Benchmarking Universal Machine Learning Interatomic Potentials for Elastic Property Prediction

Pengfei Gao
School of Intelligence Science and Technology, Nanjing University of Science and Technology, Jiangyin, Jiangsu 214443, China

Haidi Wang*
School of Physics, Hefei University of Technology, Hefei, Anhui 230009, China

Universal machine learning interatomic potentials have emerged as efficient tools for materials simulation, yet their reliability for elastic property prediction remains unclear. Here, we present a systematic benchmark of four uMLIPs—MatterSim, MACE, SevenNet, and CHGNet—against first-principles data for nearly 11 000 elastically stable materials from the Materials Project database. The results show that SevenNet achieves the highest accuracy, MACE and MatterSim balance accuracy with efficiency, while CHGNet performs less effectively overall. To further improve predictive quality, we perform targeted fine-tuning on all four uMLIPs using strained configurations derived from 185 high-error materials. After fine-tuning, CHGNet exhibits the largest overall improvement, with an average mean absolute percentage error reduction of about 23%, followed by MatterSim at around 21% and SevenNet at 18%, whereas MACE shows a performance degradation of roughly 14%. This work provides quantitative guidance for model selection and data refinement, advancing uMLIPs toward reliable applications in mechanical property prediction.

I. INTRODUCTION

Elastic properties [1], as fundamental properties of materials, play an important role in governing their mechanical behavior across a wide range of applications, from structural engineering to lithium battery systems [2–4], and related fields [5]. Accurate prediction of elastic constants and their derived mechanical parameters, such as bulk modulus, shear modulus, Young's modulus, and Poisson's ratio, is a critical task of computational materials design [6]. Although the modern density functional theory (DFT) [7–10] provides reliable and reproducible predictions of elastic properties, such calculations are often associated with high computational costs in high-throughput materials screening. Owing to this computational bottleneck, the systematic exploration of large chemical spaces is strictly constrained [11], which in turn hinders the efficient evaluation of elastic mechanical properties and delays materials design and discovery.

In recent years, machine learning interatomic potentials (MLIPs) [12–16] have rapidly emerged as important tools in materials simulation, offering an effective balance between the high accuracy of quantum mechanical calculations and the efficiency of classical potentials. Generally, these models are obtained by learning interatomic interactions from large-scale DFT data sets, and enable predictions with near-quantum accuracy while substantially reducing computational cost for crystal structure prediction [2, 17, 18], molecular dynamics simulation [19, 20], and related tasks [21–23]. Recent advances in graph neural networks, message-passing architectures, and equivariant representations have greatly improved the capabilities of MLIPs. These developments have led to the emergence of universal MLIPs (uMLIPs) [24–27], which can accurately model a wide range of chemical compositions and crystal structures. However, accurately predicting elastic properties requires a dependable evaluation of the second derivatives of the potential energy surface (PES), which introduces stricter and qualitatively different challenges than those encountered in predicting energies and forces.

So far, many efforts have been devoted to developing uMLIPs that improve the accuracy of energy, force, and stress predictions [28]. For instance, previous studies have shown that uMLIPs on the Matbench platform [29] perform well in structural optimization, structure prediction, and molecular dynamics simulations. However, their reliability and effectiveness in predicting elastic properties remain unexplored. This is because the relationship between energy–force accuracy and second-derivative precision is not straightforward, as elastic constants are highly sensitive to slight variations in the curvature of the PES, which are often difficult to capture with conventional training strategy. Therefore, analyzing the difference between the overall predictive accuracy and property-specific performance of uMLIPs, and evaluating their capability in mechanical property predictions, is of great importance.

In this work, we conduct a systematic evaluation to address the existing gap in crystal mechanical property research. Specifically, we employ four universal machine learning interatomic potentials (uMLIPs) — Crystal Hamiltonian Graph Neural Network (CHGNet) [27], MACE [30], MatterSim [31], and Scalable EquiVariance-Enabled Neural Network (SevenNet) [32] — to calculate the elastic properties of 10 994 crystal structures from the Materials Project database [11, 33–38], and systematically compare the results with the DFT reference data provided therein. We further quantify model perfor-

* haidi@hfut.edu.cn

mance differences in key indicators such as shear modulus, bulk modulus, Young's modulus, Poisson's ratio, and mechanical stability, as well as computational efficiency. Moreover, we introduce a targeted fine-tuning scheme that augments uMLIP training with strained configurations from high-error materials, enabling a direct evaluation of how incorporating non-equilibrium data affects mechanical predictive accuracy. Building on these analyses, we propose evidence-based guidelines for the suitable selection of uMLIPs and demonstrate that targeted fine-tuning can further enhance their accuracy and reliability in predicting elastic properties.

II. METHODOLOGY

A. Dataset Construction and Analysis

In this work, we collected 10,994 structures with reported elastic properties from the Materials Project database. Among them, 10,871 structures were mechanically stable at the DFT level, and these were used as our benchmark dataset. In Fig. 1(a), we present the distribution of elements. The nonmetals such as B, C, N, and O, main-group metals like Li and Mg, and transition metals including Ni, Cu, Zn, and Ti appear most often. Heavy and radioactive elements are rare. From a crystallographic perspective (see Fig. 1(b)), the dataset covers seven crystal systems. Cubic structures are the most common (23%), followed by tetragonal (20%) and orthorhombic (19%). Trigonal and monoclinic systems make up 16% and 12%, while hexagonal (7%) and triclinic (3%) systems are less frequent. In total, 169 space groups are represented, giving wide crystallographic diversity. Finally, from the distribution of the number of atoms in Fig. 1(c), we find that most structures have fewer than 20 atoms per unit cell, with 5-10 atoms being the most typical, and structures with more than 30 atoms are uncommon.

In Fig. 2, we present the distribution of the band gap, formation energy, and basic mechanical properties of the dataset, showing that these quantities exhibit a broad and diverse distribution. The statistical analysis shows that 3,248 materials (29.9%) are semiconductors or insulators, with an average band gap of 0.69 eV, while the remaining 7,623 materials (70.1%) are metallic. For the semiconductor subset, as illustrated in Fig. 2(a), the majority of structures possess negative formation energies (mean: -0.90 ± 0.98 eV/atom) and energy above hull values (mean: 0.03 ± 0.10 eV/atom) close to zero, indicating good thermodynamic stability. Regarding mechanical properties in Fig. 2(b), the dataset shows that the bulk moduli range from 0.33 to 491.33 GPa (mean: 104.41 ± 73.73 GPa), shear moduli from 0.45 to 525.42 GPa (mean: 50.93 ± 44.22 GPa), and Poisson's ratios from -0.48 to 0.80 (mean: 0.29 ± 0.07). Overall, the dataset demonstrates strong representativeness in electronic, thermodynamic, and mechanical domains, providing a reliable sample for evaluating elastic properties in real materials.

B. Brief Description of Evaluated uMLIPs

In this work, four state-of-the-art uMLIPs were selected for comprehensive evaluation based on their elastic applications.

In CHGNet [27], the total potential energy is expressed as

$$E_{\text{tot}} = \sum_i L_3 \circ g \circ L_2 \circ g \circ L_1\big(\mathbf{v}_i^{(4)}\big) \tag{1}$$

where L_1, L_2, and L_3 are successive linear transformations, and g is a nonlinear activation function (typically the SiLU function). The vector $\mathbf{v}_i^{(4)}$ represents the final latent feature of atom i, obtained after four message-passing layers that aggregate both local bonding environments and longer-range structural correlations. Through this hierarchical transformation, each atomic environment is mapped to a high-dimensional representation that captures the coupling between geometric and electronic degrees of freedom. The total energy E_{tot} is then constructed as a smooth, differentiable function of all atomic positions, ensuring physical consistency between predicted energies, forces, and stresses. CHGNet enhances this framework by embedding charge information into the latent space via magnetic moment constraints, which effectively incorporate electronic-structure effects into the learned potential. This charge-informed representation enables the model to distinguish between different ionic states and electronic configurations, a capability essential for accurately describing materials where charge redistribution and orbital occupancy govern structural stability, phase behavior, and transport properties.

MACE [30] advances interatomic potential modeling by combining the systematic completeness of Atomic Cluster Expansion (ACE) with the higher-order equivariant message passing of modern graph neural networks. Unlike conventional message-passing neural networks that primarily encode two-body interactions and rely on deep stacking to capture higher-order correlations, MACE constructs explicit many-body messages within each layer through a hierarchical expansion,

$$\begin{aligned} m_i^{(t)} = &\sum_j u_1\big(\sigma_i^{(t)};\sigma_j^{(t)}\big) + \sum_{j_1,j_2} u_2\big(\sigma_i^{(t)};\sigma_{j_1}^{(t)},\sigma_{j_2}^{(t)}\big) + \cdots \\ &+ \sum_{j_1,\ldots,j_\nu} u_\nu\big(\sigma_i^{(t)};\sigma_{j_1}^{(t)},\ldots,\sigma_{j_\nu}^{(t)}\big) \end{aligned} \tag{2}$$

where $m_i^{(t)}$ is the message received by atom i at layer t, $\sigma_i^{(t)} = (\mathbf{r}_i, z_i, h_i^{(t)})$ denotes its geometric, chemical, and latent state, and u_ν are learnable tensorial functions encoding correlations up to body order $(\nu+1)$. This formulation embeds the full hierarchy of local many-body interactions directly into each message-passing step, making

FIG. 1. Crystal structure analysis of the dataset. (a) Periodic table heatmap indicating element occurrence. (b) Sunburst plot illustrating the distribution of crystal systems and space groups, with integers placed at the outer margins representing the corresponding space-group numbers. (c) Histogram of atom counts per unit cell.

the representation both equivariant under $E(3)$ transformations and systematically improvable by increasing the correlation order ν. As a result, MACE achieves the accuracy of high-order ACE potentials with only two network layers, while maintaining linear scaling, GPU-friendly parallelism, and full physical symmetry. This fusion of traditional many-body theory and modern equivariant learning enables MACE to deliver quantum-level precision and computational efficiency, bridging the gap between explicit many-body potentials and scalable neural force-field architectures.

The MatterSim potential [31] is a large-scale, symmetry-preserving machine-learning force field that combines the M3GNet architecture with a periodic-aware Graphormer backbone. Each atomic structure is represented as a graph $G = (Z, V, R, [L, S])$, where atomic nodes Z carry feature vectors v_i, edges V connect atom pairs (i, j) within a cutoff radius r_c, $R = \{r_i\}$ are atomic coordinates, and $[L, S]$ encodes the global lattice and thermodynamic state. In the M3GNet message-passing block, each edge feature e_{ij} represents the bond between atoms i and j, including pairwise information such as chemical type and interatomic distance $r_{ij} = \|r_i - r_j\|$. To incorporate three-body geometry, e_{ij} is refined through a spherical-Bessel / spherical-harmonic expansion of its neighboring environment:

$$\tilde{e}_{ij} = \sum_k j_\ell\Big(\tfrac{z_{\ell n}\|r_{ik}\|}{r_c}\Big) Y_\ell^0(\theta_{jik}) \otimes \sigma(W_v v_k + b_v) f_c(\|r_{ij}\|) f_c(\|r_{ik}\|) \tag{3}$$

where $r_{ij} = r_i - r_j$ and θ_{jik} denotes the angle between bonds e_{ij} and e_{ik}; the functions j_ℓ and Y_ℓ^0 correspond to spherical Bessel functions and spherical harmonics with roots $z_{\ell n}$,

$$f_c(r) = 1 - 6(r/r_c)^5 + 15(r/r_c)^4 - 10(r/r_c)^3$$

is a smooth cutoff ensuring continuity at r_c, and W, b are learnable weights and biases. The intermediate term $\tilde{e}_{ij}$ aggregates angular information from neighboring atoms

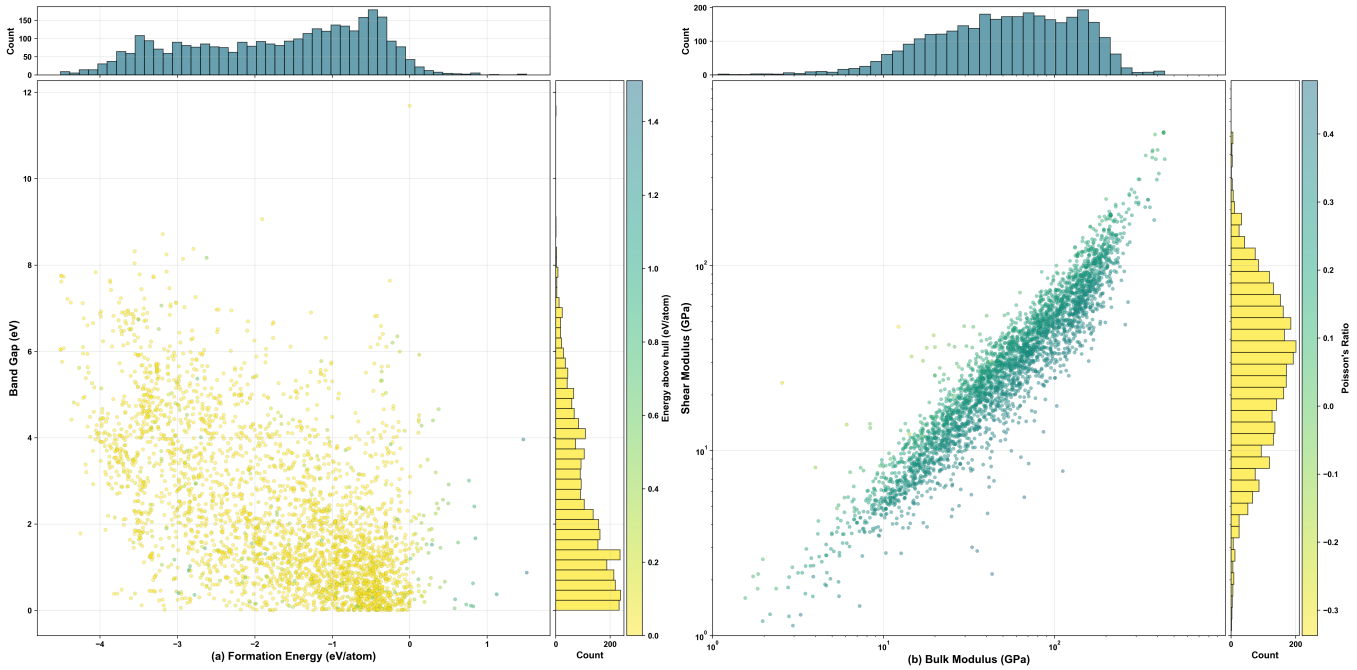

FIG. 2. Scatter plot of formation energy versus band gap, color-coded by the energy above hull. Marginal histograms illustrate the distribution of formation energies and band gaps. (b) Elastic property correlations. Scatter plot of bulk modulus versus shear modulus (log scale), color-coded by Poisson's ratio. Marginal histograms show the distributions of bulk and shear moduli.

k, and the updated edge feature e'_{ij} is obtained through nonlinear mixing:

$$e'_{ij} = e_{ij} + g(\tilde{W}_2\tilde{e}_{ij} + \tilde{b}_2) \otimes \sigma(\tilde{W}_1\tilde{e}_{ij} + \tilde{b}_1) \qquad (4)$$

where σ is the sigmoid activation and $g(x) = x\,\sigma(x)$. Here, e_{ij} encodes pairwise interactions, $\tilde{e}_{ij}$ introduces three-body angular geometry, and e'_{ij} forms the refined many-body bond embedding passed to the next layer. Built upon these physically grounded descriptors and extended with long-range, periodic-aware attention, MatterSim achieves robust generalization and order-of-magnitude accuracy improvements over previous universal machine-learning force fields, trained on more than 17 million first-principles structures spanning diverse compositions and thermodynamic conditions.

SevenNet [32] (Scalable EquiVariance-Enabled Neural NETwork) follows the atom-decomposed energy formalism widely used in machine-learned interatomic potentials. This locality ensures that the computational cost scales linearly with the number of atoms, $\mathcal{O}(N)$, enabling large-scale molecular dynamics with thousands to millions of atoms. At each message-passing layer t, atomic features are updated by

$$m_v^{(t+1)} = \sum_{w\in\mathcal{N}(v)} M_t\big(h_v^{(t)}, h_w^{(t)}, e_{vw}^{(t)}\big), \qquad (5)$$

$$h_v^{(t+1)} = U_t\big(h_v^{(t)}, m_v^{(t+1)}\big). \qquad (6)$$

where M_t and U_t are learnable equivariant mappings that propagate geometric information between atoms while preserving rotational and permutational symmetry. The edge feature $e_{vw}^{(t)}$ is constructed from the relative displacement vector $r_{vw} = r_w - r_v$ and encodes both its magnitude and orientation, ensuring proper transformation under three-dimensional rotations. SevenNet extends the NequIP architecture by reorganizing its forward and reverse communication for efficient spatial domain decomposition.

C. Elastic Property Calculations

The second-order elastic constants (C_{ij}) were calculated using the stress-strain method [39]. According to Hooke's law, the relationship between stress σ_{ij} and strain ε_{kl} with Voigt notation can be expressed as:

$$\sigma_i = C_{ij}\varepsilon_j \quad (i,j = \{1,2,3,4,5,6\}) \qquad (7)$$

The elastic tensor components are determined by applying systematic deformations to the equilibrium crystal structure and computing the resulting stress response. Based on the model-predicted stress and applied strain, the elastic constants C_{ij} are obtained through linear fitting.

For structure optimization and elastic-property calculations, we employ the Atomic Simulation Environment (ASE) [40, 41] and Pymatgen [42]. The FIRE algorithm [43] is used for energy minimization and the Fretch-CellFilter [40] is applied to preserve space group symmetry during relaxation. The force convergence criterion

was set to 0.1 eV/Åfor structure relaxation, ensuring mechanical equilibrium before strain application.

Once the elastic tensor is obtained, the bulk modulus, shear modulus, Young's modulus, Poisson's ratio and other derived mechanical properties can be calculated. In this work, all derived mechanical properties are Voigt-Reuss-Hill average values via MechElastic [44] analysis module. The Young's modulus E and Poisson's ratio ν are obtained based on the bulk modulus K and shear modulus G as follows:

$$E = \frac{9KG}{3K + G}, \tag{8}$$

$$\nu = \frac{3K - 2G}{2(3K + G)}. \tag{9}$$

III. RESULTS AND ANALYSIS

A. Model Performance Analysis

In this section, we systematically evaluate the performance of different uMLIP models in predicting elastic properties and classifying material stability, using DFT results as the reference. By incorporating both distributional comparisons and point-wise analyses, the models are assessed from the perspectives of global trends and local accuracy. Furthermore, we also analyzed the stability classification results to provide a comprehensive picture of model applicability in elasticity-related tasks.

As shown in Fig. 3, we compare the distributions of bulk, shear, and Young's moduli and Poisson's ratio obtained from DFT and from the four uMLIP models. Throughout this paragraph, values given in parentheses denote mean model predictions. Overall, all models reproduce the macroscopic DFT trends, but systematic deviations remain in their absolute magnitudes. For the bulk modulus, the DFT mean is 104.1 GPa, while the model means lie in the range 100–115 GPa, indicating robust performance in capturing volumetric compressibility. In contrast, discrepancies are more pronounced for the shear and Young's moduli: the DFT means are 47.3 and 122.5 GPa, whereas CHGNet yields the lowest values (28.6 and 77.5 GPa), systematically underestimating rigidity; MACE (58.2 and 148.1 GPa) and SevenNet (56.3 and 143.9 GPa) overestimate both moduli, reflecting a tendency to over-enhance stiffness; MatterSim (50.8 and 130.5 GPa) gives intermediate results that remain closest to the DFT benchmarks. For Poisson's ratio, the DFT mean is 0.291: CHGNet substantially overestimates it (0.371), implying artificially high ductility, MACE and SevenNet slightly underestimate it (0.279 and 0.282), and MatterSim (0.294) is nearly indistinguishable from DFT. Taken together, these distributions indicate that, while all models capture the overall elastic trends, notable systematic biases persist, particularly for the shear and Young's moduli and for Poisson's ratio.

To enable quantitative evaluation, we present point-wise comparisons of the primary elastic properties in Fig. 4. For the bulk modulus, SevenNet and MACE exhibit the highest consistency with DFT, achieving correlation coefficients of approximately $R \approx 0.94$ and mean absolute errors (MAE) around 15 GPa, outperforming both CHGNet ($R = 0.909$) and MatterSim ($R = 0.924$). For the shear modulus, MACE attains the highest correlation ($R = 0.896$), followed by SevenNet ($R = 0.895$), while MatterSim yields intermediate accuracy ($R = 0.847$) and CHGNet remains significantly weaker ($R = 0.546$). Regarding the Young's modulus, MatterSim yields the mean closest to DFT, but in this correlation-centric assessment MACE attains the higher correlation (R = 0.901) than MatterSim (R = 0.860); SevenNet is lower (R = 0.791), and CHGNet remains the weakest (R = 0.546). For the Poisson's ratio, a different trend emerges: MACE and MatterSim achieve significantly higher correlations ($R \approx 0.65$) than CHGNet ($R = 0.301$) and SevenNet ($R = 0.374$), indicating their robustness in capturing ratio-type properties. Overall, MACE and SevenNet alternate in leading performance depending on the property considered, while MatterSim also exhibits consistently reliable behavior, achieving mean values closest to DFT and competitive correlations across most properties. The relative superiority of these models remains task-dependent, reflecting the varying accuracy of current universal machine-learning interatomic potentials across different elastic property regimes.

Beyond elasticity, stability classification provides another essential benchmark for evaluating model performance. Fig. 5 compares the stability predictions of the four uMLIPs against DFT references. SevenNet and MACE achieve the highest performance, with accuracies of 98.3% and 98.1%, respectively, and F1 scores approaching 0.99, reflecting well-balanced capability in identifying both stable and unstable materials. MatterSim ranks closely behind, while CHGNet reaches only 93.4% accuracy, significantly lower than the others and showing a higher rate of missed unstable samples. The confusion matrix analysis further indicates that both MACE and SevenNet exhibit consistently high precision (≈ 0.997) and recall (>0.98), underscoring their robustness and reliability in large-scale stability screening tasks. Cross-model agreement analysis indicates strong overlap, with more than 10,700 materials consistently classified by both MACE and SevenNet in line with DFT. This highlights their superior generalization ability in stability classification across diverse material systems.

In addition, analysis of the computational efficiency for elastic property evaluations, as shown in Fig. S1, reveals that MACE achieves the best overall performance, with an average processing time of 1.132 seconds per structure and the lowest standard deviation of 0.061 seconds. CHGNet follows closely, with an average of 1.212 seconds per structure. MatterSim has an average processing time of 1.853 seconds per structure but exhibits high

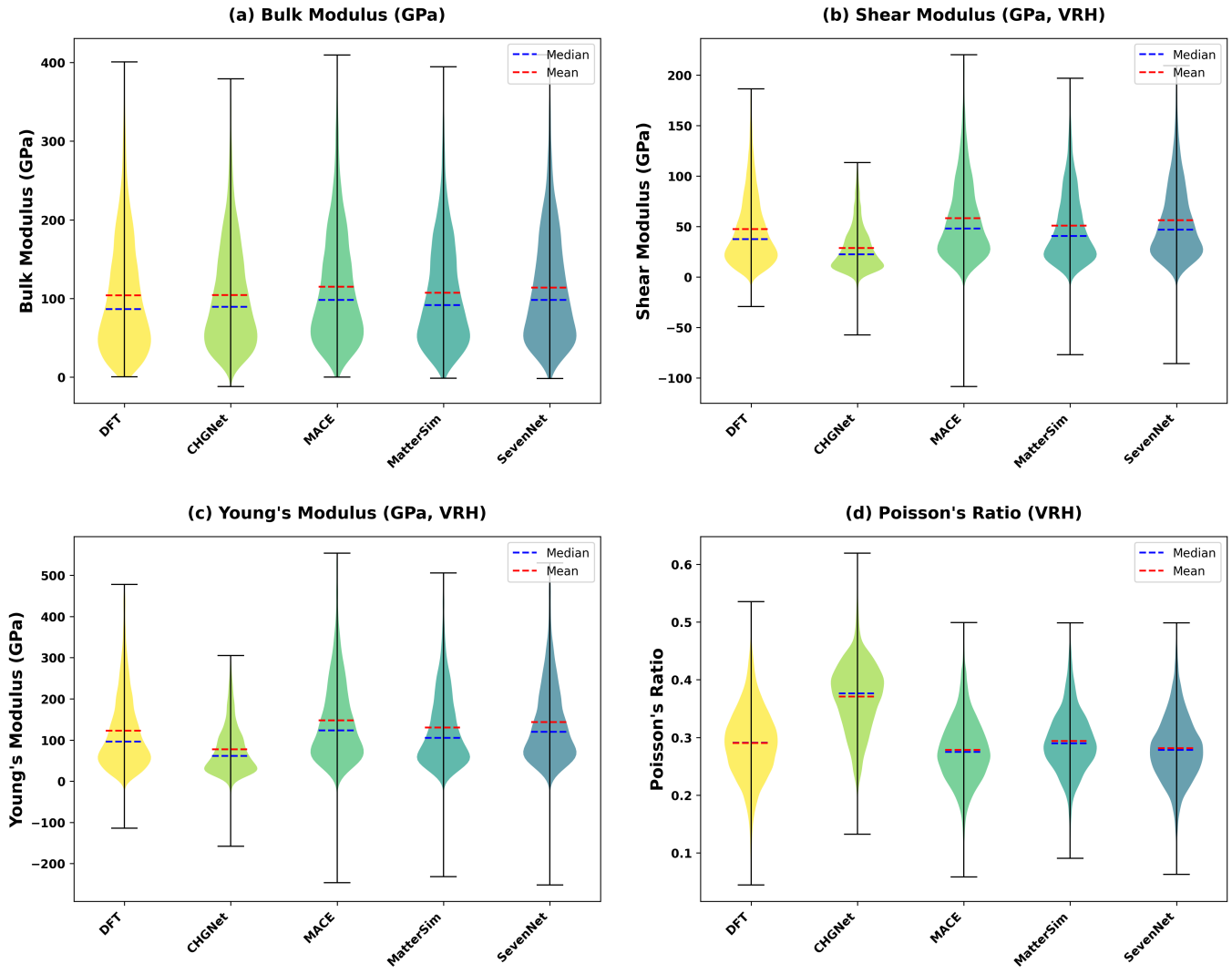

FIG. 3. Distributions of (a) bulk modulus, (b) shear modulus, (c) Young's modulus, and (d) Poisson's ratio, computed as Voigt–Reuss–Hill (VRH) averages from DFT and the four uMLIP models. Each violin plot shows the overall distribution, with blue and red dashed lines marking the median and mean, and short lines denoting the extrema.

standard deviation of 0.710 seconds, likely influenced by material complexity. Due to its large number of parameters, SevenNet has the highest computational cost, with an average processing time of 2.770 seconds per structure, 2.4 times that of the fastest model.

B. Systematic Error Analysis

To gain deeper insight into the systematic biases of different machine-learning potentials in predicting elastic properties, this section conducts a comprehensive evaluation by combining relative error distributions with mean absolute percentage error (MAPE). The joint analysis of boxplots and heatmaps reveals both the bias patterns in individual property predictions and the overall performance trends across models.

Fig. 6 presents the relative error distributions of the four uMLIPs with respect to DFT values across various elastic descriptors and the values in parentheses in this paragraph correspond to median relative errors. CHGNet exhibits pronounced systematic deviations across most properties. In bulk modulus predictions, it shows a median error of −2.61%, indicating a tendency toward underestimation, whereas MACE and SevenNet display slight overestimations (4.16% and 2.88%, respectively), and MatterSim remains close to zero bias (−1.98%). For the shear modulus and Young's modulus, CHGNet strongly underestimates both (−48.02% and −44.20%), in sharp contrast to the overestimations observed for MACE (13.83% and 12.43%) and SevenNet (9.79% and 8.89%), while MatterSim again yields nearly symmetric distributions (−2.12% and −2.24%). For Poisson's ratio, CHGNet systematically overesti-

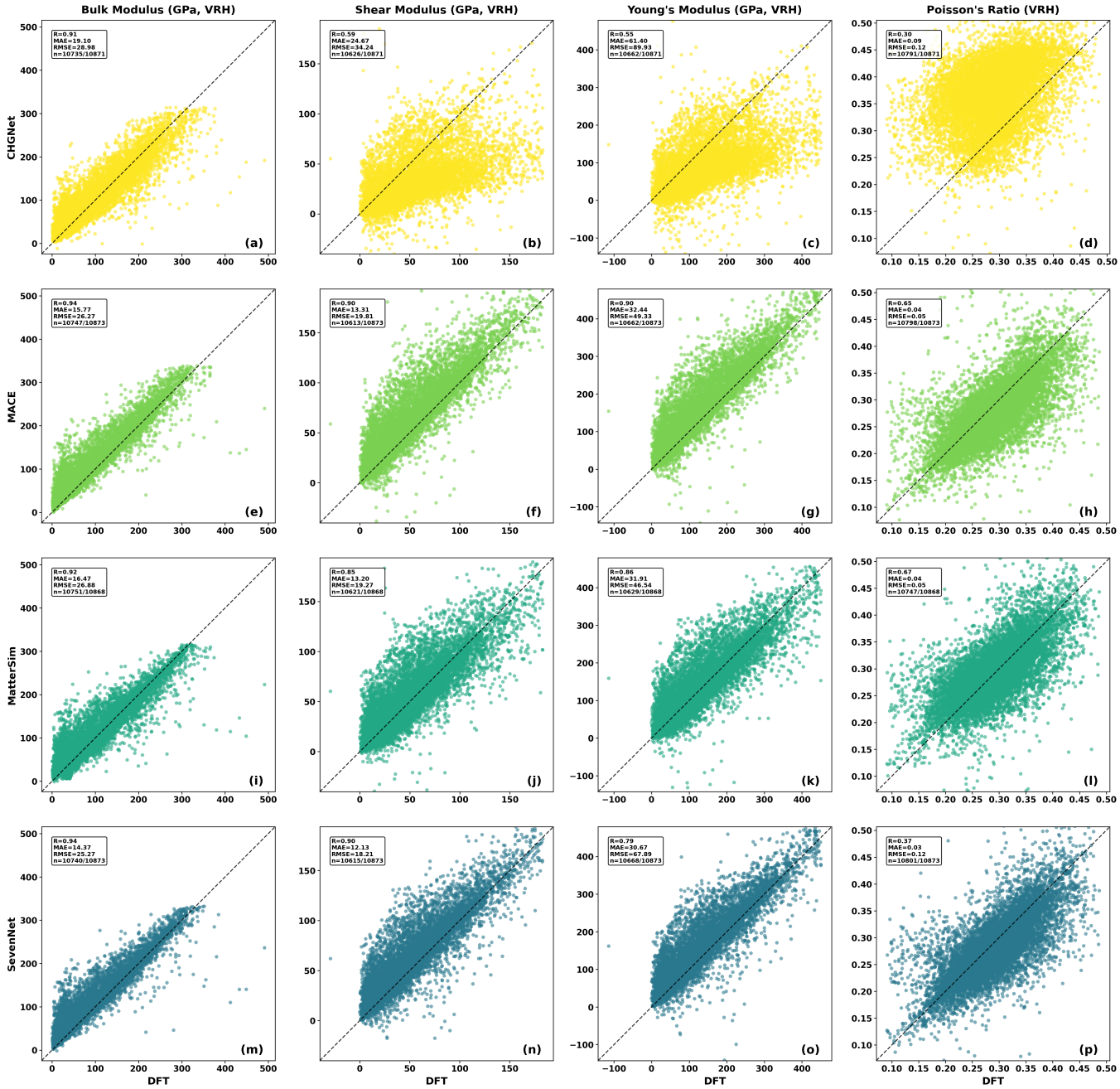

FIG. 4. (a)–(h) Scatter-plot comparison of four uMLIPs against DFT reference values for primary elastic properties—bulk modulus, shear modulus, Young's modulus, and Poisson's ratio, all computed as VRH averages. Each panel shows DFT values on the x-axis and model predictions on the y-axis, with the dashed line representing perfect agreement.

mates (27.25%), opposite to the mild underestimations of MACE (−4.35%) and SevenNet (−3.40%), whereas MatterSim remains almost unbiased (0.70%). The bulk/shear ratio further highlights CHGNet's strong positive bias (77.05%), while the other models show values close to zero. CHGNet also shows especially high variability in anisotropy metrics, suggesting instability in capturing complex anisotropic behavior. For Cauchy pressure, CHGNet exhibits a systematic positive bias, whereas the others lean toward negative deviations. The large deviations in the predicted anisotropy and Cauchy pressure mainly reflect the high sensitivity of these quantities to small differences among elastic constants. In this work, most materials exhibit a relatively small degree of elastic anisotropy, with a DFT average of 1.97. The Cauchy pressure, being a difference quantity defined as $(C_{12} - C_{44})$, has a DFT average value of 17.9 GPa, which is much smaller than the bulk and Young's mod-

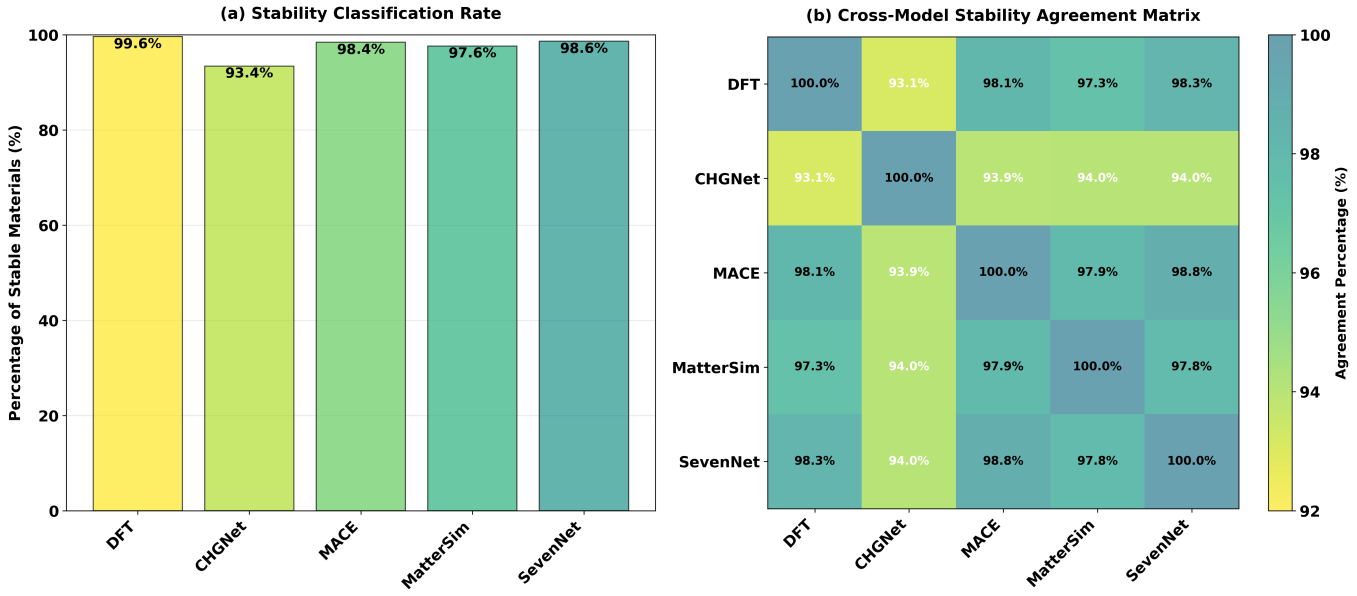

FIG. 5. Elastic stability classification analysis comparing DFT and uMLIP models. (a) Stability classification rate showing the percentage of materials predicted as stable by each model. (b) Pairwise stability-agreement matrix, showing the percentage of materials assigned the same stability outcome across model pairs. Stability was determined using reported elastic stability flags or Born mechanical stability criteria.

ulus that are generally on the order of 100 GPa. Consequently, even moderate relative errors in the stiffness components can lead to large percentage deviations in these derived quantities, indicating that further development of uMLIPs to improve their accuracy in elastic property predictions is essential. Finally, in Debye temperature predictions, CHGNet again underestimates (−25.89%), while MACE (6.55%) and SevenNet (4.89%) perform closer to DFT, and MatterSim achieves the most balanced performance (−0.69%). Overall, CHGNet displays consistent systematic biases across multiple properties, whereas MACE, MatterSim, and SevenNet yield more symmetric, near-zero error distributions, reflecting higher robustness.

To provide a clearer comparison of overall performance, Fig. 7 summarizes the MAPE values of different models across all elastic properties. It is evident that CHGNet systematically yields the highest error levels, with an average MAPE of 71.8%, underscoring its structural deficiencies in elastic property prediction. In contrast, SevenNet consistently achieves the lowest error, with an average MAPE of only 27.53%, highlighting its superior overall accuracy. Further analysis reveals that differences among models are relatively small for bulk and Young's modulus, whereas much larger discrepancies arise in the bulk/shear ratio, universal anisotropy index and Cauchy pressure—properties closely linked to mechanical stability and anisotropy. Particularly noteworthy is that CHGNet's MAPE exceeds 90% for these metrics, reflecting structural limitations in capturing the complex couplings within elastic tensors. While MACE and MatterSim outperform CHGNet, their overall accuracy remains inferior to SevenNet, reinforcing the conclusion that SevenNet exhibits stronger generalization capability in modeling the nonlinear interdependencies among elastic properties.

C. Fundamental Limitations and the Impact of Fine-Tuning

The limited performance of current uMLIPs in elastic property prediction can be partially attributed to training datasets dominated by equilibrium configurations, which do not provide sufficient coverage of strained states that are important for learning accurate mechanical responses. Fine-tuning [45] provides a practical route to mitigating such systematic biases. To evaluate this strategy, we constructed a targeted fine-tuning dataset comprising 185 materials with the largest baseline errors, together with DFT-computed energies of their deformed structures, thereby explicitly introducing non-equilibrium information into the training domain. The DFT calculation settings are detailed in the Supporting Information. Fine-tuning on this targeted dataset yields well-controlled training mean absolute errors (MAEs): CHGNet achieved energy/force/stress MAEs of 24.089 meV/atom, 26.831 meV/Å, and 1.960 meV/Å^3; MACE obtained 6.17 meV/atom, 26.38 meV/Å, and 5.67 meV/Å^3; MatterSim reached 17.89 meV/atom, 9.574 meV/Å, and 0.836 meV/Å^3; and SevenNet yielded 1.692 meV/atom, 5.101 meV/Å, and 0.528 meV/Å^3.

All four uMLIPs—CHGNet, MACE, MatterSim, and SevenNet—were fine-tuned on this dataset. The updated

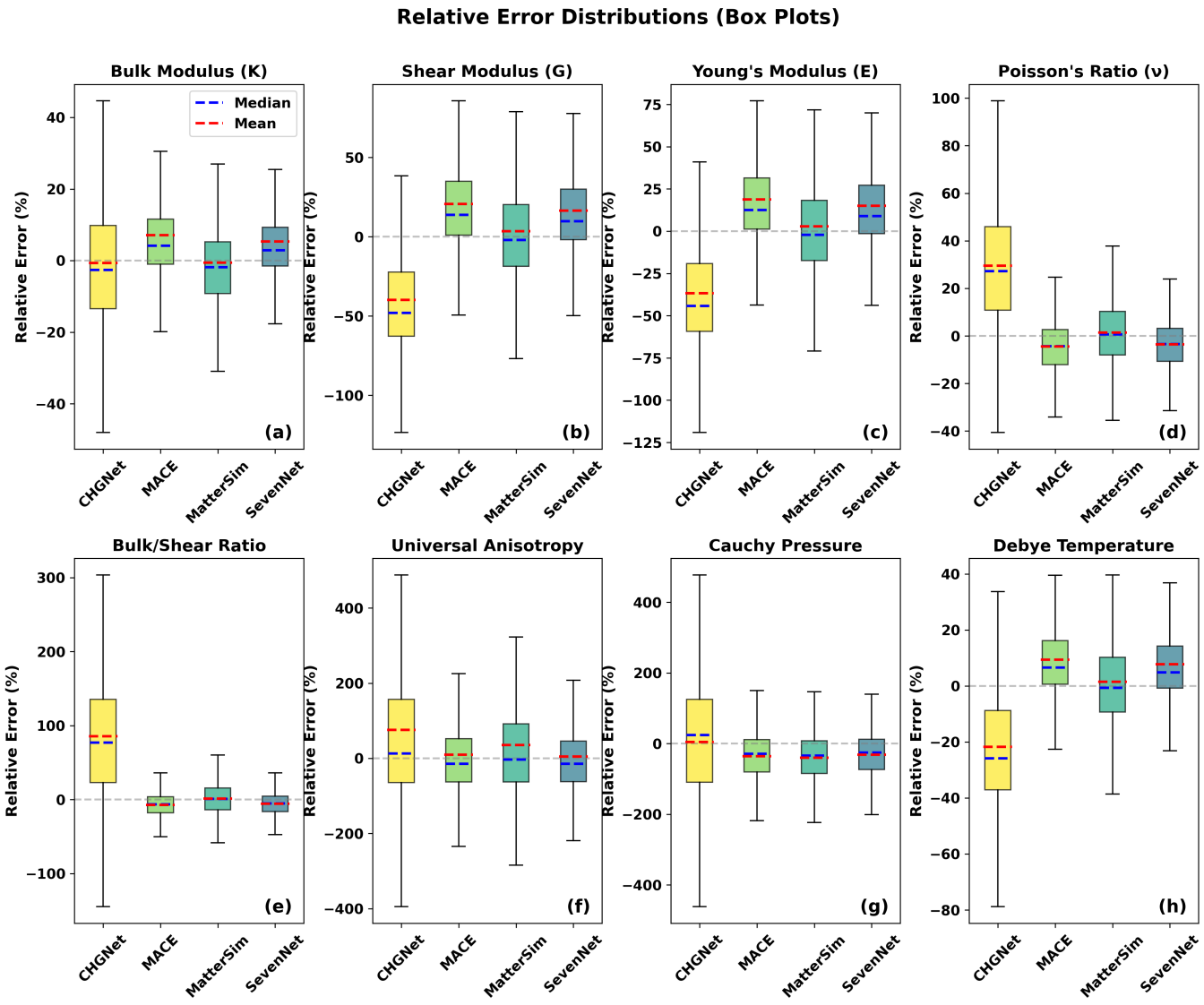

FIG. 6. (a)–(h) Distribution of relative errors (%) for CHGNet, MACE, MatterSim, and SevenNet compared with DFT reference values across eight elastic properties. Each boxplot shows the median (blue dashed line), mean (red dashed line), interquartile range (colored box), and the overall error range (short lines marking the extrema), with outliers omitted for clarity. The horizontal dashed line denotes zero error.

potentials were then reassessed on the same 185 materials. Fig. S2 and Fig. S3 display heatmaps of the MAPE before and after fine-tuning, while Fig. S4 and Fig. S5 present the corresponding distributions of relative errors. To quantify these effects, Table I summarizes the relative change in MAPE, defined as the difference between the fine-tuned MAPE and the original MAPE, normalized by the original value. CHGNet shows the largest improvement, with an average MAPE reduction of 23.2%, followed by MatterSim at 20.7% and SevenNet at 18.0%. In contrast, MACE's average MAPE increases by 13.8%, indicating a less favorable response to the added deformation information. Pronounced improvements occur for the two dimensionless ratios—the Poisson's ratio and the bulk-to-shear modulus ratio—whose MAPE reductions exceed 40% for CHGNet, MatterSim, and SevenNet. In comparison, the six dimensional elastic constants (e.g., bulk modulus, shear modulus, Young's modulus) typically show more modest reductions around 10–20%, highlighting a differential sensitivity to the inclusion of strained configurations. Table II reports the corresponding interquartile ranges (IQR), showing that fine-tuning leads to IQR reductions in seven of the eight elastic properties for CHGNet, and in all eight properties for both MatterSim and SevenNet. By contrast, MACE shows IQR reduction in only two properties and increased dispersion in the remaining six. These results demonstrate a clear difference in fine-tuning consistency among the four uMLIPs. For CHGNet, MatterSim, and SevenNet, the simultaneous improvements in both MAPE and IQR

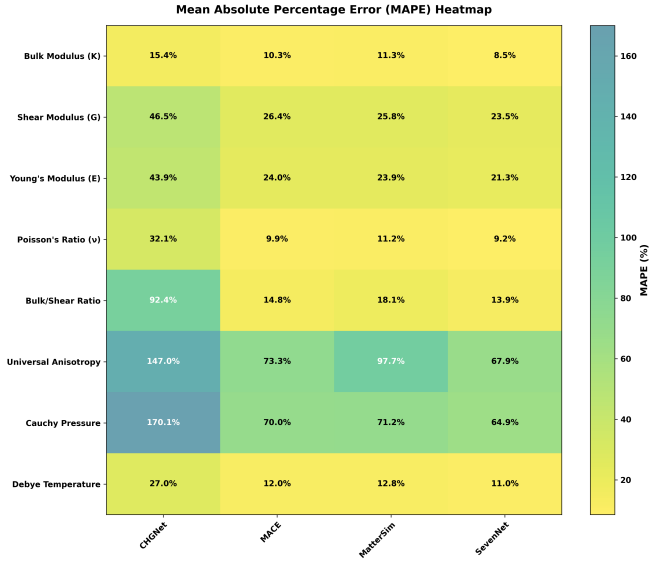

FIG. 7. MAPE heatmap for elastic properties predicted by CHGNet, MACE, MatterSim, and SevenNet relative to DFT reference values. The properties analyzed include bulk modulus, shear modulus, Young's modulus, Poisson's ratio, bulk/shear ratio, universal anisotropy index, Cauchy pressure, and Debye temperature. Darker colors indicate higher errors, with values annotated in each cell.

across all eight mechanical quantities indicate a coherent propagation of the energy–force–stress refinements into uniformly improved elastic responses. This reflects a high degree of physical consistency, where the optimization of the fundamental quantities translates directly into better predictions of derived mechanical observables. In contrast, the inconsistent behavior of MACE—improving in only a subset of properties while degrading in others—suggests weaker coupling between its learned potential-energy surface and its strain-dependent responses, pointing to limited fine-tuning consistency and reduced robustness to strained configurations. Such behavior illustrates that fine-tuning not only reduces systematic bias for CHGNet, MatterSim, and SevenNet but also yields more stable and concentrated error distributions. The degradation in MACE's performance may reflect overfitting or architectural sensitivities that warrant further investigation.

These findings confirm that targeted fine-tuning with non-equilibrium configurations provides an effective and physically consistent route for improving the mechanical predictive capability of CHGNet, MatterSim, and SevenNet, whereas MACE shows limited and inconsistent gains.

IV. DISCUSSION

A. Implications for Materials Design Applications

The benchmark and fine-tuning results discussed above provide clear guidance for selecting suitable uMLIPs according to specific application requirements. For tasks requiring highly accurate predictions of elastic properties, SevenNet should be prioritized; although its computational cost is somewhat higher, it offers more reliable performance. For high-throughput screening workflows, MACE and MatterSim strike a favorable balance between accuracy and efficiency, making them better suited for large-scale applications. While CHGNet shows comparatively weaker overall performance, it remains a viable option for simulations involving magnetic systems, where its specialized capabilities can be advantageous.

Systematic bias patterns observed across all models warrant careful consideration in practical applications. In particular, consistent tendencies toward underestimation or overestimation of elastic modulus highlight the need for bias-correction strategies. For quantitative materials design, it is recommended that final results be validated against high-accuracy DFT calculations to ensure reliability.

B. Future Directions

As demonstrated in our fine-tuning study, incorporating strained configurations into the training process can effectively improve the predictive accuracy of uMLIPs for elastic properties, underscoring the importance of dataset diversity. Future developments should focus on systematically incorporating deformed structures into training datasets, for instance through active learning strategies that aim to improve mechanical property accuracy [46]. For elastic property predictions within specific chemical spaces, constructing domain-specific fine-tuning [28] datasets and adapting pretrained models accordingly could effectively mitigate systematic biases in those regions.

Improving computational efficiency remains essential for the broader adoption of uMLIPs in materials design workflows. Although current models deliver significant speedups compared to DFT, computational demands remain high when scaling to datasets containing hundreds of thousands of materials. Further optimization is therefore critical. Future developments should focus on developing hybrid frameworks that couple large-scale, low-cost screening with targeted high-accuracy calculations to ensure both efficiency and reliability in practical applications.

TABLE I. Relative percentage change in the MAPE of eight elastic properties for the four uMLIPs after fine-tuning.

Property	**CHGNet (%)**	**MACE (%)**	**MatterSim (%)**	**SevenNet (%)**
Bulk Modulus K	11.8	-5.2	-16.5	-2.7
Shear Modulus G	-5.3	-14.6	-15.5	-5.7
Young's Modulus E	-9.2	-9.9	-18.2	-4.5
Poisson's Ratio ν	-56.6	18.4	-42.4	-48.5
Bulk/Shear Ratio	-55.4	89.4	-39.8	-41.3
Universal Anisotropy Index	-31.6	37.6	-11.8	-13.9
Cauchy Pressure	-22.4	-1.0	-9.6	-18.6
Debye Temperature	-17.1	-4.3	-11.3	-9.1
Average	-23.2	13.8	-20.7	-18.0

TABLE II. IQR of the relative error for eight elastic properties before and after fine-tuning. Arrows indicate whether the IQR decreases (↓) or increases (↑).

Property	**CHGNet (%)**	**MACE (%)**	**MatterSim (%)**	**SevenNet (%)**
Bulk Modulus K	236.2 → 242.4 (↑)	268.7 → 250.4 (↓)	270.6 → 208.6 (↓)	262.6 → 260.2 (↓)
Shear Modulus G	231.1 → 213.7 (↓)	227.1 → 236.0 (↑)	249.4 → 193.9 (↓)	232.4 → 219.9 (↓)
Young's Modulus E	272.7 → 236.6 (↓)	251.0 → 258.5 (↑)	267.6 → 217.1 (↓)	258.6 → 241.9 (↓)
Poisson's Ratio ν	69.4 → 29.0 (↓)	36.1 → 52.4 (↑)	46.2 → 20.6 (↓)	37.5 → 19.5 (↓)
Bulk/Shear Ratio	88.2 → 53.0 (↓)	51.5 → 104.2 (↑)	60.8 → 35.5 (↓)	54.2 → 28.8 (↓)
Universal Anisotropy Index	146.4 → 144.9 (↓)	127.1 → 162.2 (↑)	116.6 → 113.2 (↓)	127.2 → 119.7 (↓)
Cauchy Pressure	174.1 → 134.0 (↓)	80.1 → 147.4 (↑)	138.2 → 125.6 (↓)	145.4 → 116.0 (↓)
Debye Temperature	99.5 → 78.9 (↓)	85.2 → 99.9 (↑)	91.4 → 77.0 (↓)	86.6 → 81.6 (↓)

V. CONCLUSIONS

Our benchmark study establishes the first systematic evaluation framework for applying uMLIPs to elastic property prediction, validated across nearly 11,000 crystalline materials. The results demonstrate clear differences in model suitability: SevenNet delivers the highest overall accuracy, MatterSim and MACE achieve a favorable balance between accuracy and computational efficiency, while CHGNet, constrained by its architectural design, performs relatively less effectively. These results provide quantitative, evidence-based guidance for selecting suitable uMLIPs in mechanical property calculations, ensuring that model performance is properly matched to application requirements.

Comprehensive analyses of systematic biases further reveal common limitations among current uMLIPs, including the consistent under- or overestimation of elastic modulus and a training-data bias toward equilibrium configurations. Our fine-tuning results, incorporating strained configurations into model training, show that even limited data augmentation can effectively mitigate such biases and enhance model robustness. CHGNet, MatterSim, and SevenNet show uniform improvements across nearly all mechanical quantities, indicating coherent and physically consistent fine-tuning behavior, whereas MACE displays smaller and more variable gains. Building on these insights, we identify several promising directions for future development, such as incorporating strained structures through active learning, implementing property-specific fine-tuning protocols, and establishing systematic error-correction schemes. We anticipate that such advances will further improve the reliability of uMLIPs for quantitative and high-throughput materials design, while also laying the groundwork for the next generation of universal interatomic potentials.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (22203026, 22303040) and the Fundamental Research Funds for the Central Universities (JZ2024HGTB0162). We also acknowledge the Materials Project team for providing comprehensive elastic property data and maintaining the open-access materials database.

DECLARATIONS

During manuscript preparation, we utilized GPT-5 LLM to enhance sentence structure, readability, and coherence. We also used Claude 4 LLM for optimizing the Python plotting and simulation scripts.

AUTHOR CONTRIBUTIONS

Pengfei Gao performed the calculations, analyzed the results, and contributed to manuscript preparation.

Haidi Wang conceived the study, performed the calculations, analyzed the results, and wrote the manuscript. All authors contributed to discussions and approved the final version of the paper.

COMPETING INTERESTS

The authors declare no competing financial or non-financial interests.

DATA AVAILABILITY

All 10,994 crystal structures and the corresponding DFT-computed elastic property data used in this study are publicly available through the Materials Project database (`https://materialsproject.org`).

CODE AVAILABILITY

All analysis code, model evaluation scripts, and data processing workflows are freely available at `https://gitee.com/haidi-hfut/umlip-elastic`. The repository includes the parameter files for the four evaluated uMLIPs and the usage instructions for the scripts for reproducing the analyses and figures.

[1] E. Schreiber, O. L. Anderson, N. Soga, and J. F. Bell, Journal of Applied Mechanics **42**, 747 (1975).
[2] C. Wu, T. Kim, S.-B. Lee, M.-K. Um, S.-K. Lee, W.-Y. Lai, J.-H. Byun, and T.-W. Chou, Composites Science and Technology **229**, 10.1016/j.compscitech.2022.109714 (2022).
[3] M. Kim, Z. Yang, and I. Bloom, Journal of The Electrochemical Society **168**, 010523 (2021).
[4] H. Xie, B. Han, H. Song, X. Li, Y. Kang, and Q. Zhang, Journal of the Mechanics and Physics of Solids **156**, 104602 (2021).
[5] A. Loew, D. Sun, H.-C. Wang, S. Botti, and M. A. Marques, npj Computational Materials **11**, 178 (2025).
[6] R. Yu, J. Zhu, and H. Ye, Computer Physics Communications **181**, 671 (2010).
[7] J. Hafner, C. Wolverton, and G. Ceder, MRS Bulletin **31**, 659 (2006).
[8] P. Hohenberg and W. Kohn, Physical Review **136**, B864 (1964).
[9] W. Kohn and L. J. Sham, Physical Review **140**, A1133 (1965).
[10] S. Curtarolo, D. Morgan, and G. Ceder, Calphad **29**, 163 (2005).
[11] M. De Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. Van Der Zwaag, J. J. Plata, *et al.*, Scientific Data **2**, 1 (2015).
[12] T. Mueller, A. Hernandez, and C. Wang, The Journal of Chemical Physics **152**, 050902 (2020).
[13] P. Friederich, F. Häse, J. Proppe, and A. Aspuru-Guzik, Nature Materials **20**, 750 (2021).
[14] H. Gokcan and O. Isayev, Wiley Interdisciplinary Reviews: Computational Molecular Science **12**, e1564 (2022).
[15] V. L. Deringer, M. A. Caro, and G. Csányi, Advanced Materials **31**, 1902765 (2019).
[16] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, The Journal of Physical Chemistry C **121**, 511 (2017).
[17] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, Physical Review B **99**, 064114 (2019).
[18] M. L. Paleico and J. Behler, The Journal of Chemical Physics **153**, 054704 (2020).
[19] Z. Xu, H. Duan, Z. Dou, M. Zheng, Y. Lin, Y. Xia, H. Zhao, and Y. Xia, npj Computational Materials **9**, 105 (2023).
[20] L. Zhang, H. Wang, R. Car, and W. E, Physical Review Letters **126**, 236001 (2021).
[21] C. W. Rosenbrock, K. Gubaev, A. V. Shapeev, L. B. Pártay, N. Bernstein, G. Csányi, and G. L. Hart, npj Computational Materials **7**, 24 (2021).
[22] M. Kulichenko, B. Nebgen, N. Lubbers, J. S. Smith, K. Barros, A. E. Allen, A. Habib, E. Shinkle, N. Fedik, Y. W. Li, *et al.*, Chemical Reviews **124**, 13681 (2024).
[23] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, *et al.*, The Journal of Physical Chemistry A **124**, 731 (2020).
[24] B. Focassio, L. P. M. Freitas, and G. R. Schleder, ACS Applied Materials & Interfaces **17**, 13111 (2025), https://doi.org/10.1021/acsami.4c03815.
[25] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, Chemistry of Materials **31**, 3564 (2019).
[26] C. Chen and S. P. Ong, Nature Computational Science **2**, 718 (2022).
[27] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, Nature Machine Intelligence **5**, 1031 (2023).
[28] B. Deng, Y. Choi, P. Zhong, J. Riebesell, S. Anand, Z. Li, K. Jun, K. A. Persson, and G. Ceder, npj Computational Materials **11**, 9 (2025).
[29] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, npj Computational Materials **6**, 138 (2020).
[30] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csányi, Advances in Neural Information Processing Systems **35**, 11423 (2022).
[31] H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S. Chen, C. Zeni, *et al.*, arXiv:2405.04967 (2024), arXiv:2405.04967.
[32] Y. Park, J. Kim, S. Hwang, and S. Han, Journal of Chemical Theory and Computation **20**, 4857 (2024).
[33] M. K. Horton, P. Huck, R. X. Yang, J. M. Munro, S. Dwaraknath, A. M. Ganose, R. S. Kingsbury, M. Wen, J. X. Shen, T. S. Mathis, *et al.*, Nature Materials **24**, 1522 (2025).

[34] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, APL Materials **1**, 011002 (2013).
[35] A. Jain, G. Hautier, S. P. Ong, C. J. Moore, C. C. Fischer, K. A. Persson, and G. Ceder, Physical Review B **84**, 045115 (2011).
[36] A. Wang, R. Kingsbury, M. McDermott, M. Horton, A. Jain, S. P. Ong, S. Dwaraknath, and K. A. Persson, Scientific Reports **11**, 15496 (2021).
[37] M. Aykol, S. S. Dwaraknath, W. Sun, and K. A. Persson, Science Advances **4**, eaaq0148 (2018).
[38] S. P. Ong, L. Wang, B. Kang, and G. Ceder, Chemistry of Materials **20**, 1798 (2008).
[39] T. H. K. Barron and M. L. Klein, Proceedings of the Physical Society **85**, 523 (1965).
[40] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, Journal of Physics: Condensed Matter **29**, 273002 (2017).
[41] S. R. Bahn and K. W. Jacobsen, Comput. Sci. Eng. **4**, 56 (2002).
[42] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Computational Materials Science **68**, 314 (2013).
[43] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, and P. Gumbsch, Physical Review Letters **97**, 170201 (2006).
[44] S. Singh, L. Lang, V. Dovale-Farelo, U. Herath, P. Tavadze, F.-X. Coudert, and A. H. Romero, Computer Physics Communications **267**, 108068 (2021).
[45] X. Liu, K. Zeng, Y. Wang, and T. Zhao, arXiv:2506.07401 (2025).
[46] Y. Zhang, H. Wang, W. Chen, J. Zeng, L. Zhang, H. Wang, *et al.*, Computer Physics Communications **253**, 107206 (2020).
[47] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).
[48] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).
[49] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical review letters **77**, 3865 (1996).
[50] G. Kresse and D. Joubert, Physical review b **59**, 1758 (1999).
[51] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).

Appendix: Supporting Information

The Supporting Information provides DFT computational details together with supplementary benchmarking analyses for the four uMLIPs evaluated in this work.

DFT Computational Details

All deformed configurations were computed using density functional theory (DFT) [9] within the Vienna Ab initio Simulation Package (VASP) [47, 48]. The Perdew–Burke–Ernzerhof (PBE) generalized gradient approximation [49] and projector augmented-wave (PAW) [50] pseudopotentials were employed. To ensure strict consistency with the standardized Materials Project workflow, we adopted the same computational parameters: a plane-wave cutoff energy of 520 eV, electronic convergence set to 10^{-5} eV, and a Γ-centered Monkhorst–Pack k-point mesh [51] generated with 1000/(number of atoms per cell) k-point. Spin polarization was enabled for all materials. For each strained configuration, a single-point DFT calculation was carried out to obtain total energies, forces, and stresses for use in fine-tuning the uMLIPs.

Supplementary Figures

Fig. S1 compares the processing-time distributions of the four uMLIPs for elastic-property calculations. Figs. S2 and S3 compare the mean absolute percentage errors for eight elastic properties before and after fine-tuning. Figs. S4 and S5 present the corresponding relative-error distributions.

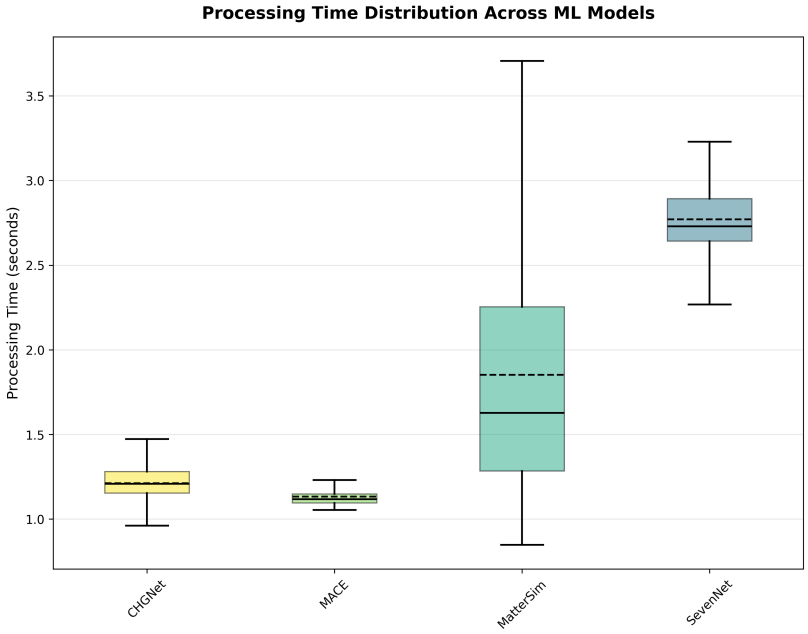

FIG. S1. Processing time distribution comparison across uMLIPs for elastic-property calculations. Each boxplot shows the median (solid line), mean (dashed line), and interquartile range (colored box), with short lines marking the extrema of the non-outlier data. Outliers were filtered using a z-score method (factor = 3.0) during preprocessing and are not displayed.

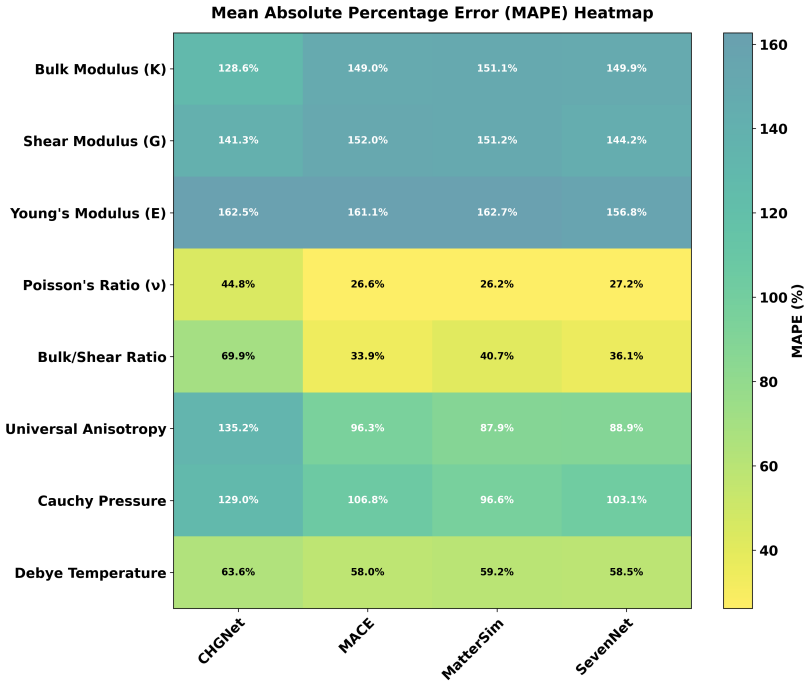

FIG. S2. MAPE heatmap for elastic properties predicted by CHGNet, MACE, MatterSim, and SevenNet relative to DFT reference values before fine-tuning, for the selected 185 materials with high prediction errors. The properties analyzed include bulk modulus, shear modulus, Young's modulus, Poisson's ratio, bulk/shear ratio, universal anisotropy index, Cauchy pressure, and Debye temperature. Darker colors indicate higher errors, with values annotated in each cell.

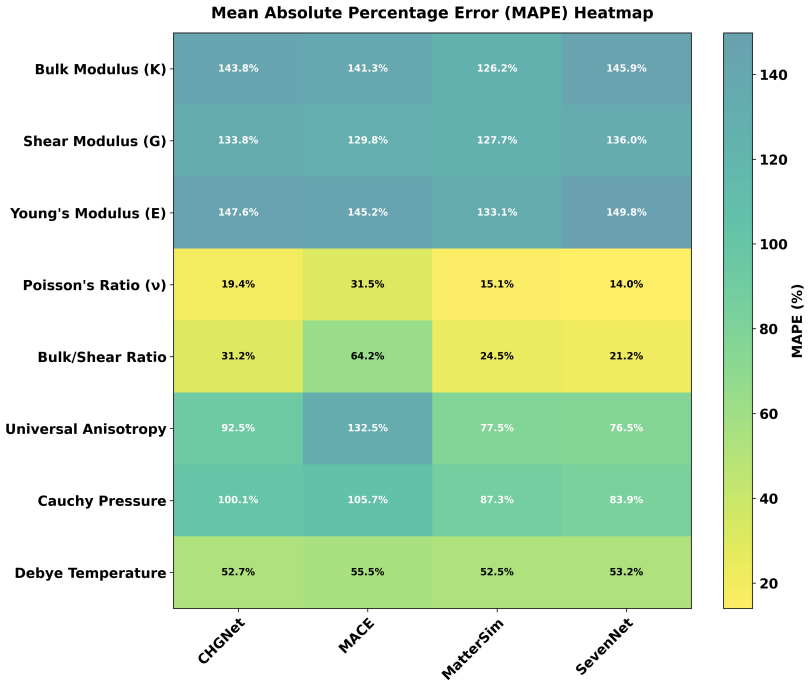

FIG. S3. MAPE heatmap for elastic properties predicted by CHGNet, MACE, MatterSim, and SevenNet relative to DFT reference values after fine-tuning, for the selected 185 materials with high prediction errors. The properties analyzed include bulk modulus, shear modulus, Young's modulus, Poisson's ratio, bulk/shear ratio, universal anisotropy index, Cauchy pressure, and Debye temperature. Darker colors indicate higher errors, with values annotated in each cell.

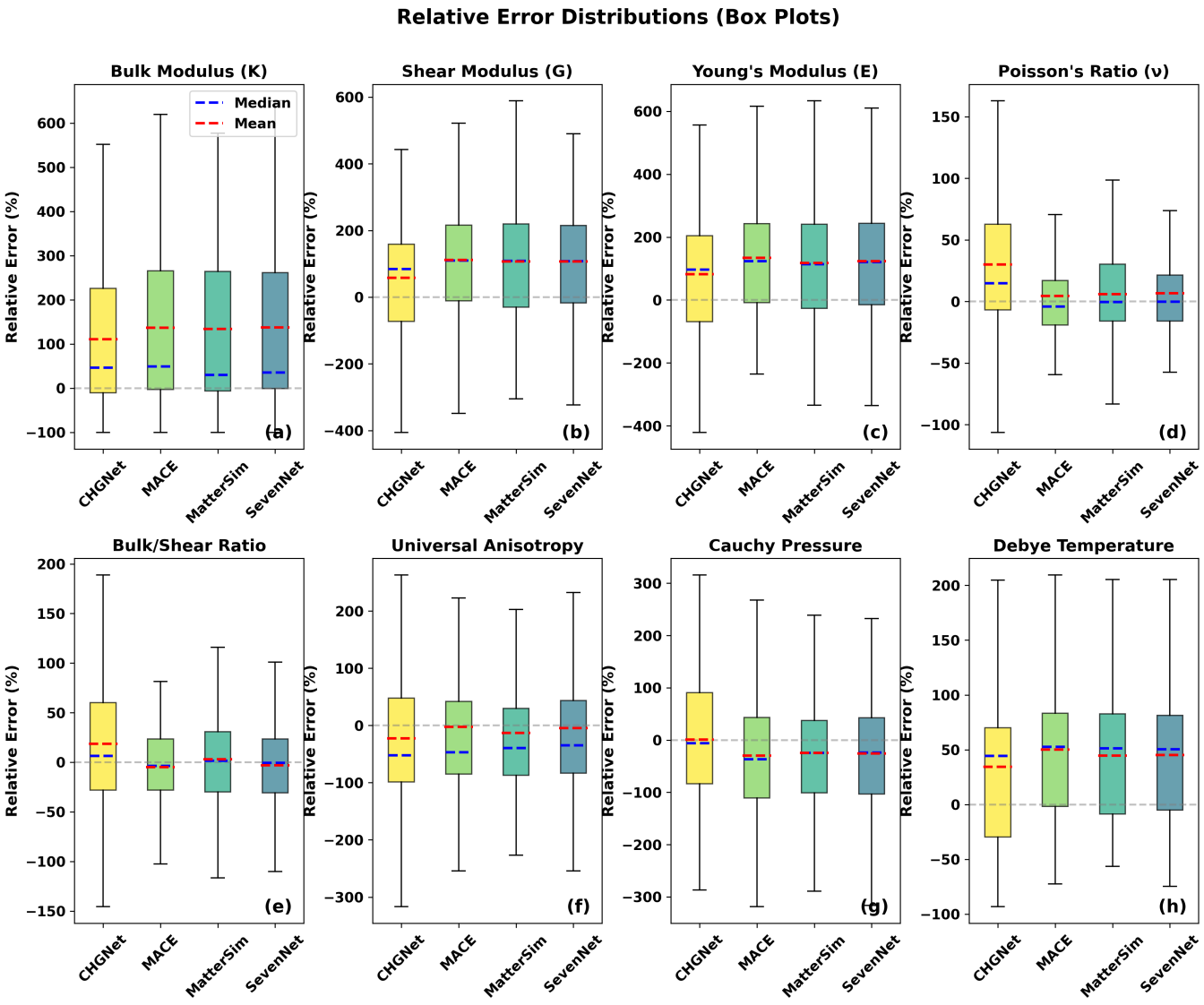

FIG. S4. (a)-(h) Distribution of relative errors (%) for CHGNet, MACE, MatterSim, and SevenNet compared with DFT reference values across eight elastic properties for the selected 185 materials with high prediction errors before fine-tuning. Each boxplot shows the median (blue dashed line), mean (red dashed line), interquartile range (colored box), and the overall error range (short lines marking the extrema), with outliers omitted for clarity. The horizontal dashed line denotes zero error.

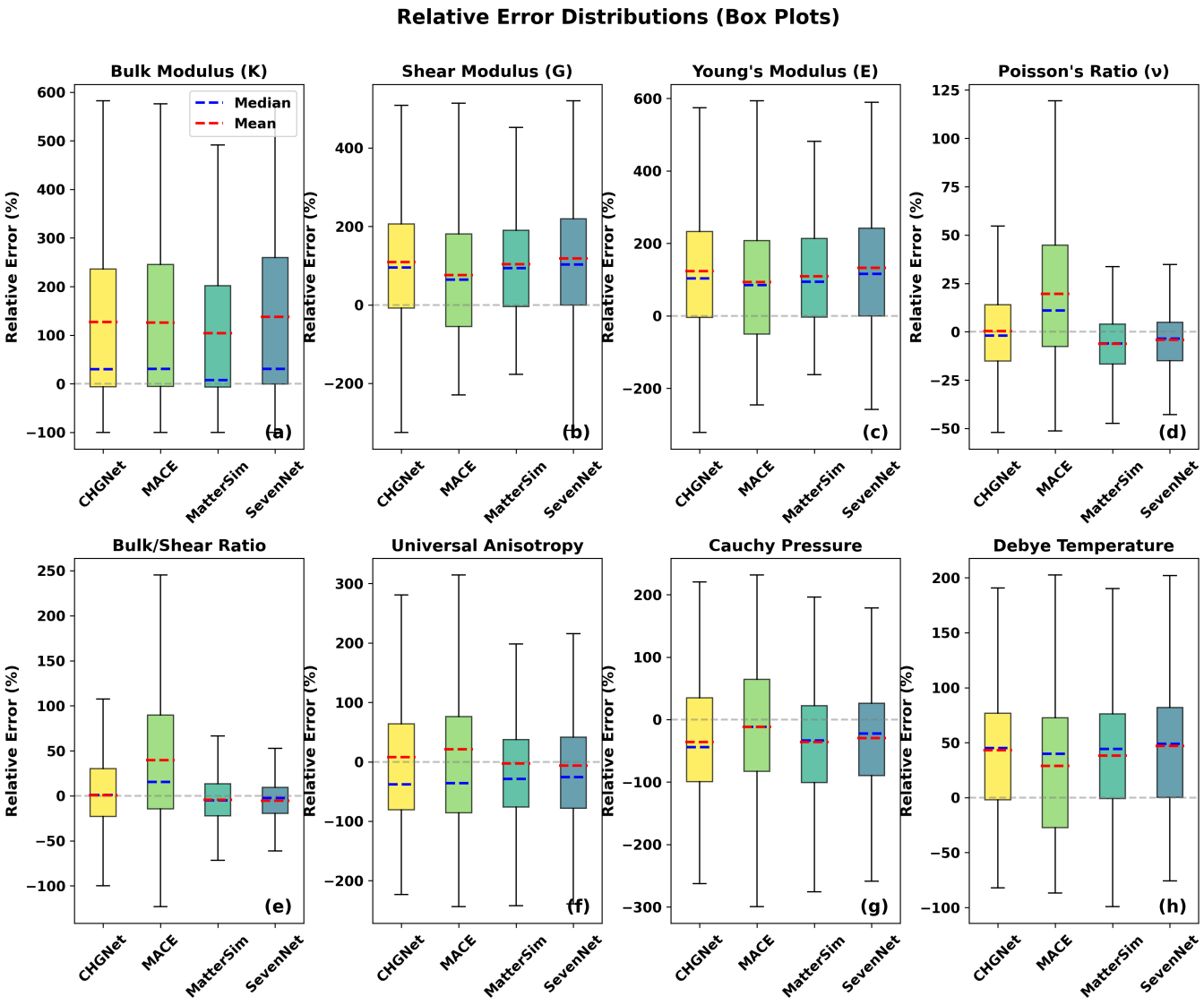

FIG. S5. (a)-(h) Distribution of relative errors (%) for CHGNet, MACE, MatterSim, and SevenNet compared with DFT reference values across eight elastic properties for the selected 185 materials with high prediction errors after fine-tuning. Each boxplot shows the median (blue dashed line), mean (red dashed line), interquartile range (colored box), and the overall error range (short lines marking the extrema), with outliers omitted for clarity. The horizontal dashed line denotes zero error.