EXFIT-BENCH: WHAT TRANSFERS WHEN UNIVERSAL INTERATOMIC POTENTIALS ARE FINE-TUNED ON EXPERIMENTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Universal machine learning interatomic potentials (uMLIPs) are now widely adopted to enable simulations across diverse materials, which makes their evaluation increasingly important. Most benchmarks measure agreement with density functional theory (DFT), while some compare properties predicted by pretrained models with experiment. When measurements are available, however, a potential can be fine-tuned to match them before further use. We introduce ExFiT-Bench to evaluate uMLIPs in this setting. The benchmark focuses on ensemble-derived properties, for which experimental measurements provide no labels for individual configurations. Under a shared protocol, we fine-tune six uMLIPs on five tasks spanning elastic constants, a phase transition and a liquid structure factor in titanium, diamond and MgO. We evaluate the target observable and transfer to held-out temperatures, another polymorph, other observables and compositions. Fine-tuning reduces the on-target error for every model on every task, but the pretrained ranking does not consistently carry over. The correction often transfers to new conditions of the target observable, but transfer across observables and compositions is less reliable and can reduce accuracy. We release the benchmark data, code and a leaderboard for evaluating new uMLIPs and fine-tuning methods.

1 INTRODUCTION

Machine learning interatomic potentials (MLIPs) make molecular dynamics (MD) simulations feasible at scales that are difficult to reach with density functional theory (DFT) calculations. Universal MLIPs (uMLIPs) extend this capability across diverse elements and compositions using a single pretrained model. By providing reusable potentials for new systems, they reduce the need for system-specific model development and support applications across solids, liquids, and interfaces (Batatia et al., 2025; Yang et al., 2024). As their applications and the number of available models expand, benchmarks become increasingly important for selecting a potential for a given simulation.

Most uMLIP benchmarks compare predictions with held-out DFT calculations or DFT simulation-derived properties, including crystal stability, phonons, thermal transport and elasticity (Riebesell et al., 2025; Loew et al., 2025; Han & Cheng, 2025; Póta et al., 2024; Gao & Wang, 2025; Peng et al., 2025; Wines & Choudhary, 2024). Others compare properties predicted by simulations with experimental measurements (Mannan et al., 2025; Stracke et al., 2026). These benchmarks evaluate pretrained models. However, when measurements are available for a material, a researcher may fine-tune a uMLIP to improve agreement with them before using it to predict properties at other conditions or in other materials. This raises a different model-selection question: which potential performs well after experimental fine-tuning?

In this work, we focus on experimental observables estimated from MD ensembles, including finite-temperature elastic constants, phase equilibria, and liquid structure. These properties connect atomistic simulations with measured material behavior through the distribution of atomic configurations at a given temperature and pressure. Agreement with DFT energies and forces alone does not ensure agreement with experiment, motivating direct fitting to these observables. Such fitting must account for how changes to the potential alter the sampled ensemble, without experimental energy or force labels for individual configurations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

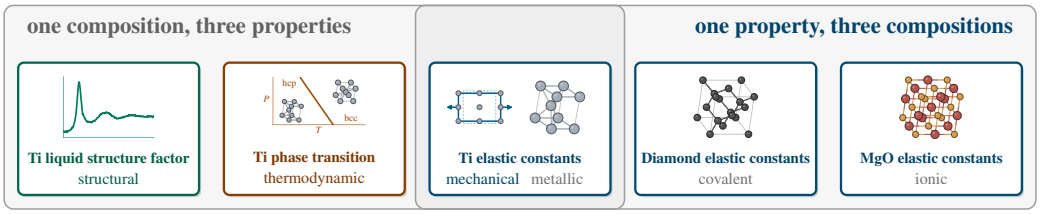


Figure 1: The five tasks. Titanium (left) varies the property fine-tuned across liquid structure, the hcp-bcc phase transition and hcp elastic constants. Elastic constants (right) vary the composition across titanium, diamond and MgO, spanning metallic, covalent and ionic bonding.

Existing benchmarks evaluate pretrained predictions, but do not capture how effectively a uMLIP incorporates experimental measurements or how the resulting updates affect predictions beyond the fitted data. In DiffTRe-based fine-tuning, these updates depend on the configurations sampled by each model and their response to changes in the potential, neither of which is characterized by pretrained prediction errors alone. Models with similar initial accuracy can therefore differ in both on-target error and transfer, motivating a benchmark that evaluates them after experimental fine-tuning.

Previous studies have demonstrated experimental fitting for system-specific potentials and examined its effects on selected properties beyond the fitted measurements (Thaler & Zavadlav, 2021; Röcken & Zavadlav, 2024; Fuchs & Zavadlav, 2026), mainly relying on differentiable trajectory reweighting (DiffTRe) to optimize an ensemble-level loss without backpropagating through long MD trajectories (Thaler & Zavadlav, 2021). How these effects vary across uMLIPs, target observables, and compositions remains insufficiently characterized since a small set of experimental measurements does not uniquely determine a potential, and models that reproduce those measurements may make different predictions elsewhere. To fill this gap, we introduce the experimental fine-tuning benchmark (ExFiT-Bench) to evaluate uMLIPs. Our contributions are twofold:

- **A benchmark of experimental fine-tuning for uMLIPs.** We introduce ExFiT-Bench, fine-tune six uMLIPs under a shared DiffTRe-based protocol on five tasks spanning elastic constants, a phase transition and a liquid structure factor across titanium, diamond and MgO, and evaluate the target observables and predictions at held-out temperatures, on another polymorph, and for other observables and compositions.
- **Selective transfer and changing model rankings.** Prior studies have shown that fitting material-specific potentials to experimental observables can improve predictions beyond the fitted measurements, while leaving other properties uncertain. We find that on-target error improves consistently, but downstream changes depend on the choice of target and model. These changes also alter model rankings, making performance after experimental fine-tuning a distinct model-selection criterion.

2 ExFiT-Bench

ExFiT-Bench fine-tunes uMLIPs on experimental observables and scores the result on and off the target. This section defines the five tasks (Section 2.1), the on-target and off-target evaluation (Section 2.2) and the fine-tuning method (Section 2.3).

2.1 Experimental tasks across properties and compositions

To cover the experimental measurements of many properties and materials, we target five tasks of diverse properties and compositions (Figure 1, Table 4 of Appendix A.1). Properties span the elastic constants, the phase transition temperature and the liquid structure factor of titanium. Compositions span titanium, diamond and magnesium oxide (MgO). Their bonding is metallic, covalent and ionic. We fine-tune six uMLIPs that differ in architecture, equivariance, pretraining data and size (Table 6): MACE (Batatia et al., 2022; 2025), MatterSim (Yang et al., 2024), PET-MAD (Mazitov et al., 2025a), SevenNet (Park et al., 2024; Kim et al., 2025a), Orb (Rhodes et al., 2025) and UMA (Wood et al., 2025).

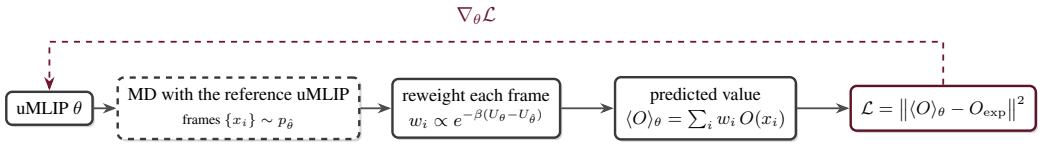


Figure 2: Fine-tuning on an experimental observable by trajectory reweighting. The trajectory regeneration is dashed, since it runs only when the effective sample size falls below a threshold.

Elastic constants. The diamond and MgO elastic tasks target three cubic elastic constants, second derivatives of the free energy with respect to strain, near room temperature. The titanium elastic task targets five hexagonal constants at four temperatures of the experimental series. The three elastic tasks also hold out temperatures: four on titanium, and a single temperature on diamond and on MgO (Appendix A.1). In the diamond and titanium elastic tasks we reproduce Thaler & Zavadlav (2021) and Röcken & Zavadlav (2024), and compare with the system-specific potential of each study, the DimeNet++ baseline, fitted to the same measurements. For MgO we use the measurement of Sumino et al. (1983). Errors are symmetric relative mean errors (SRME) over target elastic constants (Makridakis, 1993; Póta et al., 2024). We take the measurement disagreement as the SRME between the experimental values and an independent measurement of the same constants, on MgO (Fan et al., 2019) and on diamond (Vogelgesang et al., 1996).

Phase transition. The phase transition task follows Fuchs & Zavadlav (2026) and targets the transition line between the hexagonal close-packed (hcp) and bcc phases of titanium under pressure. We fit the hcp-bcc Gibbs free-energy difference ΔG at six points on the measured line, where the experimental value is zero. The error is the mean absolute error (MAE) of ΔG over these points, and we compute ΔG by thermodynamic integration (Appendix A.3). The transition temperature of a model at a given pressure is the temperature at which ΔG crosses zero.

Liquid structure factor. The liquid structure factor task extends the benchmark from the mechanical and thermodynamic properties of crystals to the structure of a liquid. Diffraction measures liquid structure as the static structure factor $S(q)$ (Fischer et al., 2006). The task targets $S(q)$ of liquid titanium at 1965 K, measured by neutron diffraction (Holland-Moritz et al., 2007), where a shoulder on the second peak marks icosahedral short-range order, in a cell at the experimental density (Ozawa et al., 2017). The experiment also measures the undercooled melt at 1745 K, which the task holds out. We report the root-mean-square error (RMSE) of $S(q)$ against the measurement.

2.2 On-target and off-target scoring

We score the fine-tuned models on target and off target. The on-target error is the error of the target observable at the fitted conditions, in the metric defined with the task above. The target is an ensemble average over the model's own Boltzmann distribution. Reaching the experimental value is not guaranteed.

The off-target errors are measured on quantities the loss does not constrain, which for each task are the other columns of Tables 1 and 3. These are the held-out temperatures of the fitted observable where the task has any, bcc titanium as the held-out polymorph, the transition line, the structure factor and the phonons as other observables of titanium, and the elastic constants of the two other compositions.

2.3 Fine-tuning by differentiable trajectory reweighting

Here, we describe our approach to fine-tuning uMLIPs on experimental observables.

Experimental loss. An experimental observable is an ensemble average over the configurations a model samples in MD, not a property of any single configuration. Direct regression to DFT labels cannot fit an ensemble average: a DFT label belongs to one configuration, and no configuration carries a label for the average itself. The loss instead compares the average with the measurement.

The experimental loss is the squared residual between the predicted and experimental values,

$$\mathcal{L}_{\text{exp}}(\theta) = \|\langle O\rangle_\theta - O_{\text{exp}}\|^2 , \tag{1}$$

where $\langle O\rangle_\theta$ is the ensemble average of the target observable O under the model with weights θ, and O_{exp} is its experimental value. Naïvely backpropagating this loss through the MD integrator needs the whole trajectory's computation graph in memory, which is intractable beyond a few hundred atoms and steps, and the gradient explodes over longer trajectories (Thaler & Zavadlav, 2021).

Differentiable trajectory reweighting. To fine-tune the uMLIPs to minimize Equation (1), researchers employ differentiable trajectory reweighting (Figure 2), which differentiates the ensemble average without differentiating the integrator. We write EXP-ONLY for fine-tuning on this loss alone, compared with four regularization strategies in Appendix A.5. From a trajectory $x_1, \dots, x_n$ generated at reference weights $\hat{\theta}$, the average at nearby θ is

$$\langle O\rangle_\theta = \sum_{i=1}^{n} w_i(\theta)\, O(x_i), \qquad w_i(\theta) = \frac{\exp\left[-\beta\left(U_\theta(x_i) - U_{\hat{\theta}}(x_i)\right)\right]}{\sum_j \exp\left[-\beta\left(U_\theta(x_j) - U_{\hat{\theta}}(x_j)\right)\right]}, \tag{2}$$

with U_θ the potential energy, T the temperature, k_{B} the Boltzmann constant and $\beta = 1/k_{\text{B}}T$. The estimate holds only near $\hat{\theta}$, so we regenerate the trajectory, and with it $\hat{\theta}$, when the effective sample size $N_{\text{eff}} = \exp\left[-\sum_i w_i \log w_i\right]$ (Thaler & Zavadlav, 2021) falls below 0.90 of the frames. We train for 40 epochs at a constant learning rate selected per model and task from the training loss (Appendix A.4). The task estimators, measurement sources and MD settings are in Appendix A.

3 RESULTS

We first ask how closely each model matches the target observable, and then whether the correction transfers beyond that observable. Fine-tuning lowers the on-target error of every model on all five tasks (the gray entries of Tables 1 and 3), but the pretrained ranking does not carry to the fine-tuned models. For an off target, (i) does fine-tuning on an observable transfer to other observables and conditions (Section 3.1)? (ii) Does fine-tuning on a composition transfer to other compositions (Section 3.2)?

3.1 PROPERTY-DEPENDENT TRANSFER AND FORGETTING IN TITANIUM

This subsection examines the three titanium tasks, the elastic constants, the phase transition and the structure factor (Table 1). Off-target performance depends on the pretrained uMLIPs and which observable we fit.

Fine-tuned on the elastic constants. At the four held-out temperatures, the error matches that at the fitted ones (Table 1, and Figure 11 for the full temperature series). Against both Fisher & Dever (1970) and Ledbetter et al. (2004), fine-tuning also lowers the bcc error (Table 11). Most models end below the disagreement between the two determinations (Figure 3a). The improvement comes from reducing the gap between the predicted and target elastic constants, which most pretrained models underestimate (C_{11}, C_{33}, C_{44}) or overestimate (C_{12}, C_{13}) on titanium (Figure 8 for the signed error per constant). Deng et al. (2025) report the same underestimate of the potential energy surface curvature on uMLIPs trained on MPtrj, so part of what the experimental loss corrects here can be a bias in the pretraining data.

Beyond the elastic constants, the transition line and the liquid structure both move, for more models toward the measurement than away from it (Table 1, Figure 3b,c), without the free energy or a liquid configuration entering the fine-tuning. The liquid error falls, and the second peak shows the short-range order is not what the correction reaches (Appendix B.3).

The phonon error falls for MatterSim, PET-MAD and UMA and stays within one standard deviation of the pretrained value for the other models. The elastic constants determine only the long-wavelength limit of the acoustic branches, and fine-tuning leaves the rest of the dispersion unconstrained (Figure 3d). MatterSim matches the DimeNet++ baseline of Röcken & Zavadlav (2024), which was pretrained on titanium DFT data.

Table 1: The three titanium tasks on various properties, pretrained and fine-tuned: one block per task below a block of the pretrained models, one row per model and one column per observable of Section 2.2, so that a column holds the three different tasks. A fine-tuned entry is the mean over three seeds with the standard deviation in parentheses, and the diagonal is the block's own observable at its own conditions, gray and not ranked. Among the other cells of a block, the lowest error in a column is bold and the second lowest underlined, and the last two columns are the mean rank of a row over the ranked columns, within its block and over the whole table, with one observable's columns averaged first so that no observable counts twice. Table 9 carries the pretrained value beside each entry, and Table 2 the orientation dependence of Orb and PET-MAD.

	Ti elastic (SRME)			ΔG (MAE, meV/atom)	$S(q)$ (RMSE)		phonon (RMSE, THz)	mean rank	
	fitted T	held-out T	bcc		fitted T	held-out T		block	all
Pretrained, no fine-tuning									
MACE	0.157	0.139	0.119	2.9	0.052	0.067	1.209	3.8	12.6
MatterSim	0.387	0.337	0.132	2.6	0.051	0.069	1.328	4.5	15.3
PET-MAD	0.322	0.288	**0.097**	9.8	0.060	0.080	0.786	4.7	15.9
SevenNet	0.162	0.152	0.108	**1.5**	0.046	0.064	0.582	**2.3**	8.7
Orb	0.159	0.160	0.150	6.8	0.040	**0.057**	**0.365**	3.0	9.6
UMA	**0.090**	**0.097**	0.106	4.5	**0.038**	**0.057**	1.176	2.6	10.0
Fine-tuned on the titanium elastic constants									
MACE	0.024 (0.004)	0.026 (0.005)	0.039 (0.004)	1.6 (0.8)	0.038 (0.001)	**0.056 (0.001)**	1.249 (0.196)	3.1	7.9
MatterSim	0.016 (0.002)	0.017 (0.002)	0.063 (0.009)	6.7 (0.9)	0.054 (0.001)	0.077 (0.003)	**0.324 (0.006)**	4.0	9.4
PET-MAD	0.032 (0.003)	0.032 (0.003)	0.044 (0.026)	**1.5 (0.5)**	0.048 (0.002)	0.065 (0.003)	0.391 (0.006)	3.3	6.4
SevenNet	0.033 (0.001)	0.040 (0.000)	0.067 (0.002)	10.9 (0.2)	0.042 (0.000)	0.063 (0.001)	0.540 (0.097)	5.0	10.9
Orb	0.013 (0.002)	0.013 (0.003)	**0.025 (0.010)**	3.7 (3.0)	0.041 (0.001)	0.060 (0.001)	0.371 (0.037)	**2.4**	**5.9**
UMA	0.011 (0.002)	**0.011 (0.004)**	0.044 (0.004)	4.4 (1.3)	**0.037 (0.001)**	0.057 (0.001)	1.019 (0.030)	3.2	8.1
Fine-tuned on the phase transition									
MACE	0.168 (0.001)	0.151 (0.001)	0.117 (0.001)	0.8 (0.0)	0.052 (0.000)	0.066 (0.000)	1.238 (0.004)	4.0	15.9
MatterSim	0.356 (0.003)	0.311 (0.003)	0.113 (0.003)	1.1 (0.0)	0.053 (0.000)	0.072 (0.001)	1.431 (0.018)	5.2	19.1
PET-MAD	0.523 (0.004)	0.366 (0.005)	0.109 (0.005)	0.8 (0.0)	0.059 (0.000)	0.078 (0.001)	0.898 (0.002)	4.7	17.8
SevenNet	0.170 (0.001)	0.158 (0.001)	0.103 (0.001)	0.5 (0.0)	0.046 (0.000)	0.064 (0.001)	0.603 (0.006)	2.6	11.7
Orb	0.181 (0.002)	0.182 (0.003)	0.144 (0.004)	0.5 (0.0)	0.040 (0.000)	0.058 (0.001)	**0.477 (0.010)**	2.6	10.7
UMA	**0.084 (0.001)**	**0.086 (0.002)**	**0.094 (0.004)**	0.8 (0.1)	**0.039 (0.001)**	**0.057 (0.001)**	1.198 (0.011)	**2.0**	10.0
Fine-tuned on the structure factor									
MACE	0.336 (0.024)	0.329 (0.025)	**0.060 (0.018)**	4.6 (0.8)	0.021 (0.001)	**0.042 (0.001)**	0.804 (0.080)	3.2	9.9
MatterSim	0.360 (0.005)	0.323 (0.006)	0.233 (0.025)	3.9 (2.1)	0.022 (0.001)	**0.042 (0.001)**	1.470 (0.093)	3.9	13.6
PET-MAD	0.179 (0.036)	0.157 (0.043)	0.152 (0.021)	8.3 (1.3)	0.033 (0.002)	0.049 (0.005)	0.563 (0.053)	4.3	11.3
SevenNet	**0.122 (0.010)**	**0.113 (0.009)**	0.167 (0.017)	7.4 (2.5)	0.021 (0.000)	**0.042 (0.000)**	0.354 (0.012)	**2.5**	7.6
Orb	0.163 (0.027)	0.167 (0.031)	0.173 (0.039)	**3.1 (2.0)**	0.023 (0.000)	0.043 (0.002)	**0.314 (0.071)**	2.6	6.9
UMA	0.146 (0.007)	0.143 (0.007)	0.132 (0.007)	9.7 (4.4)	0.022 (0.000)	0.043 (0.000)	0.881 (0.030)	4.4	11.3

Elastic constants are comparatively cheap to measure, while the transition line, the structure factor, and phonon dispersion need many state points or a scattering facility. Fine-tuning on elastic constants alone can reduce errors in these other observables, which is particularly useful when their experimental measurements are unavailable or expensive.

Fine-tuned on the phase transition or the structure factor. Unlike the elastic-constant correction, matching the phase boundary raises the phonon error for all six models (Table 1, Figure 9). While the transition line constrains a free-energy difference between two phases, the loss holds the volume of each phase near the pretrained value, and the fitted slope stays far from the experimental value even on target (Appendix B.1). Fine-tuning on the liquid structure factor extends the correction to the undercooled melt for every model (Table 1, Figure 10), and lowers the phonon error for most models, including MACE and SevenNet, which the elastic fine-tuning leaves unchanged.

Equivariance. Fine-tuning raises the force equivariance error of Orb and PET-MAD, the two models that learn rotational symmetry from data (Table 2). The elastic constants these two predict can therefore shift with the orientation of the cell, by up to 3.1%. The force error of the equivariant models stays at numerical precision. We report the constants in the cell as constructed, without averaging over orientations.

We also tested three representative regularization strategies: they lower different off-target errors, but only as the on-target error rises (Appendix A.5).

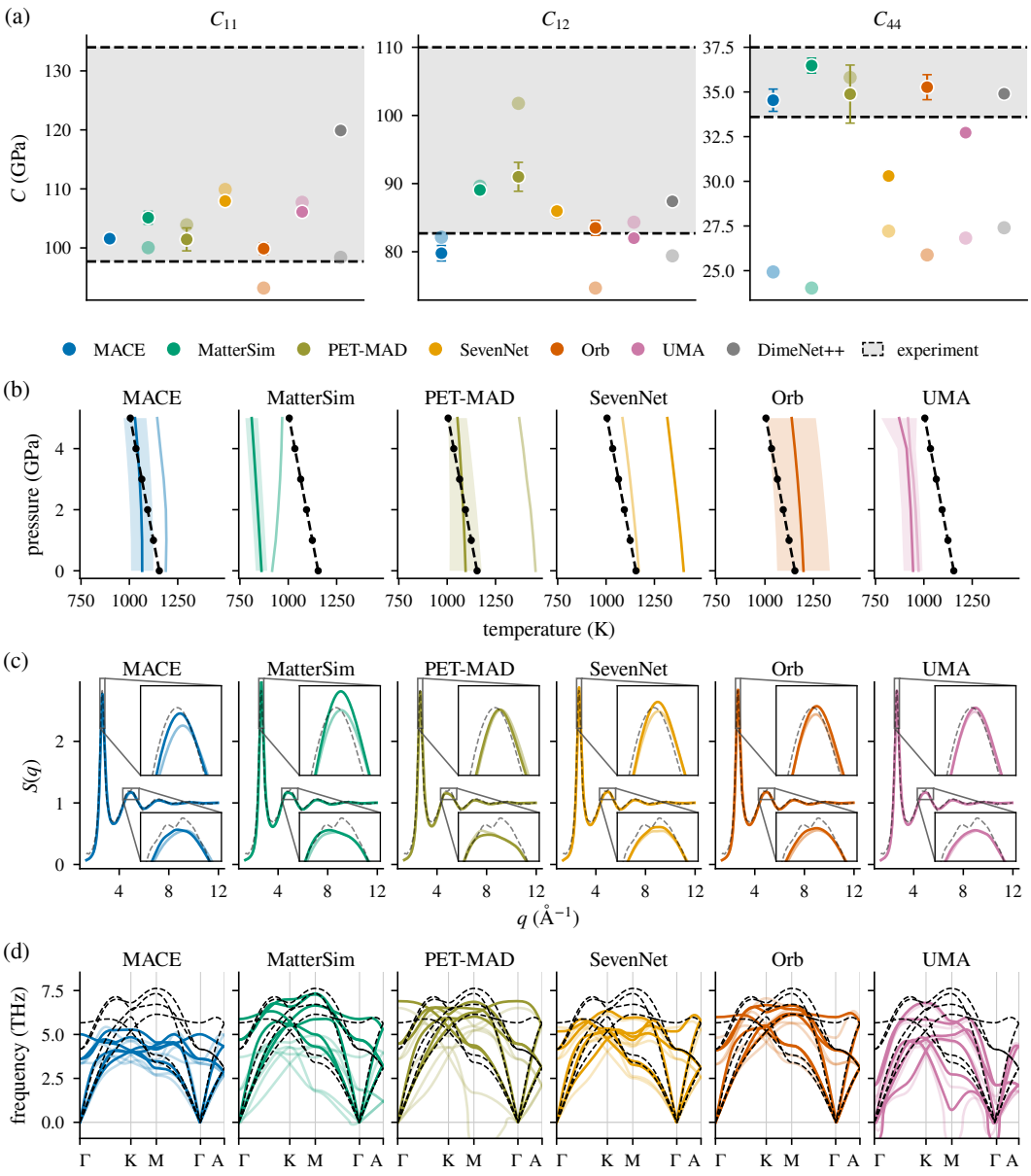


Figure 3: Off-target observables of the models fine-tuned on the titanium elastic constants. Pretrained models are faded and fine-tuned models solid, and panels (b), (c) and (d) are the three measurements of a different observable. (a) Held-out polymorph: bcc elastic constants at 1273 K, the mean over three seeds with the standard deviation as error bar. The shaded band spans three experimental determinations near this temperature (Fisher & Dever, 1970; Ledbetter et al., 2004; Petry et al., 1991), of which Fisher & Dever (1970) scores the column (Appendix B.2), and gray points are the DimeNet++ baseline of Röcken & Zavadlav (2024). (b) The hcp-bcc transition line under pressure, one panel per model, with the mean over the seeds that crossed at a pressure and the standard deviation as a shaded band. Black points are the measured line, and a model with no faded line has no pretrained crossing inside the swept windows. (c) $S(q)$ of the liquid at 1965 K (one seed) against the measurement of Holland-Moritz et al. (2007), drawn as cubic splines, with the first and second peaks magnified in the insets. (d) hcp phonon dispersions (one seed) against the Born–von Kármán fit to the measured frequencies (Stassis et al., 1979; Arch, 1980) as a black dashed line. Axis limits follow the measured frequencies, so an imaginary branch can fall below the frame.

Table 2: Rotational equivariance under fine-tuning on the titanium elastic task, pretrained → fine-tuned under EXP-ONLY. The force error is the mean relative force equivariance error $\|F(Rx) - RF(x)\|/\|F(x)\|$ over 16 random rotations R of cell and positions. The orientation dependence is $(\max_r C_r - \min_r C_r)/|\bar{C}|$ averaged over the five elastic constants, where C_r is the constant in the r-th of six randomly rotated copies of the 323 K cell, each returned to the crystal frame, and $\bar{C}$ its mean over the six. For an exactly equivariant model this column reports the variation the molecular dynamics itself produces. The equivariance column names the architecture's claim, exact by construction or learned from data.

	equivariance	force error (10^{-3})	orientation dependence (%)
MACE	exact	0.019 → 0.025	0.6 → 0.6
MatterSim	exact	0.016 → 0.011	0.9 → 0.8
PET-MAD	learned	24 → 55	1.4 → 3.1
SevenNet	exact	0.0066 → 0.0065	0.7 → 1.0
Orb	learned	3.5 → 59	0.8 → 3.0
UMA	exact	0.013 → 0.012	0.7 → 0.5

3.2 COMPOSITION-DEPENDENT TRANSFER AND FORGETTING IN ELASTIC CONSTANTS

This subsection examines the three elastic tasks, titanium, diamond and MgO (Table 3), whose bonding is metallic, covalent and ionic. The fine-tuning data of a task contain its own composition alone, so a composition changes only through weights shared across elements.

Across compositions, the correction transfers between diamond and MgO which are both cubic, and fine-tuning on hcp titanium changes neither.

Fine-tuned on titanium. The DimeNet++ baseline of Röcken & Zavadlav (2024), which was trained on titanium DFT data, has an SRME of 0.008 on titanium, lower than that of any fine-tuned uMLIP. For the off targets, however, the baseline improves less on bcc titanium than any uMLIP (Table 11). While the baseline starts from a lower bcc error, which the bcc structures in its DFT training data can explain, the uMLIPs end below that baseline, except SevenNet against Ledbetter et al. (2004).

Fine-tuning on titanium leaves the other two compositions no better. On MgO the error rises for all six models at the fitted temperature and for every model except Orb at the held-out one, and on diamond MatterSim alone improves, by less than the measurement disagreement (Table 3).

Fine-tuned on diamond. After fine-tuning on diamond elastic constants, MatterSim, UMA and SevenNet fall within the measurement disagreement on diamond (Table 3, Figure 8a). The DimeNet++ baseline of Thaler & Zavadlav (2021), starting from a classical Stillinger–Weber potential (Stillinger & Weber, 1985) with no DFT data, has an SRME of 0.041 after fine-tuning on diamond elastic constants, and every model except MACE is more accurate, unlike the titanium elastic task, where its DFT data are available. For an off target, the correction transfers to the held-out temperature for all six models, and to MgO and the diamond phonons for most, while the titanium error rises for MACE alone (Table 3). The correction transfers to the other cubic composition far more reliably than to the hexagonal one.

Fine-tuned on MgO. After fine-tuning on MgO elastic constants, MACE, MatterSim and Orb fall within the measurement disagreement on MgO (Table 3, Figure 8b). For an off target, the correction transfers to the held-out temperature, the phonons and diamond for all six models, while the titanium error rises for every model except MACE and SevenNet. A correction transfers between diamond and MgO, and not to or from titanium. Fine-tuning a uMLIP on experiment across compositions therefore needs measurements on more than one of them.

4 RELATED WORK

uMLIP benchmarks. Existing uMLIP benchmarks evaluate released models without fine-tuning. Most measure agreement with DFT for crystal stability (Riebesell et al., 2025), phonons (Loew

Table 3: The three elastic tasks, pretrained and fine-tuned. The first block is the pretrained models and the three below it one block per task, one row per model, in the conventions of Table 1. The diagonal is the composition the block was fitted to, scored on target, and the phonon column is each block's own composition. The pretrained block has no composition of its own and carries all three there, titanium / diamond / MgO, ranked within a composition rather than across them. The mean-rank columns average one composition's columns first, so that every composition counts once. Table 10 carries the changes, and Table 2 the orientation dependence of Orb and PET-MAD.

	titanium (SRME)		diamond (SRME)		MgO (SRME)		phonon (RMSE, THz)	mean rank	
	fitted T	held-out T	fitted T	held-out T	fitted T	held-out T		block	all
Pretrained, no fine-tuning									
MACE	0.157	0.139	0.287	0.280	0.071	0.060	1.209 / 1.96 / 1.28	4.2	13.1
MatterSim	0.387	0.337	0.170	0.166	0.079	0.065	1.328 / 1.03 / 0.63	4.5	14.4
PET-MAD	0.322	0.288	0.145	0.160	0.054	0.067	0.786 / 1.20 / **0.59**	3.5	12.0
SevenNet	0.162	0.152	0.222	0.222	0.080	0.069	0.582 / 2.11 / 0.86	4.7	14.2
Orb	0.159	0.160	**0.068**	**0.059**	**0.053**	0.055	**0.365** / 0.70 / 0.69	**2.0**	7.8
UMA	**0.090**	**0.097**	0.086	0.082	0.068	**0.054**	1.176 / **0.52** / 0.70	**2.0**	7.1
Fine-tuned on the titanium elastic constants									
MACE	0.024 (0.004)	0.026 (0.005)	0.305 (0.002)	0.297 (0.002)	0.074 (0.002)	0.063 (0.002)	1.249 (0.196)	4.9	12.5
MatterSim	0.016 (0.002)	0.017 (0.002)	0.169 (0.000)	0.164 (0.000)	0.081 (0.000)	0.067 (0.000)	**0.324 (0.006)**	3.0	8.7
PET-MAD	0.032 (0.003)	0.032 (0.003)	0.158 (0.015)	0.204 (0.030)	0.065 (0.003)	0.075 (0.002)	0.391 (0.006)	3.9	9.8
SevenNet	0.033 (0.001)	0.040 (0.000)	0.222 (0.000)	0.222 (0.000)	0.082 (0.000)	0.071 (0.000)	0.540 (0.097)	5.1	12.1
Orb	0.013 (0.002)	0.013 (0.003)	0.098 (0.001)	0.090 (0.001)	**0.060 (0.001)**	**0.053 (0.002)**	0.371 (0.037)	**1.8**	5.3
UMA	0.011 (0.002)	**0.011 (0.004)**	**0.086 (0.001)**	**0.082 (0.001)**	0.069 (0.001)	0.056 (0.001)	1.019 (0.030)	2.4	7.1
Fine-tuned on the diamond elastic constants									
MACE	0.173 (0.001)	0.155 (0.001)	0.059 (0.004)	0.047 (0.007)	0.052 (0.000)	0.056 (0.000)	1.05 (0.01)	4.2	8.8
MatterSim	0.387 (0.000)	0.337 (0.000)	0.002 (0.000)	**0.005 (0.000)**	0.080 (0.000)	0.066 (0.000)	0.85 (0.01)	3.5	10.2
PET-MAD	0.308 (0.003)	0.278 (0.003)	0.018 (0.007)	0.052 (0.036)	**0.040 (0.000)**	0.055 (0.001)	1.01 (0.08)	4.2	8.9
SevenNet	0.157 (0.000)	0.147 (0.000)	0.004 (0.001)	0.009 (0.001)	0.082 (0.000)	0.071 (0.000)	1.57 (0.00)	4.2	10.4
Orb	0.157 (0.001)	0.159 (0.001)	0.017 (0.011)	0.026 (0.015)	0.049 (0.000)	0.053 (0.000)	0.91 (0.06)	3.1	6.3
UMA	**0.090 (0.000)**	**0.096 (0.000)**	0.002 (0.000)	0.009 (0.000)	0.064 (0.000)	**0.051 (0.000)**	**0.55 (0.00)**	**1.8**	3.9
Fine-tuned on the MgO elastic constants									
MACE	0.149 (0.000)	0.132 (0.000)	0.276 (0.000)	0.270 (0.000)	0.010 (0.001)	**0.025 (0.001)**	0.56 (0.02)	3.5	8.0
MatterSim	0.394 (0.000)	0.342 (0.000)	0.166 (0.000)	0.162 (0.000)	0.005 (0.004)	0.031 (0.006)	0.57 (0.03)	4.6	10.5
PET-MAD	0.326 (0.001)	0.286 (0.000)	0.139 (0.002)	0.178 (0.002)	0.012 (0.000)	0.032 (0.001)	**0.29 (0.00)**	3.6	8.9
SevenNet	0.159 (0.000)	0.150 (0.000)	0.202 (0.000)	0.201 (0.000)	0.029 (0.001)	0.045 (0.001)	0.35 (0.00)	4.2	8.8
Orb	0.178 (0.000)	0.179 (0.000)	**0.040 (0.001)**	**0.035 (0.000)**	0.010 (0.001)	0.032 (0.001)	0.41 (0.01)	3.4	6.6
UMA	**0.093 (0.000)**	**0.102 (0.000)**	0.057 (0.000)	0.052 (0.000)	0.012 (0.001)	0.026 (0.002)	**0.29 (0.00)**	**1.6**	**3.6**

et al., 2025; Han & Cheng, 2025), thermal transport (Póta et al., 2024), elasticity (Gao & Wang, 2025), or several of these at once (Peng et al., 2025; Wines & Choudhary, 2024). Others test simulation behavior (Chiang et al., 2025; Gawkowski et al., 2026) or compare predicted properties with experiment (Mannan et al., 2025; Stracke et al., 2026). These benchmarks use experimental data only for testing.

Experimental fine-tuning. Most studies of experimental fine-tuning concern system-specific potentials, and fitting such potentials to experimental data predates differentiable simulation (Soper, 1996; Wang et al., 2014). Thaler & Zavadlav (2021) fitted neural potentials to ensemble averages, and later work extended this to elastic constants, transition temperatures, spectra and liquid densities (Röcken & Zavadlav, 2024; Fuchs & Zavadlav, 2026; Han & Yu, 2025; Gong et al., 2025; Gumber et al., 2025; Greener, 2025). Related studies fit classical potentials to phase diagrams and biomolecular ensembles (Jin et al., 2026; Krueger et al., 2026), or keep the pretrained model fixed and learn a correction around it (Matin et al., 2024; Nam et al., 2026). Few of these studies evaluate off-target observables, and those that report no forgetting start from a potential pretrained on the same system (Röcken & Zavadlav, 2024; Fuchs & Zavadlav, 2026; Gumber et al., 2025).

5 Conclusion

On ExFiT-Bench, fine-tuning corrects every model at the fitted conditions, and what changes beyond them depends on the observable and the composition fitted. The correction reaches the same constants at the held-out temperatures, in the other polymorph, and between the two cubic compositions for most models. The pretrained ranking does not carry to the fine-tuned models.

These findings hold within the scope of the benchmark. The titanium elastic task has no second determination of its fitted constants, and the on-target precision claim therefore covers diamond and MgO alone. Every task targets an ensemble-averaged property, which leaves out dynamical properties that require a loss differentiating through the trajectory (Han & Yu, 2025). We fine-tune all weights, and the replay set of the strategies we compare consists of DFT data of the target system. Parameter-efficient fine-tuning (Wang et al., 2025; Grandel et al., 2026) would instead update only a subset of the weights. Replaying each model's own pretraining data (Tompa et al., 2026) preserves that distribution where the target system's DFT data do not, which Appendix A.5 measures on the two models pretrained on MPtrj.

The main open problem is to correct a single uMLIP against experiment across many materials and properties without forgetting. Fine-tuning on one task falls short of that goal: the experimental values of a task constrain the potential energy surface only within the composition and conditions the task samples (Appendix A.5). Regularization ties the undetermined weights to the pretrained model or to DFT data, and the on-target error rises with the strength. Fine-tuning on many tasks at once would instead constrain the weights with experimental data, and the correction should then transfer to more materials and properties. Trajectory regeneration (Section 2) is a second limit, and trust regions and off-policy corrections address that cost in reinforcement learning (Schulman et al., 2015; Munos et al., 2016). ExFiT-Bench provides the single-task reference and the off-target evaluation for such models and methods, whether they change the loss, the data or the reweighting.

AI USE STATEMENT

We used an LLM coding agent, under our direction, for implementation, data cleaning and analysis: the training and evaluation code, the scripts that produce the numbers, tables and figures of this paper from the stored run outputs, and the analysis scripts behind the appendix. It also produced result interpretation, drafting candidate readings of the tables that we then checked against the data. Generative AI did not generate synthetic data sets, develop the theoretical models or conceptual framework, formulate mathematical claims, refine the hypotheses, design the methodology, or assist with proof or translation. The agent additionally supported literature analysis, paper drafting and editing, structure suggestions, figure creation and code editing. We reviewed all AI-assisted work, and the reported numbers come from the stored run outputs. References the agent suggested were checked against the cited papers, and the authors edited and verified the text. We take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We release the code, the settings of each task, the experimental values in machine-readable form with their sources, units and conditions, and the predictions of all three seeds, at `https://anonymous.4open.science/r/exfit-bench`. The code carries a test suite and a conformance gate against each model's reference calculator. A leaderboard carries the fine-tuned models of the body tables and scores new entries by the same protocol. Its address is withheld for review. Appendix A.1 gives the measurement sources and MD settings, and Appendix A.3 the estimator definitions. The code's README documents the submission interface.

REFERENCES

D. Arch. *Lattice dynamics in the α-phase of the IVB transition elements*. PhD thesis, Iowa State University, 1980.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C. Lawrence Zitnick, and Zachary W. Ulissi. Open materials 2024 (OMat24) inorganic materials dataset and models, 2024.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11423–11436, 2022.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, et al. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18):184110, 2025. doi: 10.1063/5.0297006.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021. doi: 10.1021/acscatal.0c04525.

Yuan Chiang, Tobias Kreiman, Christine Zhang, Matthew C. Kuner, Elizabeth Weaver, Ishan Amin, Hyunsoo Park, Yunsung Lim, Jihan Kim, Daryl Chrzan, Aron Walsh, Samuel M. Blau, Mark Asta, and Aditi S. Krishnapriyan. MLIP Arena: advancing fairness and transparency in machine learning interatomic potentials via an open, accessible benchmark platform, 2025.

Maurice de Koning, A. Antonelli, and Sidney Yip. Optimized free-energy evaluation using a single reversible-scaling simulation. *Physical Review Letters*, 83:3973–3977, 1999. doi: 10.1103/PhysRevLett.83.3973.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5:1031–1041, 2023. doi: 10.1038/s42256-023-00716-3.

Bowen Deng, Yunyeong Choi, Peichen Zhong, Janosh Riebesell, Shashwat Anand, Zhuohan Li, KyuJung Jun, Kristin A. Persson, and Gerbrand Ceder. Systematic softening in universal machine learning interatomic potentials. *npj Computational Materials*, 11, 2025. doi: 10.1038/s41524-024-01500-6. arXiv:2405.07105.

A. I. Epishin and D. S. Lisovenko. Comparison of isothermal and adiabatic elasticity characteristics of the single crystal nickel-based superalloy CMSX-4 in the temperature range between room temperature and 1300°c. *Mechanics of Solids*, 58(5):1587–1598, 2023. doi: 10.3103/S0025654423601301.

Dawei Fan, Suyu Fu, Jing Yang, Sergey N. Tkachev, Vitali B. Prakapenka, and Jung-Fu Lin. Elasticity of single-crystal periclase at high pressure and temperature: the effect of iron on the elasticity and seismic parameters of ferropericlase in the lower mantle. *American Mineralogist*, 104(2):262–275, 2019. doi: 10.2138/am-2019-6656.

Henry E. Fischer, Adrian C. Barnes, and Philip S. Salmon. Neutron and x-ray diffraction studies of liquids and glasses. *Reports on Progress in Physics*, 69(1):233–299, 2006. doi: 10.1088/0034-4885/69/1/R05.

E. S. Fisher and D. Dever. The single crystal elastic moduli of beta-titanium and titanium-chromium alloys. In R. I. Jaffee and N. E. Promisel (eds.), *The Science, Technology and Application of Titanium*, pp. 373–381. Pergamon, 1970.

Rodrigo Freitas, Mark Asta, and Maurice de Koning. Nonequilibrium free-energy calculation of solids using LAMMPS. *Computational Materials Science*, 112:333–341, 2016. doi: 10.1016/j.commatsci.2015.10.050.

Daan Frenkel and Anthony J. C. Ladd. New Monte Carlo method to compute the free energy of arbitrary solids. *The Journal of Chemical Physics*, 81:3188–3193, 1984. doi: 10.1063/1.448024.

Paul Fuchs and Julija Zavadlav. Refining machine learning potentials through thermodynamic theory of phase transitions. *npj Computational Materials*, 12(1), 2026. doi: 10.1038/s41524-026-02195-7.

Pengfei Gao and Haidi Wang. Benchmarking universal machine learning interatomic potentials for elastic property prediction, 2025.

Mikołaj J. Gawkowski, Nongnuch Artrith, Silvia Bonfanti, Abhijeet Sadashiv Gangan, Hendrik H. Heenen, Joseph Kioseoglou, Ivor Lončarić, Hemanadhan Myneni, Janosh Riebesell, Mariana Rossi, Matthias Rupp, Jonathan Schmidt, Shubham Sharma, Benjamin X. Shi, Antoni Wadowski, Lukas Hörmann, and Venkat Kapil. Dyna-Mat: end-to-end benchmarking of foundation machine learning interatomic potentials in finite-temperature ensembles, 2026.

Vahe Gharakhanyan, Luis Barroso-Luque, Yi Yang, Muhammed Shuaibi, Kyle Michel, Daniel S. Levine, Noa Marom, Zachary W. Ulissi, Anuroop Sriram, et al. Open molecular crystals 2025 (OMC25) dataset and models, 2025.

Sheng Gong, Yumin Zhang, Zhenliang Mu, Zhichen Pu, Hongyi Wang, Xu Han, Zhiao Yu, Mengyi Chen, Tianze Zheng, Zhi Wang, Lifei Chen, Zhenze Yang, Xiaojie Wu, Shaochen Shi, Weihao Gao, Wen Yan, and Liang Xiang. A predictive machine learning force-field framework for liquid electrolyte development. *Nature Machine Intelligence*, 7:543–552, 2025. doi: 10.1038/s42256-025-01009-7.

Jonas Grandel, Philipp Benner, and Janine George. Parameter-efficient fine-tuning of machine-learning interatomic potentials for phonon and thermal properties, 2026.

Joe G. Greener. Reversible molecular simulation for training classical and machine-learning force fields. *Proceedings of the National Academy of Sciences*, 122(22):e2426058122, 2025. doi: 10.1073/pnas.2426058122.

Shriya Gumber, Lorena Alzate-Vargas, Benjamin T. Nebgen, Arjen van Veelen, Smit Kadvani, Tammie Gibson, and Richard Messerly. Going beyond density functional theory accuracy: leveraging experimental data to refine pre-trained machine learning interatomic potentials, 2025.

Bin Han and Kuang Yu. Refining potential energy surface through dynamical properties via differentiable molecular simulation. *Nature Communications*, 16:816, 2025. doi: 10.1038/s41467-025-56061-z.

Bowen Han and Yongqiang Cheng. Benchmarking universal machine learning interatomic potentials for real-time analysis of inelastic neutron scattering data, 2025.

Dirk Holland-Moritz, Oliver Heinen, Robert Bellissent, and Thomas Schenk. Short-range order of stable and undercooled liquid titanium. *Materials Science and Engineering: A*, 449–451:42–45, 2007. doi: 10.1016/j.msea.2005.12.093.

Jaekyun Hwang, Taehun Lee, Yonghyuk Lee, and Su-Hyun Yoo. Fine-tuning bulk-oriented universal interatomic potentials for surfaces: accuracy, efficiency, and forgetting control, 2025.

Bin Jin, Bin Han, Wei Feng, Kuang Yu, and Shenzhen Xu. Automatic refinement of force fields based on phase diagrams. *Journal of Chemical Theory and Computation*, 22(8):3872–3880, 2026. doi: 10.1021/acs.jctc.5c02032.

Yuna Kato, Jürgen Brillo, Thomas Voigtmann, and Linnea Heitmeier. Simulation data for "Molecular Dynamics simulations of Al-Ti metallic alloy melts using a transferable machine-learning potential". Zenodo dataset, 2026.

Jaesun Kim, Jisu Kim, Jaehoon Kim, Jiho Lee, Yutack Park, Youngho Kang, and Seungwu Han. Data-efficient multifidelity training for high-fidelity machine learning interatomic potentials. *Journal of the American Chemical Society*, 2025a.

Jisu Kim, Jiho Lee, Sangmin Oh, Yutack Park, Seungwoo Hwang, Seungwu Han, Sungwoo Kang, and Youngho Kang. An efficient forgetting-aware fine-tuning framework for pretrained universal machine-learning interatomic potentials. *npj Computational Materials*, 2025b. doi: 10.1038/s41524-025-01895-w. arXiv:2506.15223; volume to confirm at camera-ready.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Lucas P. Kreuzer, Fan Yang, Andreas Meyer, and Noël Jakse. Impact of local structure on melt dynamics in Cu-Ti alloys: Insights from ab initio molecular dynamics simulations. *Physical Review B*, 111:144107, 2025. doi: 10.1103/PhysRevB.111.144107.

Ryan K. Krueger, Megan C. Engel, Ryan Hausen, and Michael P. Brenner. Fitting coarse-grained models to macroscopic experimental data via automatic differentiation. *Proceedings of the National Academy of Sciences*, 123(14):e2508255123, 2026. doi: 10.1073/pnas.2508255123.

Hassel Ledbetter, Hirotsugu Ogi, Satoshi Kai, Sudook Kim, and Masahiko Hirao. Elastic constants of body-centered-cubic titanium monocrystals. *Journal of Applied Physics*, 95(9):4642–4644, 2004.

Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Samuel M. Blau, Brandon M. Wood, et al. The open molecules 2025 (OMol25) dataset, evaluations, and models, 2025.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2830–2839. PMLR, 2018.

Antoine Loew, Dewen Sun, Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Universal machine learning interatomic potentials are ready for phonons. *npj Computational Materials*, 11:178, 2025. doi: 10.1038/s41524-025-01650-1. arXiv:2412.16551.

Spyros Makridakis. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9:527–529, 1993. doi: 10.1016/0169-2070(93)90079-3.

Sajid Mannan, Vaibhav Bihani, Carmelo Gonzales, Kin Long Kelvin Lee, Nitya Nand Gosvami, Sayan Ranu, Santiago Miret, and N. M. Anoop Krishnan. UniFFBench: Evaluating universal machine learning force fields against experimental measurements, 2025.

Sakib Matin, Alice E. A. Allen, Justin Smith, Nicholas Lubbers, Ryan B. Jadrich, Richard Messerly, Benjamin Nebgen, Ying Wai Li, Sergei Tretiak, and Kipton Barros. Machine learning potentials with the iterative Boltzmann inversion: training to experiment. *Journal of Chemical Theory and Computation*, 20(3):1274–1281, 2024. doi: 10.1021/acs.jctc.3c01051.

Arslan Mazitov, Filippo Bigi, Matthias Kellner, Paolo Pegolo, Davide Tisi, Guillaume Fraux, Sergey Pozdnyakov, Philip Loche, and Michele Ceriotti. PET-MAD, a lightweight universal interatomic potential for advanced materials modeling. arXiv:2503.14118, 2025a.

Arslan Mazitov, Sofiia Chorna, Guillaume Fraux, Marnik Bercx, Giovanni Pizzi, Sandip De, and Michele Ceriotti. Massive atomic diversity: a compact universal dataset for atomistic machine learning. *Scientific Data*, 12:1857, 2025b.

H. J. McSkimin, P. Andreatch, and P. Glynn. The elastic stiffness moduli of diamond. *Journal of Applied Physics*, 43(3):985–987, 1972. doi: 10.1063/1.1661318.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 1054–1062, 2016.

Juno Nam, Bowen Deng, Xiaochen Du, Luis Barroso-Luque, Benjamin Kurt Miller, and Rafael Gómez-Bombarelli. Universal thermodynamic interatomic potentials for crystalline materials, 2026.

S. Ozawa, K. Kudo, K. Kuribayashi, Y. Watanabe, and T. Ishikawa. Thermophysical properties of molten titanium measured by an electrostatic levitation method. *Materials Transactions*, 58(12): 1664–1668, 2017. doi: 10.2320/matertrans.L-M2017835.

Yutack Park, Jaesun Kim, Seungwoo Hwang, and Seungwu Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 20(11):4857–4868, 2024.

Anyang Peng, Chun Cai, Mingyu Guo, Duo Zhang, Chengqian Zhang, Wanrun Jiang, Yinan Wang, Antoine Loew, Chengkun Wu, Weinan E, Linfeng Zhang, and Han Wang. LAMBench: a benchmark for large atomistic models, 2025.

W. Petry, A. Heiming, J. Trampenau, M. Alba, C. Herzig, H. R. Schober, and G. Vogl. Phonon dispersion of the bcc phase of group-IV metals. I. bcc titanium. *Physical Review B*, 43(13): 10933–10947, 1991. doi: 10.1103/PhysRevB.43.10933.

Balázs Póta, Paramvir Ahlawat, Gábor Csányi, and Michele Simoncelli. Thermal conductivity predictions with foundation atomistic models, 2024.

Robert R. Reeber, Kai Goessel, and Kai Wang. Thermal expansion and molar volume of MgO, periclase, from 5 to 2900 K. *European Journal of Mineralogy*, 7:1039–1047, 1995.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale. arXiv:2504.06231, 2025.

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7:836–847, 2025. doi: 10.1038/s42256-025-01055-1.

Sebastian Röcken and Julija Zavadlav. Accurate machine learning force fields via experimental and simulation data fusion. *npj Computational Materials*, 10(1):69, 2024.

Subir Sachdev and David R. Nelson. Theory of the structure factor of metallic glasses. *Physical Review Letters*, 53:1947–1950, 1984. doi: 10.1103/PhysRevLett.53.1947.

V. Sánchez-Gil, E. G. Noya, L. Temleitner, and L. Pusztai. Reverse Monte Carlo modeling: The two distinct routes of calculating the experimental structure factor. *Journal of Molecular Liquids*, 207: 211–215, 2015. doi: 10.1016/j.molliq.2015.02.044.

M. J. L. Sangster, G. Peckham, and D. H. Saunderson. Lattice dynamics of magnesium oxide. *Journal of Physics C: Solid State Physics*, 3(5):1026–1036, 1970. doi: 10.1088/0022-3719/3/5/017.

Jonathan Schmidt, Noah Hoffmann, Hai-Chen Wang, Pedro Borlido, Pedro J. M. A. Carriço, Tiago F. T. Cerqueira, Silvana Botti, and Miguel A. L. Marques. Machine-learning-assisted determination of the global zero-temperature phase diagram of materials. *Advanced Materials*, 35(22):2210788, 2023. doi: 10.1002/adma.202210788.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, 2015.

G. Simmons and H. Wang. *Single Crystal Elastic Constants and Calculated Aggregate Properties: A Handbook*. MIT Press, Cambridge, MA, 1971.

A. K. Soper. Empirical potential Monte Carlo simulation of fluid structure. *Chemical Physics*, 202 (2-3):295–306, 1996. doi: 10.1016/0301-0104(95)00357-6.

P. Souvatzis, O. Eriksson, and M. Katsnelson. Anomalous thermal expansion in α-titanium. *Physical Review Letters*, 99(1):015901, 2007. doi: 10.1103/PhysRevLett.99.015901.

D. R. Squire, A. C. Holt, and W. G. Hoover. Isothermal elastic constants for argon: theory and Monte Carlo calculations. *Physica*, 42:388–397, 1969. doi: 10.1016/0031-8914(69)90031-7.

Anuroop Sriram, Sihoon Choi, Xiaohan Yu, Logan M. Brabson, Abhishek Das, Zachary Ulissi, Matt Uyttendaele, Andrew J. Medford, and David S. Sholl. The open DAC 2023 dataset and challenges for sorbent discovery in direct air capture. *ACS Central Science*, 10(5):923–941, 2024. doi: 10.1021/acscentsci.3c01629.

C. Stassis, D. Arch, B. N. Harmon, and N. Wakabayashi. Lattice dynamics of hcp Ti. *Physical Review B*, 19(1):181–188, 1979. doi: 10.1103/PhysRevB.19.181.

Frank H. Stillinger and Thomas A. Weber. Computer simulation of local order in condensed phases of silicon. *Physical Review B*, 31(8):5262–5271, 1985. doi: 10.1103/PhysRevB.31.5262.

Konstantin Stracke, Connor W. Edwards, and Jack D. Evans. Evaluating mechanical property prediction across material classes using molecular dynamics simulations with universal machine-learned interatomic potentials. *Communications Chemistry*, 9(1), 2026. doi: 10.1038/s42004-026-02057-9. arXiv:2511.22885.

Yoshio Sumino, Orson L. Anderson, and Isao Suzuki. Temperature coefficients of elastic constants of single crystal MgO between 80 and 1,300 K. *Physics and Chemistry of Minerals*, 9(1):38–47, 1983. doi: 10.1007/BF00309468.

Stephan Thaler and Julija Zavadlav. Learning neural network potentials from experimental data via Differentiable Trajectory Reweighting. *Nature Communications*, 12(1):6884, 2021.

Atsushi Togo. First-principles phonon calculations with Phonopy and Phono3py. *Journal of the Physical Society of Japan*, 92:012001, 2023. doi: 10.7566/JPSJ.92.012001.

Atsushi Togo and Isao Tanaka. First principles phonon calculations in materials science. *Scripta Materialia*, 108:1–5, 2015.

Tamás Lajos Tompa, Eszter Varga-Umbrich, Ilyes Batatia, Alin M. Elena, Noam Bernstein, and Gábor Csányi. Fine-tuning MLIP foundation models: strategies for accuracy and transferability, 2026.

E. Yu. Tonkov and E. G. Ponyatovsky. *Phase Transformations of Elements under High Pressure*. CRC Press, Boca Raton, FL, 2004.

R. Vogelgesang, A. K. Ramdas, S. Rodriguez, M. Grimsditch, and T. R. Anthony. Brillouin and Raman scattering in natural and isotopically controlled diamond. *Physical Review B*, 54:3989–3999, 1996. doi: 10.1103/PhysRevB.54.3989.

Chen Wang, Siyu Hu, Guangming Tan, and Weile Jia. ELoRA: Low-rank adaptation for equivariant GNNs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63113–63135, 2025.

Lee-Ping Wang, Todd J. Martinez, and Vijay S. Pande. Building force fields: an automatic, systematic, and reproducible approach. *The Journal of Physical Chemistry Letters*, 5(11):1885–1891, 2014. doi: 10.1021/jz500737m.

J. L. Warren, J. L. Yarnell, G. Dolling, and R. A. Cowley. Lattice dynamics of diamond. *Physical Review*, 158(3):805–808, 1967. doi: 10.1103/PhysRev.158.805.

Masakatsu Watanabe and William P. Reinhardt. Direct dynamical calculation of entropy and free energy by adiabatic switching. *Physical Review Letters*, 65:3301–3304, 1990. doi: 10.1103/PhysRevLett.65.3301.

T. Wen, R. Wang, L. Zhu, L. Zhang, H. Wang, D. J. Srolovitz, and Z. Wu. Specialising neural network potentials for accurate properties and application to the mechanical response of titanium. *npj Computational Materials*, 7:206, 2021. doi: 10.1038/s41524-021-00661-y.

Daniel Wines and Kamal Choudhary. CHIPS-FF: evaluating universal machine learning force fields for material properties, 2024.

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, et al. UMA: A family of universal models for atoms. arXiv:2506.23971, 2025.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, et al. MatterSim: A deep learning atomistic model across elements, temperatures and pressures. arXiv:2405.04967, 2024.

Robert W. Zwanzig. High-temperature equation of state by a perturbation method. I. Nonpolar gases. *The Journal of Chemical Physics*, 22(8):1420–1426, 1954. doi: 10.1063/1.1740409.

APPENDIX CONTENTS

A Benchmark details **17**
A.1 Tasks and measurements 17
A.2 Models 18
A.3 The experimental loss 18
A.4 Training 19
A.5 Regularization strategies 21

B Additional results **24**
B.1 On-target detail 24
B.2 Off-target errors 24
B.3 Liquid structure 26

A BENCHMARK DETAILS

A.1 TASKS AND MEASUREMENTS

Table 4 lists the target observable and the measurement source of each task.

Table 4: The five tasks, with the target observable of each at its fitted conditions and the study it reproduces. The MgO elastic and liquid structure factor tasks have no source study, so their source column names the measurement, and the MgO task takes the settings of the diamond elastic task. Table 5 lists the settings.

task	target observable	source
Diamond elastic	3 cubic constants, $\sigma = 0$, 298 K	Thaler & Zavadlav, 2021
MgO elastic	3 cubic constants, $\sigma = 0$, 300 K	this work, Sumino et al., 1983
Ti elastic	5 hexagonal constants, $P = 0$, 4 T	Röcken & Zavadlav, 2024
Ti phase transition	hcp-bcc free energy at coexistence	Fuchs & Zavadlav, 2026
Ti liquid $S(q)$	liquid $S(q)$, 1965 K	this work, Holland-Moritz et al., 2007

Measurements. The diamond elastic task targets the elastic moduli of McSkimin et al. (1972) at 298 K, as in Thaler & Zavadlav (2021). For MgO we use the resonance measurements of Sumino et al. (1983) at the lattice constant of Reeber et al. (1995) and cross-check them against the independent Brillouin measurement of Fan et al. (2019). The MgO task uses the diamond elastic task's settings, at 300 K. The experimental values for titanium are the single-crystal constants that Röcken & Zavadlav (2024) take from the compilation of Simmons & Wang (1971), in cells built from the experimental thermal expansion reported by Souvatzis et al. (2007). The experimental structure factor comes from the neutron diffraction data of Holland-Moritz et al. (2007) in the original reduction deposited by Kato et al. (2026), and we compute it in a cell at the density of Ozawa et al. (2017). The deposited data carry no uncertainty, so we estimate the noise of each curve from the smooth high-q tail. The phase transition task targets the hcp-bcc line that Fuchs & Zavadlav (2026) read from Wen et al. (2021), who took the boundary from the compilation of Tonkov & Ponyatovsky (2004).

Temperatures. The four fitted temperatures of the titanium elastic task, 23, 323, 623 and 923 K, span the hcp stability field and are those of Röcken & Zavadlav (2024). The held-out temperatures, 173, 473, 773 and 973 K, interleave with them, and every table and figure here scores those four. The experimental series runs over more temperatures than these, and Figure 11 draws the whole of it. Not every fine-tuned model is measured over the whole series, and a column averaging some rows over the series and others over four temperatures would compare temperature sets rather than models. The two cubic tasks hold out one temperature each, 100 K on diamond and 600 K on MgO, and the liquid structure factor task the undercooled melt at 1745 K.

Timesteps. Table 5 lists the settings of fine-tuning. We fine-tune at 2.0 fs on diamond and titanium, four times the timestep of the source studies. An NVE run at that timestep conserves energy: with MACE the drift stays under 0.005 meV/atom/ps, and the total energy fluctuates by 0.15% of the thermal energy on diamond and 0.006% on titanium, against the 1% we accept. The reported constants come from a separate evaluation, which runs at the timestep of Thaler & Zavadlav (2021) on diamond and at twice the timestep of Röcken & Zavadlav (2024) on titanium.

The two tasks with no source study. Table 5 lists the three tasks that reproduce one. The MgO elastic task takes the diamond task's settings. The liquid structure factor task runs 1024 atoms in the fixed cell at the measured density, at a 1.0 fs timestep under a Langevin thermostat at 4 ps^{-1}, with 20 ps of equilibration before the first trajectory and 5 ps after each regeneration, 40 ps of production and a 100 fs frame interval, at the N_{eff}/n threshold of the diamond task. These are the settings of the titanium liquid chain that the elastic task already runs, since no published study simulates this measurement.

Table 5: MD settings for fine-tuning on the three tasks that reproduce a source study. Bold marks a deviation from the published protocol. Figure 4 shows the training loss over the 40 epochs on the titanium elastic task.

	Diamond elastic (Thaler & Zavadlav, 2021)		Ti elastic (Röcken & Zavadlav, 2024)		Ti phase transition (Fuchs & Zavadlav, 2026)	
	source	ours	source	ours	source	ours
fitted coexistence lines	—	—	—	—	hcp-bcc, bcc-liquid	**hcp-bcc**
timestep	0.5 fs	**2.0 fs**	0.5 fs	**2.0 fs**	4.0 fs	same
equilibration / production	10 / 60 ps	same	10 / 70 ps	same	20 / 100 ps	same
frame interval	25 fs	**24 fs**	100 fs	same	500 fs	same
system	1000 atoms	**216 atoms**	150 atoms	same	576 atoms	same
thermostat	Langevin 4 ps^{-1}	same	Langevin 4 ps^{-1}	same	Nosé–Hoover	same
N_{eff}/n threshold	0.9	same	1.0	**0.9**	0.9	same
epochs	700	**40**	200	**40**	250	**40**

A.2 Models

The estimators of Appendix A.3 require conservative forces and stress, the analytic derivatives of the predicted energy. We exclude models that predict forces directly.

Table 6: The six uMLIPs, one released checkpoint each. Parameters are the trainable tensors of the checkpoint, counted directly. The count for UMA, a mixture of linear experts, includes all experts. Training sets: MPtrj, the Materials Project relaxation trajectories (Deng et al., 2023); sAlex, the subsampled Alexandria set (Schmidt et al., 2023) released with OMat24 (Barroso-Luque et al., 2024); OMol25 (Levine et al., 2025), OC20 (Chanussot et al., 2021), ODAC23 (Sriram et al., 2024) and OMC25 (Gharakhanyan et al., 2025); MAD, Massive Atomic Diversity (Mazitov et al., 2025b).

model	checkpoint	architecture	params (M)	training set	license
MACE	`mace-mpa-0-medium.model`	equivariant MP (e3nn)	9.1	MPtrj + sAlex	MIT
MatterSim	`MatterSim-v1.0.0-5M.pth`	invariant three-body (M3GNet)	4.5	proprietary (17M structures)	MIT
PET-MAD	`pet-mad v1.0.2`	learned symmetry, point-edge transformer	2.9	MAD	BSD-3-Clause
SevenNet	`7net-mf-ompa`	equivariant MP, multi-fidelity	25.7	OMat24 + MPtrj + sAlex	GPL-3.0
Orb	`orb-v3-conservative-inf-omat`	learned symmetry, GNN	25.5	OMat24	Apache-2.0
UMA	`uma-s-1p2.pt`	equivariant MP, MoLE (64 experts)	291.5	OMat24, OMol25, OC20, ODAC23, OMC25	FAIR Chemistry v1

Conditioning. UMA and SevenNet take a label for the pretraining data set. We use the OMat24 task label of UMA with zero charge and spin, and the SevenNet label for MPtrj + sAlex, the pretraining data of MACE. UMA mixes its linear experts into a single set of weights per composition, and fine-tuning updates every expert.

A.3 The experimental loss

Residual. In the experimental loss of Equation (1), $\langle O \rangle_\theta$ is the ensemble average of Equation (2) and the squared norm is a mean over the components of O. As in the source studies (Thaler & Zavadlav, 2021; Röcken & Zavadlav, 2024; Fuchs & Zavadlav, 2026), we add a stress term of the same form. In the cell built from the measured lattice constants or density, it holds every component of the virial stress at zero on the diamond and MgO tasks and the scalar pressure at zero on the titanium elastic and liquid structure factor tasks. For the phase transition task it holds the pressure at its coexistence value.

Elastic constants. For the elastic tasks, O is the stress-fluctuation estimator of the elastic constants (Squire et al., 1969; Thaler & Zavadlav, 2021),

$$C_{\alpha\beta} = \left\langle \frac{\partial \sigma_\alpha}{\partial \epsilon_\beta} \right\rangle_\theta - \frac{V}{k_\mathrm{B}T} \operatorname{cov}_\theta (\sigma_\alpha, \sigma_\beta) + \frac{N k_\mathrm{B} T}{V} \kappa_{\alpha\beta}, \qquad \sigma_\alpha = \frac{1}{V} \frac{\partial U_\theta}{\partial \epsilon_\alpha}, \tag{3}$$

where $\alpha, \beta \in \{1, \ldots, 6\}$ index the Voigt components, ϵ_α is a homogeneous strain of the cell, and σ_α is the instantaneous stress from the potential energy. Here V is the cell volume, N the number of atoms, and κ a diagonal matrix with entries 2 for normal and 1 for shear components. The average and the covariance $\operatorname{cov}_\theta$ use the weights of Equation (2), and we evaluate the first term, the Born term, by central finite differences in strain.

Adiabatic correction. The estimator gives isothermal constants, while the measurements are adiabatic (Epishin & Lisovenko, 2023), and we fit the adiabatic values as the source studies did. We bound the difference from above with the experimental thermal expansion and the Dulong–Petit heat capacity, and the bound is 1.6 GPa averaged over the five titanium constants at 923 K. As a relative error this is 0.008 over the four fitted temperatures, comparable to the fine-tuned on-target SRME on titanium (Table 1).

Free-energy difference. For the phase transition task, O is the Gibbs free-energy difference between the hcp and bcc phases at six state points on the experimental coexistence line, so $O_\mathrm{exp} = 0$. We follow Fuchs & Zavadlav (2026), who adjust ΔG through the Helmholtz free energy: the Gibbs and Helmholtz changes are equal when the fine-tuned model predicts the same molar volume as the pretrained one, and the stress term holds that volume. In the fixed cells of the constant-volume (NVT) trajectories,

$$\Delta G(\theta) = \Delta G(\theta_0) + \Delta F^\mathrm{hcp}(\theta) - \Delta F^\mathrm{bcc}(\theta), \tag{4}$$

with $\Delta G(\theta_0)$ the pretrained difference and $\Delta F^p(\theta)$ the free-energy perturbation estimate (Zwanzig, 1954) for phase p, accumulated across trajectory regenerations. As in that study, each phase is fitted with the other held at its latest prediction, ΔF^hcp to $\Delta F^\mathrm{bcc} - \Delta G(\theta_0)$ and ΔF^bcc to $\Delta F^\mathrm{hcp} + \Delta G(\theta_0)$, so that Equation (4) gives $\Delta G = 0$.

Thermodynamic integration. We compute ΔG by Frenkel–Ladd thermodynamic integration (Frenkel & Ladd, 1984), with the spring constant of Freitas et al. (2016) and the switching of Watanabe & Reinhardt (1990), and extend it in temperature by reversible scaling (de Koning et al., 1999). Runs with independent seeds reproduce ΔG within 0.1–0.7 meV/atom, comparable to the differences between the fine-tuned models in Table 1. Equation (4) rests on the two models holding the same cell at the same pressure. The residual pressure of either phase stays at or below 0.06 GPa before and after fine-tuning.

Structure factor. For the liquid structure factor task, O is the static structure factor. We compute it per frame on the reciprocal lattice of the periodic cell, which requires no cutoff in r,

$$S(\mathbf{k}; x) = \frac{1}{N} \left| \sum_{j=1}^{N} \exp\left(i\, \mathbf{k} \cdot \mathbf{r}_j\right) \right|^2, \qquad \mathbf{k} = 2\pi \left(\frac{n_x}{L_x}, \frac{n_y}{L_y}, \frac{n_z}{L_z} \right), \tag{5}$$

with $\mathbf{r}_j$ the atomic positions, L_x, L_y, L_z the cell edges and n_x, n_y, n_z integers. Since the liquid is isotropic, we average S over the vectors whose magnitude $q = |\mathbf{k}|$ falls in each 0.10 Å^{-1} bin from 2.30 to 12.0 Å^{-1}, and the measurement over the same bins. The lower edge is where $S(q)$ first rises to one, since below it the four measured temperatures of Holland-Moritz et al. (2007) differ by up to 0.15 in S. We divide the residual of Equation (1) by the estimated noise of the measured curve. This is the weighting reverse Monte Carlo gives a structure-factor residual (Sánchez-Gil et al., 2015). Fine-tuning moves the first peak by less than one bin, so Section 3.1 reads its position from the parabola through the highest bin and its neighbors, on the measurement as on the models.

A.4 Training

We train each model and task at one constant learning rate, shared by its three seeds (Table 7). We sweep 1×10^{-4}, 3×10^{-5} and 1×10^{-5}, adding 1×10^{-6} for the phase transition task, and take the

rate whose experimental loss is lowest, averaged over the final five epochs of the 40-epoch budget, which smooths an oscillating loss.

Table 7: The learning rate of each model and task, selected by the sweep described above. All other tables report models trained at this rate.

	Diamond elastic	MgO elastic	Ti elastic	Ti phase transition	Ti liquid $S(q)$
MACE	3×10^{-5}	3×10^{-5}	1×10^{-4}	1×10^{-5}	1×10^{-4}
MatterSim	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	3×10^{-5}
PET-MAD	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-6}	3×10^{-5}
SevenNet	3×10^{-5}	3×10^{-5}	3×10^{-5}	1×10^{-5}	1×10^{-4}
Orb	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-6}	3×10^{-5}
UMA	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	3×10^{-5}

Optimization. We optimize with Adam (Kingma & Ba, 2015) at $\beta = (0.1, 0.4)$ and a constant rate. The short moment memory follows the shape of the experimental loss: an epoch contributes one update per state point, four on the titanium elastic task and one on each cubic and liquid task, while the phase transition batches its six coexistence points three to an update and so steps two times. A run of 40 epochs takes tens of updates rather than thousands, and averages over hundreds of past gradients would reach back beyond the trajectory they were taken on. There is no minibatch on the experimental branch, since the loss is an average over the reweighted trajectory rather than over a sampled batch. The replay branch of the three replay strategies of Appendix A.5 carries its own Adam state at $\beta = (0.9, 0.999)$ and 1.0×10^{-4}, on minibatches of five configurations. Every run, training or evaluation, uses one GPU, most often an NVIDIA H200.

Figure 4 shows the training loss and the number of trajectory regenerations under EXP-ONLY.

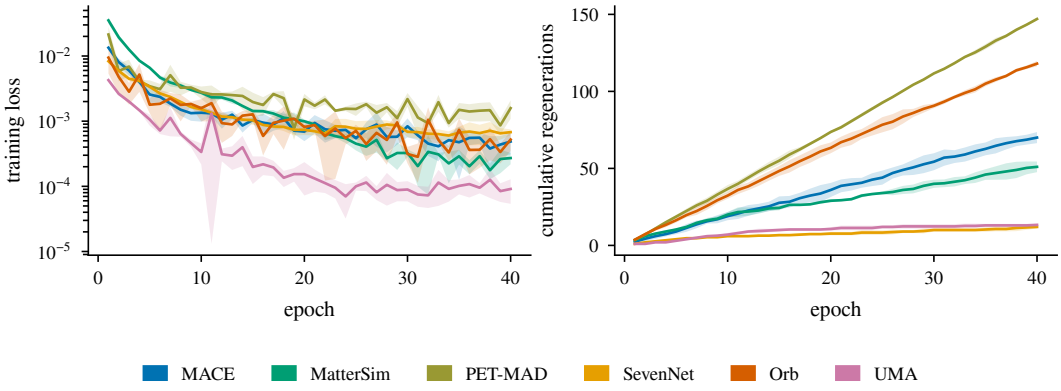


Figure 4: Training on the titanium elastic task under EXP-ONLY: the training loss and the cumulative count of trajectory regenerations, per epoch. Each line is the mean over three seeds at the selected learning rate, with the standard deviation as a shaded band and one color per model. A regeneration reruns MD when the effective sample size falls below its threshold (Section 2).

Compute. The cost of a run does not follow the cost of generating a trajectory (Table 8). PET-MAD generates one faster than any other model and still has the second longest run, because a regeneration reruns the whole trajectory and PET-MAD drops below the effective sample size threshold on most updates. Widening the region over which the reweighting estimate holds would therefore save more than a faster model would, for every model except UMA, which spends more on updates than on molecular dynamics. The regeneration rate spreads by 12 times between the cheapest and the most expensive model here, the widest of the five tasks, and by 2 times on the structure factor, where most models regenerate on nearly every update.

Table 8: The wall clock of a 40-epoch run on the titanium elastic task, measured on an idle NVIDIA H200. The trajectory column is the cost of generating one trajectory, the mean over measurement windows with the standard deviation in parentheses, and the regeneration column the number a run generates, the mean over three seeds. The MD column is their product, and the update column the gradient passes together with the staleness check before each. The MD and update hours sum to the total, up to rounding.

	trajectory (min)	regenerations	MD (h)	update (h)	total (h)
MACE	36.8 (0.5)	70	42.9	1.7	44.6
MatterSim	33.8 (0.8)	51	28.7	12.1	40.8
PET-MAD	24.3 (0.3)	147	59.6	1.7	61.2
SevenNet	43.3 (0.9)	12	8.7	2.5	11.1
Orb	32.8 (0.0)	118	64.5	5.4	69.9
UMA	50.7 (0.1)	13	11.3	16.3	27.6

A.5 Regularization strategies

Regularization. The experimental loss leaves most directions in weight space undetermined (Section 1), and changes along them can cause forgetting off target, the same failure weight regularization (Li et al., 2018; Kirkpatrick et al., 2017) and replay of pretraining data or pseudo-labels (Tompa et al., 2026; Kim et al., 2025b; Hwang et al., 2025) address when fine-tuning on DFT labels. We therefore add a regularization term that acts as a prior on these directions:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{exp}}(\theta) + \lambda R(\theta), \tag{6}$$

where $\mathcal{L}_{\text{exp}}$ is the experimental loss of Equation (1), R the added term and λ its strength. The replay strategies alternate an experimental update with replay updates carried by a second optimizer, and the weight penalty enters as a decoupled decay toward θ_0. In both, λ is a replay budget or a decay coefficient rather than a prefactor. On the titanium elastic task, where the source study's DFT data are public, we compare Exp-Only with four regularization strategies: three based on replay and one a weight penalty.

Replay. Replay penalizes the errors in energy E, forces F and virial W on the replay set $\mathcal{D}_{\text{replay}}$,

$$R(\theta) = \sum_{x \in \mathcal{D}_{\text{replay}}} \Big[w_E \big(E_\theta(x) - E_{\text{ref}}(x)\big)^2 + w_F \big\|F_\theta(x) - F_{\text{ref}}(x)\big\|^2 + w_W \big\|W_\theta(x) - W_{\text{ref}}(x)\big\|^2 \Big], \tag{7}$$

with replay weights w_E, w_F and w_W. Replaying pretraining data is a standard remedy for forgetting when fine-tuning on DFT data (Tompa et al., 2026; Kim et al., 2025b), but these data differ across the six models and are not public for MatterSim (Table 6). We therefore take $\mathcal{D}_{\text{replay}}$ from the DFT data of Wen et al. (2021), in the curated form that Röcken & Zavadlav (2024) released, so that all six models are compared on the same replay set. The three replay strategies differ in that set and in the reference labels. DFT-Replay replays it under the DFT labels, following that study, and PL-Replay under the pseudo-labels of the pretrained model with weights θ_0. Pretrain-Replay replays the corpus the model was pretrained on instead, under that corpus's DFT labels and on a set of the same size; a model carries that corpus only where it is public, which is MACE and SevenNet here, both pretrained on MPtrj (Table 6).

Weight penalty. L2-SP penalizes the distance to the pretrained weights (Li et al., 2018),

$$R(\theta) = \tfrac{1}{2}\|\theta - \theta_0\|^2, \tag{8}$$

which acts on the weights alone and holds the directions that the experimental loss leaves undetermined at their pretrained values. The penalty therefore needs no replay set, and applies to a task with no reference data.

Strengths and schedule. DFT-Replay and PL-Replay replay the set of Röcken & Zavadlav (2024) and Pretrain-Replay the pretraining corpus, all three with the replay weights w_E, w_F and w_W of Equation (7) taken from that study. PL-Replay sets $E_{\text{ref}} = E_{\theta_0}$ and likewise for

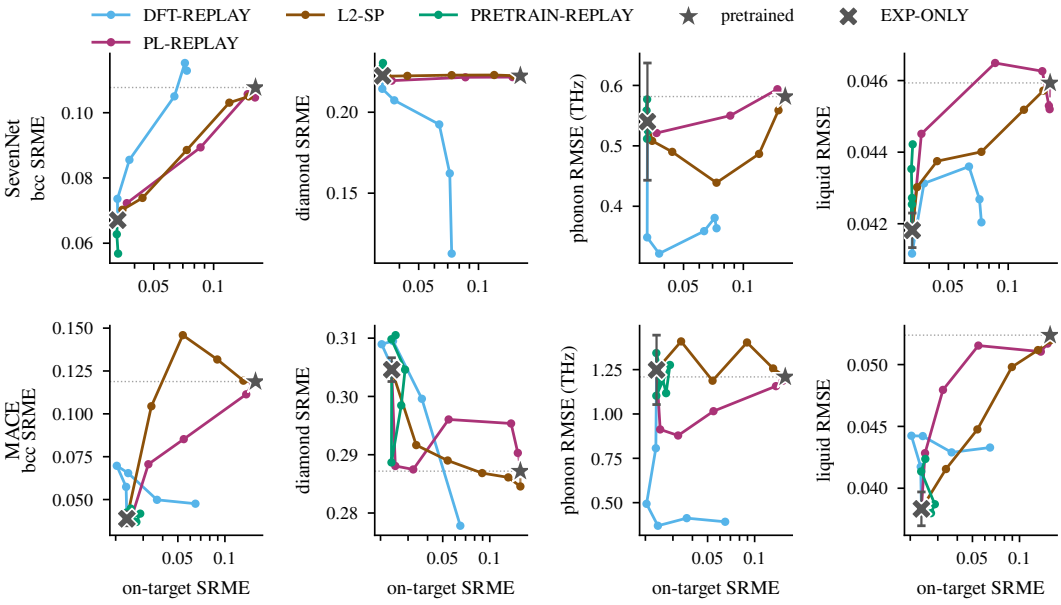


Figure 5: Regularization trade-off on the titanium elastic task. One row per model and one column per off-target property, plotted against the on-target SRME (logarithmic axis). The four properties are scored against experimental measurements, which no strategy replays, and the vertical scale is a panel's own. The gray star is the pretrained model and the gray cross EXP-ONLY, the zero-strength limit of the strategies, drawn with the standard deviation as error bar. Colored lines trace the strategies from that cross, weak to strong, and a point is the mean over three seeds. A strategy appears on a panel where its points are measured. A dotted horizontal line marks the pretrained value. A dashed segment joins L2-SP and PL-REPLAY to the star, their infinite-strength limit.

forces and virial. The three optimize Equation (6) alike, by alternating an experimental epoch with a replay epoch under a separate optimizer state, capped at 50 minibatches of five configurations so the replay updates do not dominate the four experimental ones, and the replay step count and learning rate therefore set λ, not an explicit prefactor. We apply L2-SP as decoupled weight decay, $\theta \leftarrow \theta - \eta\lambda\,(\theta - \theta_0)$ after each update with η the learning rate, and set its strength, shared across models, to $\lambda = 1111$, the largest tested value at which the experimental loss still decreases while the distance to θ_0 stops growing. On SevenNet and MACE, Figure 5 sweeps five strengths for each strategy from EXP-ONLY at zero strength.

Trade-off. On a strength sweep of SevenNet, stronger regularization raises the on-target error, and the strategies lower different off-target errors (Figure 5). The strengths of the strategies are not commensurable, and we read the trade-off at equal on-target error along the sweep. At a matched on-target error, DFT-REPLAY gives the lowest phonon and diamond errors but the highest bcc error, and at the two largest strengths this bcc error rises to or above the pretrained value. PL-REPLAY and L2-SP leave the diamond error at the pretrained value across the whole sweep. Which property a strategy costs is not the same on the two models: on MACE the bcc error rises furthest under L2-SP instead. PRETRAIN-REPLAY moves little from EXP-ONLY on these four properties, and Figure 6 reads it on the distribution it replays. The strategies therefore preserve different properties, and retention is evaluated against the property a study needs.

Replay source. PRETRAIN-REPLAY and DFT-REPLAY replay different distributions under the same DFT labels, which separates what each one preserves.

On MACE and SevenNet, PRETRAIN-REPLAY holds the force error on the pretraining distribution at its pretrained value at every strength, and the on-target error stays within one standard deviation of the EXP-ONLY value (Figure 6). DFT-REPLAY raises that force error as the strength grows, and at its weakest strengths it brings the phonon error of SevenNet close to that of the DimeNet++ baseline,

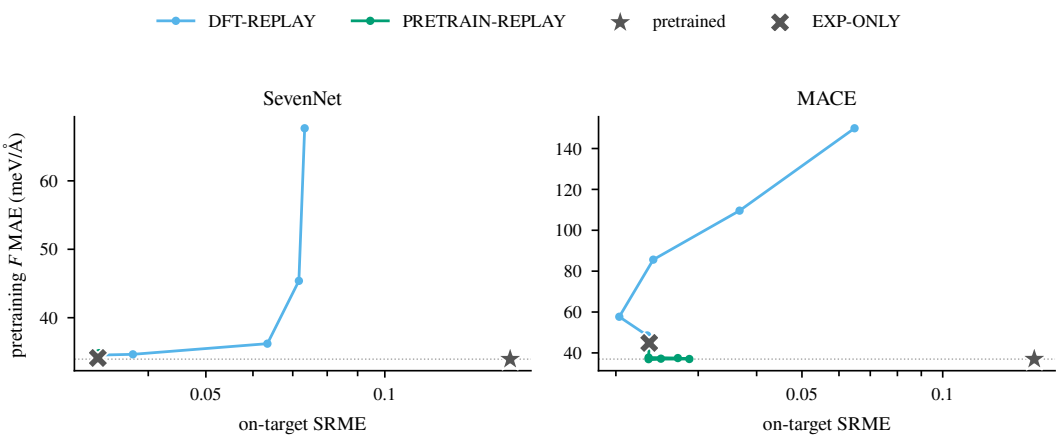


Figure 6: The two replay sources on the pretraining distribution, one panel per model. The force error is measured on 3994 structures of the corpus that model was pretrained on, MPtrj for both, against the on-target SRME (logarithmic axis). DFT-REPLAY replays the titanium DFT data and PRETRAIN-REPLAY the pretraining distribution, on the same strength axis. The gray star is the pretrained model and the gray cross EXP-ONLY, the zero-strength limit of both, and a dotted horizontal line marks the pretrained value. A point is the mean over three seeds. Each panel carries its own corpus, so a curve is read against the others in its panel.

which was trained on the same DFT data. The replay set therefore decides which distribution a strategy preserves.

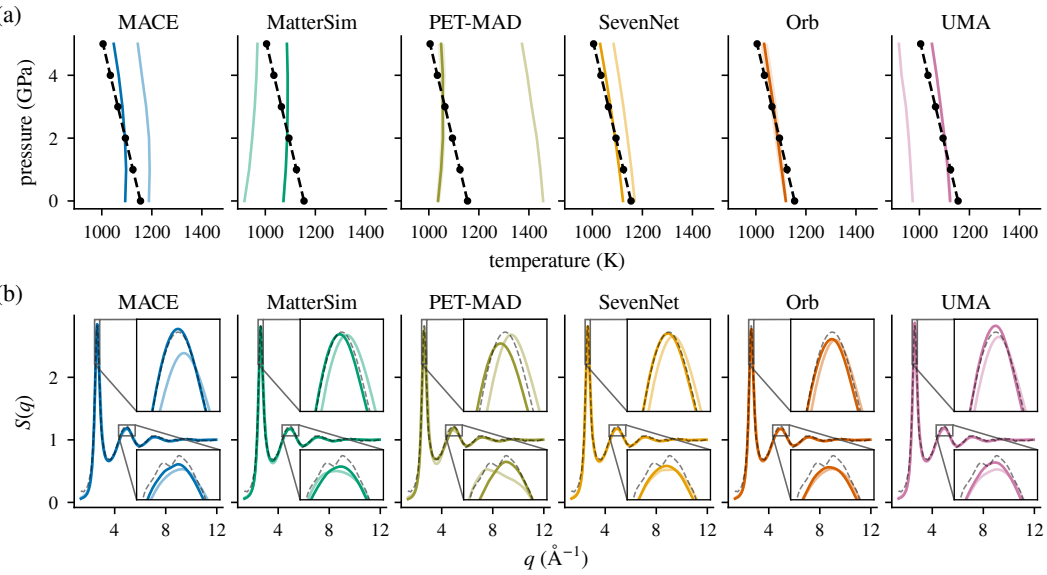


Figure 7: On-target accuracy on the phase transition and the liquid structure factor tasks. Pretrained models are faded and fine-tuned models solid. (a) Phase transition task: the hcp-bcc transition line under pressure, drawn as in Figure 3b. Where a panel carries no faded line, the pretrained transition falls below the temperature range of the reversible-scaling sweeps at every pressure. (b) Liquid structure factor task: $S(q)$ at 1965 K (one seed) against the measurement of Holland-Moritz et al. (2007), drawn as in Figure 3c.

B Additional results

B.1 On-target detail

Elastic constants. Every pretrained model underestimates C_{44} on diamond and MgO, and every model but UMA does on titanium (Figure 8).

Phase transition. The transition lines shift toward the measurement, but the slope of the line, $\mathrm{d}T/\mathrm{d}P$, stays far from the measured -30 K/GPa (Table 9, Figure 7a). Except for PET-MAD, the slopes change by at most 8 K/GPa. Fuchs & Zavadlav (2026) trace this small change to a pressure term that holds the volume of each phase near the pretrained value, and our loss carries the same term (Appendix A.3). By the Clausius–Clapeyron relation, the slope is the ratio of the volume and entropy differences between the phases. Adding the volumes to the loss could correct the slope, provided the entropy difference is already right. For PET-MAD, by contrast, the slope moves away from the experimental value, from -17 to 1 K/GPa.

Liquid structure factor. The RMSE of $S(q)$ falls as the first peak moves toward the experimental position at lower q and the second peak rises (Figure 7b, and Appendix B.3 for the two peaks under both fits).

B.2 Off-target errors

The phonon error is the RMSE over the measured titanium frequencies (Stassis et al., 1979; Arch, 1980), each scored against the nearest model branch, since the measurements carry no branch labels. The Born–von Kármán model of these frequencies (Arch, 1980) is a least-squares fit to them, and its residual on the same points, 0.13 THz, is the floor under the column. We compute the dispersions with Phonopy (Togo & Tanaka, 2015; Togo, 2023) in the supercell and displacement of Röcken & Zavadlav (2024), displacing in both directions and imposing the acoustic sum rule, at each model's zero-pressure cell.

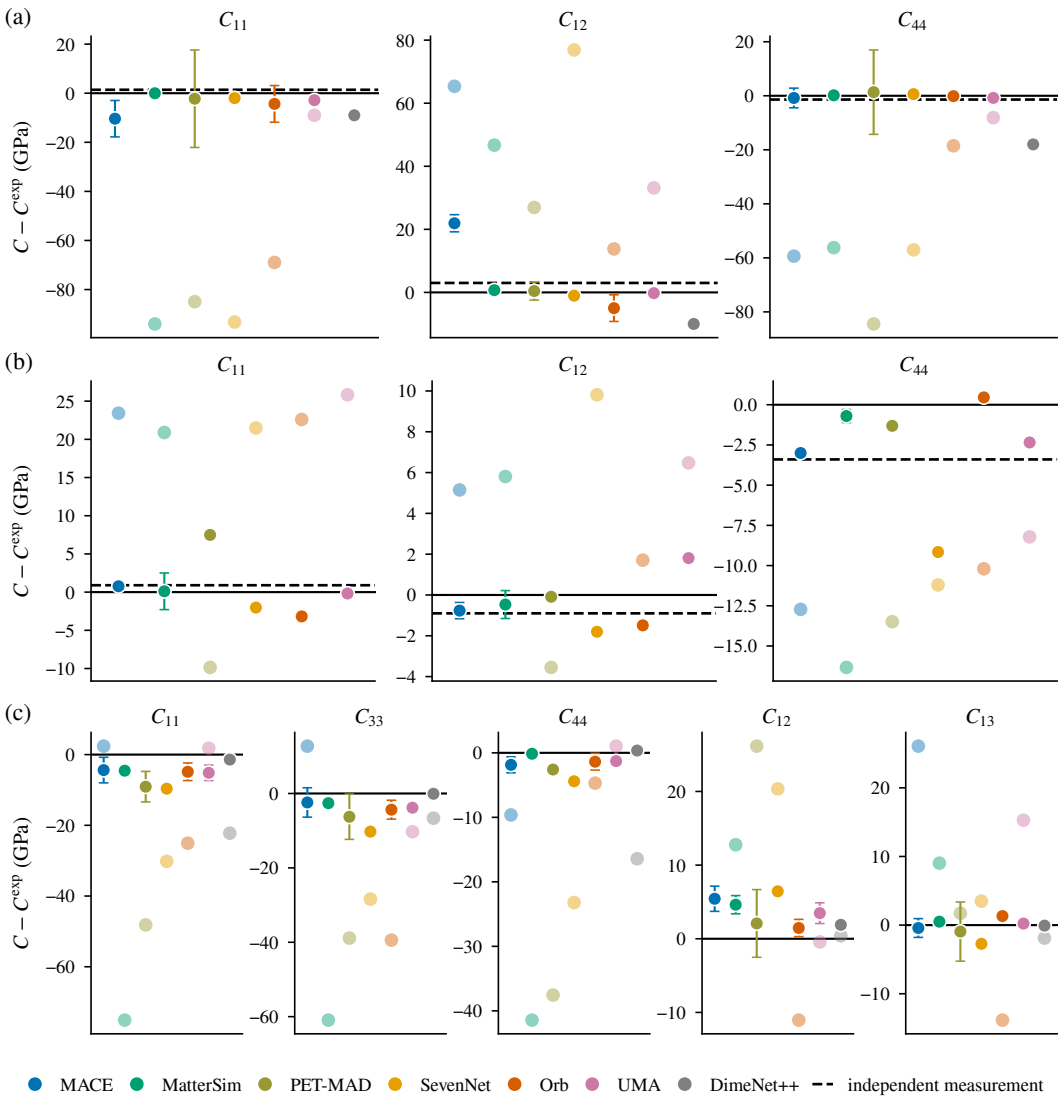


Figure 8: On-target signed error per constant (model − experiment). (a) Diamond at 298 K (McSkimin et al., 1972). (b) MgO at 300 K (Sumino et al., 1983). (c) Titanium hcp at the first fitted temperature, with the full series in Figure 11. Pretrained models are faded and fine-tuned models solid, with the mean over three seeds and the standard deviation as error bar. Gray points are the DimeNet++ baselines of Thaler & Zavadlav (2021) on diamond and of Röcken & Zavadlav (2024) on titanium, and a black dashed line marks an independent measurement, by Vogelgesang et al. (1996) on diamond and Fan et al. (2019) on MgO. A pretrained error more than twice as large as any other in its panel is not drawn (MACE on diamond C_{11}, −310 GPa; MACE on titanium C_{12}, 56 GPa).

Held-out polymorph. Three determinations of the bcc elastic constants lie near 1273 K, and we score the column against the ultrasonic measurement of Fisher & Dever (1970), extrapolated from bcc-stabilized titanium-chromium monocrystals. Ledbetter et al. (2004), who invert polycrystal resonance through an assumed anisotropy, agree to 3.9 GPa on all three constants, and the SRME between the two determinations is 0.050, the reference floor for this column. Petry et al. (1991) read their constants from neutron phonon-dispersion slopes, and we leave them out of the scoring.

Against Petry et al. (1991), the third determination, fine-tuning lowers the error of every model but PET-MAD.

Table 9: The titanium tasks, pretrained → fine-tuned: the entries of Table 1 with the pretrained value beside each, so that a cell carries the starting point of its own change. Blue marks a lower and red a higher mean, darker for a larger change within each column, and the diagonal is gray and untinted, since a change scored on target does not share a scale with the transfer beside it. The bold and underline marks of Table 1 fall on the fine-tuned value, and a mean-rank cell reads as the pretrained model's rank among the six pretrained models → the row's rank among the rows of its block, and the row's rank among all rows of the table.

	Ti elastic (SRME)			ΔG (MAE, meV/atom)	$S(q)$ (RMSE)		phonon (RMSE, THz)	mean rank	
	fitted T	held-out T	bcc		fitted T	held-out T		block	all
Fine-tuned on the titanium elastic constants									
MACE	0.157 → 0.024 (0.004)	0.139 → 0.026 (0.005)	0.119 → 0.039 (0.004)	2.9 → 1.6 (0.8)	0.052 → 0.038 (0.001)	0.067 → **0.056 (0.001)**	1.209 → 1.249 (0.196)	3.9 → 3.1	7.9
MatterSim	0.387 → 0.016 (0.002)	0.337 → 0.017 (0.002)	0.132 → 0.063 (0.009)	2.6 → 6.7 (0.9)	0.051 → 0.054 (0.001)	0.069 → 0.077 (0.003)	1.328 → **0.324 (0.006)**	4.5 → 4.0	9.4
PET-MAD	0.322 → 0.032 (0.003)	0.288 → 0.032 (0.003)	0.097 → 0.044 (0.026)	9.8 → **1.5 (0.5)**	0.060 → 0.048 (0.002)	0.080 → 0.065 (0.003)	0.786 → 0.391 (0.006)	4.5 → 3.3	6.4
SevenNet	0.162 → 0.033 (0.001)	0.152 → 0.040 (0.000)	0.108 → 0.067 (0.002)	1.5 → 10.9 (0.2)	0.046 → 0.042 (0.000)	0.064 → 0.063 (0.001)	0.582 → 0.540 (0.097)	2.2 → 5.0	10.9
Orb	0.159 → 0.013 (0.002)	0.160 → 0.013 (0.003)	0.150 → **0.025 (0.010)**	6.8 → 3.7 (3.0)	0.040 → 0.041 (0.001)	0.057 → 0.060 (0.001)	0.365 → 0.371 (0.037)	3.2 → **2.4**	**5.9**
UMA	0.090 → 0.011 (0.002)	0.097 → **0.011 (0.004)**	0.106 → 0.044 (0.004)	4.5 → 4.4 (1.3)	0.038 → **0.037 (0.001)**	0.057 → 0.057 (0.001)	1.176 → 1.019 (0.030)	2.7 → 3.2	8.1
Fine-tuned on the phase transition									
MACE	0.157 → 0.168 (0.001)	0.139 → 0.151 (0.001)	0.119 → 0.117 (0.001)	2.9 → 0.8 (0.0)	0.052 → 0.052 (0.000)	0.067 → 0.066 (0.000)	1.209 → 1.238 (0.004)	4.1 → 4.0	15.9
MatterSim	0.387 → 0.356 (0.003)	0.337 → 0.311 (0.003)	0.132 → 0.113 (0.003)	2.6 → 1.1 (0.0)	0.051 → 0.053 (0.000)	0.069 → 0.072 (0.001)	1.328 → 1.431 (0.018)	5.4 → 5.2	19.1
PET-MAD	0.322 → 0.523 (0.004)	0.288 → 0.366 (0.005)	0.097 → 0.109 (0.005)	9.8 → 0.8 (0.0)	0.060 → 0.059 (0.000)	0.080 → 0.078 (0.001)	0.786 → 0.898 (0.002)	4.2 → 4.7	17.8
SevenNet	0.162 → 0.170 (0.001)	0.152 → 0.158 (0.001)	0.108 → 0.103 (0.001)	1.5 → 0.5 (0.0)	0.046 → 0.046 (0.000)	0.064 → 0.064 (0.001)	0.582 → 0.603 (0.006)	2.8 → 2.6	11.7
Orb	0.159 → 0.181 (0.002)	0.160 → 0.182 (0.003)	0.150 → 0.144 (0.004)	6.8 → 0.5 (0.0)	0.040 → 0.040 (0.000)	0.057 → 0.058 (0.001)	0.365 → **0.477 (0.010)**	2.4 → 2.6	10.7
UMA	0.090 → **0.084 (0.001)**	0.097 → **0.086 (0.002)**	0.106 → **0.094 (0.004)**	4.5 → 0.8 (0.1)	0.038 → **0.039 (0.001)**	0.057 → **0.057 (0.001)**	1.176 → 1.198 (0.011)	2.2 → **2.0**	10.0
Fine-tuned on the structure factor									
MACE	0.157 → 0.336 (0.024)	0.139 → 0.329 (0.025)	0.119 → **0.060 (0.018)**	2.9 → 4.6 (0.8)	0.052 → 0.021 (0.001)	0.067 → **0.042 (0.001)**	1.209 → 0.804 (0.080)	3.7 → 3.2	9.9
MatterSim	0.387 → 0.360 (0.005)	0.337 → 0.323 (0.006)	0.132 → 0.233 (0.025)	2.6 → 3.9 (2.1)	0.051 → 0.022 (0.001)	0.069 → **0.042 (0.001)**	1.328 → 1.470 (0.093)	4.7 → 3.9	13.6
PET-MAD	0.322 → 0.179 (0.036)	0.288 → 0.157 (0.043)	0.097 → 0.152 (0.021)	9.8 → 8.3 (1.3)	0.060 → 0.033 (0.002)	0.080 → 0.049 (0.005)	0.786 → 0.563 (0.053)	4.7 → 4.3	11.3
SevenNet	0.162 → **0.122 (0.010)**	0.152 → **0.113 (0.009)**	0.108 → 0.167 (0.017)	1.5 → 7.4 (2.5)	0.046 → 0.021 (0.000)	0.064 → **0.042 (0.000)**	0.582 → 0.354 (0.012)	2.3 → **2.5**	7.6
Orb	0.159 → 0.163 (0.027)	0.160 → 0.167 (0.031)	0.150 → 0.173 (0.039)	6.8 → **3.1 (2.0)**	0.040 → 0.023 (0.000)	0.057 → 0.043 (0.002)	0.365 → **0.314 (0.071)**	3.0 → 2.6	6.9
UMA	0.090 → 0.146 (0.007)	0.097 → 0.143 (0.007)	0.106 → 0.132 (0.007)	4.5 → 9.7 (4.4)	0.038 → 0.022 (0.000)	0.057 → 0.043 (0.000)	1.176 → 0.881 (0.030)	2.7 → 4.4	11.3

Table 10: The three elastic tasks, pretrained → fine-tuned: the entries of Table 3 with the pretrained value beside each, in the conventions of Table 9, marks and mean-rank columns included. The diagonal is the composition the block was fitted to, scored on target, gray and untinted.

	titanium (SRME)		diamond (SRME)		MgO (SRME)		phonon (RMSE, THz)	mean rank	
	fitted T	held-out T	fitted T	held-out T	fitted T	held-out T		block	all
Fine-tuned on the titanium elastic constants									
MACE	0.157 → 0.024 (0.004)	0.139 → 0.026 (0.005)	0.287 → 0.305 (0.002)	0.280 → 0.297 (0.002)	0.071 → 0.074 (0.002)	0.060 → 0.063 (0.002)	1.209 → 1.249 (0.196)	4.1 → 4.9	12.5
MatterSim	0.387 → 0.016 (0.002)	0.337 → 0.017 (0.002)	0.170 → 0.169 (0.000)	0.166 → 0.164 (0.000)	0.079 → 0.081 (0.000)	0.065 → 0.067 (0.000)	1.328 → **0.324 (0.006)**	5.1 → 3.0	8.7
PET-MAD	0.322 → 0.032 (0.003)	0.288 → 0.032 (0.003)	0.145 → 0.158 (0.015)	0.160 → 0.204 (0.030)	0.054 → 0.065 (0.003)	0.067 → 0.075 (0.002)	0.786 → 0.391 (0.006)	3.6 → 3.9	9.8
SevenNet	0.162 → 0.033 (0.001)	0.152 → 0.040 (0.000)	0.222 → 0.222 (0.000)	0.222 → 0.222 (0.000)	0.080 → 0.082 (0.000)	0.069 → 0.071 (0.000)	0.582 → 0.540 (0.097)	4.0 → 5.1	12.1
Orb	0.159 → 0.013 (0.002)	0.160 → 0.013 (0.003)	0.068 → 0.098 (0.001)	0.059 → 0.090 (0.001)	0.053 → **0.060 (0.001)**	0.055 → **0.053 (0.002)**	0.365 → 0.371 (0.037)	1.9 → **1.8**	5.3
UMA	0.090 → 0.011 (0.002)	0.097 → **0.011 (0.004)**	0.086 → **0.086 (0.001)**	0.082 → **0.082 (0.001)**	0.068 → 0.069 (0.001)	0.054 → 0.056 (0.001)	1.176 → 1.019 (0.030)	2.2 → 2.4	7.1
Fine-tuned on the diamond elastic constants									
MACE	0.157 → 0.173 (0.001)	0.139 → 0.155 (0.001)	0.287 → 0.059 (0.004)	0.280 → 0.047 (0.007)	0.071 → 0.052 (0.000)	0.060 → 0.056 (0.000)	1.96 → 1.05 (0.01)	4.1 → 4.2	8.8
MatterSim	0.387 → 0.387 (0.000)	0.337 → 0.337 (0.000)	0.170 → 0.002 (0.000)	0.166 → **0.005 (0.000)**	0.079 → 0.080 (0.000)	0.065 → 0.066 (0.000)	1.03 → 0.85 (0.01)	4.4 → 3.5	10.2
PET-MAD	0.322 → 0.308 (0.003)	0.288 → 0.278 (0.003)	0.145 → 0.018 (0.007)	0.160 → 0.052 (0.036)	0.054 → **0.040 (0.000)**	0.067 → 0.055 (0.001)	1.20 → 1.01 (0.08)	3.9 → 4.2	8.9
SevenNet	0.162 → 0.157 (0.000)	0.152 → 0.147 (0.000)	0.222 → 0.004 (0.001)	0.222 → 0.009 (0.001)	0.080 → 0.082 (0.000)	0.069 → 0.071 (0.000)	2.11 → 1.57 (0.00)	5.1 → 4.2	10.4
Orb	0.159 → 0.157 (0.001)	0.160 → 0.159 (0.001)	0.068 → 0.017 (0.011)	0.059 → 0.026 (0.015)	0.053 → 0.049 (0.000)	0.055 → 0.053 (0.000)	0.70 → 0.91 (0.06)	2.0 → 3.1	6.3
UMA	0.090 → **0.090 (0.000)**	0.097 → **0.096 (0.000)**	0.086 → 0.002 (0.000)	0.082 → 0.009 (0.000)	0.068 → 0.064 (0.000)	0.054 → **0.051 (0.000)**	0.52 → **0.55 (0.00)**	1.5 → **1.8**	3.9
Fine-tuned on the MgO elastic constants									
MACE	0.157 → 0.149 (0.000)	0.139 → 0.132 (0.000)	0.287 → 0.276 (0.000)	0.280 → 0.270 (0.000)	0.071 → 0.010 (0.001)	0.060 → **0.025 (0.001)**	1.28 → 0.56 (0.02)	4.2 → 3.5	8.0
MatterSim	0.387 → 0.394 (0.000)	0.337 → 0.342 (0.000)	0.170 → 0.166 (0.000)	0.166 → 0.162 (0.000)	0.079 → 0.005 (0.004)	0.065 → 0.031 (0.006)	0.63 → 0.57 (0.03)	4.0 → 4.6	10.5
PET-MAD	0.322 → 0.326 (0.001)	0.288 → 0.286 (0.000)	0.145 → 0.139 (0.002)	0.160 → 0.178 (0.002)	0.054 → 0.012 (0.000)	0.067 → 0.032 (0.001)	0.59 → **0.29 (0.00)**	3.5 → 3.6	8.9
SevenNet	0.162 → 0.159 (0.000)	0.152 → 0.150 (0.000)	0.222 → 0.202 (0.000)	0.222 → 0.201 (0.000)	0.080 → 0.029 (0.001)	0.069 → 0.045 (0.001)	0.86 → 0.35 (0.00)	4.9 → 4.2	8.8
Orb	0.159 → 0.178 (0.000)	0.160 → 0.179 (0.000)	0.068 → **0.040 (0.001)**	0.059 → **0.035 (0.000)**	0.053 → 0.010 (0.001)	0.055 → 0.032 (0.001)	0.69 → 0.41 (0.01)	2.4 → 3.4	6.6
UMA	0.090 → **0.093 (0.000)**	0.097 → **0.102 (0.000)**	0.086 → 0.057 (0.000)	0.082 → 0.052 (0.000)	0.068 → 0.012 (0.001)	0.054 → 0.026 (0.002)	0.70 → **0.29 (0.00)**	2.0 → **1.6**	**3.6**

Cubic tasks. Neutron work on diamond and MgO reports the zone-boundary modes and not a measured path, so we score the two dispersions against the high-symmetry frequencies each measurement tabulates (Warren et al., 1967; Sangster et al., 1970). Every measured frequency takes the nearest model branch, as on titanium, and the errors pool as an RMSE. Table 3 carries that dispersion beside the compositions of each task, and Figure 12 draws it.

MgO is polar, and the splitting between its longitudinal and transverse optic modes at the zone center comes from a macroscopic field that no short-range potential carries. We therefore supply the non-analytic correction from the measured Born effective charge and high-frequency dielectric constant, the same pair for every model, so that the column ranks each model's transverse frequency. Diamond is non-polar and takes no correction.

B.3 Liquid structure

The RMSE of $S(q)$ pools every bin, and two models with the same error can miss the measurement in different places. Table 12 therefore reads the second peak: its height $S(q_2)$, and its position relative to the first peak, q_2/q_1. Perfect icosahedral order puts that ratio near two (Sachdev & Nelson, 1984),

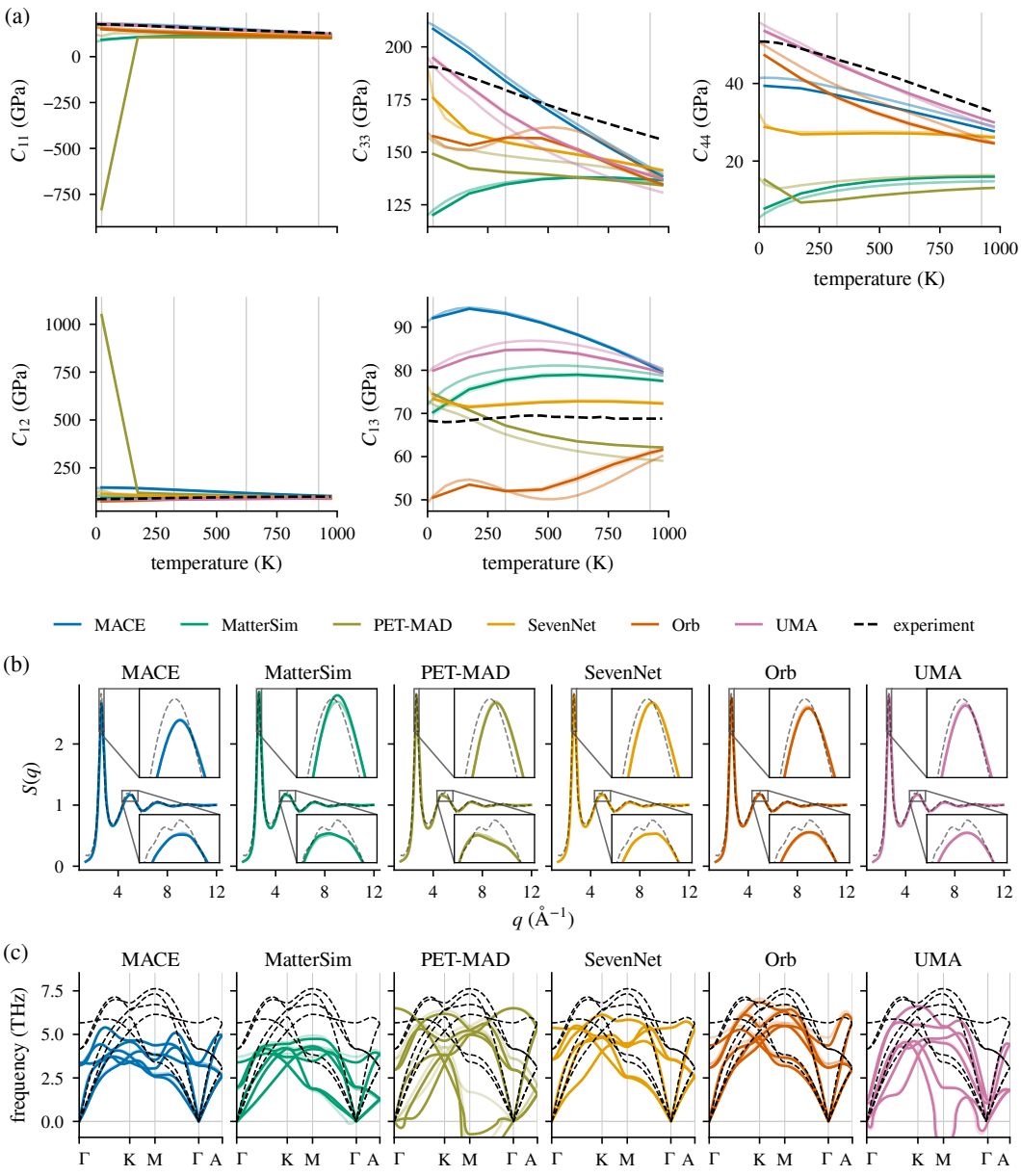


Figure 9: Off-target observables of the models fine-tuned on the phase transition. Pretrained models are faded and fine-tuned models solid. (a) hcp elastic constants at the four fitted temperatures of the titanium elastic task and the four sampled held-out ones, joined by a line, the mean over three seeds with the standard deviation as a shaded band, against the experimental series as a black dashed line. Vertical lines mark the fitted temperatures. (b) $S(q)$ of the liquid at 1965 K (one seed) against the measurement of Holland-Moritz et al. (2007), drawn as in Figure 3c. (c) hcp phonon dispersions (one seed) against the Born–von Kármán fit to the measured frequencies, drawn as in Figure 3d.

and Kreuzer et al. (2025) read liquid alloys with it. The table covers the models fine-tuned on $S(q)$ and the models fine-tuned on the titanium elastic constants (Section 3.1). Both peaks are read the same way (Appendix A.3).

On target, both peaks move toward the measurement. The first peak moves to lower q for every model, and the second rises for every model, closing 27–46% of the height gap while staying below the measured value. The ratio q_2/q_1 lies more than one bin from the measured value for MatterSim and PET-MAD before fine-tuning and for no model after it.

Table 11: The held-out polymorph under the two determinations of Appendix B.2, pretrained → fine-tuned under EXP-ONLY. The first column repeats the bcc column of Table 9. A fine-tuned entry is the mean over three seeds with the standard deviation in parentheses. The last row is the DimeNet++ baseline of Röcken & Zavadlav (2024), from the study's published values.

	Held-out polymorph, bcc Ti (SRME)	
	Fisher & Dever (1970)	Ledbetter et al. (2004)
MACE	0.119 → 0.039 (0.004)	0.150 → 0.052 (0.008)
MatterSim	0.132 → 0.063 (0.009)	0.181 → 0.058 (0.002)
PET-MAD	0.097 → 0.044 (0.026)	0.105 → 0.069 (0.010)
SevenNet	0.108 → 0.067 (0.002)	0.157 → 0.117 (0.001)
Orb	0.150 → 0.025 (0.010)	0.172 → 0.031 (0.003)
UMA	0.106 → 0.044 (0.004)	0.150 → 0.076 (0.004)
DimeNet++	0.092 → 0.086	0.120 → 0.110

Off target the peaks do not follow the error. The ratio stays more than one bin from the measured value for the same two models, and the second peak of PET-MAD moves down, away from the measurement, while its $S(q)$ error falls below the pretrained value (Table 1). The pooled error can fall while the structure the measurement resolves does not improve. A study that needs the short-range order reads the peaks rather than the RMSE.

Table 12: The second peak of the liquid structure factor at 1965 K, pretrained → fine-tuned: its height $S(q_2)$ and its position relative to the first peak, q_2/q_1, for the models fine-tuned on $S(q)$ (on target) and for the models fine-tuned on the titanium elastic constants (off target). A fine-tuned entry is the mean over three seeds with the standard deviation in parentheses. The last row is the measurement of Holland-Moritz et al. (2007) on the same bins.

	on target		off target	
	$S(q_2)$	q_2/q_1	$S(q_2)$	q_2/q_1
MACE	1.171 → 1.185 (0.002)	1.88 → 1.88 (0.00)	1.171 → 1.170 (0.004)	1.88 → 1.86 (0.01)
MatterSim	1.165 → 1.186 (0.005)	1.83 → 1.88 (0.01)	1.165 → 1.172 (0.003)	1.83 → 1.79 (0.01)
PET-MAD	1.172 → 1.191 (0.007)	1.76 → 1.89 (0.02)	1.172 → 1.155 (0.002)	1.76 → 1.80 (0.00)
SevenNet	1.171 → 1.186 (0.004)	1.88 → 1.87 (0.00)	1.171 → 1.183 (0.001)	1.88 → 1.86 (0.00)
Orb	1.172 → 1.184 (0.006)	1.87 → 1.87 (0.01)	1.172 → 1.178 (0.004)	1.87 → 1.86 (0.01)
UMA	1.170 → 1.192 (0.004)	1.88 → 1.88 (0.01)	1.170 → 1.172 (0.001)	1.88 → 1.87 (0.01)
measurement	1.217	1.89	1.217	1.89

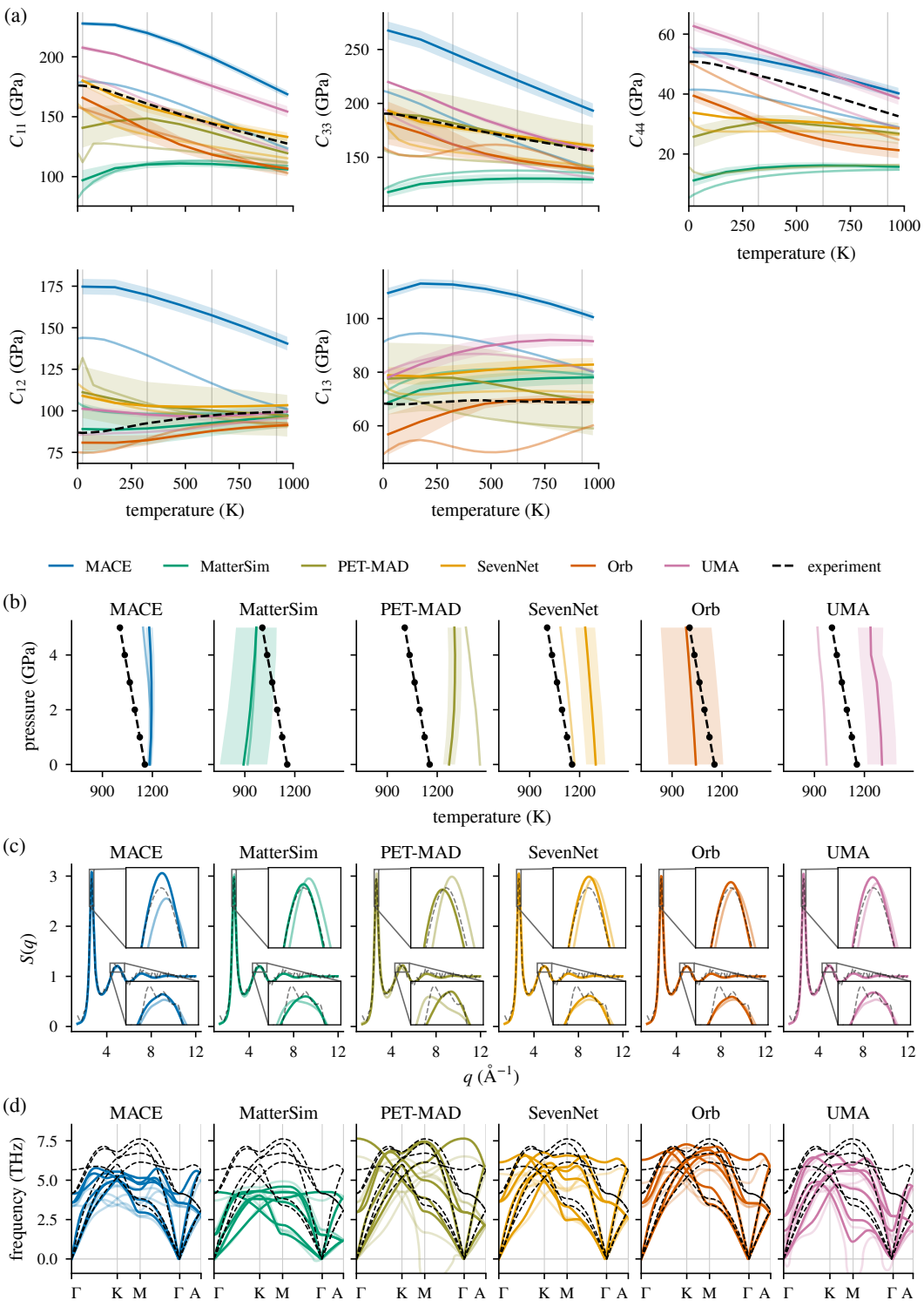


Figure 10: Off-target observables of the models fine-tuned on the liquid structure factor, drawn as in Figure 9. (a) hcp elastic constants at the fitted and sampled held-out temperatures. (b) The hcp-bcc transition line under pressure, one panel per model, drawn as in Figure 3b. (c) The task's own held-out temperature: $S(q)$ of the undercooled melt at 1745 K (one seed) against the measurement of Holland-Moritz et al. (2007), drawn as in Figure 3c. (d) hcp phonon dispersions.

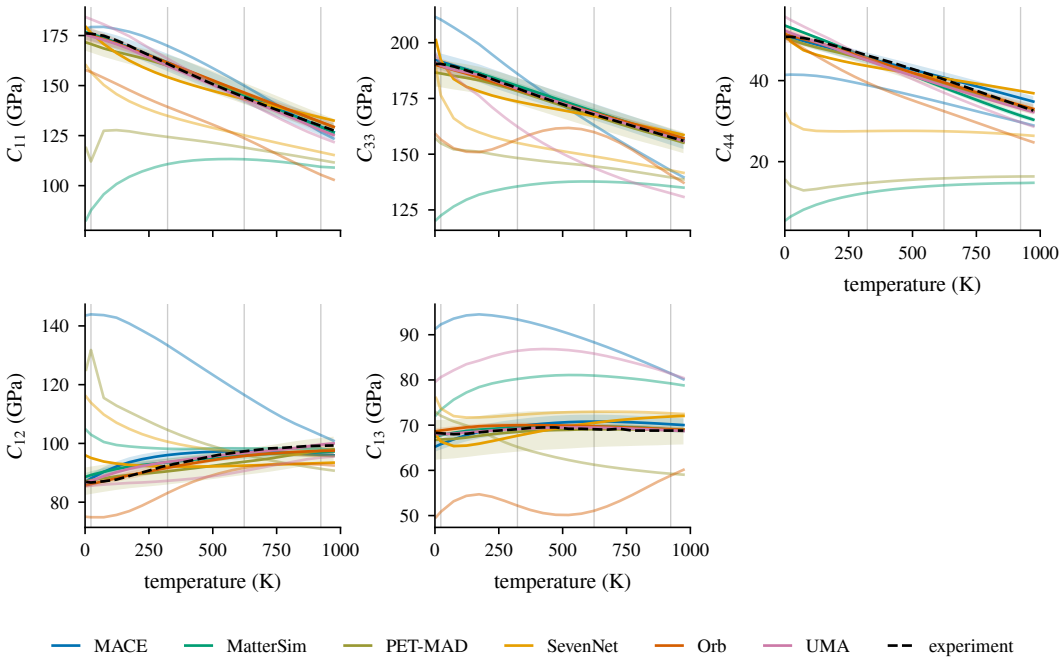


Figure 11: Fitted and held-out temperatures: titanium hcp elastic constants at the 22 measured temperatures, 4 K, 298 K and every 50 K from 23 to 973 K. Pretrained models are faded and fine-tuned models solid, and a fine-tuned line is the mean over three seeds with the standard deviation as a shaded band. A black dashed line marks the experimental series. Each temperature is an independent MD estimate in the cell built from the measured lattice constants at that temperature. Vertical lines mark the four fitted temperatures, and the tables score the held-out temperatures 173, 473, 773 and 973 K.

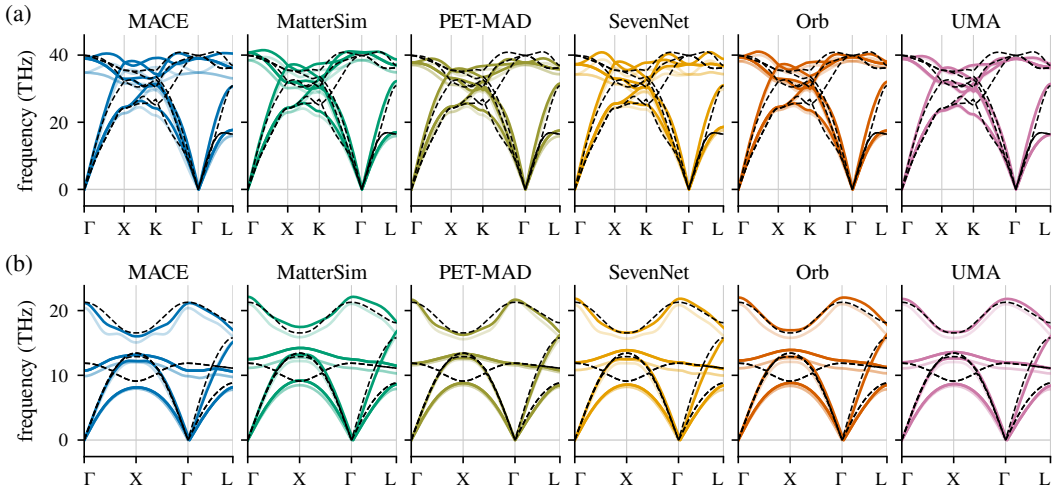


Figure 12: Phonon dispersions of the cubic tasks at the model's zero-pressure cell, one panel per model. Pretrained models are faded and fine-tuned models solid, and a black dashed line is the fit to the measured frequencies (Warren et al., 1967; Sangster et al., 1970). Axis limits follow those frequencies, so an imaginary branch can fall below the frame. The MgO panel carries the non-analytic correction, which splits the zone-center optic modes. (a) Diamond. (b) MgO.