RESEARCH ARTICLE | SEPTEMBER 23 2026

Structural benchmarks overlook density-of-states errors in machine-learned interatomic potentials

Shao-Yu Tseng ; Guangshuai Han ; Tianhao Li ; Xiao Xu ; Jiayue Hu ; Gengyao Qiu ; Corey Oses

APL Mach. Learn. 4, 036117 (2026)
https://doi.org/10.1063/5.0339982

Structural benchmarks overlook density-of-states errors in machine-learned interatomic potentials

Cite as: APL Mach. Learn. 4, 036117 (2026); doi: 10.1063/5.0339982
Submitted: 22 April 2026 • Accepted: 21 August 2026 •
Published Online: 23 September 2026

Shao-Yu Tseng,[1,2] Guangshuai Han,[1,2] Tianhao Li,[1,2] Xiao Xu,[1,2] Jiayue Hu,[1,2] Gengyao Qiu,[1] and Corey Oses[1,2,3,4,a)]

AFFILIATIONS

[1] Department of Materials Science and Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA
[2] Johns Hopkins Data Science and AI Institute, Johns Hopkins University, Baltimore, Maryland 21218, USA
[3] Ralph O'Connor Sustainable Energy Institute, Johns Hopkins University, Baltimore, Maryland 21218, USA
[4] William H. Miller III Department of Physics and Astronomy, Johns Hopkins University, Baltimore, Maryland 21218, USA

[a)] **Author to whom correspondence should be addressed:** corey@jhu.edu

ABSTRACT

Universal machine-learned interatomic potentials are widely used as fast surrogates for density functional theory in structural modeling. Standard benchmarks emphasize energies, forces, and relaxed geometries. Whether the resulting structures also preserve the density of states remains largely unexplored. We examine this link across oxides, iodides, and alloys by relaxing structures with four machine-learned interatomic potentials, recalculating the density of states with a common density-functional protocol, and comparing against fully density-functional references. Across all three material classes, structural benchmark performance does not predict electronic agreement. Nevertheless, the hybrid workflow remains useful: single-point density-functional calculations on machine-learned geometries recover Fermi energies, reduce convergence cost, and approximate the ensemble-averaged density of states of a disordered high-entropy alloy from a small subset of ordered representatives. These results show that structural benchmarks alone are insufficient for assessing density-of-states agreement and identify a practical route for reducing the cost of density-functional workflows.

INTRODUCTION

Machine-learned interatomic potentials (MLIPs) have become a foundational tool for computational materials science,[1,2] approaching density functional theory (DFT) accuracy for energies and forces at a fraction of the computational cost. Graph neural network architectures and large-scale training datasets have broadened the range of systems MLIPs can describe and improved how they handle conditions further from their training data.[3–5] MLIPs are now routine in structure relaxation, molecular dynamics, and large-scale materials discovery workflows.[6,7] This success invites an overdue question: as MLIPs become standard infrastructure, are the metrics we use to evaluate them keeping pace with the properties we ultimately want to predict?

The community evaluation paradigm for MLIPs is overwhelmingly structural—energy mean absolute errors (MAEs), force errors, root-mean-square deviations (RMSDs), and geometric benchmarks dominate leaderboards and comparative studies.[7] This paradigm assumes that the accuracy of predicted energies, forces, and relaxed structures is a sufficient proxy for downstream properties. For properties set by the potential energy surface alone—phonon spectra, elastic constants, and mechanical response[4]—it largely is. However, an expanding set of use cases, from catalyst screening[8] to optoelectronic design,[9] depends on the density of states (DOS) and derived electronic quantities, which MLIPs do not compute directly. Whether low errors in structural metrics translate into DOS agreement remains unclear.

Two complementary approaches address the electronic prediction problem. The first trains ML models to predict the DOS directly from structure.[10–13] A recent example, PET-MAD-DOS, applies a universal transformer-based architecture to the DOS across a chemically diverse dataset.[14] Its authors pair the model with

a complementary toolkit—denoising, fine-tuning, and uncertainty quantification—to support extraction of secondary quantities such as bandgaps and Fermi levels.[14] Direct DOS models are constrained by training data: electronic structure datasets remain orders of magnitude smaller than those available for MLIP training.[10] The second approach is hybrid: an MLIP relaxes the structure, and a single-point DFT calculation on the relaxed geometry produces the electronic structure. We adopt this hybrid protocol as an evaluation framework: running the same single-point DFT calculation on MLIP-relaxed and DFT-relaxed geometries isolates the electronic consequences of MLIP-induced geometric deviations. To quantify those consequences, we employ two complementary DOS-based metrics—cosine distance and Wasserstein distance—developed previously in the context of high-entropy catalyst screening,[8] which capture shape similarity and spectral redistribution, respectively.

We implement this framework across three representative material classes—oxides, iodides, and alloys—chosen to span bonding character and electronic complexity. Oxides are among the most technologically consequential materials for energy applications, including catalysis, batteries, and thermoelectrics.[15,16] Their electronic structure is governed by localized *d*-states and crystal-field effects whose sensitivity to local coordination makes them a stringent test for whether relaxed geometries preserve the DOS. Iodides occupy the opposite end of the bonding spectrum: large, polarizable anions with significant spin–orbit coupling and soft lattice dynamics. Their complex bonding environments expose DFT to systematic errors—rooted in dispersion and relativistic effects—that propagate into datasets used to train surrogate models,[17,18] making them a sensitive probe of whether MLIPs reproduce structural features relevant to the electronic landscape. Alloys, and in particular high-entropy alloys (HEAs), provide a metallic baseline where delocalized bonding is expected to be more forgiving of small structural perturbations;[8,15] configurational disorder nonetheless introduces a statistical dimension to the electronic structure that must be treated explicitly.

RESULTS

Models and evaluation protocol

We evaluate four MLIPs: the Message-passing Atomic Cluster Expansion (MACE),[19] the Crystal Hamiltonian Graph Neural Network (CHGNet),[20] the Smooth Expressive Neural Network (eSEN),[21] and the Universal Models for Atoms (UMA).[22] These models were selected to span a range of leaderboard performance[7] and differ in scale by nearly three orders of magnitude, from CHGNet (~0.4M parameters) to UMA (~150M). This spread lets us test whether increasing model capacity, which typically improves energy and force benchmarks, also improves DOS agreement. For each of the 1500 systems (500 per material class), we compare the two pathways of Fig. 1: the reference pathway uses a DFT relaxation followed by a single-point DFT evaluation, while the hybrid pathway replaces the relaxation with each MLIP in turn and holds the single-point DFT step identical. Architectural details, model versions, and relaxation settings are provided in the section titled Methods.

Density-of-states agreement

The full DOS is our primary probe for how MLIP-induced structural deviations propagate into the electronic structure. Discrepancies between the hybrid-pathway and reference-pathway DOS are quantified by the cosine distance $(1 - \cos\theta)$, which measures shape similarity, and the Wasserstein distance (d_w), which captures the redistribution of spectral weight along the energy axis. Both metrics are defined in the section titled Methods; smaller values indicate better agreement. Representative examples of large and small discrepancies are shown in Figs. 2(a) and 2(b).

The distributions of these metrics across material classes [Figs. 2(c)–2(h)] reveal the central finding of this work: DOS discrepancies occupy a similar range across all four MLIPs. Although the models exhibit varying accuracy in energy and RMSD (Fig. S1 of the supplementary material), those differences do not translate into proportional gains in DOS agreement: the cosine distance shows only weak Pearson correlation with RMSD (r = 0.03–0.24; symmetry- and topology-resolved versions are shown in Figs. S4 and S5 of the supplementary material). The largest DOS discrepancies are, therefore, not captured by RMSD alone. Across material classes, alloys show the smallest deviations, while oxides and iodides exhibit broader distributions. MACE tends to yield slightly smaller DOS discrepancies for oxides and iodides, although the differences between models remain modest. A complete summary of metric values is provided in Table I.

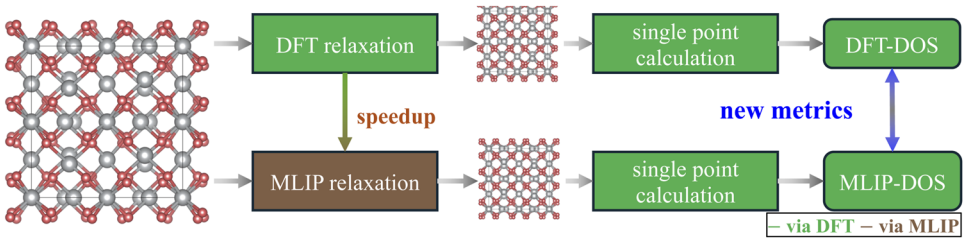

FIG. 1. Evaluation workflow for assessing the impact of MLIPs on electronic structure. In the reference pathway (top), crystal structures are relaxed using DFT, followed by single-point DFT to obtain the ground-truth DOS (DFT-DOS). In the accelerated pathway (bottom), structures are first relaxed using MLIPs, then evaluated using identical single-point DFT calculations to obtain the corresponding DOS (MLIP-DOS). Comparing DFT-DOS and MLIP-DOS isolates the effect of MLIP-induced structural deviations on electronic properties, independent of differences in electronic structure methods.

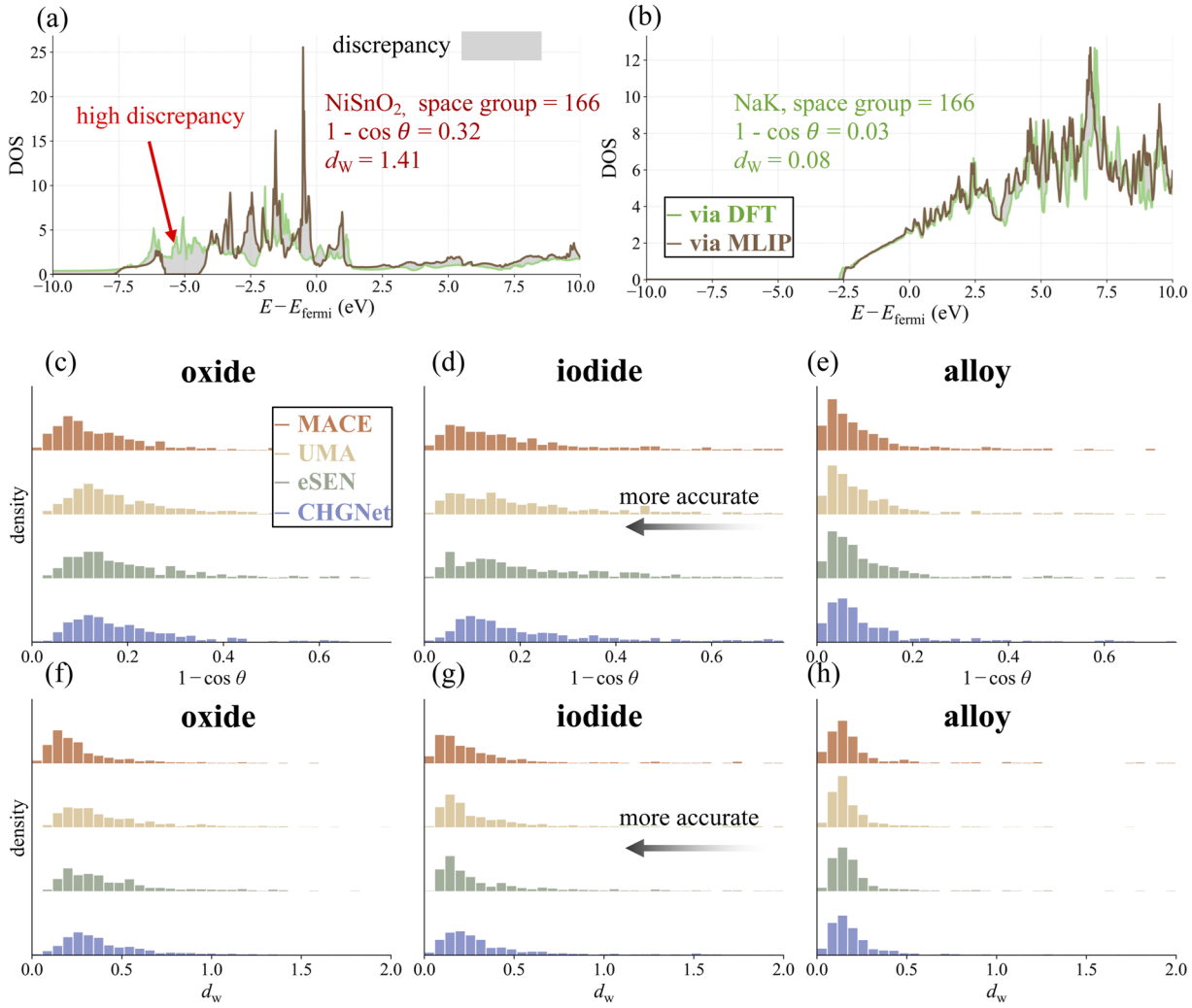

FIG. 2. DOS discrepancies between the reference and hybrid workflows. The reference workflow uses full DFT relaxation followed by a single-point DFT evaluation; the hybrid workflow uses MLIP relaxation followed by the same single-point DFT evaluation (Fig. 1). (a) Representative large DOS discrepancy: $NiSnO_2$ [space group 166, $(1 - \cos\theta) = 0.32$, $d_w = 1.41$]. (b) Representative small DOS discrepancy: NaK [space group 166, $(1 - \cos\theta) = 0.03$, $d_w = 0.08$]. (c)–(h) Distribution of the cosine distance $(1 - \cos\theta)$ and Wasserstein distance (d_w) across iodides, oxides, and alloys.

Fermi-energy residuals

Within the full DOS, the Fermi energy E_F sets transport, carrier statistics, and level alignment. Figures 3(a)–3(c) report the Fermi-energy residuals (hybrid minus reference) for iodides, oxides, and alloys across the four MLIPs.

Residuals are narrowly distributed and comparable in width between models, indicating that the single-point DFT step on an MLIP-relaxed geometry recovers E_F reliably and with little dependence on which MLIP performed the relaxation. System-dependent trends nonetheless persist. Alloys exhibit the narrowest residuals, consistent with delocalized metallic bonding that is less sensitive to small structural perturbations. Oxides and iodides show broader distributions with a tendency toward underestimation: their electronic states come from localized atomic orbitals—transition-metal d states in oxides and heavy-halogen p states in iodides—whose energies shift when the local geometry deviates from the reference. The Wasserstein distance tracks the magnitude of these residuals ($r = 0.60$ for oxides, 0.29–0.33 for iodides and alloys; Figs. S6–S8 of the supplementary material).

Computational efficiency

Beyond DOS agreement, the hybrid workflow delivers a practical benefit: MLIP-relaxed geometries accelerate single-point DFT convergence. As shown in Figs. 3(d)–3(f), convergence costs follow a log-normal distribution across all material classes and relaxation

TABLE I. Summary of structural and electronic metrics across material classes. Values are reported as mean ± standard deviation. Speedup factors (in parentheses) are relative to full DFT relaxation. Here, $\mathbf{R}_{\mathrm{MLIP}}$ and $\mathbf{R}_{\mathrm{DFT}}$ denote MLIP- and DFT-relaxed geometries, respectively. The term $E_{\mathrm{DFT}}(\mathbf{R}_{\mathrm{MLIP}})$ is a single-point DFT energy evaluated on the MLIP-relaxed geometry.

MLIP	$E_{\mathrm{F}}(\mathbf{R}_{\mathrm{MLIP}}) - E_{\mathrm{F}}(\mathbf{R}_{\mathrm{DFT}})$(eV)	$E_{\mathrm{MLIP}}(\mathbf{R}_{\mathrm{MLIP}}) - E_{\mathrm{DFT}}(\mathbf{R}_{\mathrm{DFT}})$(eV)	$E_{\mathrm{DFT}}(\mathbf{R}_{\mathrm{MLIP}}) - E_{\mathrm{DFT}}(\mathbf{R}_{\mathrm{DFT}})$(eV)	RMSD (Å)	CPU hours (speedup)	$1-\cos\theta$	d_{w}
			Alloys				
CHGNet	−0.42 ± 0.64	−0.03 ± 0.31	0.02 ± 0.27	0.02 ± 0.04	18.8 ± 58.6 (×4.0)	0.13 ± 0.15	0.25 ± 0.38
eSEN	−0.44 ± 0.70	−0.02 ± 0.33	0.00 ± 0.75	0.01 ± 0.03	23.0 ± 84.7 (×3.2)	0.12 ± 0.13	0.25 ± 0.38
MACE	−0.46 ± 0.45	0.02 ± 0.40	0.01 ± 0.13	0.01 ± 0.02	7.0 ± 20.1 (×10.6)	0.12 ± 0.15	0.25 ± 0.42
UMA	−0.47 ± 0.46	0.09 ± 0.29	−0.01 ± 0.16	0.01 ± 0.03	21.8 ± 81.7 (×3.4)	0.12 ± 0.14	0.24 ± 0.42
			Iodides				
CHGNet	−0.62 ± 1.36	−0.33 ± 0.15	0.01 ± 0.12	0.06 ± 0.06	54.1 ± 100.1 (×4.5)	0.24 ± 0.21	0.39 ± 0.49
eSEN	−0.91 ± 1.26	−0.05 ± 0.12	−0.04 ± 0.13	0.03 ± 0.05	46.3 ± 72.0 (×5.2)	0.24 ± 0.20	0.41 ± 0.56
MACE	−0.39 ± 1.12	−0.01 ± 0.15	0.00 ± 0.09	0.04 ± 0.05	20.4 ± 32.9 (×11.9)	0.21 ± 0.20	0.36 ± 0.57
UMA	−0.91 ± 1.19	−0.02 ± 0.12	−0.04 ± 0.11	0.03 ± 0.06	47.0 ± 80.5 (×5.2)	0.25 ± 0.21	0.41 ± 0.55
			Oxides				
CHGNet	−0.68 ± 1.07	−0.39 ± 0.28	−0.01 ± 0.23	0.05 ± 0.06	23.3 ± 34.4 (×6.8)	0.20 ± 0.13	0.43 ± 0.35
eSEN	−0.75 ± 1.10	0.42 ± 0.58	0.01 ± 0.24	0.04 ± 0.06	23.9 ± 35.1 (×6.6)	0.20 ± 0.13	0.45 ± 0.33
MACE	−0.45 ± 0.87	0.00 ± 0.19	0.02 ± 0.14	0.04 ± 0.06	10.8 ± 17.6 (×14.6)	0.16 ± 0.13	0.30 ± 0.32
UMA	−0.75 ± 1.09	0.45 ± 0.59	0.01 ± 0.21	0.04 ± 0.07	22.3 ± 38.3 (×7.1)	0.20 ± 0.13	0.43 ± 0.29

methods, reflecting the stochastic nature of self-consistent field convergence.

Among the models, MACE achieves the largest acceleration, with speedups of 11×–15×, compared to 3×–7× for the other MLIPs (Table I). One contributing factor is that MACE-relaxed structures tend to preserve more crystallographic symmetry and topological integrity (Table II), which reduces the size of the irreducible Brillouin zone and the cost of electronic convergence. MACE is better able to preserve the connectivity of the crystal structure, maintaining topological preservation for 1133 systems compared to 891–984 for the other models, while keeping the count of broken topologies to a minimum (245 vs 345–400). This comparison shows that symmetry and topology preservation are useful diagnostics for the hybrid workflow. It does not identify why one model preserves them better than another.

Disordered systems

We examine the implications of these findings for disordered systems, adopting a supercell ensemble approach[23,24] in which the electronic structure is obtained by averaging over a set of ordered representative configurations. As a test case, we consider equiatomic Co–Cr–Fe–Mo–Ni, a well-studied HEA whose corrosion resistance, mechanical properties, and phase stability have been characterized experimentally.[25–27]

The convergence behavior is illustrated in Fig. 4. The DOS converges with an increasing number of sampled configurations, with only minor changes beyond ~10 snapshots [Fig. 4(a)]. Both $(1-\cos\theta)$ and d_{w} decrease and reach a plateau at small sample sizes [Figs. 4(b) and 4(c)]. Using only 10 out of 49 configurations yields minimal deviation from the fully converged reference obtained from DFT. The dip in d_{w} between five and ten samples appears to be a sampling effect. In that range, the selected configurations happen to bring the running ensemble-averaged DOS closer to the full 49-configuration reference. Adding lower-weight configurations then moves the curve toward its final plateau.

These results indicate that, for disordered systems, the ensemble electronic structure can be approximated using a limited number of representative configurations. In this benchmark, the MLIP does not replace the final DFT electronic calculation; it ranks configurations so that the dominant representatives can be evaluated first. Using MLIPs both to pre-screen the ordered supercell representatives and to accelerate individual relaxations enables a substantial reduction in computational cost while retaining accuracy in the resulting electronic properties.

DISCUSSION

Geometric agreement does not guarantee DOS agreement

The central observation of this work is that MLIP accuracy on conventional structural benchmarks—energy, forces, RMSD—does not predict DOS agreement. The four models span nearly three orders of magnitude in parameter count. Energy, force, and RMSD performance vary substantially across them, while DOS discrepancies remain of similar magnitude (Fig. 2). This is not an artifact of poor structural models; all four MLIPs achieve competitive accuracy on established structural benchmarks.[7] The implication is that

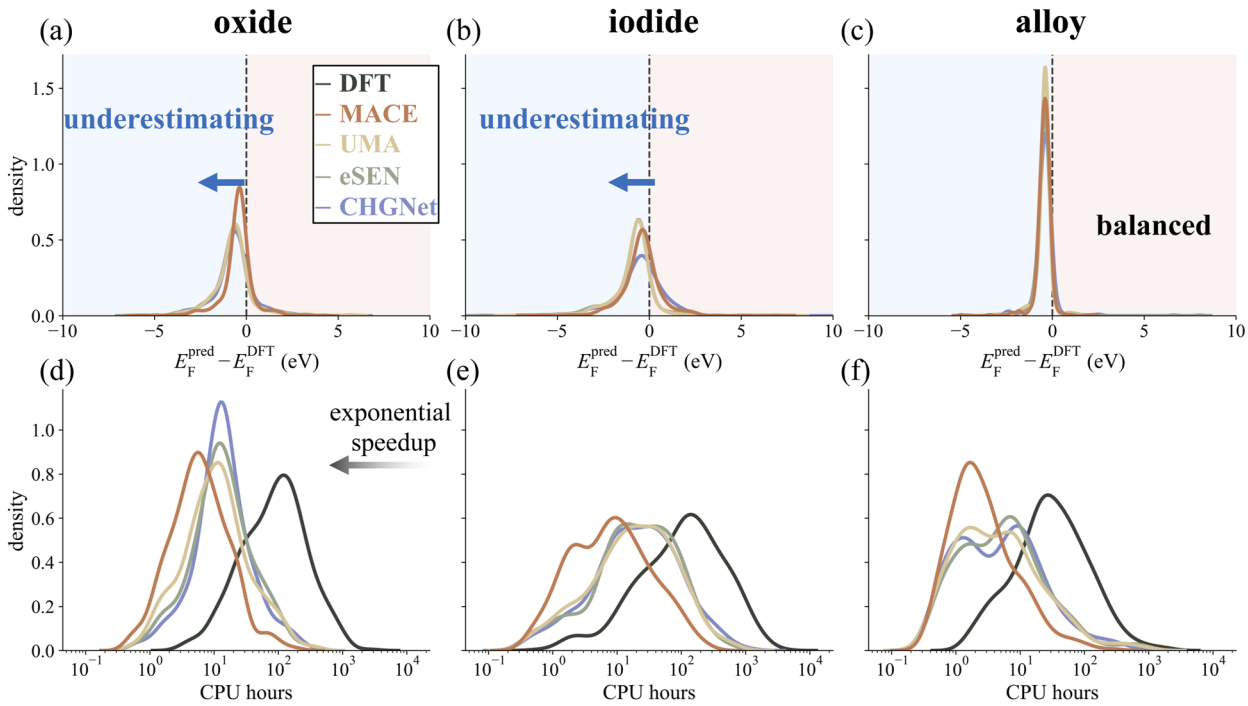

FIG. 3. Fermi-energy residuals and DFT convergence cost for the hybrid workflow. (a) Iodide, (b) oxide, and (c) alloy Fermi-energy E_F residuals (hybrid workflow minus reference). (d) Iodide, (e) oxide, and (f) alloy DFT convergence time in core-hours.

TABLE II. Symmetry preservation and topology classification after MLIP relaxation. Number of systems (out of 1500) whose symmetry-operation cardinality matches that of the DFT-relaxed reference. Topology categories indicate whether the MLIP-relaxed structure preserves, degrades, or breaks the connectivity of the DFT-relaxed reference.

MLIP	Symmetry-preserved count	Topology preserved	Topology degraded	Topology broken
CHGNet	404	891	138	400
eSEN	324	962	115	356
UMA	415	984	110	345
MACE	1260	1133	104	245

geometric agreement is necessary for reliable electronic-structure screening, but it is not sufficient to guarantee DOS agreement.

A likely explanation is that the DOS is sensitive to geometric features that are not well captured by global metrics such as total energy or average force residuals. Small changes in bond angles near band edges, shifts in coordination environment, or subtle distortions of octahedral or tetrahedral cages can redistribute electronic states without large energetic penalties.[9] MLIPs trained on energy and force labels have no direct incentive to preserve these electronically relevant features if they contribute little to the potential energy surface. This is not a failure of the models—it reflects the objectives they were trained on. Extending those objectives to include electronic-structure-aware loss terms or evaluation protocols is a natural next step and one that the metrics employed here could support. Conventional structural metrics and DOS agreement are largely decoupled across all three material classes, so RMSD accuracy cannot substitute for a DOS-level check. The Wasserstein distance, in contrast, correlates directly with the Fermi-level error—most strongly for oxides—which anchors the metric[8] to a physically interpretable observable rather than a purely mathematical shape comparison.

Symmetry preservation

Among the four models tested, MACE preserves crystallographic symmetry in 84% of systems, compared to 22%–28% for the other MLIPs (Table II). This difference correlates with the largest DFT convergence speedups (Table I) and with modestly better DOS agreement for oxides and iodides.

The connection between symmetry and electronic structure runs deeper than computational cost. A change in space group alters the irreducible Brillouin zone and can modify allowed electronic transitions, shift degeneracies, and open or close gaps at

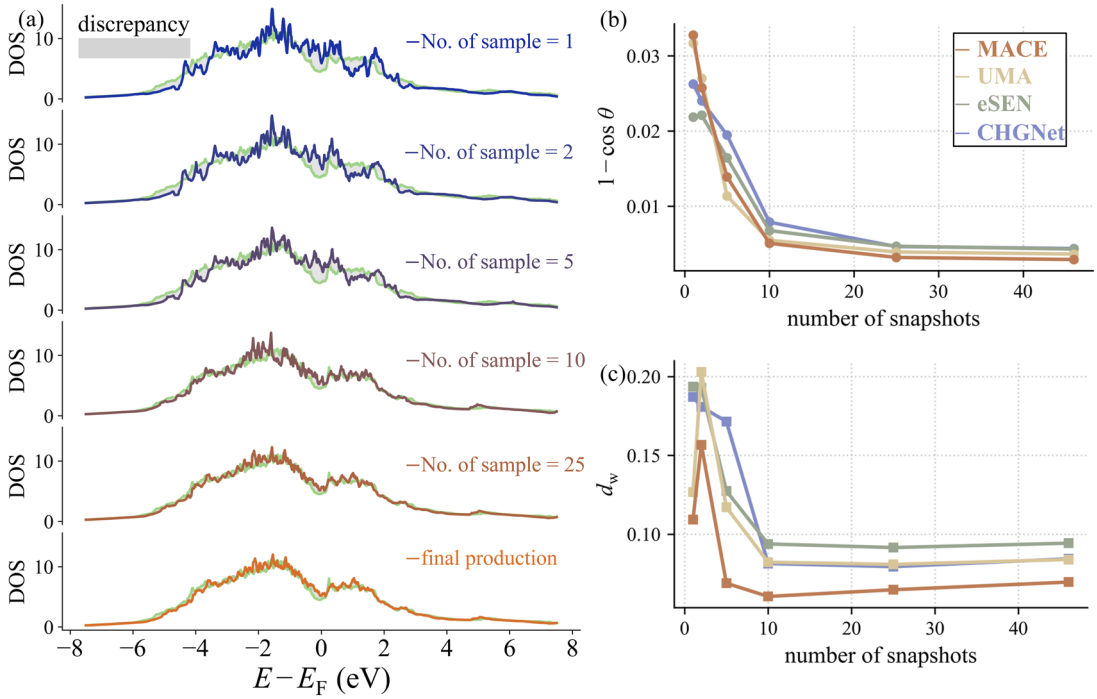

FIG. 4. Ensemble DOS convergence for disordered systems. (a) DOS convergence with increasing number of sampled configurations. (b) and (c) Convergence of $(1 - \cos\theta)$ and d_w with increasing number of representative configurations.

high-symmetry points.[28,29] In materials where electronic properties depend sensitively on crystallographic symmetry—for example, Jahn–Teller distortions in transition-metal oxides,[30] spin–orbit-induced band splittings in heavy iodides, or symmetry-enforced band crossings in Heusler-type alloys[31]—an MLIP that breaks that symmetry during relaxation may introduce qualitative, not just quantitative, errors in the electronic structure.

The stronger symmetry preservation observed for MACE is unlikely to have a single cause. Compared to eSEN-OAM and UMA-S, the MACE-MATPES-PBE-0 checkpoint used here is additionally fine-tuned on MatPES, another materials-focused dataset with greater control on convergence and coverage of non-equilibrium environments.[19,21,22] Greater exposure to periodic crystals in the training data may, therefore, contribute to the stronger symmetry preservation observed for MACE. The checkpoints and workflows differ in more than training data, including model architecture, model size, and relaxation protocol. These coupled differences prevent us from assigning MACE's stronger symmetry preservation to a single factor, such as maximum angular momentum or parameter count.

The topology analysis shows that coarse structural failures explain part of the DOS heavy tail. Figures S4 and S5 of the supplementary material show that topology changes mark many cases with large Wasserstein distances. Some high-error cases preserve both symmetry and topology. These cases show that binary structural labels do not capture all electronic errors. A structure can retain the same symmetry and topology while still shifting DOS peaks or redistributing spectral weight.

Symmetry breaking during relaxation is not unique to MLIPs. DFT relaxation can also change the space group of a prototype structure when the initial configuration sits near a symmetry-breaking instability. The difference is how the geometry and the electronic state are resolved. DFT relaxation updates them together, so the final structure and the electronic state it carries are mutually consistent. An MLIP relaxation, by contrast, sets the geometry on its own. It does not solve for wavefunctions or the Kohn–Sham Hamiltonian, even when the model incorporates charge or magnetic features during inference. The subsequent single-point DFT calculation must then build the electronic structure from scratch on whatever geometry the MLIP provides. Any symmetry the MLIP fails to preserve, therefore, has to be recovered by the downstream DFT step, which is a stronger demand than when the two stages are run self-consistently. Symmetry-aware evaluation deserves inclusion alongside energy and force metrics in future MLIP benchmarks.[32]

Material-class dependence

The three material classes respond differently to MLIP-induced structural perturbations. Alloys show the smallest electronic deviations across all metrics, consistent with metallic bonding that distributes the DOS over a broad energy range and reduces sensitivity to local geometric changes. Oxides and iodides show larger and broader distributions, reflecting the role of *d*-orbital localization (oxides) and soft, polarizable bonding environments (iodides)

in shaping the electronic spectrum. We expect similarly broad DOS-error distributions in other material classes when the target property depends on local coordination, orbital character, spin–orbit effects, disorder averaging, or symmetry. The size of the DOS error will remain chemistry- and dataset-dependent, and energy, force, and RMSD errors should be checked against electronic metrics before they are used as surrogates for DOS agreement.

This hierarchy has practical implications. For metallic systems, including the HEAs studied here and in related work,[8,15] the hybrid MLIP-plus-DFT workflow is viable: structural errors are small, electronic deviations are modest, and ensemble averaging over a limited number of configurations recovers accurate electronic properties (Fig. 4). For oxides and iodides, the same workflow remains useful—the convergence speedup is material-class independent—but the resulting electronic structure should be interpreted with greater caution, and the DOS metrics employed here provide a quantitative basis for deciding when that caution is warranted.

The iodide case is instructive. These systems are known to challenge DFT itself due to dispersion and relativistic effects,[17,18] and the MLIP models inherit those limitations through their training data. The broader DOS distributions observed for iodides may, therefore, reflect not only MLIP-induced structural errors but also the propagation of systematic biases in the underlying DFT reference. Disentangling these contributions is an important direction for future work.

Practical implications and outlook

Two results from this study have immediate practical value. First, the hybrid workflow delivers consistent convergence speedups across all material classes and models, making it a low-risk way to accelerate electronic structure calculations in high-throughput settings. Second, the ensemble convergence result for disordered systems (Fig. 4) provides a concrete protocol for reducing the computational cost of electronic structure evaluation in HEAs and other configurationally complex materials: a small fraction of configurations recovers the full electronic structure to within the resolution of our DOS metrics.

Looking ahead, several extensions are worth exploring. Incorporating electronic-structure-aware labels or descriptors into MLIP training could narrow the gap between geometric agreement and DOS agreement. The DOS metrics employed here—cosine distance and Wasserstein distance[8]—are not limited to benchmarking; they could serve as loss components or validation criteria during MLIP training, providing direct feedback on electronic relevance. Similarly, the observation that direct DOS prediction models[12–14] and the hybrid approach studied here face complementary limitations suggests that combining both strategies—using MLIP-relaxed geometries as inputs to lightweight DOS models rather than full DFT—may offer a path toward fast, accurate, and interpretable electronic structure evaluation at scale. Domain-specific fine-tuning is likely to improve energy, force, and geometric metrics when the target chemistry is under-represented and may reduce DOS errors when those errors are geometry mediated. The impact of chemistry-specific fine-tuning remains an open question. Fine-tuning only on energies and forces is not expected, by itself, to guarantee DOS-level accuracy. Electronic labels, descriptors, or validation metrics would still be needed to assess whether the relaxed geometries preserve the DOS.

CONCLUSION

We examined the relationship between conventional structural metrics and density-of-states agreement across four MLIPs spanning nearly three orders of magnitude in parameter count, applied to oxides, iodides, and alloys. By combining MLIP-based relaxation with identical single-point DFT calculations, we isolated the effect of structural approximations on the resulting DOS.

Three findings emerge. First, improvements in conventional structural metrics—energy, forces, and RMSD—do not translate into proportional gains in DOS agreement. DOS discrepancies remain comparable across models of vastly different scales, showing that geometric agreement is necessary but not sufficient for electronic-structure screening. Second, symmetry preservation during relaxation correlates with DFT convergence speedups, suggesting that symmetry-aware evaluation deserves inclusion alongside energy and force benchmarks. Third, for disordered systems, the ensemble-averaged DOS converges with only a handful of representative configurations, so the full electronic structure is accessible at a fraction of the cost.

Several directions remain open. Incorporating electronic-structure-aware labels or descriptors into MLIP training could narrow the gap between conventional structural metrics and DOS agreement. The DOS metrics employed here could serve as validation criteria or loss components during training. Combining MLIP-relaxed geometries with lightweight DOS prediction models may offer a route toward fast electronic structure evaluation at scale.

METHODS

Data selection

We selected 500 systems from each of three families of ordered materials—alloys, oxides, and iodides—to build a representative sample, drawn from the CHAOS database (Combinatorial High-throughput Analysis & Optimization for Synthesis).[33] Table S1 of the supplementary material lists the compositions and crystallographic prototype identifiers.[34] Five HEA systems were separately selected for evaluation within the supercells ensemble formalism.[23,24]

MLIP models

We evaluate MACE-MATPES-PBE-0 (~9.06M parameters),[19] CHGNet (~0.413M),[20] eSEN-OAM (~30.2M),[21] and UMA-S (~150M).[22] MACE, eSEN, and UMA are E(3)-equivariant graph neural network models; eSEN and UMA further incorporate transformer-based architectures. CHGNet is an invariant architecture that includes charge and magnetic moment features derived from DFT training data. The models span nearly three orders of magnitude in parameter count, providing a spectrum of architectural design and capacity.

Processing pipeline

For each system, a crystal prototype is generated from its space group and Wyckoff positions.[23,32] Each prototype is then relaxed using each of the four MLIPs via the `ASE` Python interface,[35] with the maximum residual force set to 0.001 eV/Å and a limit of 500 relaxation steps. For a given system, all MLIP and DFT relaxations start from this same unrelaxed prototype, so differences in the final geometries arise from the relaxation method rather than from different starting configurations. The resulting relaxed structures are used as inputs to single-point DFT calculations performed with VASP,[36,37] employing PAW pseudopotentials[38] and the PBE generalized-gradient approximation.[39] All remaining parameters—including plane wave cutoffs, convergence criteria, and Fourier-transform meshes—follow the AFLOW standard.[40] Only electronic convergence is performed; no ionic relaxation is carried out in VASP.

After relaxation, RMSD and maximum displacement are computed with the `StructureMatcher` class in `pymatgen`, which aligns each MLIP-relaxed structure to the DFT-relaxed reference before calculating distances.[41,42] We also use `StructureMatcher` to classify topology as preserved, degraded, or broken. In Figs. S4 and S5 of the supplementary material, degraded and broken cases are grouped as topology changes. Symmetry preservation is assessed separately by comparing the cardinality of the symmetry-operation set—as reported in the VASP OUTCAR—between the MLIP-relaxed structure and the DFT-relaxed reference. A structure is classified as symmetry-preserved if the two cardinalities match. This criterion reflects the symmetry that VASP uses internally to construct the irreducible Brillouin zone, which directly affects the cost of electronic convergence. For a detailed treatment of crystallographic symmetry analysis, see Ref. 32.

DOS evaluation

The DOS is computed using the AFLOW static calculation protocol: 10 000 KPPRA with the tetrahedron method and Blöchl corrections for Brillouin zone integration, evaluated at 5000 points spanning −30 to 45 eV. Because the Fermi-level alignment shifts the energy grid of each structure independently, the resulting DOS values are not directly comparable across systems. We, therefore, interpolate the aligned DOS using SciPy's[43] `RBFInterpolator` with a Gaussian kernel ($\sigma = 0.05$, smoothing $= 10^{-5}$) and resample at 1000 points from −10 to 10 eV relative to the Fermi level. This restricts the comparison to the chemically relevant valence and low-energy conduction states and places all structures on a common grid. We denote the interpolated DOS as $\mathrm{DOS}_{\mathrm{int}}(E)$.

DOS metrics

For each structure, the Fermi-aligned DOS vector is

$$\mathrm{DOS}(i) = \mathrm{DOS}_{\mathrm{int}}(E(i) + E_{\mathrm{F}}), \tag{1}$$

where E_{F} is the Fermi level and $E(i)$ spans −10 to 10 eV. Given the DOS vectors from the DFT-relaxed structure ($\mathrm{DOS}_{\mathrm{DFT}}$) and the MLIP-relaxed structure ($\mathrm{DOS}_{\mathrm{MLIP}}$), we define the cosine distance,

$$d_{\mathrm{c}} = 1 - \cos\theta = 1 - \frac{\mathrm{DOS}_{\mathrm{DFT}} \cdot \mathrm{DOS}_{\mathrm{MLIP}}}{\|\mathrm{DOS}_{\mathrm{DFT}}\| \|\mathrm{DOS}_{\mathrm{MLIP}}\|}, \tag{2}$$

where $\|\cdot\|$ denotes the L_2 norm, and the Wasserstein distance,

$$F_k(x) = \int_{-\infty}^{x} \mathrm{DOS}_k(x')\, dx', \quad k \in \{\mathrm{DFT}, \mathrm{MLIP}\}, \tag{3}$$

$$d_{\mathrm{w}} = \int_{-\infty}^{\infty} |F_{\mathrm{DFT}}(x) - F_{\mathrm{MLIP}}(x)|\, dx, \tag{4}$$

where F_{DFT} and F_{MLIP} are the cumulative distribution functions of the two DOS distributions. The cosine distance measures shape similarity and is insensitive to absolute magnitude; the Wasserstein distance captures spectral redistribution along the energy axis. Both integrals are evaluated using the trapezoidal rule.[8]

High-entropy ensemble averaging

For HEA systems, the electronic DOS is obtained as a Boltzmann-weighted ensemble average over the ordered supercell representatives.[23,24] In most such systems, the majority of the probability mass is concentrated in a small fraction of the representatives—configurations with low energies, high degeneracies, or both—motivating the use of MLIPs to rank configurations by Boltzmann weight and restrict DFT evaluation to the dominant subset. For the convergence benchmark in Fig. 4, the reference ensemble is the Boltzmann-weighted DOS from all 49 unique ordered supercell representatives[23,24] after DFT relaxation and single-point DFT DOS evaluation. The truncated curves are generated by selecting the highest-weight representatives according to the MLIP ranking, evaluating only that subset with DFT, and comparing the running Boltzmann-weighted average with the all-49 DFT reference. Thus, the MLIP is used for configuration ranking and structural preconditioning; the reported electronic DOS values remain DFT calculations.

A direct average of the raw DOS values on the absolute energy grid is problematic because each finite supercell enforces its own charge neutrality condition, producing a distinct Fermi level. In a macroscopic disordered metal, however, the electrochemical potential is uniform: electrons are delocalized across all local configurations, and a single chemical potential governs the entire system. The individual supercell Fermi levels are artifacts of the finite-cell approximation, and averaging the DOS without correcting for them introduces spurious misalignment along the energy axis. To recover a physically meaningful energy reference, we define a Boltzmann-weighted Fermi level,

$$\bar{E}_{\mathrm{F}}(T) = \sum_i p_i(T)\, E_{\mathrm{F},i}, \tag{5}$$

where $p_i(T) = g_i\, e^{-E_i/k_{\mathrm{B}}T} / \sum_j g_j\, e^{-E_j/k_{\mathrm{B}}T}$ is the Boltzmann weight for the i-th representative, g_i is its degeneracy, and $E_{\mathrm{F},i}$ is its Fermi level. All individual DOS are re-aligned to $\bar{E}_{\mathrm{F}}$ before ensemble averaging. In the early-stopping protocol evaluated in this work, the full ensemble is not yet available at the point of truncation. The Boltzmann-weighted Fermi level is, therefore, computed

from the subset of structures evaluated so far, and the ensemble DOS is constructed as a running average over the available representatives.

SUPPLEMENTARY MATERIAL

The supplementary material contains energy residual distributions, energy parity plots, comparisons between structural deviations and density-of-states distances, symmetry and topology analyses, pairwise correlation matrices, and a table of compositions and crystallographic prototypes.

ACKNOWLEDGMENTS

The authors thank Bill McNamara (JHU), A. J. Donelson (JHU), Chip Cotton (JHU), Assel Aitkaliyeva (ARPA-E), Robert Ledoux (ARPA-E), Jeremiah Beam (ARPA-E), Anil Ganti (ARPA-E), Jonathan Melville (ARPA-E), and Jonathan R. Owens (GE Vernova Advanced Research) for helpful discussions. Research was sponsored by the Advanced Research Projects Agency-Energy (ARPA-E) (Grant No. DE-AR0001787). J.H. and G.H. acknowledge support from the Data Science and Artificial Intelligence (DSAI) Institute at JHU. T.L. and X.X. acknowledge support from the Institute for Data-Intensive Engineering and Science (IDIES) at JHU. C.O. acknowledges the computational resources provided by the Advanced Research Computing at Hopkins (ARCH), which is supported by the National Science Foundation (NSF) (Grant No. OAC1920103) and the Texas Advanced Computing Center (TACC) at The University of Texas at Austin.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

S.-Y.T. and G.H. contributed equally to this work.

C.O. conceived, designed, and supervised the research. S.-Y.T., G.H., and C.O. developed the DOS metrics and ensemble averaging protocol. S.-Y.T. and G.H. carried out the high-throughput MLIP relaxations, DFT calculations, and electronic structure analysis. T.L., X.X., J.H., and G.Q. contributed to data processing and analysis. S.-Y.T., G.H., and C.O. drafted the manuscript, and all authors reviewed, edited, and approved the final version.

Shao-Yu Tseng: Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Software (lead); Validation (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (equal). **Guangshuai Han**: Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Software (lead); Validation (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (equal). **Tianhao Li**: Data curation (equal); Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Writing – review & editing (equal). **Xiao Xu**: Investigation (equal); Methodology (equal); Validation (equal); Writing – review & editing (equal). **Jiayue Hu**: Investigation (equal); Methodology (equal); Validation (equal); Writing – review & editing (equal). **Gengyao Qiu**: Investigation (equal); Methodology (equal); Validation (equal); Writing – review & editing (equal). **Corey Oses**: Conceptualization (lead); Funding acquisition (lead); Project administration (lead); Resources (lead); Supervision (lead); Writing – review & editing (lead).

DATA AVAILABILITY

The data that support the findings of this study are available through the CHAOS database (Combinatorial High-throughput Analysis & Optimization for Synthesis) API at https://s4e.ai/chaos.html.[33] The data and scripts supporting this study are openly available at https://github.com/entropy4energy/2026-mlip-dos-benchmark.

REFERENCES

[1] Q. Zhao, Y. Han, D. Zhang, J. Wang, P. Zhong, T. Cui, B. Yin, Y. Cao, H. Jia, and C. Duan, "Harnessing machine learning to enhance transition state search with interatomic potentials and generative models," Adv. Sci. **12**, e06240 (2025).
[2] R. Jacobs, D. Morgan, S. Attarian, J. Meng, C. Shen, Z. Wu, C. Y. Xie, J. H. Yang, N. Artrith, B. Blaiszik, G. Ceder, K. Choudhary, G. Csanyi, E. D. Cubuk, B. Deng, R. Drautz, X. Fu, J. Godwin, V. Honavar, O. Isayev, A. Johansson, B. Kozinsky, S. Martiniani, S. P. Ong, I. Poltavsky, K. J. Schmidt, S. Takamoto, A. P. Thompson, J. Westermayr, and B. M. Wood, "A practical guide to machine learning interatomic potentials - Status and future," Curr. Opin. Solid State Mater. Sci. **35**, 101214 (2025).
[3] B. Focassio, L. P. M. Freitas, and G. R. Schleder, "Performance assessment of universal machine learning interatomic potentials: Challenges and directions for materials surfaces," ACS Appl. Mater. Interfaces **17**, 13111–13121 (2024).
[4] A. Loew, D. Sun, H.-C. Wang, S. Botti, and M. A. L. Marques, "Universal machine learning interatomic potentials are ready for phonons," npj Comput. Mater. **11**, 178 (2025).
[5] C. Jiang, C. A. Marianetti, M. Khafizov, and D. H. Hurley, "Machine learning potential assisted exploration of complex defect potential energy surfaces," npj Comput. Mater. **10**, 21 (2024).
[6] M. Kulichenko, B. Nebgen, N. Lubbers, J. S. Smith, K. Barros, A. E. A. Allen, A. Habib, E. Shinkle, N. Fedik, Y. W. Li, R. A. Messerly, and S. Tretiak, "Data generation for machine learning interatomic potentials and beyond," Chem. Rev. **124**, 13681–13714 (2024).
[7] J. Riebesell, R. E. A. Goodall, P. Benner, Y. Chiang, B. Deng, G. Ceder, M. Asta, A. A. Lee, A. Jain, and K. A. Persson, "A framework to evaluate machine learning crystal stability predictions," Nat. Mach. Intell. **7**, 836–847 (2025).
[8] G. Han, T. Li, X. Xu, J. Lee, G. Qiu, S. Sequeira, A. Ajith, and C. Oses, "The search for high-entropy fuel-cell catalysts using disorder descriptors," Nano Futures **9**, 045001 (2025).
[9] S. Huo, S. Zhang, Q. Wu, and X. Zhang, "Feature-assisted machine learning for predicting band gaps of binary semiconductors," Nanomaterials **14**, 445 (2024).
[10] V. Fung, P. Ganesh, and B. G. Sumpter, "Physically informed machine learning prediction of electronic density of states," Chem. Mater. **34**, 4848–4855 (2022).
[11] A. Aryanpour and A. Sadeghi, "Machine learning the local electronic density of states," arXiv:2504.16052 (2025).
[12] S. Kong, F. Ricci, D. Guevarra, J. B. Neaton, C. P. Gomes, and J. M. Gregoire, "Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings," Nat. Commun. **13**, 949 (2022).
[13] C. Ben Mahmoud, A. Anelli, G. Csányi, and M. Ceriotti, "Learning the electronic density of states in condensed matter," Phys. Rev. B **102**, 235130 (2020).
[14] W. B. How, P. Febrer, S. Chong, A. Mazitov, F. Bigi, M. Kellner, S. Pozdnyakov, and M. Ceriotti, "A universal machine learning model for the electronic density of states," arXiv:2508.17418 (2025).
[15] C. Oses, C. Toher, and S. Curtarolo, "High-entropy ceramics," Nat. Rev. Mater. **5**, 295–309 (2020).

[16]G. Qiu, T. Li, X. Xu, Y. Liu, M. Niyogi, K. Cariaga, and C. Oses, "High entropy powering green energy: Hydrogen, batteries, electronics, and catalysis," npj Comput. Mater. **11**, 145 (2025).
[17]X. Xu, G. Han, G. Qiu, T. Li, S.-Y. Tseng, J. Hu, J. R. Owens, and C. Oses, "Learning experiment-grounded thermodynamics in data-scarce chemistries" (Unpublished) (2026).
[18]X. Xu, G. Han, T. Li, J. Hu, G. Qiu, S.-Y. Tseng, G. H. Han, A. D. Dupuy, A. S. Hall, and C. Oses, "Stabilizing iodine in pyrochlore: Toward new nuclear waste forms," Precis. Chem. **4**, 1211–1224 (2026).
[19]I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csanyi, "MACE: Higher order equivariant message passing neural networks for fast and accurate force fields," Adv. Neural Inf. Process. Syst. **35**, 11423–11436 (2022).
[20]B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, "CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling," Nat. Mach. Intell. **5**, 1031–1041 (2023).
[21]X. Fu, B. M. Wood, L. Barroso-Luque, D. S. Levine, M. Gao, M. Dzamba, and C. L. Zitnick, "Learning smooth and expressive interatomic potentials for physical property prediction," arXiv:2502.12147 (2025).
[22]B. M. Wood, M. Dzamba, X. Fu, M. Gao, M. Shuaibi, L. Barroso-Luque, K. Abdelmaqsoud, V. Gharakhanyan, J. R. Kitchin, D. S. Levine, K. Michel, A. Sriram, T. Cohen, A. Das, A. Rizvi, S. J. Sahoo, Z. W. Ulissi, and C. L. Zitnick, "UMA: A family of universal models for atoms," arXiv:2506.23971 (2026).
[23]C. Oses, M. Esters, D. Hicks, S. Divilov, H. Eckert, R. Friedrich, M. J. Mehl, A. Smolyanyuk, X. Campilongo, A. van de Walle, J. Schroers, A. G. Kusne, I. Takeuchi, E. Zurek, M. B. Nardelli, M. Fornari, Y. Lederer, O. Levy, C. Toher, and S. Curtarolo, "aflow++: AC++ framework for autonomous materials design," Comput. Mater. Sci. **217**, 111889 (2023).
[24]K. Yang, C. Oses, and S. Curtarolo, "Modeling off-stoichiometry materials with a high-throughput ab-initio approach," Chem. Mater. **28**, 6484–6492 (2016).
[25]G. Cui, B. Han, Y. Yang, Y. Wang, and H. Chunyang, "Microstructure and tribological property of CoCrFeMoNi high entropy alloy treated by ion sulfurization," J. Mater. Res. Technol. **9**, 2598–2609 (2020).
[26]S. J. Brito-Garcia, J. C. Mirza-Rosca, C. Jiménez-Marcos, and I. Voiculescu, "Impact of Ti doping on the microstructure and mechanical properties of CoCrFeMoNi high-entropy alloy," Metals **13**, 854 (2023).
[27]M. Ma, C. Zhu, Z. Wang, Y. Xi, and B. Wang, "Effect of Si addition on phase structure and wear resistance of CoCrFeMoNi alloy coatings," J. Meas. Eng. **13**(3), 534–546 (2025).
[28]M. Z. Hasan and C. L. Kane, "*Colloquium*: Topological insulators," Rev. Mod. Phys. **82**, 3045–3067 (2010).
[29]B. A. Bernevig and T. L. Hughes, *Topological Insulators and Topological Superconductors* (Princeton University Press, 2013).
[30]I. Bersuker, *The Jahn-Teller Effect* (Cambridge University Press, 2006).
[31]I. Belopolski, K. Manna, D. S. Sanchez, G. Chang, B. Ernst, J. Yin, S. S. Zhang, T. Cochran, N. Shumiya, H. Zheng, B. Singh, G. Bian, D. Multer, M. Litskevich, X. Zhou, S.-M. Huang, B. Wang, T.-R. Chang, S.-Y. Xu, A. Bansil, C. Felser, H. Lin, and M. Z. Hasan, "Discovery of topological Weyl fermion lines and drumhead surface states in a room temperature magnet," Science **365**, 1278–1281 (2019).
[32]D. Hicks, C. Oses, E. Gossett, G. Gomez, R. H. Taylor, C. Toher, M. J. Mehl, O. Levy, and S. Curtarolo, "AFLOW-SYM: Platform for the complete, automatic and self-consistent symmetry analysis of crystals," Acta Crystallogr., Sect. A: Found. Adv. **74**, 184–203 (2018).
[33]J. Hu, S.-Y. Tseng, M. Brownrigg, P. Boctor, G. Han, T. Li, W. Shiber, X. Xu, B. Lim, M. Niyogi, K. Clark, and C. Oses, "From quantum descriptors to natural-language queries: The CHAOS database for high-entropy oxide discovery" (Unpublished) (2026).
[34]D. Hicks, C. Toher, D. C. Ford, F. Rose, C. D. Santo, O. Levy, M. J. Mehl, and S. Curtarolo, "AFLOW-XtalFinder: A reliable choice to identify crystalline prototypes," npj Comput. Mater. **7**, 30 (2021).
[35]S. R. Bahn and K. W. Jacobsen, "An object-oriented scripting interface to a legacy electronic structure code," Comput. Sci. Eng. **4**, 56–66 (2002).
[36]G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," Phys. Rev. B **54**, 11169–11186 (1996).
[37]G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," Comput. Mater. Sci. **6**, 15–50 (1996).
[38]G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," Phys. Rev. B **59**, 1758–1775 (1999).
[39]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865–3868 (1996).
[40]C. E. Calderon, J. J. Plata, C. Toher, C. Oses, O. Levy, M. Fornari, A. Natan, M. J. Mehl, G. Hart, M. Buongiorno Nardelli, and S. Curtarolo, "The AFLOW standard for high-throughput materials science calculations," Comput. Mater. Sci. **108**(Part A), 233–238 (2015).
[41]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, "Python materials genomics (pymatgen): A robust, open-source python library for materials analysis," Comput. Mater. Sci. **68**, 314–319 (2013).
[42]pymatgen Development Team, pymatgen.analysis.structure_matcher source code, https://github.com/materialsproject/pymatgen/blob/main/src/pymatgen/analysis/structure_matcher.py, 2026; accessed 03 July 2026.
[43]P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. P. Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko, Y. Vázquez-Baeza, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental algorithms for scientific computing in python," Nat. Methods **17**, 261–272 (2020).