

Evaluating Mechanical Property Prediction across Material Classes using Molecular Dynamics Simulations with Universal Machine-Learned Interatomic Potentials

Konstantin Stracke¹, Connor W. Edwards¹, and Jack D. Evans^{*1}

¹School of Physics, Chemistry and Earth Sciences, The University of Adelaide, North Terrace, Adelaide, 5005, South Australia, Australia

December 1, 2025

Abstract

We assess the accuracy of six universal machine-learned interatomic potentials (MLIPs) for predicting the temperature and pressure response of materials by molecular dynamics simulations. Accuracy is evaluated across 13 diverse materials (nine metal-organic frameworks and four inorganic compounds), computing bulk modulus, thermal expansion, and thermal decomposition. These MLIPs employ three different architectures (graph neural networks, graph network simulators, and graph transformers) with varying training datasets. We observe qualitative accuracy across these predictions but systematic underestimation of bulk modulus and overestimation of thermal expansion across all models, consistent with potential energy surface softening. From all tested models, three top performers arise; ‘MACE-MP-0a’, ‘fairchem_OMAT’, and ‘Orb-v3’, with average error across metrics and materials of 41 %, 44 %, and 47 %, respectively. Despite strong overall performance, questions arise about the limits of model transferability: dataset homogeneity and structural representation dominate model accuracy. Our results show that certain architectures can compensate for biases, a step closer to truly universal MLIPs.

Keywords: Materials Science, Machine-learned Interatomic Potentials, Thermal Expansion, Bulk Modulus, Thermal Stability

1 Introduction

The fundamental physics governing interatomic forces are well established, yet their application to complex systems yields equations that render quantum-mechanical methods such as density functional theory (DFT) computationally prohibitive[1]. Machine-learning offers a promising alternative by learning the underlying potential energy surface (PES) from highly accurate reference data (often DFT), enabling near-ab initio accuracy at dramatically reduced cost[2]. The field has progressed toward an ambitious goal: the development of universal, general-purpose models capable of accurately capturing the essential physics of complex systems, without requiring system-specific parameterization or fine-tuning[3]. This vision mirrors earlier efforts in classical force field development, from systematic system-specific parametrization approaches for coarse-grained models[4] or atomistic force fields like QuickFF[5] and MOF-FF[6], toward

^{*}Corresponding author: j.evans@adelaide.edu.au

broadly transferable frameworks exemplified by the Universal Force Field[7]. While recent studies have demonstrated that system specific fine-tuning of machine-learned interatomic potentials (MLIPs) can enhance accuracy for particular applications, for instance in high-temperature dynamics[8], there is promise that universal MLIPs can offer robust ‘out-of-the-box’ performance across a wide range of chemical environments and conditions.

Over time, graph-based MLIP architectures have become dominant[9], including graph neural networks (GNN) that estimate the atom (node) environment by passing messages from neighboring atoms within a cutoff[10]. Graph network simulators (GNS) extend this approach by incorporating dynamic properties such as velocities into the node embeddings[11]. Graph transformers (GTs) further enhance message passing by weighting (‘attention’) contributions so that information can propagate beyond the local cutoff and potentially capture global effects[12].

Mechanical and thermal properties are critical descriptors that quantify how materials respond to fundamental environmental stimuli. Specifically, the bulk modulus quantifies resistance to compression, thermal expansion coefficients describe dimensional response to temperature changes, and decomposition temperatures define the stability to thermal stress. Together, these properties define both the scope of practical applications and the operational limits of a material. The computational speedup of MLIPs relative to DFT has enabled high-throughput materials screening[3]. However, another transformative potential of these methods lies in the production of long-timescale molecular dynamics (MD) simulations. This capability comes with a trade-off that MLIPs are physically agnostic[13], making benchmarking and validation a necessity. Recently, Kraß et al. provided a systematic benchmark of mechanical properties, comparing MLIPs against DFT and classical potentials for metal-organic frameworks (MOFs)[14]. They demonstrated that static property validation is insufficient and dynamic modeling must also be assessed, which was investigated by *NPT* MD, including coordination environment and cell parameter analysis, focusing on MD stability.

In this work, we have evaluated the ability of universal MLIPs to predict key mechanical properties across a variety of material classes. Six MLIP models were selected that span three graph-based architectures (GNS, GNN, GT) and trained on different datasets to investigate the interplay between model architecture, training data, and predictive performance. This selection allows systematic trends to be assessed and the effects of architecture versus data coverage to be isolated. For this study, a database comprising 13 materials with diverse and structurally complex characteristics was constructed (Table 1). The comprehensive scope of this study encompasses 13 materials with densities ranging from 0.19 to 2.93 g cm⁻³ across multiple material classes and compositions, demonstrating the breadth and complexity of the systems investigated. A series of *NPT* and *NVT* MD simulations were used to evaluate the temperature- and pressure-dependent response of each material, including negative thermal expansion, bulk modulus, and thermal decomposition. These simulation-derived properties were subsequently validated against reference data, prioritizing experimental measurements and supplemented with quantum mechanical calculations.

2 Methods

All *NVT* and *NPT* MD simulations were performed within the Atomic Simulation Environment[15] with six different MLIP models (Table 2). The models include the ‘Orb-v3’ MLIP (Orb-v3-conservative-inf-omat)[16] from the Orb framework[17], which is trained on the OMat dataset[18]. Multiple versions of the MACE framework (version 0.3.13) were tested, including ‘MACE-1’ (MACE-MP-0a)[3], trained on the Materials Project trajectories dataset[19], and ‘MACE-2’ (MACE-MPA-0), an updated version trained on an expanded dataset that includes the sub-sampled Alexandria dataset[20]. Additionally, the ‘MACE-MOF’ was included, which is the MACE-MP-0b (medium) model fine-tuned for metal-organic frameworks (MOF) and fine-tuned with a dataset of 127 MOFs[21]. Finally, from the Fairchem framework[22], two pre-trained

Table 1: Materials investigated in this study, including their chemical formula, density, and classification, with MOFs further subdivided by structural or compositional family.

Material	Chemical Formula	Density / g cm^{-3}	Class
Zn-based MOFs (carboxylate linkers)			
MOF-5	$\text{C}_{24}\text{H}_{12}\text{O}_{13}\text{Zn}_4$	0.35	Zn-BDC (terephthalate)
MOF-10	$\text{C}_{42}\text{H}_{24}\text{O}_{13}\text{Zn}_4$	0.19	Zn-BPDC (biphenyl dicarboxylate)
TRUMOF	$\text{C}_{24}\text{H}_{12}\text{O}_{13}\text{Zn}_4$	0.84	Zn-truxene-BDC
Zn-based MOFs (azolate linkers)			
ZIF-8	$\text{C}_6\text{H}_6\text{N}_4\text{Zn}$	0.49	ZIF (imidazolate)
CALF-20	$\text{C}_3\text{H}_2\text{N}_3\text{O}_2\text{Zn}$	0.94	Zn-triazolate-oxalate
Zr-based MOFs (UiO/NU family)			
UiO-66	$\text{C}_{24}\text{H}_{14}\text{O}_{16}\text{Zr}_3$	0.72	Zr-BDC (terephthalate)
UiO-66-NH ₂	$\text{C}_{24}\text{H}_{17}\text{N}_3\text{O}_{16}\text{Zr}_3$	0.76	Zr-NH ₂ -BDC
UiO-67	$\text{C}_{42}\text{H}_{26}\text{O}_{16}\text{Zr}_3$	0.43	Zr-BPDC (biphenyl dicarboxylate)
NU-1000	$\text{C}_{44}\text{H}_{30}\text{O}_{16}\text{Zr}_3$	0.29	Zr-TBAPy (pyrene tetrabenzoate)
Inorganic materials			
Zn(CN) ₂	$\text{C}_2\text{N}_2\text{Zn}$	1.16	Coordination polymer
SiO ₂	SiO ₂	1.30	Inorganic oxide, α/β -Cristobalite
Zr(WO ₄) ₂	W_2ZrO_8	2.92	Inorganic oxide
CaMn ₇ O ₁₂	$\text{CaMn}_7\text{O}_{12}$	2.87	Perovskite oxide

models were tested; the 'OMat' model trained on the OMat dataset[18] and the 'ODAC' model specifically trained for MOFs and direct air capture applications[23].

Table 2: Summary of the MLIP models employed with their respective DOI. Further the MLIPs class of Graph Network Simulator (GNS), Graph Neural Network (GNN) or Graph Transformer (GT) and the training datasets are also shown.

Name	DOI	Class	Datasets
Orb-v3	10.48550/arXiv.2504.06231	GNS	OMat[18]
MACE-1	10.48550/arXiv.2401.00096	GNN	MPtrj[19]
MACE-2	10.48550/arXiv.2401.00096	GNN	MPtrj[19] + sAlex[20]
MACE-MOF	10.1038/s41524-025-01611-8	GNN	MPtrj[19] + 127 MOFs
fairchem_OMAT	10.5281/zenodo.15587498	GT	OMat[18]
fairchem_ODAC	10.5281/zenodo.15587498	GT	ODAC[23]

NVT simulations at ambient temperature used the cell and geometry optimized configuration which was isotropically scaled to volumes corresponding to 0.96, 0.98, 1.00, 1.02, and 1.04 times the original volume of the cell and subsequently geometry optimized. *NPT* simulations were performed at temperatures ranging from 50 to 1000 K in 50 K increments. Each temperature step was initialized from the final configuration of the preceding step, with the 50 K simulation starting from the optimized structure. Both *NVT* and *NPT* simulations ran for a total of 125 ps with a timestep of 1 fs, and system configurations were recorded every 50 steps, which were used for all subsequent analyses. The initial 25 ps of the simulation trajectory was reserved for equilibration. Temperature and pressure were controlled using the Berendsen thermostat and barostat[24]. The thermostat used a timestep of 1 fs and a temperature coupling (τ_T) of 100 fs, while the barostat time constant (τ_p) and the material compressibility were adjusted for each simulation, as described in the Supplementary Information (Section S3.2).

The bulk modulus at constant temperature K_T is defined as the volume V that changes with

a pressure change ∂P :

$$K_T = -V \left(\frac{\partial P}{\partial V} \right)_T \quad (1)$$

In practice, K_T is calculated from NPT simulations via averaged pressures from the stress tensor components. Bulk modulus was then determined using linear regression over the pressure–volume data collected at different temperatures. In contrast to NPT , the NVT simulations of bulk moduli were calculated by the Birch-Murnaghan pressure-volume relationship fit. For calculating the bulk modulus values, only NVT and NPT simulations at 300 K were used to compare to literature values also measured at 300 K. To show these simulations are equilibrated, autocorrelation functions for volume (NPT) and pressure (NVT) at 300 K are provided in the Supplementary Information (Sections S2.1 and S3.1).

The volumetric coefficient of thermal expansion (CTE) α_V , is defined as the change in volume with respect to temperature at constant pressure.

$$\alpha_V = \frac{1}{V} \left(\frac{\partial V}{\partial T} \right)_P \quad (2)$$

For each material and model, α_V was obtained from the slope of the volume–temperature relationship, computed from the NPT simulations by evaluating the temperature dependence of the average cell volume and averaged over the reference temperature range.

The temperature of decomposition ($T_{\text{decomp.}}$) was defined by coordination bond breakage between the metal-ligand bond. First, organic atoms bonded to the metals are identified from the optimized structures using a predefined cutoff corresponding to the equilibrated bond length. An additional threshold of 1 Å was added to the equilibrium bond length to accommodate typical MD fluctuations at elevated temperatures. The entire trajectory was then scanned and if any bond exceeds this threshold for 250 fs, the material is classified as collapsed at that temperature. Reference values for bulk modulus, thermal expansion, and thermal decomposition were compiled from experimental measurements and, where experimental values were unavailable, quantum mechanical calculations (Table 3). These reference values are used to validate and evaluate the model accuracy of property predictions.

Table 3: Reference values for all materials for the bulk modulus, (K_T), thermal expansion coefficient (α_V), with temperature range, and the thermal decomposition temperature.

Material Name	Bulk Modulus K_T / GPa	Thermal Expansion α_V / MK ⁻¹ ΔT / K		Decomp. $T_{\text{decomp.}}$ / K
MOF-10	6.0 [25]	-65 [26]	80-500	650 [27]
MOF-5	17.0 [28]	-48 [29]	4-600	673-783 [30]
ZIF-8	6.5 [31]	27.6 [32]	200-500	550 [33]
NU-1000	8.2 [34]	-42 [35]	430-450	700 [36]
CALF-20	29.1 [37]	11.54 [37]	10-400	380 [38]
TRUMOF	7.5 [39]	-	-	450 [40]
UiO-66	41.0 [41]	-35 [42]	50-250	650 [43]
UiO-66-NH ₂	17.0 [44]	-	-	400 [45]
UiO-67	17.4 [46]	-26.1 [47]	550-800	670 [47]
Zn(CN) ₂	34.2 [48]	-51 [49]	10-370	375 [49]
Zr(WO ₄) ₂	61.3 [50]	-26.7 [51]	0.3-1050	1050 [51]
CaMn ₇ O ₁₂	190 [52]	19.5 [53]	377-488	550 [53]
SiO ₂ (β -cristobalite)	37 [54]	32.7 [55]	100-500	1986 [56]

The densities in Table 1 are obtained at 300 K, averaging over all models and their corresponding MD trajectories. Because CTE values depend strongly on temperature, comparisons

between reference data and simulations are performed only within the temperature range specified by the reference source. When comparing models using the mean absolute error (MAE, %), $\text{CaMn}_7\text{O}_{12}$ and $\text{Zr}(\text{WO}_4)_2$ were omitted because Ca and W are not represented within MACE-MOF.

3 Results and Discussion

The behavior of materials, from microscopic atomic rearrangements to macroscopic deformations, is fundamentally shaped by temperature and pressure, the most basic external stimuli. Bulk modulus, the material response to pressure, has macroscopic implications for engineering, among other fields, and it directly arises from the atomistic structure. With its strong dependence on the atomistic structure and the physical and chemical environment, the bulk modulus is a central property in materials science. In Figure 1 the calculated bulk moduli K_T from two different approaches, NVT and NPT simulations, with respective reference values are presented.

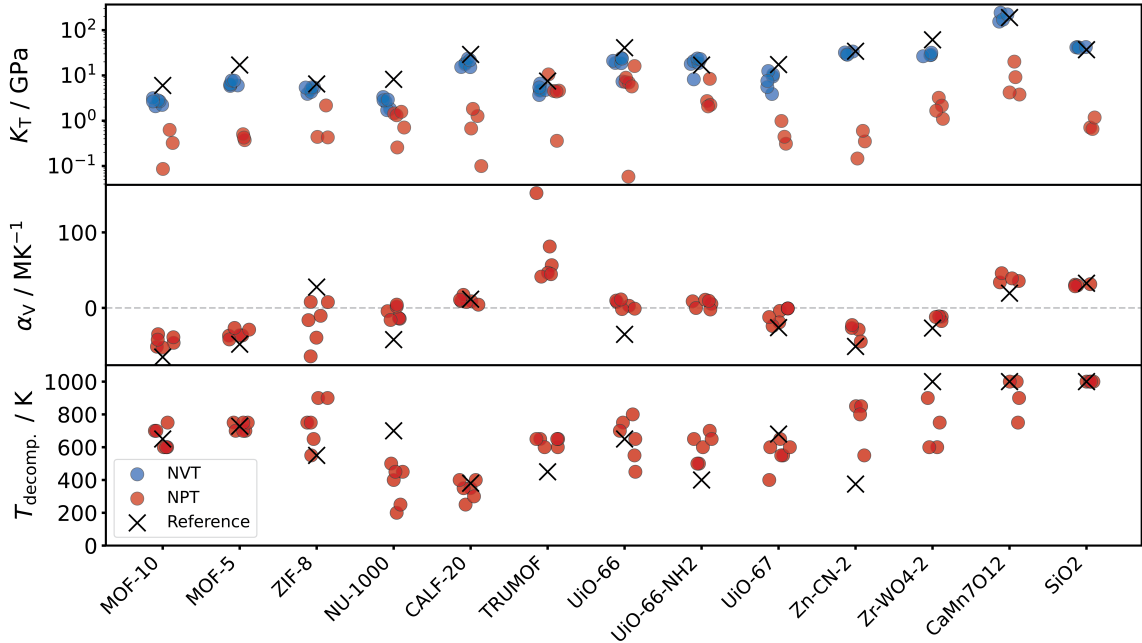


Figure 1: Summary of the bulk modulus (K_T), volumetric thermal expansion coefficient (α_V), and decomposition temperature for all materials, as predicted by the models. Corresponding literature values (Table 3) are highlighted for direct comparison.

Overall, the bulk modulus from NVT simulations matches the qualitative trends between materials while also achieving an MAE of $43.8\% \pm 6.9\%$, relative to reference values. As outlined in the Supplementary Information (Sections S2.1 and S3.1) the autocorrelation analysis of the MD trajectories show that volume relaxes substantially more slowly than pressure. Averaged over all materials at 300 K, the characteristic correlation time of volume is 6.72 ps, compared to 59.5 fs for pressure, indicating that volume relaxes approximately 114 times slower. Thus, the bulk modulus was more reliably obtained from short NVT trajectories[57]. This is clearly illustrated by the example of $\text{CaMn}_7\text{O}_{12}$, where the NPT simulations yield an average value of 9.3 GPa, while the NVT simulations give 197.8 GPa, which is much closer to the experimentally-determined value of 190 GPa[52].

Many MOFs exhibit fascinating negative thermal expansion (NTE), which can be largely attributed to porosity and flexible linkers[58]. However, this behavior is not unique to MOFs and also occurs in $\text{Zr}(\text{WO}_4)_2$, which has rigid units allowing flexible rotational motion (‘hinging’) around these corners (Table 3)[59, 51]. Complicating matters, many materials undergo various forms of structural reordering, resulting in complex thermal behavior with different CTEs across temperature ranges, including transitions from positive to negative thermal expansion[60]. Even when comparing within the same temperature range ambiguities remain. For example, Vornholt et al. observed substantial thermal hysteresis in UiO-66, with CTE values ranging between 45 and -80 MK^{-1} depending whether the material is heated or cooled over the same temperature range. In the same work, they determined the ‘intrinsic’ equilibrated CTE to be at -35 MK^{-1} [42]. Additionally, universal MLIPs are not trained on higher temperature dynamics, which may improve their CTE predictions[8]. As expected, predicting CTEs is more difficult compared to the bulk modulus (43.8 %) with a MAE of $76.2 \% \pm 25.2 \%$. Yet overall, the qualitative trend between materials and positive to negative thermal expansion is captured across models, with the exceptions of UiO-66, and for ZIF-8 where only MACE-1 and fairchem_OMAT correctly reproduce this behavior (Figure 1).

Beyond a volumetric response to temperature, thermal decomposition temperature is a fundamental property, determining the suitability of a material for practical applications. It reflects the ability to maintain structural integrity when subjected to heat, the resistance to bond breakage, and allows interatomic forces and bond-breaking to be analyzed using MLIPs. Within similar types of materials, MOF-10 is predicted to be less stable than MOF-5, while the UiO family is correctly predicted to be stable within a very similar temperature range. Considering the intrinsic material complexity, this level of agreement is remarkable. This in turn reinforces our approach of metal-organic bond breaking as an indicator for decomposition and reveals the ability of all tested MLIPs to accurately capture bond-breaking processes, in contrast to many classical force fields.

Beyond examining individual predictions, assessing the average accuracy across these predictions is essential and can highlight systematic errors. In Figure 2 a violin plot is displayed summarizing the distribution of deviations for each metric from the reference. The median predictions across all models are consistently close to the reference values, underscoring the overall accuracy of these models. The median deviations from reference values, averaged across all models, are -6.92 GPa for bulk modulus, 11.38 MK^{-1} for CTE and 18.50 K for decomposition temperature. To put those absolute values in context, different metrics span different ranges, namely CTE -64.10 to 152.09 MK^{-1} ($\Delta 216.19 \text{ MK}^{-1}$) and for K_T from -11.21 to 243.02 GPa ($\Delta 254.23 \text{ GPa}$) and for decomposition temperature 200 to 1000 K^{-1} ($\Delta 800 \text{ K}^{-1}$). This shows high accuracy when averaged across materials, but also reveals a systematic bias: all models underestimate the bulk modulus while overestimating the thermal expansion coefficient. This behavior is consistent with a systematic softening of the learned PES[61]. A softened PES reduces the predicted pressure, since pressure is the first derivative of energy with respect to volume, and at the same time permits larger volume fluctuations as thermal energy encounters smaller barriers.

A closer look of our simulations illustrates that specific materials challenge these models. For bulk modulus, $\text{CaMn}_7\text{O}_{12}$ emerges as an outlier and is overestimated by Orb-v3 and MACE-1 while underestimated by fairchem_ODAC. In this case, the material has an exceptionally large bulk modulus, which amplifies these absolute deviations. Regarding CTE, ZIF-8 is consistently the most underestimated material across all models. For decomposition temperature, NU-1000 and $\text{Zr}(\text{WO}_4)_2$ appear to be the least accurately predicted materials. While NU-1000 is an anisotropic framework featuring a large tetradentate linker, structural changes in $\text{Zr}(\text{WO}_4)_2$ stem from the aforementioned ‘hinging’ of rigid units. These results demonstrate that universal MLIPs have reached a level of maturity where they can reliably predict complex thermomechanical properties across diverse material classes. Although individual predictions can show deviations,

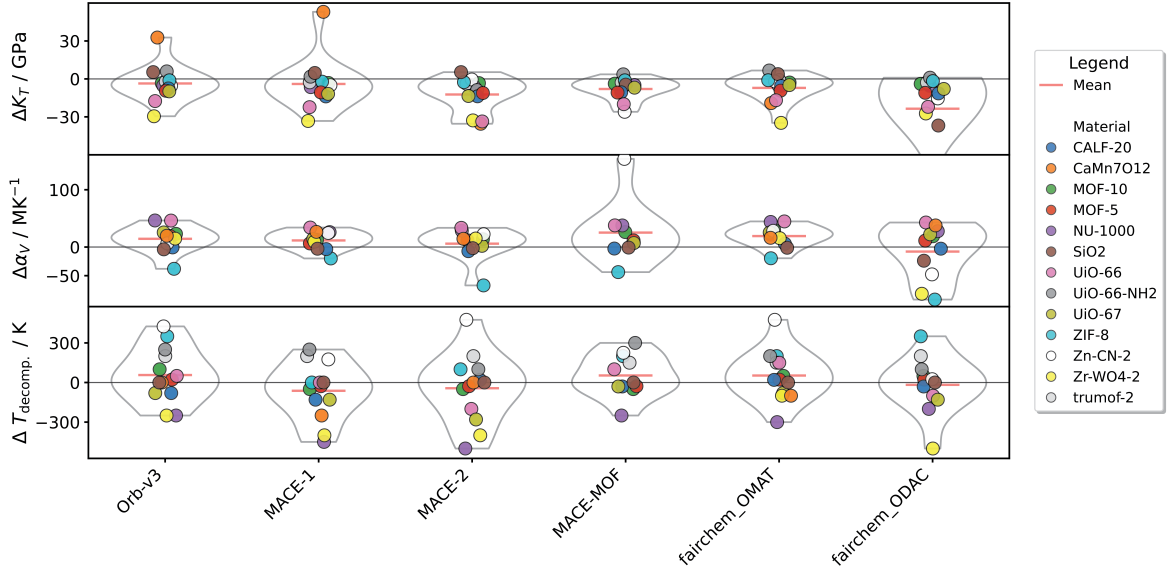


Figure 2: Prediction deviations for bulk modulus K_T , CTE α_V , and decomposition temperature across models, with materials indicated in the legend. Materials for which no reference value is available are omitted.

the models exhibit exceptional qualitative accuracy when compared across material averages. The systematic nature of observed biases, including an underestimation in bulk modulus and an overestimation in thermal expansion suggests clear pathways for future model refinement, possibly through targeted training data augmentation.

Having established the predictive capabilities of MLIPs across the material set, the overall accuracy of the models was evaluated. Figure 3 compares the MAE (%) of all models across the evaluated metrics in addition to efficiency and the average accuracy across all metrics. The three best performing models are MACE-1, fairchem_OMAT, and Orb-v3. Fairchem_OMAT has the highest accuracy for bulk, Orb-v3 is the fastest model, and MACE-1 has the highest accuracy across metrics.

Among these top three models (MACE-1, fairchem_OMAT, and Orb-v3), the architectures are distinct, corresponding to a GNN, GT, and GNS, respectively. Despite these architectural differences, all three achieve comparable overall accuracy. The consistent accuracy of these three models across all materials, despite architectural diversity, underscores that data curation and coverage exert a greater influence on average performance than architectural approach. We note this observation aligns with findings of Kraß et al.[14]. Interestingly, two of the three best performers were trained on OMAT, while the top-performing model, MACE-1, was trained on MPtrj. The OMAT-trained models excel at bulk modulus predictions but underperform on CTE. This accuracy trade-off reflects the inherent characteristics of each training dataset. OMAT contains non-equilibrium structures well-suited for molecular dynamics and bulk property predictions[14], whereas MPtrj has relaxation trajectories which better capture vibrational modes. Although fairchem_ODAC performs poorly on vibrational properties and bulk modulus, its superior thermal stability predictions reflect training emphasis on accurate interatomic forces from MOF-guest structures.

We now examine the different MACE models: MACE-1, the ‘multi-head’ fine-tuned MACE model[21], MACE-MOF, and the MACE model that was trained on an enlarged dataset combining MPtrj and sAlex[3], MACE-2. Counterintuitively, the initial release MACE model (MACE-1) performs better than either. This is surprising as the later model, MACE-2, was trained on more data and not fine-tuned. Critically, both models share identical architectures[3], elim-

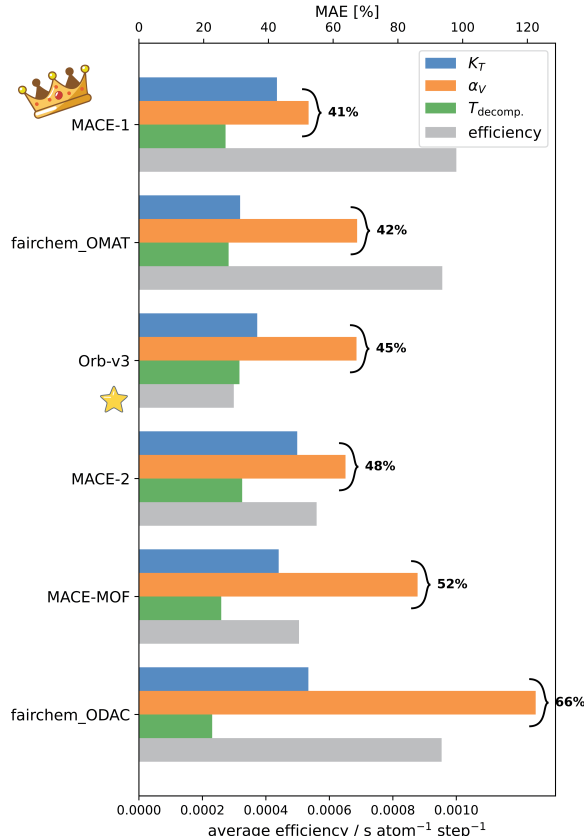


Figure 3: Mean absolute error (MAE, %) of all models across bulk modulus K_T , CTE α_V , decomposition temperature, and efficiency. The average accuracy across metrics (excluding efficiency) is indicated.

inating architectural differences as an explanation. Although MACE-2 has a lower accuracy than MACE-1, it achieves a higher computational efficiency, a trade-off that is also reported for MOF deformation by guests[62]. The speed-up of MACE-2 stems from optimizations to how the architecture is implemented rather than architectural simplification. Additionally, both training sets were generated by PBE/PBE+U functionals, which suggests it is not necessarily the level of theory that leads to less accurate results. The sAlex structures (10.4M of 1D, 2D, and 3D inorganic materials[63]) vastly outnumber MPtrj structures (1.58M) in the combined training set. The inclusion of additional structures from the sAlex dataset reduced predictive accuracy for structures already well represented in the MPtrj dataset. Comparing MACE-MOF with MACE-1, the results show that they have similar accuracy when describing MOFs, however MACE-1 can also perform on non-MOF structures (Supplementary Figure S1). For MACE-MOF, the MPtrj workflow is used to generate DFT training data for the curated MOF dataset, with only D3 dispersion correction added, as this correction is deemed particularly important for MOFs. While there are different approaches to fine-tuning[64], the MACE-1 versus MACE-2 comparison demonstrates that mixing datasets can compromise performance. In summary, underrepresented structural characteristics will lead to reduced model accuracy. Homogeneous datasets, including both equilibrium and non-equilibrium structures sampled at a consistent level of theory, are needed for accurate dynamics and property prediction.

To assess whether network architecture influences performance for specific systems rather than on average, we divided our MOF dataset based on cell symmetry into isotropic and anisotropic subsets (NU-1000, CALF-20, UiO-66-NH₂). Subsets were curated according to the anisotropic thermal expansion behavior revealed by *NPT* simulations (Supplementary Fig-

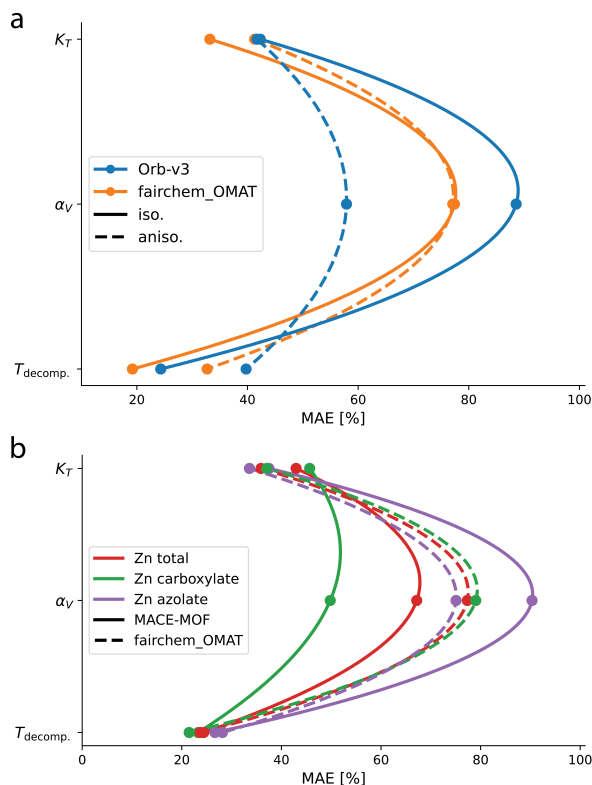


Figure 4: Model performance for predicting bulk modulus (K_T), CTE (α_V), and decomposition temperature across selected material subsets. a) Orb-v3 and fairchem_OMAT evaluated on isotropic (iso.) and anisotropic (aniso.) MOFs. b) MACE-MOF and fairchem_OMAT evaluated on zinc-based MOFs separated by linker type (Zn carboxylate vs. Zn azolate). Connecting curves added to aid visual interpretation.

ure S2). Since fairchem_OMAT and Orb-v3 were trained on the same dataset, performance differences can be attributed to their architectures or implementations. In comparison, we find that fairchem_OMAT performs slightly worse for anisotropic MOFs, a trend that holds across bulk modulus, CTE, and decomposition temperature (Figure 4a). Interestingly, Orb-v3 predicts the CTE of anisotropic MOFs more accurately than for isotropic MOFs. This implies that Orb-v3 accuracy varies between isotropic and anisotropic MOFs, whereas fairchem_OMAT shows consistent accuracy across both classes. Collectively, all models except fairchem_OMAT display this symmetry bias (Supplementary Figure S3).

Finally, the influence of material chemistry on model performance was examined by focusing on zinc-based MOFs, separated into ‘Zn carboxylate’ and ‘Zn azolate’ subsets (Table 1). For this we compare the MOF fine-tuned model MACE-MOF against fairchem_OMAT. Figure 4b shows that both models predict decomposition temperature with comparable accuracy across both linker types. Both models predict bulk modulus slightly better for azolate frameworks, but the most pronounced difference emerges for CTE, where MACE-MOF achieves substantially higher accuracy for Zn-carboxylate MOFs. The MACE-MOF model shows a strong bias toward Zn-carboxylate MOFs, achieving nearly twice the accuracy of Zn-azolate systems, with MAEs of 50 % and 90 %, respectively. This trend, with respect to CTE, is present for all models (Supplementary Fig. S4), except fairchem_OMAT, which shows near no changes between these ligands (Fig. 4). Notably, fairchem_ODAC, with a different and more specific training dataset, displays the same bias. The fairchem_OMAT model achieves more uniform accuracy across ligand types and symmetries, despite using the same training data as Orb-v3 and lacking the domain-specific fine-tuning of MACE-MOF, suggesting that thoughtful architectural design may

be as important as training data or fine-tuning in mitigating prediction bias.

4 Summary and Outlook

Our results reveal that universal MLIPs can reliably predict thermomechanical properties, though performance depends critically on both architecture and training data composition. We have demonstrated detailed analysis across isotropic and anisotropic MOFs, as well as Zn carboxylate and azolate linkers, which reveal that GNN and GNS architectures exhibit biases toward these structural and symmetry differences, while GT appears unbiased. By comparing fairchem_OMAT to Orb-v3, which are both trained on the OMAT dataset, this bias can be isolated as architecture-dependent. However, dataset quality supersedes architectural advantages: inconsistent data representation degrades predictive accuracy regardless of architecture (fairchem_OMAT versus fairchem_ODAC), while simultaneously reintroducing these structure and symmetry biases even in graph transformer models. This principle extends to dataset modification strategies, enlarging the MPtrj dataset with sAlex data (MACE-2 vs. MACE-1) and fine-tuning on MOF-specific data (MACE-MOF vs. MACE-1) both failed to improve predictions. While MACE-MOF showed no performance gains, even on its target MOF domain, both modified models exhibited degraded overall accuracy.

The top-performing models, MACE-1, fairchem_OMAT, and Orb-v3, exhibit comparable overall accuracy despite distinct architectures (GNN, GT, and GNS, respectively), highlighting that dataset composition and coverage strongly influence performance. Qualitative trends are well recreated across properties and the accuracy across materials and models is close to reference values, averaging at MAE of 44 % for our top performers. Yet, it is important to highlight that task specific trained models like fairchem_ODAC have a low overall accuracy (MAE, 66 %). However, they excel at task specific accuracy, namely predicting the decomposition temperature with a MAE of 23 %. There is a systematic underestimation of bulk modulus and overestimation of thermal expansion across all models, consistent with PES softening. Notably, all MD simulations remained stable, confirming the robustness of universal MLIPs even at elevated temperatures and deformed simulation cells.

From a practical standpoint, Orb-v3 demonstrates the fastest computational performance, while implementation optimizations in MACE-2 have achieved substantial speed improvements over its predecessor. These advances in computational efficiency, combined with our findings on dataset quality and architectural bias mitigation, point beyond semi-universal, task-specific models toward the prospect of truly universal MLIPs.

Supplementary Information

Plots showing accuracy of MACE-1 and MACE-MOF on MOF and non-MOF subsets, Model-averaged linear thermal expansion across all materials, model resolved accuracy for MOF subsets of Zn-Carboxylate and Zn-azolate, pressure and volume fluctuations (NVT and NPT), pressure and volume autocorrelation functions (NVT and NPT), barostat parameter refitting. All code and data for this research is available on Zenodo: [10.5281/zenodo.17730688](https://doi.org/10.5281/zenodo.17730688).

Acknowledgements

J.D.E is the recipient of an Australian Research Council Discovery Early Career Award (project number DE220100163) funded by the Australian Government. Phoenix HPC service at the University of Adelaide is thanked for providing high-performance computing resources. This research was supported by the Australian Government’s National Collaborative Research Infrastructure

Strategy (NCRIS), with access to computational resources provided by Pawsey Supercomputing Research Centre through the National Computational Merit Allocation Scheme.

References

- [1] Paul Adrien Maurice Dirac. Quantum mechanics of many-electron systems. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 123(792):714–733, 1929.
- [2] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.
- [3] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schrieffer, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O’Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18):184110, 11 2025.
- [4] Huong TL Nguyen and David M Huang. Systematic bottom-up molecular coarse-graining via force and torque matching using anisotropic particles. *The Journal of Chemical Physics*, 156(18), 2022.
- [5] Louis Vanduyfhuys, Steven Vandenbrande, Jelle Wieme, Michel Waroquier, Toon Verstraelen, and Veronique Van Speybroeck. Extension of the quickff force field protocol for an improved accuracy of structural, vibrational, mechanical and thermal properties of metal-organic frameworks, 2018.
- [6] Sareeya Bureekaew, Saeed Amirjalayer, Maxim Tafipolsky, Christian Spickermann, Tapta Kanchan Roy, and Rochus Schmid. Mof-ff—a flexible first-principles derived force field for metal-organic frameworks. *physica status solidi (b)*, 250(6):1128–1141, 2013.
- [7] Anthony K Rappé, Carla J Casewit, KS Colwell, William A Goddard III, and W Mason Skiff. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American chemical society*, 114(25):10024–10035, 1992.
- [8] Connor W. Edwards and Jack D. Evans. Exploring foundational machine learned potentials for treating the high temperature dynamics of metal-organic frameworks. *Advanced Theory and Simulations*, October 2025.

- [9] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.
- [10] Oliver T Unke and Markus Meuwly. Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of chemical theory and computation*, 15(6):3678–3693, 2019.
- [11] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International conference on machine learning*, pages 8459–8468. PMLR, 2020.
- [12] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023.
- [13] Joe D Morrow, John LA Gardner, and Volker L Deringer. How to validate machine-learned interatomic potentials. *The Journal of chemical physics*, 158(12), 2023.
- [14] Hendrik Kraß, Ju Huang, and Seyed Mohamad Moosavi. Mofsimbench: Evaluating universal machine learning interatomic potentials in metal–organic framework molecular modeling. *arXiv preprint arXiv:2507.11806*, 2025.
- [15] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dulak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, et al. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.
- [16] Benjamin Rhodes, Sander Vandenhoute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale. *arXiv preprint arXiv:2504.06231*, 2025.
- [17] Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential. *arXiv preprint arXiv:2410.22570*, 2024.
- [18] 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.
- [19] Materials Project Trajectory (MPtrj) dataset. https://figshare.com/articles/dataset/Materials_Project_Trajectory_MPtrj_Dataset/23713842, 2023. Figshare dataset (MPtrj) — accessed 2025-10-30.
- [20] sAlex: a Matbench-Discovery compliant subsample of the Alexandria dataset. <https://matbench-discovery.materialsproject.org/data/salex>, 2024. sAlex dataset used in MACE second-generation models. Accessed 2025-10-30.
- [21] Alin Marin Elena, Pratham Divakar Kamath, Théo Jaffrelot Inizan, Andrew S Rosen, Federica Zanca, and Kristin A Persson. Machine learned potential for high-throughput phonon calculations of metal—organic frameworks. *npj Computational Materials*, 11(1):125, 2025.
- [22] FAIR Chemistry (facebookresearch/fairchem). FAIRChem: FAIR Chemistry’s library of machine learning methods for chemistry. <https://github.com/facebookresearch/fairchem>, 2025. GitHub repository. Accessed: 2025-10-30.

- [23] Anuroop Sriram, Logan M. Brabson, Xiaohan Yu, Sihoon Choi, Kareem Abdelmaqsoud, Elias Moubarak, Pim de Haan, Sindy Löwe, Johann Brehmer, John R. Kitchin, Max Welling, C. Lawrence Zitnick, Zachary Ulissi, Andrew J. Medford, and David S. Sholl. The open dac 2025 dataset for sorbent discovery in direct air capture, 2025.
- [24] Herman JC Berendsen, JPM van Postma, Wilfred F Van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.
- [25] Sandro Wieser and Egbert Zojer. Machine learned force-fields for an ab-initio quality description of metal-organic frameworks. *npj Computational Materials*, 10(1):18, 2024.
- [26] David Dubbeldam, Krista S Walton, Donald E Ellis, and Randall Q Snurr. Exceptional negative thermal expansion in isorecticular metal-organic frameworks. *Angewandte Chemie-International Edition*, 46(24):4496–4499, 2007.
- [27] Jing-Yang Chung, Chi-Wei Liao, Yi-Wei Chang, Bor Kae Chang, Hao Wang, Jing Li, and Cheng-Yu Wang. Influence of metal–organic framework porosity on hydrogen generation from nanoconfined ammonia borane. *The Journal of Physical Chemistry C*, 121(49):27369–27378, 2017.
- [28] M Mattesini, JM Soler, and F Yndurain. Ab initio study of metal-organic framework-5 zn 4 o (1, 4-benzenedicarboxylate) 3: An assessment of mechanical and spectroscopic properties. *Physical Review B—Condensed Matter and Materials Physics*, 73(9):094111, 2006.
- [29] Wei Zhou, Hui Wu, Tanner Yildirim, Jeffrey R Simpson, and AR Hight Walker. Origin of the exceptional negative thermal expansion in metal-organic framework-5 zn 4 o (1, 4-benzenedicarboxylate) 3. *Physical Review B—Condensed Matter and Materials Physics*, 78(5):054114, 2008.
- [30] Colm Healy, Komal M Patil, Benjamin H Wilson, Lily Hermanspahn, Nathan C Harvey-Reid, Ben I Howard, Carline Kleinjan, James Kolien, Fabian Payet, Shane G Telfer, et al. The thermal stability of metal-organic frameworks. *Coordination Chemistry Reviews*, 419:213388, 2020.
- [31] Karena W Chapman, Gregory J Halder, and Peter J Chupas. Pressure-induced amorphization and porosity modification in a metal- organic framework. *Journal of the American Chemical Society*, 131(48):17546–17547, 2009.
- [32] Ashleigh M Chester, Celia Castillo-Blas, Roman Sajzew, Bruno P Rodrigues, Giulio I Lampronti, Adam F Sapnik, Georgina P Robertson, Matjaž Mazaj, Daniel JM Irving, Lothar Wondraczek, et al. Loading and thermal behaviour of zif-8 metal–organic framework-inorganic glass composites. *Dalton Transactions*, 53(25):10655–10665, 2024.
- [33] Kyo Sung Park, Zheng Ni, Adrien P Côté, Jae Yong Choi, Rudan Huang, Fernando J Uribe-Romo, Hee K Chae, Michael O’Keeffe, and Omar M Yaghi. Exceptional chemical and thermal stability of zeolitic imidazolate frameworks. *Proceedings of the National Academy of Sciences*, 103(27):10186–10191, 2006.
- [34] Georgina P Robertson, Sara Mosca, Celia Castillo-Blas, Florencia A Son, Omar K Farha, David A Keen, Simone Anzellini, and Thomas D Bennett. Survival of zirconium-based metal–organic framework crystallinity at extreme pressures. *Inorganic chemistry*, 62(26):10092–10099, 2023.

- [35] Zhihengyu Chen, Gautam D Strosio, Jian Liu, Zhiyong Lu, Joseph T Hupp, Laura Gagliardi, and Karena W Chapman. Node distortion as a tunable mechanism for negative thermal expansion in metal–organic frameworks. *Journal of the American Chemical Society*, 145(1):268–276, 2022.
- [36] Joseph E Mondloch, Wojciech Bury, David Fairen-Jimenez, Stephanie Kwon, Erica J DeMarco, Mitchell H Weston, Amy A Sarjeant, SonBinh T Nguyen, Peter C Stair, Randall Q Snurr, et al. Vapor-phase metalation by atomic layer deposition in a metal–organic framework. *Journal of the American Chemical Society*, 135(28):10294–10297, 2013.
- [37] Dong Fan, Supriyo Naskar, and Guillaume Maurin. Unconventional mechanical and thermal behaviours of mof calf-20. *Nature Communications*, 15(1):3251, 2024.
- [38] Natchaya Phongsuk, Chalarat Chaemchamrat, Taya Ko Saothayanun, Nopphon Weeranoppanant, and Sareeya Bureekaew. Efficient solvent-free mechanochemical synthesis of calf-20 for carbon dioxide capture. *CrystEngComm*, 27(36):5959–5964, 2025.
- [39] Emily G Meekel, Phillippa Partridge, Robert AI Paraoan, Joshua JB Levinsky, Ben Slater, Claire L Hobday, and Andrew L Goodwin. Enhanced elastic stability of a topologically disordered crystalline metal–organic framework. *Nature Materials*, 23(9):1245–1251, 2024.
- [40] Emily G Meekel, Ella M Schmidt, Lisa J Cameron, A David Dharma, Hunter J Windsor, Samuel G Duyker, Arianna Minelli, Tom Pope, Giovanni Orazio Lepore, Ben Slater, et al. Truchet-tile structure of a topologically aperiodic metal–organic framework. *Science*, 379(6630):357–361, 2023.
- [41] Hui Wu, Taner Yildirim, and Wei Zhou. Exceptional mechanical stability of highly porous zirconium metal–organic framework uio-66 and its important implications. *The journal of physical chemistry letters*, 4(6):925–930, 2013.
- [42] Simon M Vornholt, Zhihengyu Chen, Jan Hofmann, and Karena W Chapman. Node distortions in uio-66 inform negative thermal expansion mechanisms: Kinetic effects, frustration, and lattice hysteresis. *Journal of the American Chemical Society*, 146(25):16977–16981, 2024.
- [43] Loredana Valenzano, Bartolomeo Civalieri, Sachin Chavan, Silvia Bordiga, Merete H Nilsen, Søren Jakobsen, Karl Petter Lillerud, and Carlo Lamberti. Disclosing the complex structure of uio-66 metal organic framework: a synergic combination of experiment and theory. *Chemistry of Materials*, 23(7):1700–1718, 2011.
- [44] Nicholas C Burtch, Jurn Heinen, Thomas D Bennett, David Dubbeldam, and Mark D Allendorf. Mechanical properties in metal–organic frameworks: emerging opportunities and challenges for device functionality and technological applications. *Advanced materials*, 30(37):1704124, 2018.
- [45] Sergio J Garibay and Seth M Cohen. Isorecticular synthesis and modification of frameworks with the uio-66 topology. *Chemical communications*, 46(41):7700–7702, 2010.
- [46] Claire L Hobday, Ross J Marshall, Colin F Murphie, Jorge Sotelo, Tom Richards, David R Allan, Tina Düren, François-Xavier Coudert, Ross S Forgan, Carole A Morrison, et al. A computational and experimental approach linking disorder, high-pressure behavior, and mechanical properties in uio frameworks. *Angewandte Chemie*, 128(7):2447–2451, 2016.
- [47] Isabella Goodenough, Venkata Swaroopa Datta Devulapalli, Wenqian Xu, Mikaela C Boyanich, Tian-Yi Luo, Mattheus De Souza, Mélissandre Richard, Nathaniel L Rosi, and Eric Borguet. Interplay between intrinsic thermal stability and expansion properties of functionalized uio-67 metal–organic frameworks. *Chemistry of Materials*, 33(3):910–920, 2020.

- [48] Karena W Chapman and Peter J Chupas. Pressure enhancement of negative thermal expansion behavior and induced framework softening in zinc cyanide. *Journal of the American Chemical Society*, 129(33):10090–10091, 2007.
- [49] Andrew L Goodwin and Cameron J Kepert. Negative thermal expansion and low-frequency modes in cyanide-bridged framework materials. *Physical Review B—Condensed Matter and Materials Physics*, 71(14):140301, 2005.
- [50] CA Perottoni and JAH da Jornada. Pressure-induced amorphization and negative thermal expansion in zrw2o8. *Science*, 280(5365):886–889, 1998.
- [51] T A Mary, JSO Evans, T Vogt, and AW Sleight. Negative thermal expansion from 0.3 to 1050 kelvin in zrw2o8. *Science*, 272(5258):90–92, 1996.
- [52] Michal Stekiel, A Girard, D Zimmer, V Monteseguro, K Glazyrin, I de Pedro, L Bayarjargal, B Winkler, and Javier Ruiz-Fuertes. Pressure-induced charge ordering transition in camn 7 o 12. *Physical Review B*, 105(6):064110, 2022.
- [53] K Gautam, DK Shukla, S Francoual, J Bednarcik, JRL Mardegan, H-P Liermann, R Sankar, FC Chou, DM Phase, and J Strempfer. Large negative thermal expansion in the cubic phase of camn 7 o 12. *Physical Review B*, 95(14):144112, 2017.
- [54] Francesco Ricci, Wei Chen, Umut Aydemir, G Jeffrey Snyder, Gian-Marco Rignanese, Anubhav Jain, and Geoffroy Hautier. An ab initio electronic transport database for inorganic materials. *Scientific data*, 4(1):1–13, 2017.
- [55] F Aumento. Stability, lattice parameters, and thermal expansion of β -cristobalite. *American Mineralogist: Journal of Earth and Planetary Materials*, 51(7):1167–1176, 1966.
- [56] William M Haynes. *CRC handbook of chemistry and physics*. CRC press, 2016.
- [57] JP Wittmer, H Xu, P Polińska, C Gillig, J Helfferich, F Weysser, and J Baschnagel. Compressibility and pressure correlations in isotropic solids and fluids. *The European Physical Journal E*, 36(11):131, 2013.
- [58] Jack D Evans, Johannes P Dürholt, Stefan Kaskel, and Rochus Schmid. Assessing negative thermal expansion in mesoporous metal–organic frameworks by molecular simulation. *Journal of materials chemistry A*, 7(41):24019–24026, 2019.
- [59] W Miller, CW Smith, DS Mackenzie, and KE Evans. Negative thermal expansion: a review. *Journal of materials science*, 44(20):5441–5451, 2009.
- [60] Konstantin Stracke and Jack D Evans. Investigating the temperature-induced expansion of mil-53 under different gas environments using molecular simulations. *The Journal of Physical Chemistry C*, 129(6):3226–3233, 2025.
- [61] Bowen Deng, Yuneong 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(1), January 2025.
- [62] Logan M Brabson, Andrew J Medford, and David S Sholl. Comparing classical and machine learning force fields for modeling deformation of metal–organic frameworks relevant for direct air capture. *The Journal of Physical Chemistry C*, 129(37):16811–16825, 2025.
- [63] Jonathan Schmidt, Tiago FT Cerqueira, Aldo H Romero, Antoine Loew, Fabian Jäger, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, 48:101560, 2024.

- [64] Alice EA Allen, Nicholas Lubbers, Sakib Matin, Justin Smith, Richard Messerly, Sergei Tretiak, and Kipton Barros. Learning together: Towards foundation models for machine learning interatomic potentials with meta-learning. *npj computational materials*, 10(1):154, 2024.