneck.

Foundational interatomic potentials seek to resolve this bottleneck by creating a single, universal model that encompasses the entire periodic table from the outset. This ambitious goal requires both enormous training datasets and an efficient method for representing the vast space of chemical interactions. A brute-force enumeration of interactions is unfeasible; parameterizing just the four-body interactions would involve on the order of 10^8 interactions.

Foundational MLIPs resolve this problem by embedding complex chemistry into a low-dimensional space; a long-standing concept in materials simulation. Early tight-binding models, for example, used the valence electron count to effectively describe chemical trends and structural stability across multiple elements[12]. Modern foundational MLIPs build upon this legacy, using multi-dimensional embeddings to leverage the inherent correlations between elements. This approach is remarkably effective, enabling the entire periodic table's chemistry to be captured in few dimensions.[13–20]

The first universal force field was published more than 30 years ago[21]. Parameterizations of MLIPs across the periodic table started to appear when large training datasets became available[22–32].

To date the development of universal MLIPs was tightly bound to the progress in message passing graph neural networks[33–49], despite the fact that universal parameterizations across the periodic table are in principle independent of MLIP architecture. The first universal MLIPs that built on the Atomic Cluster Expansion (ACE)[4] became available two years ago, within the framework of Multi-ACE[50] that employs general many-body messages and as implemented in MACE[24,51].

Here we employ the Graph Atomic Cluster Expansion (GRACE)[49]. By extending ACE to tree-graphs, GRACE stands out from graph neural networks by providing a complete basis for the parameterization of atomic interactions as a function of atomic positions and chemical species. By straightforward tensor decomposition of the GRACE expansion coefficients one directly obtains sparse representations with efficient chemical embedding that can be evaluated recursively. The recursive evaluation of graph basis functions can be understood as message passing and because of its complete basis, GRACE is able to rationalize and represent other message passing graph neural networks architectures in general. As GRACE further facilitates recursive evaluation of effective ACE on each message passing layer, it benefits from linear scaling with the number of recursion layers as well as linear scaling with the complexity of ACE messages within each layer for efficient double-recursive evaluation.

Training data for universal force fields needs to cover the periodic table comprehensively. Only few publicly available datasets are suitable, notably the Materials Project[52,53], Alexandria[54,55] and the OMat24[56] datasets that are in the focus of our work here, but also the

Open Quantum Mechanical Database[57] (OQMD) and AFLOWLIB[58] and more recent additions such as MatPES[59] and MP-ALOE[60].

Validation of universal force fields is challenging. Traditional strategies that are employed for MLIPs with only few elements and that probe test errors for specifically relevant simulation tasks are not possible because of the combinatorically many different simulations that would be required across the periodic table. Validation therefore is necessarily limited to tests that seem particularly relevant or are widely adopted in the community.

In this work we present a number of GRACE models with varying complexity that were parameterized on the OMat24[56], Alexandria[54,55] and MPTraj[53] datasets and can serve as a foundation models for atomistic modelling in materials science.

II. RESULTS

A. Foundational GRACE interatomic potentials

Developing foundational MLIPs capable of accurate predictions across a wide array of chemical elements and diverse structures requires exceptionally large and varied datasets. The GRACE framework is designed to effectively manage this inherent complexity. Our primary training source was the OMat24 dataset[56], which currently is the largest publicly available compilation for materials property prediction. It encompasses 110 million DFT calculations, primarily computed with VASP[61–63] with the GGA-PBE functional[64], including Hubbard U corrections for specific oxides and fluorides, consistent with Materials Project defaults[52]. Importantly, OMat24 extends beyond near-equilibrium structures, distinguishing it from datasets like Alexandria and MPTraj, but covers same 89 elements as those two. The dataset's diversity comes from its generation methods, which include Boltzmann sampling of structures with randomly displaced atomic positions, ab initio molecular dynamics (AIMD), and subsequent relaxations of these configurations.

We developed a number of GRACE models with one- (1L) and two-layer (2L) architectures, systematically varying their complexity through small (no suffix), medium (-M suffix), and large (-L suffix) setups. The one-layer models are built on ACE star-graphs with direct interactions, the two-layer models include semi-local interactions mediated by equivariant message passing. Both, one-layer and two-layer models employ chemical embedding for efficiently condensing chemical interactions into low rank representations. The initial parameterizations, designated as "-OMAT-base", were conducted using the OMat24 dataset and employed a loss function that equally weighted energies, forces, and stresses. Further fine-tuning, which used larger weights for the energy loss component, resulted in a series of models designated with an "-OMAT-ft-E" or just "-OMAT" suffix.

While OMat24 provides a robust foundation, its DFT and pseudopotential settings differ from those used in Alexandria and the Materials Project. To address this, we fine-tuned the OMAT-base models using a combined dataset of MPTraj and a subsampled Alexandria (sAlex) dataset[54–56]. The sAlex subset was curated to prevent data leakage with the WBM test set[65], a crucial step to ensure model compatibility with Matbench Discovery[66]. The resulting models are denoted with suffix "-OAM". These GRACE models are designed to serve as robust foundational interatomic potentials and to provide uniform accuracy across a broad range of chemical compositions and structural configurations.

B. Validation

Foundational MLIPs must demonstrate uniform accuracy across multiple application domains. In this section, we present several critical validation tests. We evaluate our models against the MatBench Discovery benchmark[66] for formation energies and thermodynamic stability and the κ-SRME[67] test for thermal conductivity, which reflects second- and third-order derivatives. We further determine the performance of our models on elastic properties and for non-equilibrium and defective configurations by predicting formation energies of grain boundaries, surfaces, and point defects in pure elements. This suite of tests provides a good assessment of the models' capabilities and limitations.

Matbench Discovery[66] serves as a benchmark for high-throughput discovery of stable inorganic crystals. It is specifically designed to evaluate the efficacy of various foundational MLIPs in predicting formation energy and thermodynamic stability of novel crystal structures. The benchmark task involves geometry optimization of structures sourced from the WBM dataset, a diverse collection of 257,000 candidate crystal structures spanning a wide range of compositions. The central goal is to predict formation energies and assess the stability of these structures relative to the original convex hull from Materials Project[52] with high fidelity. Achieving this requires models to accurately predict potential energy, atomic forces, and stress tensors, with an accuracy comparable to the Materials Project's density functional theory (DFT) calculations. Figure 1 (a) illustrates the performance of GRACE models, specifically the OAM-fine-tuned versions, against other publicly available foundational MLIPs in relation to their computational efficiency. The figure shows that GRACE models consistently occupy the Pareto optimality front, demonstrating a superior balance of performance and computational speed. This efficiency can be further enhanced by using the LAMMPS molecular dynamics code[68], with specific timings available in Table I. While the F1 metric reflects a model's accuracy for identifying stable structures, the mean absolute error (MAE) of formation energies offers

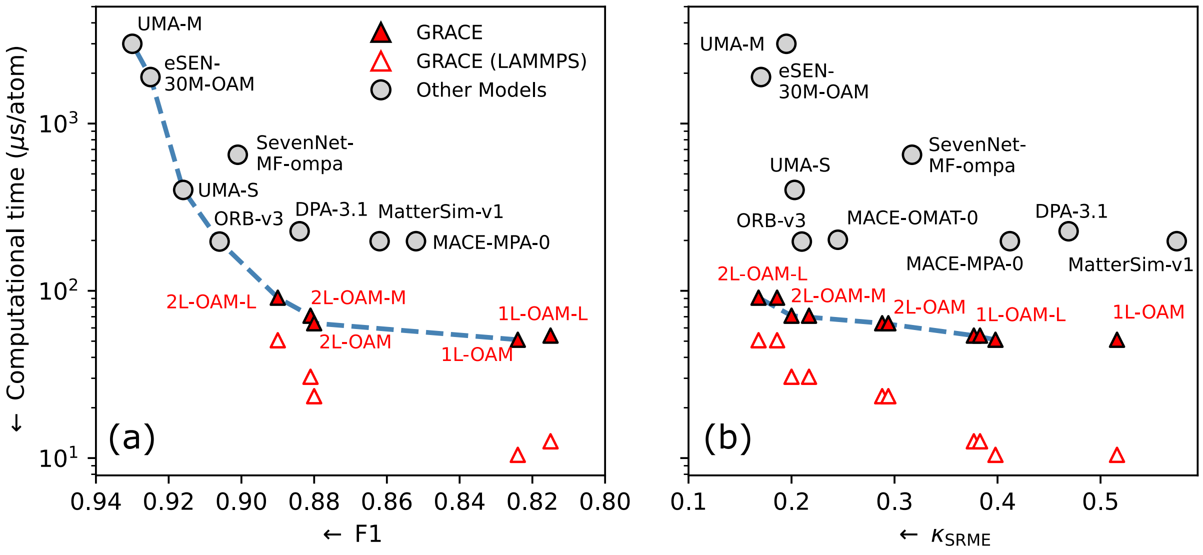

FIG. 1

TABLE I: Performance of foundational MLIPs in materials discovery task (F1), thermal conductivity task (κ_{SRME}) and computational performance (in μs/atom/step) for W-BCC crystal with 1024 atoms on A100-80GB GPU in ASE and LAMMPS molecular dynamics simulations. Top part of the table corresponds to models compliant with MatBench Discovery task, bottom part to models trained on OMat24 dataset only.

Model	F1	κ_{SRME}	t_{ASE}	t_{LAMMPS}
UMA-M-1.1	**0.930**	0.195	2981	
eSEN-30M-OAM	0.925	0.170	1897	
eqV2 M	0.917	1.771	-	
UMA-S-1.1	0.916	0.203	401	
ORB-v3	0.906	0.210	197	
SevenNet-MF-ompa	0.901	0.317	651	
GRACE-2L-OAM-L	0.890	**0.168**	91	51
DPA-3.1-3M-FT	0.884	0.469	227	
GRACE-2L-OAM-M	0.881	0.200	71	31
GRACE-2L-OAM	0.880	0.294	64	23
MatterSim-v1	0.862	0.574	198	
MACE-MPA-0	0.852	0.412	198	
GNOME	0.829	-	-	
GRACE-1L-OAM	0.824	0.516	**51**	10
GRACE-1L-OAM-L	0.815	0.377	54	13
GRACE-2L-OMAT-L	-	0.186	91	51
GRACE-2L-OMAT-M	-	0.217	71	31
MACE-OMAT-0	-	0.245	202	
GRACE-2L-OMAT	-	0.288	64	23
GRACE-1L-OMAT-L	-	0.383	54	13
GRACE-1L-OMAT	-	0.398	**51**	10

a more general measure of accuracy. The performance of the models for formation energy MAE, which shows a similar trend with GRACE models on the Pareto front, is provided in the ementary materials.

To evaluate the ability of foundational MLIPs to predict force-dependent properties like phonons and anharmonic thermal conductivity, we used the symmetric relative mean error κ_{SRME} metric[67]. This test quantifies a model's performance by predicting the thermal conductivity κ across 103 binary structures. The thermal conductivity values are calculated from forces predicted by the foundational MLIPs and subsequently analyzed using the `phono3py` software[69,70]. Accurate thermal conductivity predictions serve as a strong indicator of a model's performance for other simulation tasks, such as modeling metal-organic frameworks[71]. The results for various foundational MLIPs, including our GRACE models, are presented in Fig. 1 (b) and Table I. The family of one- and two-layer GRACE models notably form the Pareto front, achieving the best performance in thermal conductivity prediction. In particular, the GRACE-2L-OAM-L model achieved the lowest error with $\kappa_{\mathrm{SRME}} = 0.168$, underscoring its exceptional accuracy in this domain.

Predicting elastic moduli is a crucial validation test for interatomic potentials. We categorize elastic constants into three subgroups: longitudinal (C_{11}, C_{22}, C_{33}), Poisson's ratio-related (C_{12}, C_{13}, C_{23}), and shear (C_{44}, C_{55}, C_{66}). Because these groups often have varying magnitudes, we primarily focused on the symmetric relative mean error (SRME) and MAE within each subgroup. Figure 2 presents both metrics with respect to reference data from the Materials Project[72]. Among all tested models, GRACE-2L-OAM-L demonstrated the lowest C_{SRME}. Generally, most models showed comparable performance, with the notable exceptions of MatterSim and DPA3-openlam, whose training sets differed from the Materials Project DFT settings used for the reference elastic constant calculations. A consistent trend across all models is that longitudinal constants typically exhibit lower SRME but higher absolute errors (ΔC) due to their larger magnitudes. In contrast, the second and third groups of elastic constants tend to show smaller absolute errors but larger relative errors.

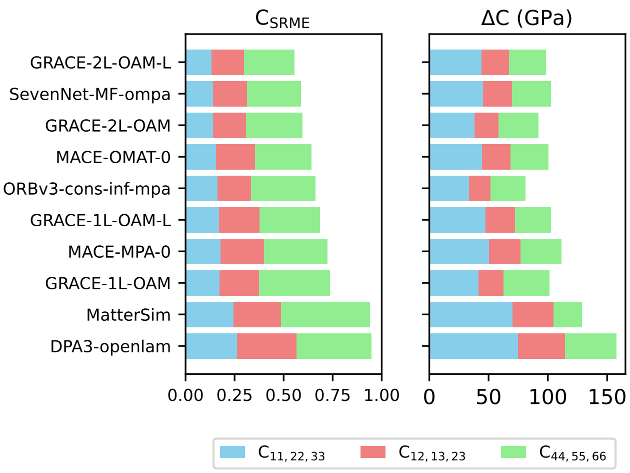

FIG. 2

To assess the performance of foundational MLIPs for bulk structural defects, we utilized an existing dataset of grain boundary formation energies (γ_{GB} computed for pure metals[73]. The models' accuracy was quantified by calculating both the γ_{GB}-SRME and the mean absolute error $\Delta\gamma_{\mathrm{GB}}$. The results are presented in Fig. 3. The relative error γ_{GB}-SRME generally ranges from 0.275 to 0.4, with MatterSim and SevenNet-MF-ompa being notable exceptions. A larger γ_{GB}-SRME for K, Rb, and Cs was consistently observed across almost all models, suggesting that the typically used 6 Å cutoff may be insufficient for these alkali elements. For most models, the absolute error $\Delta\gamma_{\mathrm{GB}}$ remains below 5 meV/Å^2, with a few exceptions observed for eSEN-30M-Omat, MatterSim, and SevenNet-MF-ompa. More detailed information can be found in the ementary Information.

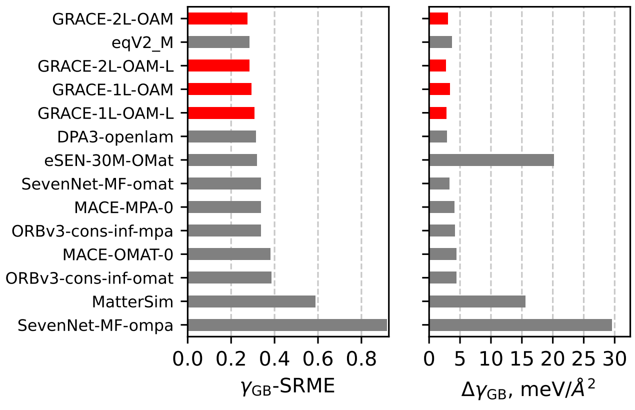

FIG. 3

We assessed the ability of foundational MLIPs to predict open structures, such as surfaces, using surface energies (γ_{surf}) for pure elements[74]. Model accuracy was quantified by calculating both the symmetric relative mean error (γ_{surf}-SRME) and the mean absolute error ($\Delta\gamma_{\mathrm{surf}}$), with results presented in Fig. 4. Here, the relative error γ_{surf}-SRME ranged from 0.168 for the ORBv3-cons-inf-mpa model to 0.279 for the MACE-OMAT-0 model. The absolute error $\Delta\gamma_{\mathrm{surf}}$ typically varied from 8 to 14 meV/Å^2, with a few exceptions noted for MACE-MPA-0, MatterSim, and eqV2_M. A consistent finding across all models was poor γ_{surf}-SRME metrics for Potassium (K), Rubidium (Rb), Cesium (Cs), and Indium (In). More detailed information can be found in the ementary Information.

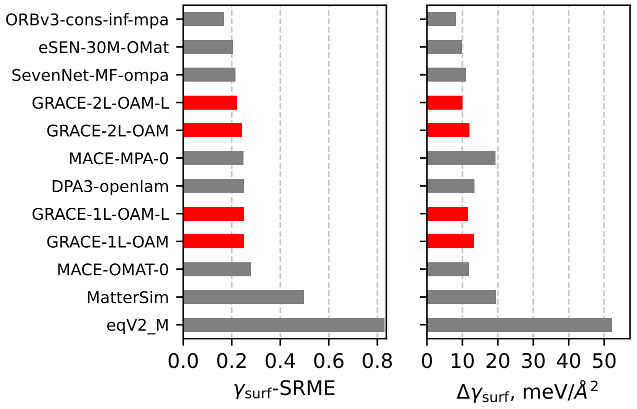

FIG. 4

We used systematically computed formation energies for self-interstitials (SIA) and vacancies in BCC[75] and FCC[76] metals available in the literature as references. Given the varying scales of these formation energies across different defect types and metals, we used the symmetric relative mean error metrics, $\mathrm{E}_{\mathrm{SIA}}$-SRME and $\mathrm{E}_{\mathrm{vac}}$-SRME, as main measure of accuracy. As shown in Fig. 5, the SRME metrics for both defect types generally fall within 0.1 to 0.3, with a few outliers such as MatterSim. This discrepancy is likely due to the different DFT settings between MatterSim's training set and the reference data, as observed in our previous analyses. In terms of absolute values, the mean absolute error (MAE) for SIA formation energies typically ranges from about 0.2 to 0.4 eV, while for vacancy formation energies, the errors are between 0.1 to 0.2 eV. Further details can be found in the ementary Information.

C. Long-time stability of MD

Long-time stability in molecular dynamics (MD) simulations is essential for accurately capturing the dynamical behavior and thermodynamic properties that cannot be fully assessed by static property benchmarks or short

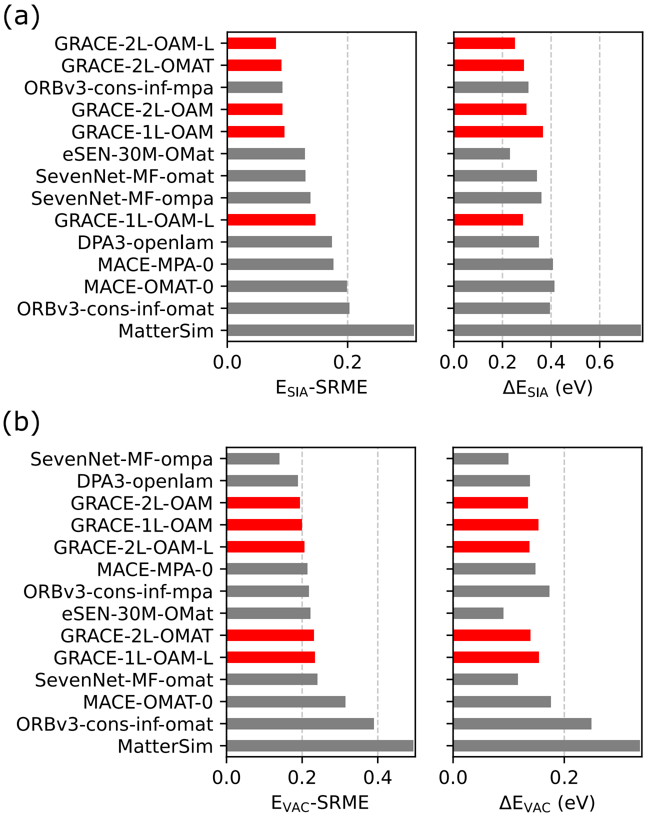

FIG. 5

simulations. To demonstrate the out-of-the-box MD stability and performance of the GRACE foundational MLIPs, we conducted a 1 ns MD simulation of a FLiBe cell containing approximately three thousand atoms at 973 K, utilizing the GRACE-2L-OMAT-L model in the NVE ensemble, observing a negligible total energy drift of $5{\cdot}10^{-9}$ eV/atom/ns (see ementary Information for more details). We compared the resulting radial distribution functions (RDFs) to reference AIMD data from Ref.[77,78], as shown in Fig. 6. Additionally, we estimated diffusion coefficients for each element from our MD simulation, obtaining values of 1.33, 1.58, and 5.86 × 10^{-5} cm^{2}s^{-1} for Be, F, and Li, respectively. These values align well with the AIMD results (0.83 ± 0.1, 1.73 ± 0.17 and 5.67 ± 0.52 × 10^{-5} cm^{2}s^{-1}, respectively[77]). This simulation confirms the model's stability and its accuracy in predicting both structural and dynamical properties over extended timescales.

D. Computational performance

Computational performance is a critical factor, especially for high-throughput calculations and large-scale, long-duration MD simulations. We evaluated the performance of the GRACE models using LAMMPS on an NVIDIA A100 GPU with 80 GB of memory. Test systems included carbon diamond, liquid water, aluminum FCC, and the molten salt FLiBe, which have different densities and numbers of atomic neighbors. All systems were simulated for a few steps in the NVT ensemble at 300 K, except for FLiBe, which was at 823 K.

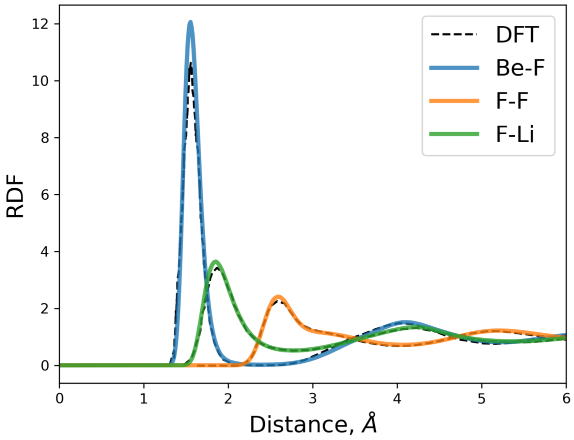

FIG. 6

To ensure accurate performance analysis, we excluded the initial MD step, which was significantly slower due to just-in-time (JIT) model compilation. System sizes were increased incrementally until an out-of-memory error occurred. The computational performance, expressed in microseconds per atom, and the maximum number of atoms fitting into memory are presented in Fig. 7.

The carbon diamond system, with the highest number of neighboring atoms within the cutoff radius, proved to be the most computationally demanding. Still, even for our most intensive model, GRACE-2L-L, up to 20,000 carbon atoms could be accommodated with a performance of approximately 124 µs per atom per step. Across different systems, two-layer GRACE models exhibited computational performance ranging from 27 to 120 µs/atom, while single-layer models ranged from 10 to 28 µs/atom, enabling efficient MD simulations.

Regarding memory usage, a single A100-80GB GPU could accommodate between 20,000 and 55,000 atoms for two-layer GRACE models and 78,000 to 215,000 atoms for one-layer GRACE models. Since the one-layer GRACE model is local (interactions are limited to a cutoff radius), it can be parallelized via domain decomposition as implemented in LAMMPS, which further boosts computational performance for very large systems with millions or billions of atoms.

We also measured the computational performance of other foundational MLIPs by running ten MD steps in ASE[79] of W-BCC supercells with different numbers of atoms. These tests utilized different GPU hardware, including the commodity RTX3060 (12 GB), and L40s (40 GB), A100 (80 GB), and H200 (141 GB). The resulting execution times are presented in Table II. GRACE models deliver excellent performance even on commodity GPUs like the RTX 3060, despite consistently operating with FP64 (double-precision) accuracy. As shown in Fig.1, the use of ASE can introduce substantial computational overhead, stemming primarily from Python's execution speed and suboptimal neighbor list construction algorithms.

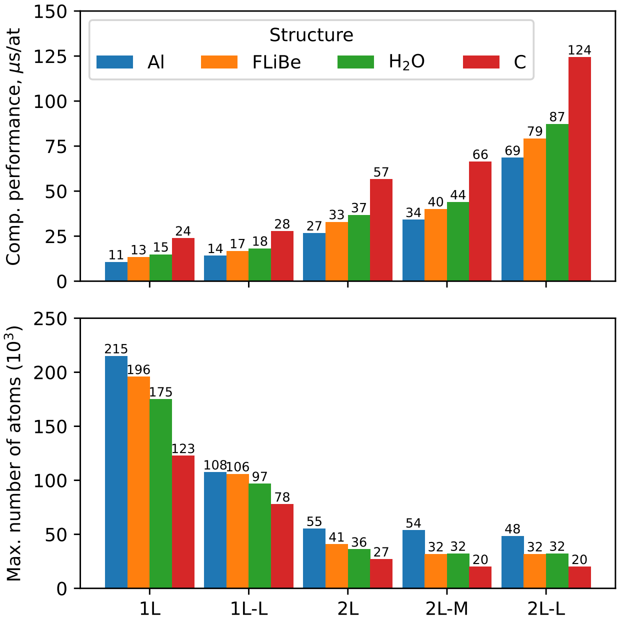

FIG. 7

E. Fine-tuning

Foundational interatomic potentials are trained to provide accurate simulations across the periodic table, but they may lack the specific precision or performance necessary for particular downstream tasks. Fine-tuning serves as an effective strategy to address these limitations by adapting a pre-existing foundational MLIP. This is done by continuing the training on a new, often small and specialized dataset. This process aims to enhance the model's performance or tailor its predictions for specific chemical systems, properties, or higher levels of theoretical accuracy, while leveraging the extensive knowledge and robust representations acquired during its initial foundational training. Here we demonstrate the fine-tuning of GRACE-2L foundational MLIP for the Al-Li binary system[80] and a hydrogen combustion dataset[81].

For fine-tuning, we utilized a dataset for the Al-Li binary system from Ref.[80]. We curated the data by removing structures from the liquid phase and those corresponding to randomly sampled space groups. This resulted in a total of three thousand structures within 1 eV/atom above the convex hull, from which we allocated 5% as a test set. To study data efficiency, we then created a series of training subsets of varying sizes (5%, 10%, 25%, 50%, and 75%) from the remaining data. Using these training and test sets, we fine-tuned the 1L-OMAT and 2L-OMAT models by updating all weights, a process we term ”naive fine-tuning.” For comparison, we also trained two models from scratch: a one-layer ACE model as implemented in the PACE software[82,83] and a GRACE-2L model, designated 2L-baseline, with a complexity identical to GRACE-2L-OMAT. As shown in Fig. 8, zero-shot predictions without fine-tuning demonstrate very good accuracy for both GRACE-1L-OMAT and GRACE-2L-OMAT. The fine-tuned 2L-OMAT-ft model consistently outperforms the other models, even with small fractions of the curated dataset, while the 2L-baseline model only reaches comparable accuracy with more data. The fine-tuned 1L-OMAT-ft model is slightly less accurate than the two-layer models for most tests and shows comparative performance to the specialized PACE model. We attribute this good performance to the small number of elements and the relatively limited Al-Li dataset, which primarily includes close-to-equilibrium structures. These results demonstrate that fine-tuning foundational GRACE potentials can be superior to training from scratch, especially in low-data regimes.

The hydrogen combustion (H2COMB) dataset[81] includes intrinsic reaction coordinate (IRC) calculations, *ab initio* MD simulations, and normal mode displacement calculations, covering 19 reaction channels for hydrogen combustion. This dataset was computed using Q-Chem with ωB97X-V/cc-pVTZ. These DFT settings differ significantly from those of the OMat24, Alexandria, and MPTraj datasets that we used to train the foundational GRACE OMAT potentials. Consequently, zero-shot predictions by all foundational models show a rather high error. For the GRACE-2L-OAM model, the force MAE is 740.5 meV/Å[84,85].

Due to the systematic difference in datasets, naive fine-tuning may result in catastrophic forgetting - the tendency to forget previously learned information when learning a new task. To fine-tune a GRACE-2L-OAM model to the H2COMB dataset without this issue for other elements, we explored several strategies: 1) naive fine-tuning: all parameters are trainable; 2) frozen-weights approach: only ACE expansion coefficients are trainable, while all other parameters (including chemical embeddings, radial functions, and the energy readout) are kept unchanged. Within this approach, we considered two cases (a) only the coefficients of the final ACE expansions before atomic energy readout are trainable and (b) the coefficients of the final ACE expansions and the ACE expansion coefficients for messages passed between the first and second layers are trainable. These coefficients depend on the central atom type; thus, parameters for elements absent in the fine-tuning dataset will not be updated.

The results of these strategies are shown in Figure 9. The baseline zero-shot model shows a very high force MAE for the H2COMB dataset but a low error on the

TABLE II: Computational performance of GRACE foundational potentials and other foundational MLIPs for MD simulations of a tungsten-BCC supercell in ASE, reported in μs/atom/step across different GPU architectures.

GPU	RTX 3060 (12 GB)		L40s (40Gb)			A100 (80Gb)			H200 (141Gb)		
Num. of atoms	256	512	512	1k	2k	512	1k	2k	512	1k	2k
GRACE-1L	111	103	82	70	68	54	51	48	53	44	42
GRACE-1L-L	159	142	89	77	79	57	54	51	55	46	45
GRACE-2L	214	200	101	92	90	68	64	62	59	50	49
GRACE-2L-M	303	292	118	109	111	74	71	69	63	54	50
GRACE-2L-L	405	394	142	133	140	95	91	95	76	66	63
ORBv3(fp32)	506	470	142	129	125	208	197	195	114	97	89
ORBv3(fp64)	15566	15328	2451	2407	2381	231	223	221	131	119	114
MatterSim	409	319	149	124	151	288	198	163	244	145	92
MACE-OMAT-0(cuEq)	683	640	177	153	140	288	202	129	296	162	99
SevenNet-MF-OMPA	2744	OOM	1283	1289	OOM	700	651	633	380	357	334
eSEN-30M-OMat	OOM	OOM	2120	OOM	OOM	1937	1897	OOM	834	779	753

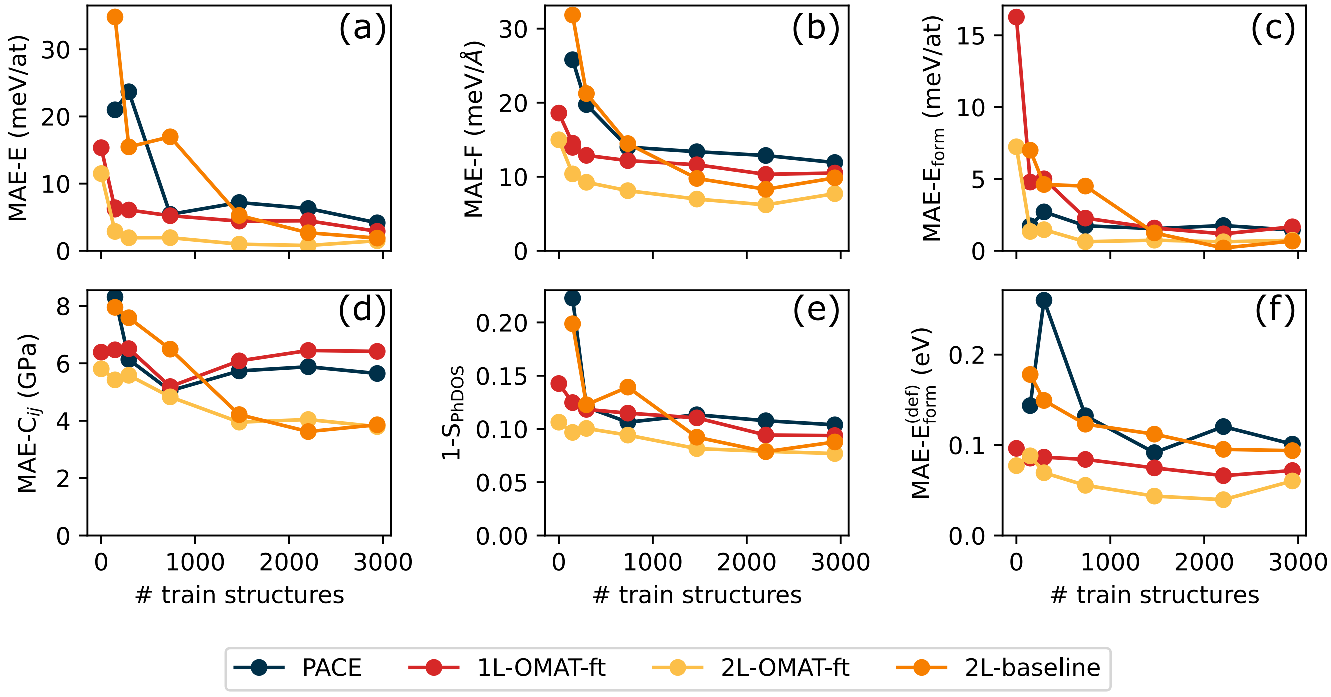

FIG. 8

original sAlex dataset. The latter's performance is computed on two subsets: H and O atoms, and all other elements. Naive fine-tuning (strategy 1) achieved the best error metrics on the downstream task (37 meV/Å), but the error on the original dataset increased drastically to 339 $\times 10^3$ meV/Åfor H/O and 6.5 $\times 10^3$ meV/Åfor the remaining elements, confirming catastrophic forgetting for all elements.

In contrast, the frozen-weights approaches show only a small increase in force MAE on the original sAlex dataset for elements excluding H and O (from 25 meV/Åfor the baseline model to 31 meV/Åand 35 meV/Åfor strategies 2a and 2b, respectively). The predictions of these models remain unchanged from the baseline if the structures contain neither H nor O atoms, indicating that the increase in metrics is associated with the presence of H/O atoms within the cutoff. For H and O atoms from the sAlex dataset, errors increase much less than with naive fine-tuning (from 49 meV/Åfor the baseline to 144 meV/Åand 174 meV/Å), while downstream task errors became 55 and 47 meV/Åfor strategies 2a and 2b, respectively.

The errors on the original versus the new dataset form a Pareto front as displayed in Fig. 9). By freezing certain

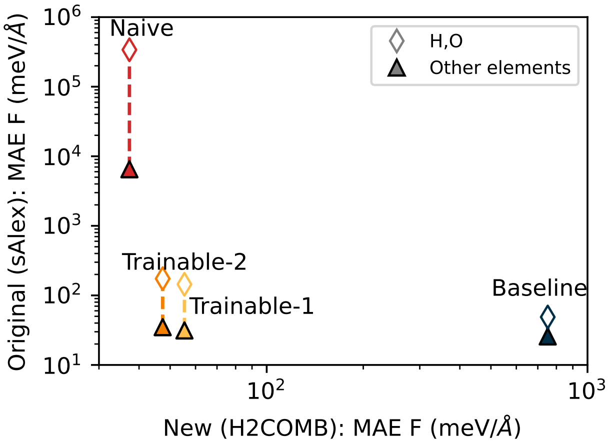

FIG. 9

parts of the model, catastrophic forgetting of the original task can be mitigated, even for the elements that are updated and despite a systematic shift in reference data. Other methods for fine-tuning foundational GRACE potentials that can mitigate catastrophic forgetting, such as low-rank adaptation (LoRA) or delta tuning, will be considered in future work.

F. Model distillation

While foundational MLIPs offer high accuracy across a vast chemical and configurational space, they often have lower computational performance and limited parallelization compared to chemistry-specific models. To bridge this gap, model distillation can be employed[86,87], a process involving the retraining of a simpler, faster "student" model on a dataset labeled by a foundational "teacher" model. We investigate different distillation and fine-tuning pathways using the combined HEA25[88] and HEA25S[89] datasets as a case study. We employ the foundational GRACE-2L-OMAT model as the "teacher" and the GRACE-FS architecture as the "student", selected for its straightforward parallelization and CPU-only inference capabilities.

We evaluated performance based on two tasks with corresponding accuracy metrics. The primary task reflects model accuracy on the new downstream application, measured by the force component MAE on the HEA25S validation set. The secondary task serves as a proxy for the model's retention of general chemical knowledge and stability, as the HEA25S dataset lacks pure unary or binary structures. This was measured by the MAE of formation energy of binaries compounds from the Materials Project comprised of non-magnetic elements covered in the HEA25 dataset.

The different pathways for the downstream task are illustrated in Fig. 10. The initial foundation model ("Foundation") exhibits a large error on the primary task due to the DFT functional mismatch between its training data (PBE) and the HEA25S target (PBEsol). However, it retains a low error (14 meV/atom) on the secondary task—comparable to the 23 meV/atom formation energy MAE on the Matbench Discovery leaderboard.

We explored three distinct approaches to address this. In the first method, we fine-tuned the foundation model on the HEA25S dataset to create a new teacher ("Fine-tuned"). This teacher, having achieved the lowest error on the primary task, was then used to re-label the HEA25S dataset for parameterizing the GRACE-FS student model called "Naive Distilled". A second pathway is opposite to the first: we re-labeled the HEA25S dataset using the original foundation model (GRACE-2L-OMAT) to parameterize a "Raw distilled" student model, which was subsequently fine-tuned using the original HEA25S dataset to produce the "Distilled/Finetuned" model. Finally, the third approach involved parameterizing the GRACE-FS model from scratch using the HEA25S dataset directly to create the "Bespoke" model. As shown in Fig. 10, all three approaches yield models with similar error metrics, though the bespoke model performs slightly better on the primary task due to the usage of the original DFT data.

To improve performance on the secondary task, while preserving good accuracy of the primary task, we generated a synthetic training dataset comprising structures from HEA25 and HEA25S, along with unary and binary structures (including both ideal and rattled configurations) for all 25 elements. This extended dataset was labeled by the GRACE-2L teacher ("Finetuned") and used to parameterize a new student model ("Extended Distilled"). This model demonstrates significantly improved performance on the secondary task, close to the teacher model, while retaining metrics on the primary task that are only slightly worse to the bespoke model.

We consistently observe a slight degradation in accuracy for student models compared to their teachers due to the simpler architecture of the former. However, this simplicity yields substantial gains in computational efficiency. On a CPU with 10 physical cores, the GRACE-2L model requires approximately 34.48 ms/atom/core. In contrast, the GRACE-FS model, using a standalone C++ implementation, achieves 0.56 ms/atom/core, representing a speedup of nearly 70×. Thus, the extended distillation strategy offers the optimal balance, recovering the generalizability lost in standard fine-tuning or bespoke training. The effect of different complexities of teacher and student models, together with other details are presented in the ementary Information.

III. DISCUSSION

We present a series of foundational machine learning interatomic potentials (MLIPs) based on the Graph Atomic Cluster Expansion (GRACE). GRACE builds

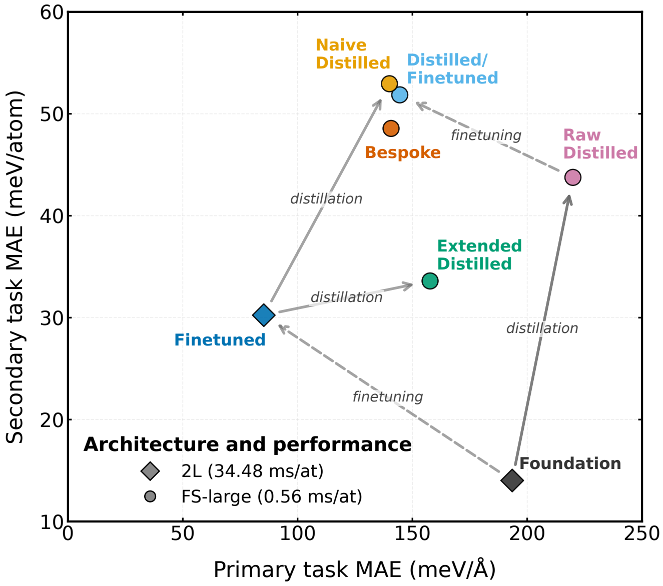

FIG. 10

on a formally complete mathematical basis of the local and semi-local atomic environment. Our foundational GRACE potentials obey fundamental physical symmetries and guarantee invariance under rotation, inversion, translation, as well as permutation of identical atomic species. Forces are conservative and computed as gradients of the energy. We further demonstrate computational efficiency at high precision (FP64). This sets GRACE apart from many other MLIPs that employ uncontrolled basis approximations, replace exact symmetries or exact gradients by numerical estimates, or lower precision to FP32 for computational efficiency.

Our foundational GRACE potentials were trained on a massive dataset of DFT calculations that encompasses a broad range of atomic configurations and chemistries. Comprehensive validation across diverse simulation tasks demonstrates excellent performance and versatility of the GRACE potentials. On the MatBench Discovery leaderboard, our models largely define the Pareto front, showcasing a superior trade-off between predicted thermodynamic stability, formation energy MAE and computational speed. Furthermore, the two-layer GRACE models exhibit leading performance in predicting thermal conductivity, a property highly sensitive to a model's ability to capture anharmonic contributions to atomic displacements and higher-order derivatives. This highlights the robustness of the GRACE framework beyond simple energy and force predictions.

Further validation on structural defects - grain boundaries, surfaces, and point defects - confirms that the GRACE models are able to describe effectively non-equilibrium and open structures. Some outliers, for example, for the alkali metals K, Rb and Cs, are attributed simply to the cut-off radius, which is too small for these large elements.

Long-time molecular dynamics simulations of FLiBe molten salt demonstrated the GRACE models' ability to maintain stability and accurately capture dynamic properties, such as radial distribution functions and diffusion coefficients, over extended timescales.

Beyond the core validation, we explored the practical application of foundational GRACE potentials through fine-tuning and distillation. We show that fine-tuning GRACE models on small, specialized datasets for Al-Li significantly improves their accuracy for specific downstream tasks. This approach is particularly effective in low-data regimes, outperforming models trained from scratch. Moreover, our investigation into catastrophic forgetting demonstrates that freezing specific model layers can mitigate a model's tendency to lose general knowledge when learning new, distinct tasks, as illustrated for a dataset specific for hydrogen combustion. Finally, we successfully distilled a complex GRACE model into a simpler, more computationally efficient one, while also improving its accuracy on a wider configurational space in comparison to a model trained from scratch. This shows that foundational potentials can act as powerful "teachers" and opens a path for creating specialized, high-performance ACE and GRACE potentials for specific applications without the need for extensive new DFT calculations.

IV. METHODS

A. Models architecture and parameterization

All GRACE models were implemented in the gracetensorpotential package, which is based on the TensorFlow library[90]. The models implement a graph extension of the atomic cluster expansion (ACE) method, utilizing both star- and tree-like many-body basis functions, that form an orthonormal and complete basis set[49]. The configurations of the GRACE foundational potentials are schematically depicted in Fig. 11 and provided in more detail in Table III.

The atomic structure is represented by chemical species types μ_i and atomic positions $\mathbf{r}_i$, which are transformed into bond vectors $\mathbf{r}_{\rm ij}$ between neighboring atoms within a 6 Å cutoff radius. Geometric information is encoded using a Chebyshev radial basis and spherical harmonics $Y_{\rm lm}$ up to $l_{\rm max} = 4$[4,82,83]. The Chebyshev polynomials are transformed into radial functions $R^{(\dots)}_{\rm nl}$ using a multi-layer perceptron (MLP) with two hidden layers, each containing 64 units. Chemical species are embedded into a 128-dimensional space Z_i. Single-particle basis functions are constructed from these inputs and summed into atomic bases $A_{\rm i,nl}$[4,49]. Many-body basis functions up to the fourth product order are constructed via recursive ACE evaluation using sparse coupling operations (see Fig. 11, bottom panel). During the recursive ACE basis evaluation, we utilize an equivariant sum operation $\oplus$, summing equivariant quantities with the same

l-character, and a sparse equivariant coupling $\otimes$ operation which employs Clebsch-Gordan coefficients. We employ a specific coupling order to avoid degenerate product functions. The maximum angular momentum, $L_{\max}$, varies depending on the product order. In the one-layer model, this procedure yields a star-graph ACE basis. For two-layer models, equivariant basis functions are linearly combined to define equivariant atomic representations $I_{\mathrm{i,nL}}$[48]. These representations carry geometric information about the atomic environment, effectively extending the interaction range to $2r_{\mathrm{cut}}$. They serve as input for a second recursive ACE evaluation, resulting in a tree graph GRACE basis[49]. Finally, all invariant basis functions from all product orders of the first and second layers are aggregated to form the complete GRACE basis, which is then linearly combined to generate a set of atomic densities $\varphi_{i,p}$. The atomic energy is computed via an embedding scheme as the sum of the first atomic density and the result of processing the remaining densities through another MLP. The total energy of the structure is the sum of all atomic energies. Forces are computed via the gradient of the total energy with respect to the atomic positions. All calculations are performed in double precision (FP64). The models are just-in-time (JIT) compiled using XLA to achieve optimal performance. Further details on the GRACE models will be published in a separate work.

The GRACE models were parameterized using the recently published OMat24[56] dataset in combination with the sAlex[54,55] dataset and the MPTraj dataset (v2022.10.28)[52,53]. OMat24 includes a wide range of structures, whereas sAlex and MPTraj include only relaxation trajectories. Raw VASP energies, forces and stresses were used for parameterizations.

We employed a loss function that consists of different parts,

$$\mathcal{L} = \alpha_{\mathrm{E}}\mathcal{L}_{\mathrm{E}} + \alpha_{\mathrm{F}}\mathcal{L}_{\mathrm{F}} + \alpha_{\mathrm{S}}\mathcal{L}_{\mathrm{S}}, \tag{1}$$

where $\mathcal{L}_{\mathrm{E}}$, $\mathcal{L}_{\mathrm{F}}$, and $\mathcal{L}_{\mathrm{S}}$ correspond to losses of energy per atom, force component, and stress component, respectively. We utilize the Huber loss for $\mathcal{L}$ with parameter $\delta = 0.01$ for all components. All models were initially trained on the OMat24 dataset for 12 epochs for 1L models and for 8 epochs for 2L models, constituting the "OMat-base" models. For this stage we use $\alpha_{\mathrm{E}} : \alpha_{\mathrm{F}} : \alpha_{\mathrm{S}} = 16 : 128 : 128$. Subsequently, we fine-tuned "OMat-base" models on the combination of MPTraj and sAlex datasets for additional 8 and 4 epochs for 1L and 2L models respectively with $\alpha_{\mathrm{E}} : \alpha_{\mathrm{F}} : \alpha_{\mathrm{S}} = 128 : 128 : 256$, leading to the "OAM" models. In addition, we fine-tuned "OMat-base" models on the same OMat24 dataset for additional 2 epochs with adjusted loss component weights to $\alpha_{\mathrm{E}} : \alpha_{\mathrm{F}} : \alpha_{\mathrm{S}} = 128 : 128 : 256$ to give more weight to energies. This models are designated with "ft-E".

For loss optimization we employed the Adam[91] optimizer with cosine learning rate reduction scheme, initial learning rate of $8{\times}10^{-3}$ and minimum learning rate of $5{\times}10^{-4}$. For fine tuning we use constant learning rate of $1{\times}10^{-4}$. To optimize data throughput we split the data into batches based on the total number of bonds rather than structures. The batch size was set to 165000 bonds per device which on average corresponds to about 200 structures. Training was performed using a single node with 8 Nvidia H100 80GB GPUs. Complete training cost of GRACE models varies between 400 and 700 GPU hours for the smallest and the most complex model, respectively.

B. Validation

Regarding the foundational machine learning interatomic potentials used for comparison, for the MACE models, the 'medium-mpa-0' checkpoint was used for MACE-MPA-0, and 'mace-omat-0-medium' for MACE-OMAT-0 (mace-torch version 0.3.12, cuequivariance version 0.3.0, cuequivariance-ops-torch-cu12 version 0.3.0, cuequivariance-torch version 0.3.0) All MACE models were using float64 precision and had the `cuEQ` option enabled. The MatterSim model utilized the 'mattersim-v1.0.0-5m' checkpoint. The ORB model was based on the 'v3-conservative-inf' modification, configured for 'float32-highest' precision, that corresponds to full float32 precision. SevenNet employed the '7net-mf-ompa' checkpoint with its 'mpa' modality. For eSEN, the 'esen_30m_omat' checkpoint with a seed of 0 was selected. UMA-M and UMA-S models corresponded to their 'uma-m-1p1' and 'uma-s-1p1' checkpoints, respectively, specifying the 'omat' task. The DPA model used the '2025-01-10-dpa3-openlam' checkpoint, whereas for computational performance tests "DPA-3.1-3M-ft" checkpoint was used. For eqV2_M model "eqV2_86M_omat_mp_salex" checkpoint was used. All GRACE models consistently operated with float64 precision.

The computational performance of foundational MLIPs, depicted in Fig. 1, was determined by measuring the averaged wall time for ten molecular dynamics (MD) steps. These simulations were performed for a 1024-atom supercell of tungsten BCC using ASE on a single NVIDIA A100 GPU with 80 GB of memory. Initial runs were excluded to account for Just-In-Time (JIT) compilation and other caching effects. Each MD simulation was independently repeated ten times and the results averaged. The final metrics are normalized by the number of atoms and MD steps.

We used elastic properties from the Materials Project[72] as our reference, employing an energy-based method[92] to compute the elastic tensor C_{ij} in Voigt notation. Out of 10073 reference elastic matrix calculations, we could not compute elastic constants for all structures due to relaxation and convergence issues with some potentials. Therefore, we relied on a common subset of 7962 structures for which elastic tensors were successfully computed by all potentials. This subset serves as a robust and representative test set for evaluating elastic matrix predictions. The eSEN model was not validated due to its high

TABLE III: Configurations of GRACE foundational potentials. See text for more details.

Configuration	1L	1L-medium	1L-large	2L	2L-medium	2L-large
r_{cut} (Å)	6	6	6	6	6	6
Radial basis func.	Cheb	Cheb	Cheb	Cheb	Cheb	Cheb
Num. radial basis	8	10	10	8	10	10
$l_{\max}$	4	4	4	4	4	4
Num.elements	89	89	89	89	89	89
Chem. embedding	128	128	128	128	128	128
Num.radial funcs.	32	32	32	32	42	42
Product order	4	4	4	4	4	4
$L_{\max}$ per product order (layer 1):	4/4/0/0	4/4/0/0	4/4/0/0	4/4/1/0	4/4/1/1	4/4/3/1
Tot. num. equivar. funcs.	-	-	-	4000	9576	13860
$L_{\max}$ per product order (layer 2)	-	-	-	4/4/0/0	3/3/0/0	3/3/0/0
Tot. num. invar. funcs.	2848	2848	5664	5696	7194	7194
Num. densities	12+1	16+1	16+1	12+1	16+1	16+1
Tot. num. params.	3447148	4461497	8953529	12597516	21764956	26394284

computational expenses.

Grain boundary structures and their corresponding reference structures were sourced from the Crystalium project[73], a dataset closely related to the Materials Project[52]. The reference structures were fully relaxed using the BFGS method with a FrechetCellFilter, applying a relaxation criterion of 0.001 eV/Å. Grain boundary relaxation was carried out using the FIRE minimization algorithm, also with a FrechetCellFilter, enforcing a maximum force threshold of 0.01 eV/Å and a limit of 500 optimization steps. For eqV2 model, relaxation criteria was loosen to 0.02 eV/Å due to numerical instabilities. In total, 327 grain boundaries were initially computed across 58 different elements. To ensure consistency and avoid discrepancies, only structures with a grain boundary plane orthogonal to the z-direction were selected, resulting in a final set of 297 grain boundary structures.

Surface structures and their corresponding reference data were sourced from the Crystalium project[74], which is associated with the Materials Project[52]. The same relaxation settings as for grain boundaries were applied. From an initial total of 1124 surface structures, only those with a surface plane orthogonal to the z-direction were selected, yielding a final set of 716 surface structures for analysis.

Reference data for self-interstitial and vacancy formation energies were taken from Ref.[75] for 13 BCC metals and Ref.[76] for 13 FCC metals. All these reference values were computed using the PBE functional.

V. DATA AVAILABILITY

Training datasets (MPTrj, sAlex and OMat24) are publicly available. GRACE foundational potentials are available at gracemaker.readthedocs.io/en/latest/gracemaker/foundation

VI. CODE AVAILABILITY

Code for GRACE potential is available at github.com/ICAMS/grace-tensorpotential

VII. FUNDING

We acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through SFB 1394 (project number 409476157).

VIII. ACKNOWLEDGEMENTS

High-performance computing (HPC) resources were provided by the Paderborn Center for Parallel Computing (PC^2) and the Elysium HPC cluster at Ruhr-University Bochum.

IX. AUTHOR CONTRIBUTIONS

Conceptualisation and Project Administration: All authors. Y.L. and A.B. developed the software and parameterized the models. Writing - original draft: Y.L. Writing-review and editing: All authors. Resources and funding acquisition: Y.L., R.D.

X. COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

[1] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).

[2] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[3] A. V. Shapeev, Multiscale Model. Simul. **14**, 1153 (2016).

[4] R. Drautz, Phys. Rev. B **99**, 014104 (2019).

[5] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, Chemical Reviews **121**, 10142 (2021).

[6] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth, and B. Kozinsky, Nature Communications **14**, 579 (2023).

[7] R. Jacobs, D. Morgan, S. Attarian, J. Meng, C. Shen, Z. Wu, C. Y. Xie, J. H. Yang, N. Artrith, B. Blaiszik, *et al.*, Current Opinion in Solid State and Materials Science **35**, 101214 (2025).

[8] M. Kulichenko, B. Nebgen, N. Lubbers, J. S. Smith, K. Barros, A. E. A. Allen, A. Habib, E. Shinkle, N. Fedik, Y. W. Li, R. A. Messerly, and S. Tretiak, Chemical Reviews **124**, 13681 (2024).

[9] G. Wang, C. Wang, X. Zhang, Z. Li, J. Zhou, and Z. Sun, iScience **27**, 109673 (2024).

[10] F. L. Thiemann, N. O'Neill, V. Kapil, A. Michaelides, and C. Schran, Journal of Physics: Condensed Matter **37**, 073002 (2024).

[11] E. C. Y. Yuan, Y. Liu, J. Chen, P. Zhong, S. Raja, T. Kreiman, S. Vargas, W. Xu, M. Head-Gordon, C. Yang, S. M. Blau, B. Cheng, A. Krishnapriyan, and T. Head-Gordon, "Foundation models for atomistic simulation of chemistry and materials," (2025), arXiv:2503.10538 [physics.chem-ph].

[12] D. G. Pettifor, Journal of Physics C: Solid State Physics **3**, 367 (1970).

[13] D. G. Pettifor, Journal of Physics C: Solid State Physics **19**, 285 (1986).

[14] D. G. Pettifor, *Bonding and Structure in Molecules and Solids* (Oxford University Press, Oxford, 1995).

[15] B. Seiser, R. Drautz, and D. Pettifor, Acta Materialia **59**, 749 (2011).

[16] A. F. Bialon, T. Hammerschmidt, and R. Drautz, Chemistry of Materials **28**, 2550 (2016), https://doi.org/10.1021/acs.chemmater.5b04299.

[17] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, International Journal of Quantum Chemistry **115**, 1094 (2015), https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.24917.

[18] B. Parsaeifard, D. Sankar De, A. S. Christensen, F. A. Faber, E. Kocer, S. De, J. Behler, O. Anatole von Lilienfeld, and S. Goedecker, Machine Learning: Science and Technology **2**, 015018 (2021).

[19] N. Lopanitsyna, G. Fraux, M. A. Springer, S. De, and M. Ceriotti, Phys. Rev. Mater. **7**, 045802 (2023).

[20] T. F. T. Cerqueira, H. Wang, S. Botti, and M. A. L. Marques, "A non-orthogonal representation of the chemical space," (2025), arXiv:2406.19761 [cond-mat.mtrl-sci].

[21] A. K. Rappé, C. J. Casewit, K. Colwell, W. A. Goddard III, and W. M. Skiff, Journal of the American chemical society **114**, 10024 (1992).

[22] C. Chen and S. P. Ong, Nature Computational Science **2**, 718 (2022).

[23] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, Nature Machine Intelligence **5**, 1031 (2023).

[24] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, *et al.*, arXiv preprint arXiv:2401.00096 (2023).

[25] L. Barroso-Luque, M. Shuaibi, X. Fu, B. M. Wood, M. Dzamba, M. Gao, A. Rizvi, C. L. Zitnick, and Z. W. Ulissi, arXiv preprint arXiv:2410.12771 (2024).

[26] H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S. Chen, C. Zeni, *et al.*, arXiv preprint arXiv:2405.04967 (2024).

[27] J. Kim, J. Kim, J. Kim, J. Lee, Y. Park, Y. Kang, and S. Han, Journal of the American Chemical Society **147**, 1042 (2024).

[28] B. Yin, J. Wang, W. Du, P. Wang, P. Ying, H. Jia, Z. Zhang, Y. Du, C. P. Gomes, C. Duan, *et al.*, arXiv preprint arXiv:2501.07155 (2025).

[29] D. Zhang, A. Peng, C. Cai, W. Li, Y. Zhou, J. Zeng, M. Guo, C. Zhang, B. Li, H. Jiang, *et al.*, arXiv preprint arXiv:2506.01686 (2025).

[30] X. Fu, B. M. Wood, L. Barroso-Luque, D. S. Levine, M. Gao, M. Dzamba, and C. L. Zitnick, arXiv preprint arXiv:2502.12147 (2025).

[31] A. Mazitov, F. Bigi, M. Kellner, P. Pegolo, D. Tisi, G. Fraux, S. Pozdnyakov, P. Loche, and M. Ceriotti, arXiv preprint arXiv:2503.14118 (2025).

[32] T. Liang, K. Xu, E. Lindgren, Z. Chen, R. Zhao, J. Liu, E. Berger, B. Tang, B. Zhang, Y. Wang, *et al.*, arXiv preprint arXiv:2504.21286 (2025).

[33] J. Gasteiger, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," (2020), arXiv:2003.03123 [cs.LG].

[34] B. Anderson, T.-S. Hy, and R. Kondor, in *Advances in Neural Information Processing Systems 32*, Neural Information Processing Systems Conference, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Neural Information Processing Systems Foundation, Inc, Vancouver, Canada, 2019) p. 9596.

[35] N. Lubbers, J. S. Smith, and K. Barros, The Journal of Chemical Physics **148**, 241715 (2018).

[36] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, arXiv preprint arXiv:1802.08219 (2018).

[37] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, Nature communications **13**, 2453 (2022).

[38] V. G. Satorras, E. Hoogeboom, and M. Welling, in *International conference on machine learning* (PMLR, 2021) pp. 9323–9332.

[39] O. T. Unke and M. Meuwly, Journal of chemical theory and computation **15**, 3678 (2019).

[40] K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, Advances in neural information processing systems **30** (2017).

[41] M. Haghighatlari, J. Li, X. Guan, O. Zhang, A. Das, C. J. Stein, F. Heidar-Zadeh, M. Liu, M. Head-Gordon, L. Bertels, *et al.*, Digital Discovery **1**, 333 (2022).

[42] K. T. Schütt, O. T. Unke, and M. Gastegger, “Equivariant message passing for the prediction of tensorial properties and molecular spectra,” (2021), arXiv:2102.03150 [cs.LG].

[43] J. Gasteiger, F. Becker, and S. Günnemann, “Gemnet: Universal directional graph neural networks for molecules,” (2022), arXiv:2106.08903 [physics.comp-ph].

[44] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Science Advances **3**, e1603015 (2017).

[45] S. Pozdnyakov and M. Ceriotti, Advances in Neural Information Processing Systems **36**, 79469 (2023).

[46] J. Nigam, S. Pozdnyakov, G. Fraux, and M. Ceriotti, J. Chem. Phys. **156**, 204115 (2022).

[47] I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. Simm, R. Drautz, C. Ortner, B. Kozinsky, and G. Csányi, Nature Machine Intelligence **7**, 56 (2025).

[48] A. Bochkarev, Y. Lysogorskiy, C. Ortner, G. Csányi, and R. Drautz, Physical Review Research **4**, L042019 (2022).

[49] A. Bochkarev, Y. Lysogorskiy, and R. Drautz, Phys. Rev. X **14**, 021036 (2024).

[50] I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. Simm, R. Drautz, C. Ortner, B. Kozinsky, and G. Csányi, Nature Machine Intelligence **7**, 56 (2025).

[51] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csányi, Advances in neural information processing systems **35**, 11423 (2022).

[52] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, APL Materials **1**, 011002 (2013).

[53] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, Nature Machine Intelligence **5**, 1031 (2023).

[54] J. Schmidt, N. Hoffmann, H.-C. Wang, P. Borlido, P. J. Carriço, T. F. Cerqueira, S. Botti, and M. A. Marques, Advanced Materials **35**, 2210788 (2023).

[55] H.-C. Wang, J. Schmidt, M. A. Marques, L. Wirtz, and A. H. Romero, 2D Materials **10**, 035007 (2023).

[56] 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].

[57] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, npj Computational Materials **1**, 1 (2015).

[58] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, *et al.*, Computational Materials Science **58**, 227 (2012).

[59] A. D. Kaplan, R. Liu, J. Qi, T. W. Ko, B. Deng, J. Riebesell, G. Ceder, K. A. Persson, and S. P. Ong, arXiv preprint arXiv:2503.04070 (2025).

[60] M. C. Kuner, A. D. Kaplan, K. A. Persson, M. Asta, and D. C. Chrzan, arXiv preprint arXiv:2507.05559 (2025).

[61] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).

[62] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).

[63] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).

[64] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[65] H.-C. Wang, S. Botti, and M. A. Marques, npj Computational Materials **7**, 12 (2021).

[66] J. Riebesell, R. E. Goodall, P. Benner, Y. Chiang, B. Deng, G. Ceder, M. Asta, A. A. Lee, A. Jain, and K. A. Persson, Nature Machine Intelligence **7**, 836 (2025).

[67] B. Póta, P. Ahlawat, G. Csányi, and M. Simoncelli, arXiv preprint arXiv:2408.00755 (2024).

[68] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. In’t Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, *et al.*, Computer physics communications **271**, 108171 (2022).

[69] A. Togo, L. Chaput, and I. Tanaka, Physical review B **91**, 094306 (2015).

[70] A. Togo, L. Chaput, T. Tadano, and I. Tanaka, Journal of Physics: Condensed Matter **35**, 353001 (2023).

[71] H. Kraß, J. Huang, and S. M. Moosavi, “Mofsimbench: Evaluating universal machine learning interatomic potentials in metal-organic framework molecular modeling,” (2025), arXiv:2507.11806 [cond-mat.mtrl-sci].

[72] 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.*, Scientific data **2**, 1 (2015).

[73] H. Zheng, X.-G. Li, R. Tran, C. Chen, M. Horton, D. Winston, K. A. Persson, and S. P. Ong, Acta Materialia **186**, 40 (2020).

[74] R. Tran, Z. Xu, B. Radhakrishnan, D. Winston, W. Sun, K. A. Persson, and S. P. Ong, Scientific data **3**, 1 (2016).

[75] P.-W. Ma and S. Dudarev, Physical Review Materials **3**, 013605 (2019).

[76] P.-W. Ma and S. Dudarev, Physical Review Materials **5**, 013601 (2021).

[77] S. T. Lam, Q.-J. Li, R. Ballinger, C. Forsberg, and J. Li, ACS Applied Materials & Interfaces **13**, 24582 (2021).

[78] A. Rodriguez, S. Lam, and M. Hu, ACS Applied Materials & Interfaces **13**, 55367 (2021).

[79] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, Journal of Physics: Condensed Matter **29**, 273002 (2017).

[80] S. Menon, Y. Lysogorskiy, A. L. Knoll, N. Leimeroth, M. Poul, M. Qamar, J. Janssen, M. Mrovec, J. Rohrer, K. Albe, *et al.*, npj Computational Materials **10**, 261 (2024).

[81] X. Guan, A. Das, C. J. Stein, F. Heidar-Zadeh, L. Bertels, M. Liu, M. Haghighatlari, J. Li, O. Zhang, H. Hao, *et al.*, Scientific data **9**, 215 (2022).

[82] Y. Lysogorskiy, M. Rinaldi, S. Menon, C. van der Oord, T. Hammerschmidt, M. Mrovec, A. Thompson, G. Csányi, C. Ortner, and R. Drautz, Npj Comput. Mater. **7**, 97 (2021).

[83] A. Bochkarev, Y. Lysogorskiy, S. Menon, M. Qamar, M. Mrovec, and R. Drautz, Physical Review Materials **6**, 013804 (2022).

[84] A. Peng, C. Cai, M. Guo, D. Zhang, C. Zhang, A. Loew, L. Zhang, and H. Wang, arXiv preprint arXiv:2504.19578 (2025).

[85] AI Squared, “OpenLAM Benchmark Introduction,” `https://www.aissquare.com/openlam?tab=Benchmark&type=Introduction` (2024), accessed: 2025-07-02.

[86] J. D. Morrow and V. L. Deringer, The Journal of Chemical Physics **157** (2022).

[87] J. L. Gardner, D. F. Toit, C. B. Mahmoud, Z. F. Beaulieu, V. Juraskova, L.-B. Paşca, L. A. Rosset, F. Duarte, F. Martelli, C. J. Pickard, *et al.*, arXiv preprint arXiv:2506.10956 (2025).

[88] N. Lopanitsyna, G. Fraux, M. A. Springer, S. De, and M. Ceriotti, Physical Review Materials **7**, 045802 (2023).

[89] A. Mazitov, M. A. Springer, N. Lopanitsyna, G. Fraux, S. De, and M. Ceriotti, Journal of Physics: Materials **7**, 025007 (2024).

[90] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, in *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (2016) pp. 265–283.

[91] D. P. Kingma and J. Ba, arXiv preprint arXiv:1412.6980 (2014).

[92] R. Golesorkhtabar, P. Pavone, J. Spitaler, P. Puschnig, and C. Draxl, Computer Physics Communications **184**, 1861 (2013).

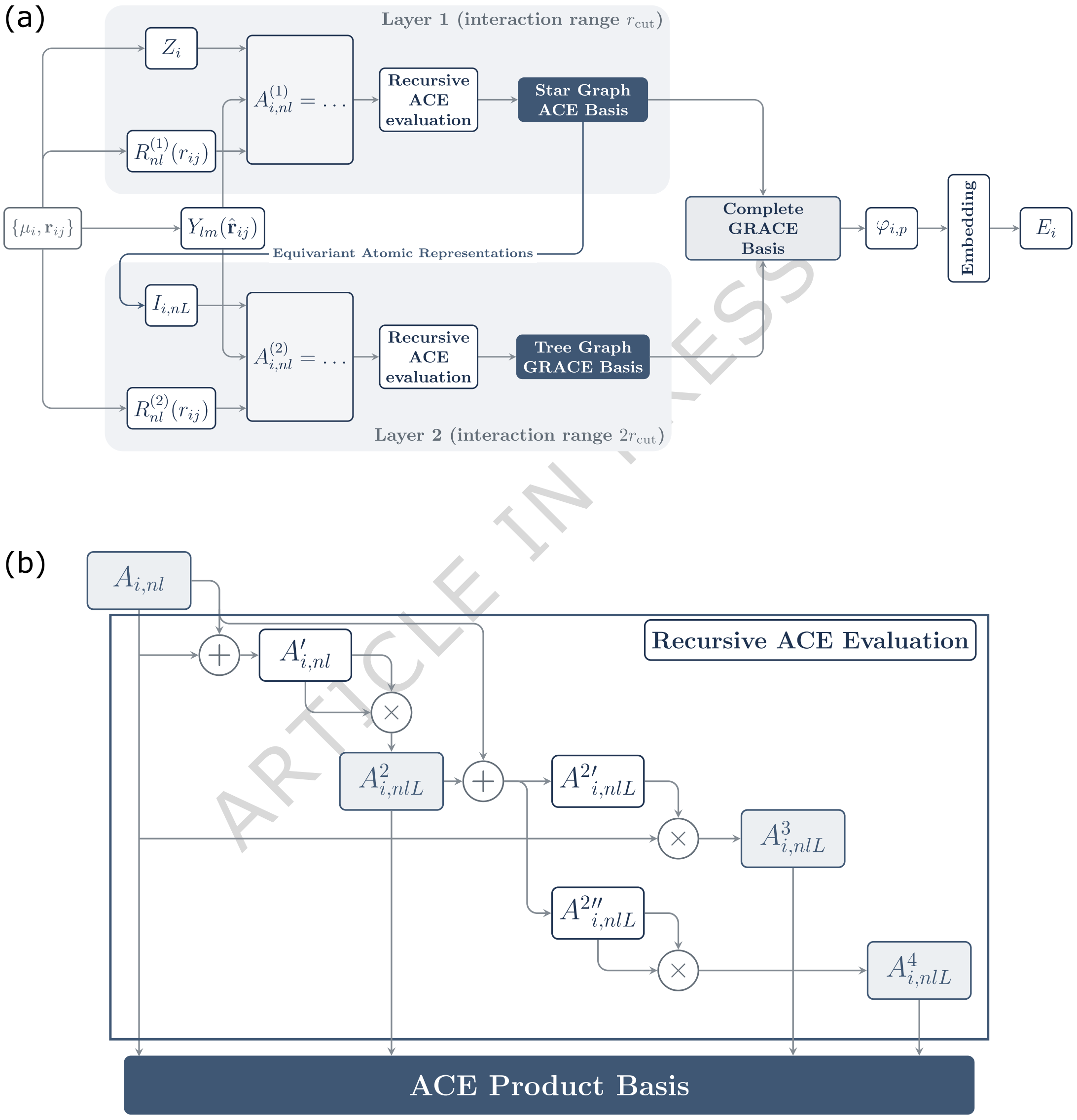
(a)
Layer 1 (interaction range r_{cut})
Z_i
$A^{(1)}_{i,nl} = \dots$
Recursive ACE evaluation
Star Graph ACE Basis
$R^{(1)}_{nl}(r_{ij})$
$\{\mu_i, \mathbf{r}_{ij}\}$
$Y_{lm}(\hat{\mathbf{r}}_{ij})$
Equivariant Atomic Representations
$I_{i,nL}$
$A^{(2)}_{i,nl} = \dots$
Recursive ACE evaluation
Tree Graph GRACE Basis
$R^{(2)}_{nl}(r_{ij})$
Layer 2 (interaction range $2r_{\mathrm{cut}}$)
Complete GRACE Basis
$\varphi_{i,p}$
Embedding
E_i
(b)
$A_{i,nl}$
Recursive ACE Evaluation
$+$
$A'_{i,nl}$
$\times$
$A^2_{i,nlL}$
$+$
$A^{2\prime}_{i,nlL}$
$\times$
$A^3_{i,nlL}$
$A^{2\prime\prime}_{i,nlL}$
$\times$
$A^4_{i,nlL}$
ACE Product Basis

FIG. 11

FIGURE LEGENDS

Figure 1. Pareto front of accuracy versus efficiency. (a) Model performance for stable structure identification (F1 score in MatBench Discovery benchmark) versus computational time. (b) Thermal conductivity prediction error (κ_{SRME}) versus computational time. A higher F1 score and lower κ_{SRME} indicate better performance. The blue dashed line links Pareto optimal models. Computational performance is estimated via ASE (filled symbols) and LAMMPS (open symbols), with GRACE models indicated in red.

Figure 2. The symmetric relative mean error (SRME) and MAE (ΔC in GPa) for elastic constants, categorized into three subgroups: longitudinal (C_{11}, C_{22}, C_{33}), Poisson's ratio-related (C_{12}, C_{13}, C_{23}), and shear (C_{44}, C_{55}, C_{66}). See text for more details.

Figure 3. Accuracy for grain boundary formation energies of unary systems: symmetric relative mean error γ_{GB}-SRME and mean absolute error $\Delta\gamma_{\mathrm{GB}}$. GRACE models are highlighted in red. GRACE models are highlighted in red.

Figure 4. Accuracy for surface formation energies of unary systems: symmetric relative mean error γ_{surf}-SRME and mean absolute error $\Delta\gamma_{\mathrm{surf}}$. GRACE models are highlighted in red.

Figure 5. Accuracy for point defect formation energies of unary systems. Error metrics for (a) self-interstitials (SIA) and (b) vacancies. GRACE models are highlighted in red.

Figure 6.RDF of FLiBe salt from MD at 973K using GRACE in comparison to AIMD-DFT results[77,78].

Figure 7. Computational performance and memory scaling (maximum number of atoms) of foundational GRACE potentials in LAMMPS on a single A100-80GB GPU, evaluated across diverse materials systems.

Figure 8. Fine-tuning performance on the Al-Li system. Mean absolute error (MAE) for (a) energies, (b) forces, (c) formation energies, (d) elastic matrix elements, (e) phonon density of states (PhDOS), and (f) vacancy formation energies. PhDOS error is measured by Tanimoto similarity $1-\mathrm{S}_{\mathrm{PhDOS}}$. Data at zero number of train structures correspond to zero-shot models.

Figure 9. Mean Absolute Error (MAE) of forces for fine-tuned GRACE-2L-OAM models. The x-axis shows MAE on the new H2COMB dataset (downstream task), the y-axis shows MAE on the original sAlex dataset. Performance for sAlex is split into H,O atoms (light grey diamonds) and other elements (dark grey triangles). 'Baseline' refers to the zero-shot model. 'Naive' corresponds to strategy 1 (naive fine-tuning). 'Trainable-1' and 'Trainable-2' correspond to strategies 2a and 2b, respectively, representing different frozen-weights approaches.

Figure 10. Efficiency and accuracy trade-offs in model distillation. (a) Trade-off between accuracy on the primary task (HEA25S force MAE) and general chemical stability (secondary task: formation energy MAE). (b) Computational cost versus primary task accuracy. (c) Computational cost versus secondary task accuracy. Computational performance for both GRACE-2L and GRACE-FS architectures was evaluated on a 10-core CPU and normalized per core. See text for details.

Figure 11. Architecture of GRACE models. (a) Overall schematic of the model architecture. (b) Recursive ACE basis evaluation details.