ELEMENTA: Toward Unbiased Scaling of Materials Data from First Principles

Kairos Materials

research@kairosmaterials.com

Scaling has become a defining route to generality in artificial intelligence. The largest advances have come from scaling computation and experience rather than increasingly handcrafted designs. The same transition is now emerging in materials science. Here we introduce **ELEMENTA**, a composition-complete and prototype-independent first-principles dataset built to provide the physical foundation for materials-data scaling. A systematically enumerated chemical space spanning unary, binary and ternary compositions with up to twelve atoms per cell is explored through random structure search and represented by density-functional-theory relaxation trajectories, yielding approximately **210 million frames** with energies, forces and stresses across broad chemical and configurational space. Selected structures are further expanded in two complementary physical dimensions. **ELEMENTA Spin** contains over **4 million frames** from systematic magnetic-moment scans, and **ELEMENTA Vib** contains approximately **8 million perturbed structures** probing local vibrational environments and energy-surface curvature. Together, these collections provide a large-scale foundation for training atomistic models and a rigorous benchmark for measuring generalization across composition, structure and physical state. Evaluations of representative materials foundation models reveal structured failure modes that are not apparent from aggregate regression metrics alone, including insufficient resolution of subtle energy differences among competing low-energy structures, limiting reliable ground-state identification, together with physically interpretable errors in ionic coordination, magnetic ordering and local bonding environments. **ELEMENTA** shows that the next stage of materials-data scaling requires not only more first-principles calculations, but broader sampling and finer resolution of materials space.

Date: July 27, 2026

KAIROS MATERIALS

1 Introduction

Scaling has become the central organizing principle of frontier artificial intelligence. Its first phase operated through pre-training, where capability improved predictably with model size, data and computation, and compute-optimal studies established that data must grow alongside model capacity (Kaplan et al., 2020; Hoffmann et al., 2022). The frontier has since expanded beyond pre-training. Large-scale reinforcement learning and inference-time search now convert additional train-time and test-time computation into stronger reasoning and verification (Lightman et al., 2024; Snell et al., 2025; DeepSeek-AI, 2025), while the emerging generation of agents seeks to acquire new experience through interaction rather than learning only from fixed human corpora (Wang et al., 2024). Across these developments, progress has come from increasingly general procedures that turn computation into broader and more informative experience. Materials science provides an unusually rigorous setting for this transition because such experience can be generated from the physical interactions that govern matter.

First-principles calculations provide the physical substrate for this form of scaling. Density-functional theory maps elemental identities and atomic configurations to energies, forces and stresses within a common quantum-mechanical framework, and can therefore evaluate structures that have neither been synthesized nor previously proposed (Hohenberg and Kohn, 1964; Kohn and Sham, 1965). Scaling these calculations has already transformed materials science. The Materials Project, OQMD and Alexandria turned high-throughput electronic-structure calculations into reusable corpora spanning broad chemical spaces (Jain et al., 2013; Horton et al., 2025; Kirklin et al., 2015; Schmidt et al., 2023, 2024). MPtrj and related relaxation datasets subsequently made atomically resolved trajectories available for learning potential-energy surfaces, supporting

transferable models such as M3GNet, CHGNet, MACE-MP and MatterSim (Deng et al., 2023; Chen and Ong, 2022; Batatia et al., 2022, 2025; Yang et al., 2024; Kaplan et al., 2025; Kuner et al., 2025). These models extended learned capabilities from property prediction to structure relaxation, atomistic simulation, lattice dynamics and thermodynamics (Jacobs et al., 2025; Fu et al., 2025; Póta et al., 2024; Loew et al., 2025; Yang et al., 2026). GNoME demonstrated learning-guided exploration of millions of candidate crystals, MatterGen established generative design across the periodic table, and OMat24 scaled open energy–force–stress supervision beyond one hundred million configurations (Merchant et al., 2023; Zeni et al., 2025; Barroso-Luque et al., 2026). Taken together, these advances establish first-principles data as a principal determinant of the capability frontier in materials AI.

The success of this scaling has depended on equally important advances in deciding what to calculate (Jinnouchi et al., 2020; Schwalbe-Koda et al., 2021; Li et al., 2023). Experimental structures, prototype substitutions, chemical constraints, stability filters, generative proposals, structural perturbations and active-learning criteria have all provided effective ways to direct finite computational resources (Jain et al., 2013; Kirklin et al., 2015; Wang et al., 2021; Mehl et al., 2017; Merchant et al., 2023; Zeni et al., 2025; Chanussot et al., 2021; Tran et al., 2023; Sriram et al., 2024; Barroso-Luque et al., 2026). These policies have made modern first-principles datasets possible. They also become part of the resulting data, because the assumptions used to propose calculations determine which compositions, configurations and physical states are available to learn. As frontier AI begins to move from fixed corpora towards generated experience, materials science faces a corresponding question. Can first-principles data generation scale through a more general search process, allowing additional computation to broaden the observed materials world without requiring a parallel expansion of hand-designed chemical and structural rules?

We present **ELEMENTA**, a first-principles dataset constructed to pursue this form of unbiased scaling. Across its supported elements, **ELEMENTA** systematically enumerates unary, binary and ternary composition representable with no more than twelve atoms per cell and explores each composition through systematic random structure search followed by density-functional-theory relaxation. The core corpus contains approximately **210 million relaxation frames** spanning diverse configurations, pathways and competing local minima. **ELEMENTA Spin** adds approximately **4 million frames** from systematic magnetic-moment scans of selected structures, while **ELEMENTA Vib** contributes approximately **8 million displaced configurations** that probe local vibrational response and energy-surface curvature. The collection is designed both as a training substrate for atomistic foundation models and as a benchmark of physical generalization. Evaluations across representative models reveal structured, physically interpretable failure modes that aggregate regression metrics can conceal, including insufficient resolution of subtle energy differences among competing local minima required for reliable ground-state identification, together with biases in ionic coordination, magnetic ordering, crystal symmetry, and local bonding environments. **ELEMENTA** tests whether first-principles scaling can deliver both a broader domain of physical experience and the resolution needed to distinguish the competing states on which materials discovery depends.

2 Scaling philosophy

The frontier of artificial intelligence has repeatedly advanced by changing where human knowledge enters the system. Pre-training scaled exposure to data (Kaplan et al., 2020; Hoffmann et al., 2022), inference-time computation scaled search and verification (Lightman et al., 2024; Snell et al., 2025; DeepSeek-AI, 2025), and emerging agents increasingly generate experience through interaction (Wang et al., 2024). In each case, human effort moved away from specifying individual solutions and towards constructing procedures that improve as computation grows. **ELEMENTA** applies this principle to first-principles data generation.

Materials data are usually produced only after deciding which compositions and structures appear sufficiently plausible to calculate (Jain et al., 2013; Kirklin et al., 2015; Merchant et al., 2023; Barroso-Luque et al., 2026). **ELEMENTA** instead uses a general search procedure over a chemically broad domain. The same eligibility and sampling logic are applied across supported elements and across all unary, binary and ternary stoichiometries with up to twelve atoms per cell. No known crystal prototype is required, and neither common elements nor familiar composition ratios receive privileged treatment. Human knowledge enters through universal physical constraints that prevent pathological structures, not through an expanding list of expected material forms.

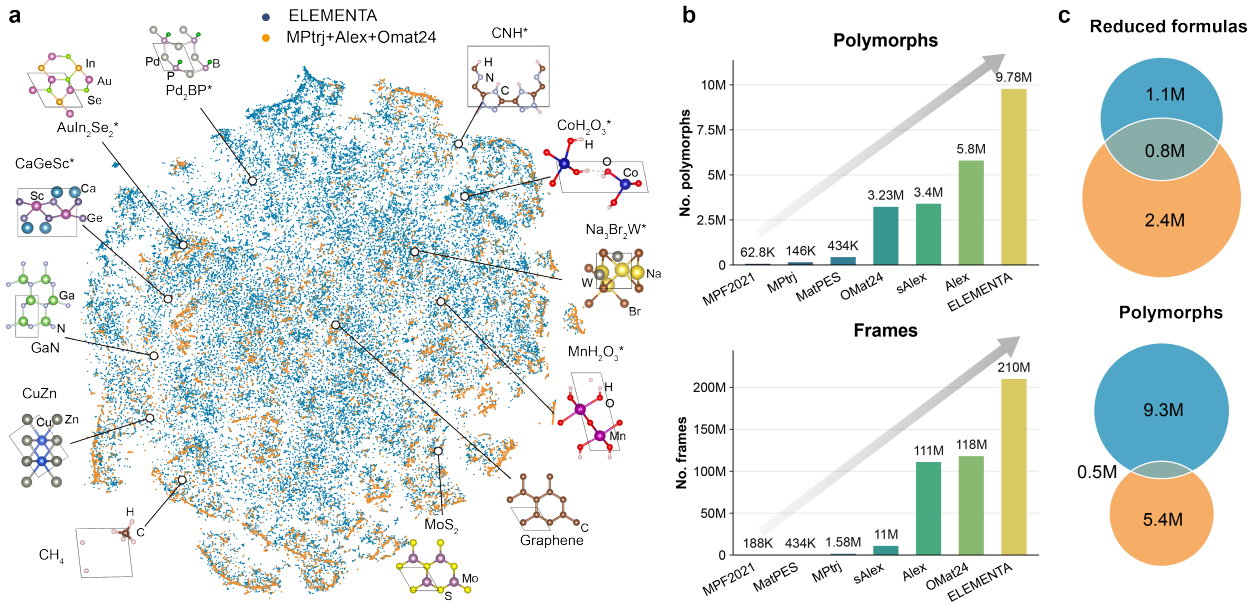

Figure 1 Overview of **ELEMENTA** dataset. **a**, Distributions of local atomic environments in **ELEMENTA** and existing datasets based on Smooth Overlap of Atomic Positions (SOAP) (Bartók et al., 2013). Blue denotes **ELEMENTA**, and orange denotes existing datasets. Representative materials are shown; an asterisk (*) denotes materials that are unique to **ELEMENTA**. **b**, Comparison of the numbers of polymorphs and frames in **ELEMENTA** and various materials datasets. **c**, Overlap of formulas and polymorphs between **ELEMENTA** and databases consisting of MPtrj, Alex, and OMat24.

This changes how computation is converted into data. Additional search can reveal new coordination environments, structural basins and bonding regimes without requiring new composition-specific rules. Figure 1a shows the resulting distribution through a projection of local atomic environments and representative relaxed structures. Known crystals coexist with alternative polymorphs, molecular and extended forms, low-dimensional motifs, database-unmatched minima and ordered defect-like environments.

ELEMENTA Spin and **ELEMENTA Vib** extend the same procedure beyond structural search. They expose magnetic initialization and local displacement as additional coordinates, and allow first-principles calculation to resolve the states and responses that follow. Scale is therefore expressed as broader physical experience across composition, configuration and state. The dataset grows by observing more of what atomic interactions permit, not by continuously enlarging a catalogue of what materials are expected to resemble.

3 Generation Method

ELEMENTA adopts an unbiased, or more precisely, low-prior data generation strategy that begins from systematic composition enumeration and independent structure generation, rather than known crystal prototypes or database-template substitutions. This design separates the exploration of chemical space from the historical distribution of known materials, allowing both familiar and rarely studied compositions to be treated under the same sampling procedure. As illustrated in Figure 2a, the workflow combines random structure search, MLFF-assisted pre-relaxation, structural deduplication and ranking, DFT refinement, and extensions to spin and vibrational degrees of freedom.

3.1 Composition enumeration

At the compositional level, **ELEMENTA** systematically enumerates unary, binary, ternary, and selected higher-order formula spaces across the periodic table covering 84 elements, using only element combinations and reduced-stoichiometry size constraints. The binary and ternary subsets are defined as follows.

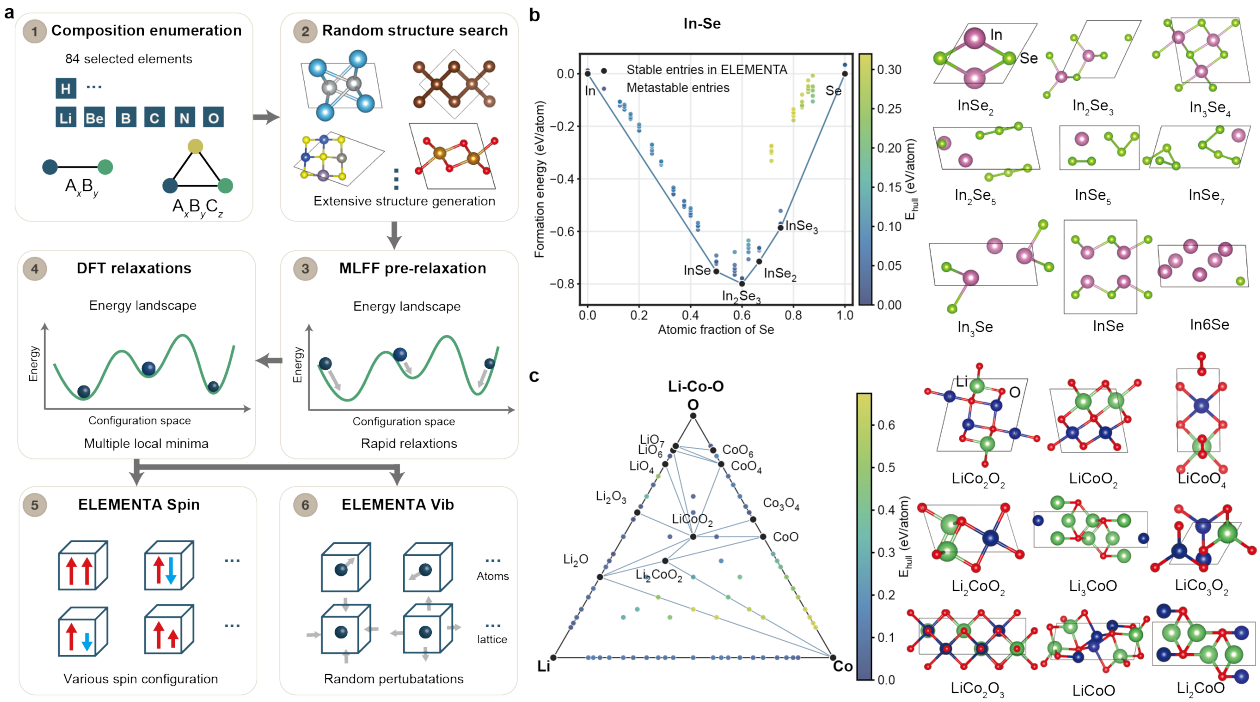

Figure 2 Generation workflow and representative chemical spaces in **ELEMENTA**. **a**, Schematic of the **ELEMENTA** data-generation pipeline. **b,c**, Representative binary In–Se and ternary Li–Co–O chemical spaces sampled by **ELEMENTA**. Stable entries are shown in black, while metastable entries are colored by energy above hull E_{hull}. Representative relaxed structures illustrate stable and metastable polymorphs covered in these chemical systems.

Binary compounds define the pairwise elemental interaction space. They are constructed from reduced stoichiometries of the form A_xB_y, subject to

$$1 \le x, y, \qquad x + y \le 8, \qquad \gcd(x, y) = 1.$$

This enumeration gives 3,486 unique binary element combinations and 73,206 reduced formulas.

Ternary compounds form the largest part of the compact search space. They are constructed from reduced stoichiometries of the form $A_xB_yC_z$, subject to

$$1 \le x, y, z, \qquad x + y + z \le 6, \qquad \gcd(x, y, z) = 1.$$

This enumeration yields approximately 95,284 unique ternary element combinations and more than 1.8 million reduced stoichiometries. This ternary subset is central to the dataset, extending beyond pairwise chemistry to capture multi-element coordination, competing bonding motifs, and the broader compositional complexity characteristic of realistic materials.

Beyond the compact ternary search space, **ELEMENTA** further includes representative ternary and quaternary systems with reduced formula sizes extending to $x + y + z + w \le 20$. These additions improve coverage of realistic materials families that cannot be adequately represented within the compact $x + y + z \le 6$ space alone. The resulting structures contain up to 40 atoms per simulation cell, enabling greater compositional and structural complexity.

3.2 Random Structure Generation and MLFF Pre-relaxation

For each enumerated composition, **ELEMENTA** generates candidate crystal structures using random structure search (RSS) (Pickard and Needs, 2011). A total of 100 initial structures are sampled for each fixed composition

without using predefined crystal prototypes or database-template substitutions. The initial cell volume is estimated as the sum of the elemental reference volumes weighted by stoichiometry. To balance structural diversity and physical plausibility, both low-symmetry and moderately symmetric cells are sampled by allowing one to four symmetry operations. During random packing, the lattice shape is allowed to adapt to the atomic-size constraints and is reduced to a compact representation. Species-dependent minimum interatomic distances are imposed to prevent unphysically short contacts, and candidate structures retaining excessive atomic overlap are rejected.

The generated structures are pre-relaxed using a machine-learning force field (MLFF), MatterSim-v1-5M, (Yang et al., 2024). A small hydrostatic external pressure is applied to suppress unphysically low-density structures. After optimization, candidate structures are ranked by their predicted energies and filtered using structural deduplication to remove equivalent and near-duplicate configurations. The five lowest-energy unique structures are then selected for first-principles refinement. This step substantially reduces the cost of large-scale DFT calculations while preserving chemically and structurally diverse candidates.

3.3 DFT Relaxation

The selected structures are refined using density functional theory (DFT) calculations performed with the Vienna *ab initio* Simulation Package (VASP) (Kresse and Furthmüller, 1996). The workflow follows the Materials Project relaxation protocol and is implemented through the `MPRelaxSet` class in `pymatgen` (Jain et al., 2013; Ong et al., 2013), ensuring consistent calculation settings across the dataset. Exchange-correlation effects are treated using the Perdew–Burke–Ernzerhof (PBE) functional within the generalized gradient approximation, together with projector augmented-wave (PAW) pseudopotentials (Perdew et al., 1996; Blöchl, 1994; Kresse and Joubert, 1999). For transition-metal oxides and fluorides, Hubbard U corrections are applied following the standard Materials Project scheme (Wang et al., 2006). Initial magnetic moments are assigned using element-dependent initialization strategies derived from the Materials Project workflow. Structural relaxations are performed using consistent electronic and ionic convergence criteria. Calculations that fail to converge or exhibit numerical instabilities are removed through automated quality-control procedures.

The resulting first-principles calculations provide total energies, atomic forces, stress tensors, and site-resolved magnetic moments as training targets. For each composition, **ELEMENTA** does not retain only a single relaxed structure. Instead, multiple low-energy candidates selected from RSS are refined by DFT, yielding distinct local minima that can differ in coordination environment, bonding motif, lattice geometry, and magnetic state despite sharing the same chemical composition.

For each DFT refinement, **ELEMENTA** also retains the relaxation trajectory, including intermediate configurations sampled along the optimization path toward local equilibrium. These trajectory frames provide off-equilibrium geometries with physically consistent energy, force, and stress labels. As a result, the dataset represents both the relaxed local minima and the nearby regions of the potential energy surface explored during relaxation, providing richer training signal than a collection of static ground-state structures alone.

Following this workflow, **ELEMENTA** covers approximately 1.9 million reduced formulas, yields over 9.7 million distinct DFT-relaxed polymorphic structures, and contains approximately 210 million DFT-labeled frames (Figure 1c). About 0.5 million structures overlap with historical materials databases, spanning known inorganic oxides, sulfides, intermetallic compounds, alloys, organic chain-like structures, and small-molecule crystals. These known structures are not introduced through database templates, crystal prototypes, or hand-crafted rules, but are rediscovered through the low-prior search procedure. The remaining approximately 9.3 million polymorphs correspond to new chemical compositions or new atomic configurations, including local minima, coordination motifs, and bonding environments that are not recorded in existing databases. These structures expand the dataset beyond known materials and provide a direct testbed for evaluating generalization across unexplored chemical and structural space. Representative phase diagrams and diverse structures for the binary In–Se and ternary Li–Co–O systems constructed from **ELEMENTA** are shown in Figure 2b,c.

4 Statistics of ELEMENTA

ELEMENTA contains broad distributions of relaxed crystal symmetries, numbers of atoms in the primitive cell, and polymorphic energy landscapes. Figure 3a shows the distribution of relaxed structures by crystal system and space group. The dataset spans a broad range of crystal symmetries, with triclinic, monoclinic, and orthorhombic systems representing the largest fractions. This trend is expected because the MLFF pre-relaxation and DFT relaxation are performed without enforcing symmetry constraints, allowing structures to lower their symmetry during optimization. Nevertheless, higher-symmetry systems, including tetragonal, trigonal, hexagonal, and cubic structures, are recovered, indicating that the search procedure can rediscover ordered symmetric minima when they remain energetically favorable after relaxation.

The primitive-cell size distribution is shown in Figure 3b. The structures span a range of primitive-cell sizes, with prominent populations at 6, 8, 10, and 12 atoms per primitive cell. This distribution follows from the enumerated stoichiometric constraints and subsequent symmetry reduction after relaxation. The resulting dataset therefore spans both small and relatively large primitive cells, increasing the diversity of local coordination environments and medium-range structural motifs available for model training.

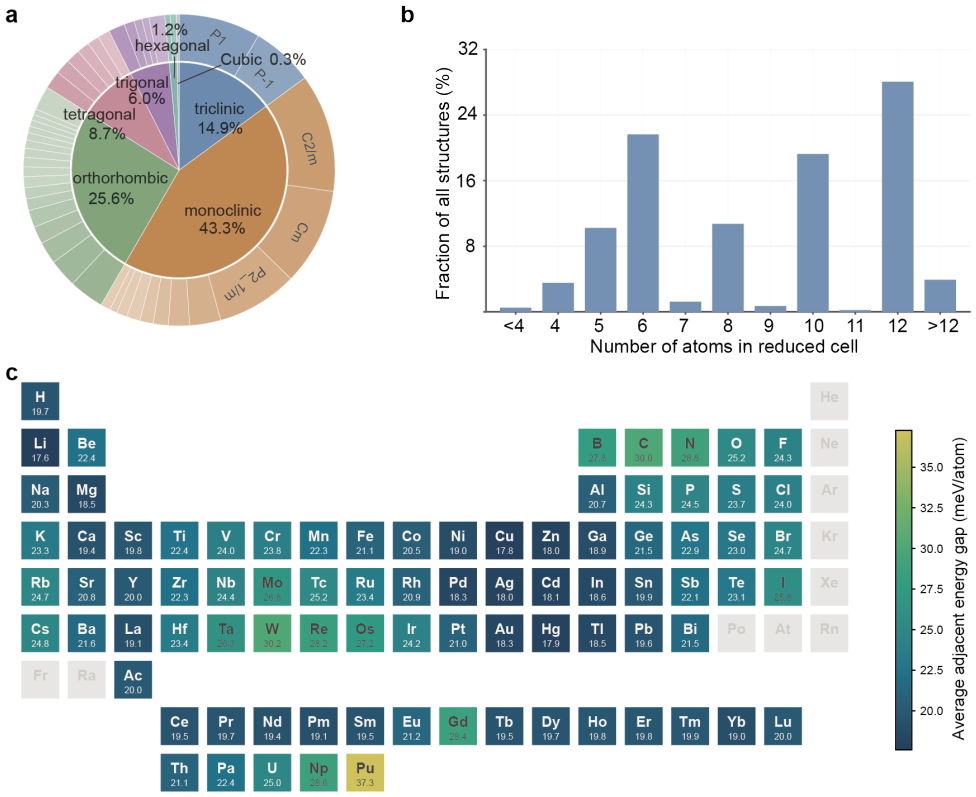

Figure 3 Structural statistics and polymorphic energy diversity in **ELEMENTA**. **a**, Distribution of relaxed structures by crystal system and space group, identified using a symmetry tolerance of `symprec` $= 0.5$. The inner ring shows the fraction of structures in each crystal system, and the outer ring shows the corresponding distribution over space groups. **b**, Distribution of the number of atoms in the primitive cell for relaxed structures, using the same symmetry tolerance. **c**, Element-resolved energy scale for polymorph ranking. For each formula, the five retained local minima are sorted by energy, and the mean adjacent energy gap is calculated from the four consecutive energy differences. Colors and numbers show the average of this quantity for formulas containing each element, indicating the energy resolution required to distinguish their relative stability. Grey tiles denote elements for which statistics are unavailable.

ELEMENTA further reveals an element-dependent energy scale for distinguishing competing polymorphs. Figure 3c reports, for each element, the average adjacent energy gap between the five retained local minima after energy sorting, averaged over formulas containing that element. This quantity characterizes the energy

resolution required to correctly rank polymorphs by stability. Larger gaps, as observed for formulas containing C, W, Re, Gd, Np, and Pu, indicate more clearly separated local minima whose relative stability can be distinguished with a comparatively moderate energy accuracy. Conversely, smaller gaps correspond to closely competing polymorphs and therefore impose a more stringent accuracy requirement on machine-learning interatomic potentials: prediction errors comparable to these gaps may alter the energetic ordering and lead to an incorrect stability ranking. The element-dependent variation of this energy scale thus provides a direct measure of the precision required from a force field for reliable polymorph ranking across different chemical systems.

To further resolve these polymorphic energy landscapes, Figure 4a–d shows the distributions of relative energies among retained local minima. For each composition, the lowest-energy structure among the retained DFT-relaxed candidates is used as the reference ($\Delta E = 0$). Throughout this work, we refer to this structure as the DFT ground state for convenience, although it represents the lowest-energy structure identified within the present search rather than a guaranteed global minimum. To focus on the energy distribution of metastable polymorphs, the reference structure ($\Delta E = 0$) is excluded, and only the remaining retained local minima are included. Across the full dataset, most retained minima lie within a low-energy window, while a long tail extends to higher relative energies. The representative material classes show related but distinct energy statistics. Binary alloys exhibit a steep, nearly monotonic decline in density with increasing ΔE, indicating that retained polymorphs are predominantly concentrated near the ground state. In contrast, transition-metal oxides display the most broadly distributed energy landscape, with substantial populations extending across the sampled energy range and a pronounced intermediate-energy shoulder, reflecting competing coordination and oxidation-state arrangements. Chlorides show an intermediate behavior, with a more gradual decay than binary alloys but a less uniform distribution than transition-metal oxides. These class-dependent distributions show that **ELEMENTA** samples not only the most stable structure for each formula, but also diverse metastable configurations that remain relevant for finite-temperature behavior and synthesis pathways (Sun et al., 2016).

Representative examples are shown in Figure 4e–h. For Bi_2Te_3, $CaMnNi_2$, $SrTiO_3$, and $CsPbCl_3$, the retained structures include distinct local minima with different coordination motifs, lattice geometries, and dimensionalities. These examples illustrate the role of polymorphism in **ELEMENTA**: a single chemical formula can correspond to multiple physically plausible atomic arrangements separated by finite energy gaps. By retaining this multiplicity, the dataset provides atomistic models with a richer view of the potential energy surface than would be obtained from ground-state structures alone.

5 Magnetism and Vibrations

In the base **ELEMENTA** dataset, each polymorph is initialized with a ferromagnetic (FM) spin configuration when magnetic moments are present. Although this provides a consistent default setting, a single FM initialization cannot sample the full magnetic energy landscape (Sanspeur and Kitchin, 2024; Xu et al., 2025). To construct **ELEMENTA Spin**, we selected approximately 50 thousand magnetic materials from **ELEMENTA**, focusing primarily on systems containing $3d$, $4d$, $4f$, and $5f$ elements, where at least one atom carries a non-zero initial magnetic moment. Starting from the corresponding **ELEMENTA** structures, we generated up to ten collinear magnetic initializations for each material using a targeted random sampling strategy. Relative to the reference FM configuration, the sampling favored configurations in which one or more local magnetic moments are reversed, thereby efficiently covering antiferromagnetic (AFM), ferrimagnetic (FiM), and other collinear spin arrangements with a limited number of initializations. First-principles calculations are then used to resolve the self-consistent magnetic states, their relative energies, and the associated relaxation trajectories, yielding over 4 million trajectory frames in **ELEMENTA Spin**.

Figure 5a and b summarize the magnetic statistics collected in **ELEMENTA Spin**. As shown in Figure 5a, the site-resolved magnetic moments of Cr, Mn, and Fe span a broad range from nearly zero to about 5 μ_B, with pronounced multimodal distributions. These features indicate that the same magnetic element can adopt multiple local-moment states depending on its chemical environment, oxidation state, and magnetic initialization. Figure 5b further compares the relative energies of sampled magnetic configurations for $SrCr_2As_2$, LiMnAs, and $AlFe_2S_3$. The sampled AFM, FiM, and FM configurations converge to distinct self-consistent states with energy separations ranging from tens to hundreds of meV/atom. In these examples,

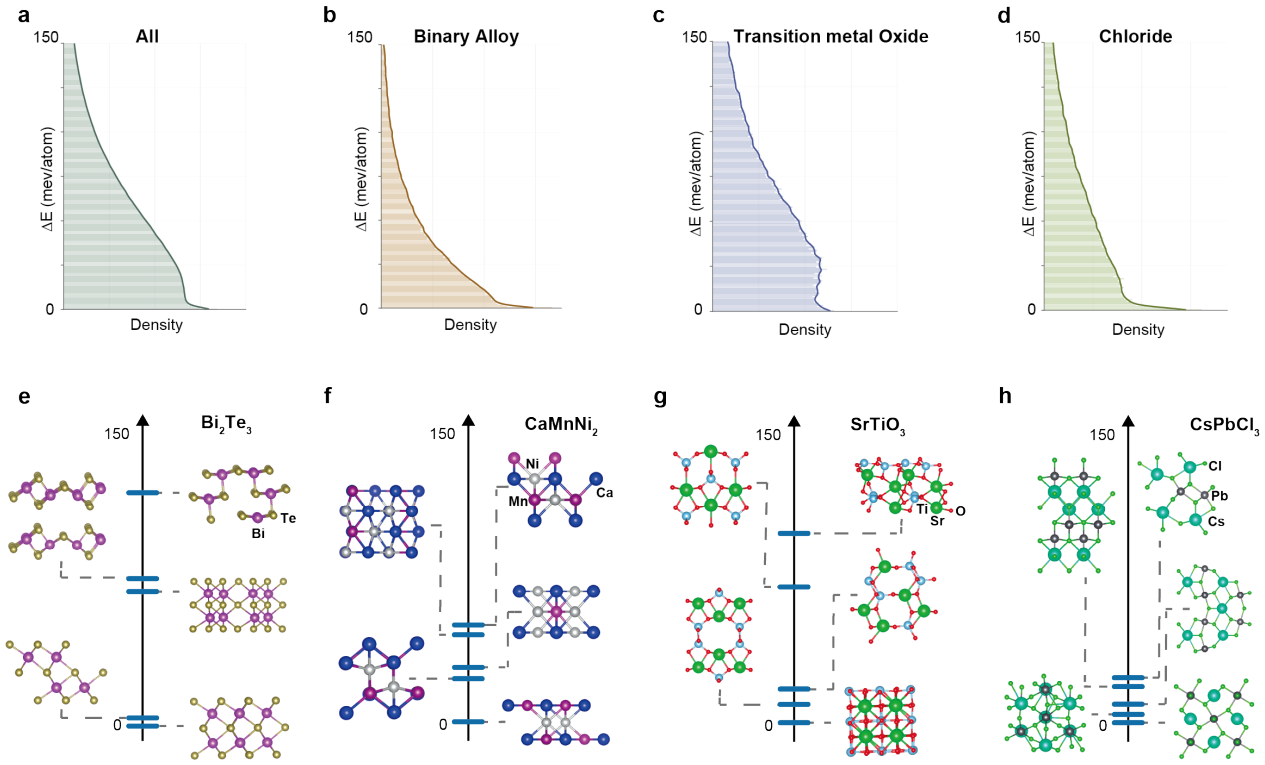

Figure 4 Polymorphic energy landscapes sampled by **ELEMENTA**. **a–d**, Energy statistics of retained local minima for all compositions, binary alloys, transition-metal oxides, and chlorides, respectively. For each formula, the lowest-energy structure defines the reference $\Delta E = 0$, and only the remaining local minima, measured relative to this reference, are included to characterize the metastable energy landscape. Histograms show the empirical density of relative energies, and curves indicate smoothed density estimates. **e–h**, Representative local minima for selected formulas, including Bi_2Te_3, $CaMnNi_2$, $SrTiO_3$, and $CsPbCl_3$. Structures are arranged along the vertical energy axis according to their relative energies, illustrating that **ELEMENTA** retains multiple distinct polymorphs for the same formula across layered, alloy, oxide, and halide systems.

the lowest-energy states are not necessarily obtained from the default FM initialization, highlighting the importance of explicitly sampling competing magnetic orderings.

These statistics highlight the intrinsic multiplicity of magnetic states in crystalline materials. The default FM configuration is insufficient to describe this behavior, because the same crystal structure may support multiple magnetic arrangements with distinct local moments and energies. By providing DFT-resolved energies and local magnetic moments for multiple spin configurations, **ELEMENTA Spin** establishes a foundation for quantitatively describing the coupling among magnetic order, crystal structure, and chemical environment. In particular, the relative energies of competing magnetic configurations encode information about effective exchange interactions and can be mapped onto spin Hamiltonians (Xiang et al., 2013), such as

$$\mathcal{H} = -\sum_{i<j} J_{ij}\, \mathbf{S}_i \cdot \mathbf{S}_j,$$

to extract exchange parameters J_{ij}. These parameters provide microscopic inputs for Monte Carlo and related statistical-mechanical simulations of magnetic transition temperatures and finite-temperature behavior. **ELEMENTA Spin** therefore supports a description of magnetic interactions beyond the single-state picture commonly provided by previous materials datasets (Deng et al., 2023; Barroso-Luque et al., 2026; Schmidt et al., 2024).

Beyond magnetic degrees of freedom, another important aspect is the lattice response to small atomic displacements (Liao et al., 2024a; Deng et al., 2025). Static equilibrium structures alone often do not fully characterize the local potential energy surface. Although a structure may converge to a DFT-relaxed

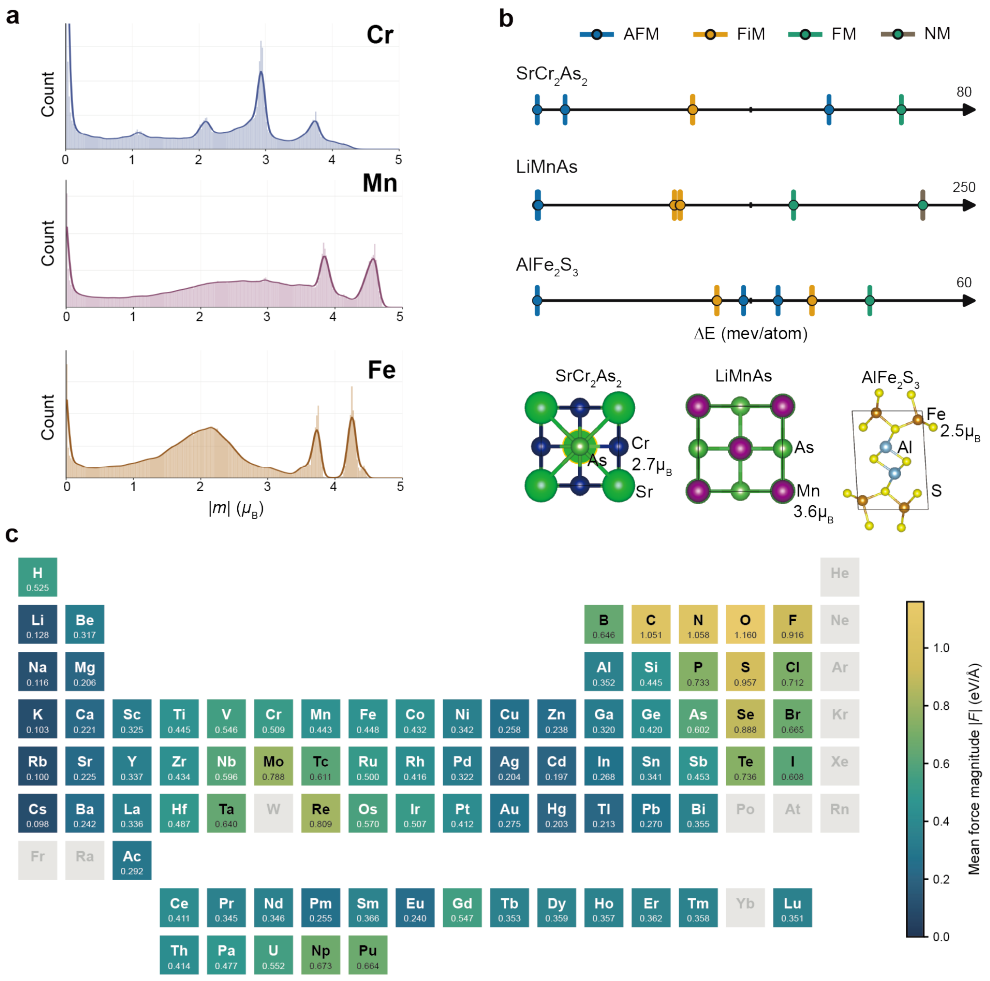

Figure 5 Statistics of **ELEMENTA Spin** and **ELEMENTA Vib**. **a**, Distributions of site-resolved magnetic moments ($|m|$) for Cr, Mn, and Fe in **ELEMENTA Spin**. **b**, Relative energies ΔE of sampled magnetic configurations for representative compounds $SrCr_2As_2$, LiMnAs, and $AlFe_2S_3$. The configurations include antiferromagnetic (AFM), ferrimagnetic (FiM), ferromagnetic (FM), and non-magnetic (NM) states. Representative crystal structures and labeled local magnetic moments are shown below the energy diagrams. **c**, Element-resolved force statistics in **ELEMENTA Vib** after local atomic perturbations. Each periodic-table tile is colored by the mean force magnitude $|F|$ on atoms of that element in the perturbed structures, with grey tiles indicating elements without available statistics.

minimum, this does not guarantee dynamical stability under small perturbations. Two polymorphs with similar relaxed energies may exhibit markedly different responses to atomic displacements: one may remain locally stable, whereas another may undergo spontaneous symmetry breaking and relax toward a lower-symmetry configuration. Such lattice instabilities are common in crystalline materials, for example the oxygen-octahedral rotations frequently observed in perovskites. To capture this missing information, **ELEMENTA Vib** systematically samples the local potential energy surface around relaxed crystal structures through controlled atomic and lattice perturbations.

Starting from **ELEMENTA**, we randomly selected structures spanning approximately 550 thousand distinct polymorphs to ensure broad chemical and structural coverage. Structures containing fewer than 20 atoms are first expanded into compact supercells containing at least 20 atoms. For each selected polymorph, the equilibrium volume is isotropically scaled to $V/V_0 = 0.95$, 1.00, and 1.05. At each volume, five perturbed configurations are generated by applying random atomic displacements with amplitudes ranging from 0.005 to 0.03 . All generated configurations are evaluated using DFT single-point calculations to obtain total energies, atomic forces, and stress tensors, resulting in more than 8 million labeled configurations. These controlled near-equilibrium samples provide direct supervision of the energetic, force, and stress responses to lattice and atomic distortions, enabling foundation models to learn the local curvature of the potential energy surface beyond equilibrium structures alone.

The element-resolved force statistics reveal systematic differences in the local restoring forces sampled by **ELEMENTA Vib**, as shown in Figure 5c. Within the harmonic approximation (Togo and Tanaka, 2015), the restoring force acting on atom i can be expressed as

$$\mathbf{F}_i \approx -\sum_j \mathbf{H}_{ij}\mathbf{u}_j,$$

where $\mathbf{H}_{ij}$ is the corresponding block of the Hessian matrix (equivalently, the second-order interatomic force-constant matrix), and $\mathbf{u}_j$ is the displacement of atom j from its equilibrium position. Because the displacement amplitudes are sampled from the same narrow range across all structures, larger mean force magnitudes generally correspond to stiffer local environments, whereas smaller forces indicate softer bonding environments. Light non-metal elements, particularly C, N, O, S, and Se, exhibit relatively large mean force magnitudes, while alkali and alkaline-earth elements generally show weaker restoring responses. Transition-metal, post-transition-metal, and rare-earth elements span a broad intermediate range. These statistics characterize the average local stiffness sampled across different chemical and coordination environments. Overall, **ELEMENTA Vib** captures diverse near-equilibrium force responses across chemical space, substantially enriching the representation of local potential energy surfaces.

6 Benchmark on ELEMENTA

Consistent with the low-prior scaling strategy introduced above, most structures in **ELEMENTA** were not intentionally sampled from known materials, model-specific training distributions, or predefined failure cases. Instead, a composition-complete, prototype-independent search strategy was applied uniformly across chemical space, allowing additional computation to broaden the diversity of sampled physical environments rather than reinforce existing human-curated or model-specific distributions. This model-agnostic construction reduces benchmark selection bias and provides a fair basis for comparing atomistic foundation models. The resulting diversity of compositions, crystal structures, and local minima further enables systematic evaluation of model generalization across both chemical and configurational spaces.

Seven representative model checkpoints are evaluated: M3GNet (Chen and Ong, 2022), CHGNet (Deng et al., 2023), MACE-MP-0 (Batatia et al., 2025), MatterSim-v1-5M (Yang et al., 2024), NequIP-OAM-XL (Batzner et al., 2022), EquFlash-v2 (Lee et al., 2025), and Equiformer-v3 (Liao et al., 2024b, 2026). These models span successive stages in the expansion of first-principles training data, from MPF.2021 (188K structures) (Chen and Ong, 2022), through MPtrj (1.58M structures) (Deng et al., 2023), to the combined OMat24, MPtrj, and sAlex datasets (113M structures) (Barroso-Luque et al., 2026; Deng et al., 2023; Schmidt et al., 2024). This progression enables systematic evaluation of how increasing data scale, diversity, and physical coverage influence model accuracy and generalization.

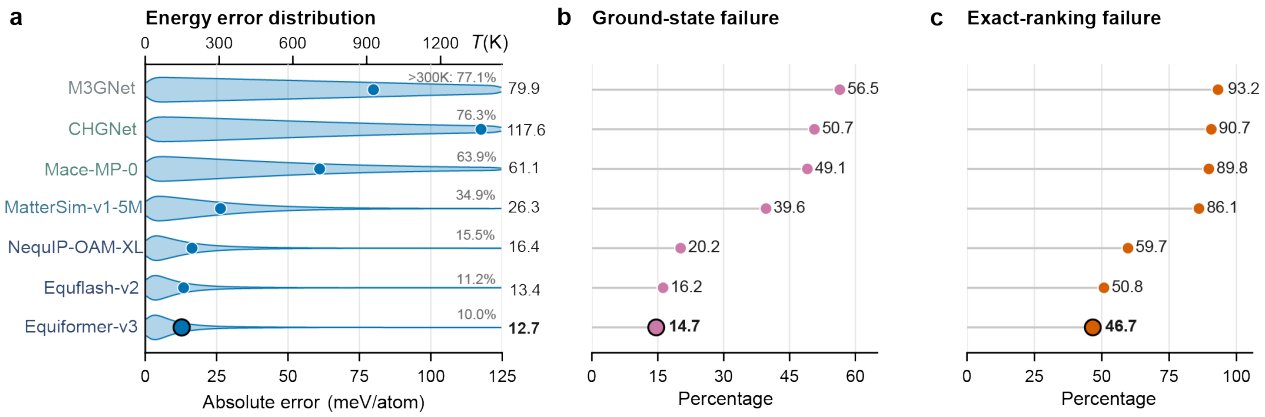

Figure 6 Benchmarking atomistic foundation models on local energy landscapes in **ELEMENTA**. **a,** Absolute energy prediction errors of seven representative atomistic foundation models (M3GNet, CHGNet, MACE-MP-0, MatterSim-v1-5M, NequIP-OAM-XL, Equflash-v2, and Equiformer-v3) evaluated on the **ELEMENTA** benchmark. Filled circles denote the mean absolute error (MAE), while percentages indicate the fraction of predictions with an absolute error exceeding 25 meV/atom (300 K).Label colors indicate training-data regimes: grey, MPF.2021; muted teal, MPtrj-based datasets; steel blue, MatterSim; and deep navy, OMat24+sAlex+MPtrj. **b,** Ground-state identification failure rates. For each composition containing multiple DFT-relaxed local minima, a prediction is considered incorrect if the model fails to identify the DFT ground state as the lowest-energy structure. **c,** Exact-ranking failure rates. A prediction is counted as failed when the complete ordering of the retained local minima does not exactly match the DFT reference ordering.

Figure 6 benchmarks atomistic foundation models from three complementary perspectives: absolute energy prediction, ground-state identification, and exact local-minima ranking. As shown in Figure 6a, models trained on MPF.2021 or MPtrj-based datasets, including M3GNet, CHGNet, and MACE-MP-0, exhibit relatively large prediction errors, with MAEs exceeding 60 meV/atom and more than 60% of predictions lying beyond the room-temperature thermal energy scale ($k_{\mathrm{B}}T \approx 25$ meV). Prediction accuracy improves consistently with increasing training-data scale and diversity, with models trained on the combined OMat24, MPtrj, and sAlex datasets achieving the best overall performance. Equiformer-v3 attains the lowest MAE of 12.7 meV/atom. Nevertheless, its error distribution remains distinctly long-tailed, with 10.0% of predictions still exceeding 25 meV.

Beyond absolute energy prediction, the multiple DFT-relaxed local minima associated with each composition in **ELEMENTA** enable direct evaluation of relative stability through ground-state identification (Figure 6b) and exact local-minima ranking (Figure 6c). These stability-oriented benchmark tasks are largely absent from existing atomistic benchmarks. As shown in Figure 6b, models trained on MPF.2021 or MPtrj-based datasets fail to identify the DFT ground state in five candidates for more than 40% of compositions. The failure rate decreases consistently with increasing training-data scale and diversity, indicating that larger and more diverse first-principles datasets substantially improve a model's ability to distinguish the lowest-energy structure among competing local minima. Nevertheless, the best-performing model, Equiformer-v3, still fails to identify the DFT ground state for 14.7% of compositions.

A more stringent evaluation is provided by exact local-minima ranking (Figure 6c), which requires recovering the complete energetic ordering of all competing local minima rather than simply identifying the ground state. For M3GNet, CHGNet, MACE-MP-0, and MatterSim-v1-5M, the ranking failure rate exceeds 80%, highlighting the difficulty of resolving the full stability hierarchy. Although the same trend of improvement is retained, the top-performing model, Equiformer-v3, still fails to reproduce the exact DFT ranking for 46.7% of compositions. This progressive increase in task difficulty demonstrates that improvements in global energy prediction do not directly translate into sufficient energetic resolution to recover the complete stability hierarchy. Exact local-minima ranking requires models to reliably resolve subtle energy differences among competing metastable structures, a capability that becomes increasingly important beyond static 0 K energy ordering, where thermal fluctuations render multiple local minima thermally accessible. Consequently, practical prediction of materials stability requires atomistic foundation models to achieve not only low average energy errors but also sufficient energetic resolution to faithfully reconstruct the underlying local energy landscape.

Exact local-minima ranking therefore provides a substantially more stringent assessment of this capability than conventional energy regression.

7 Failure Modes of Atomistic Foundation Models

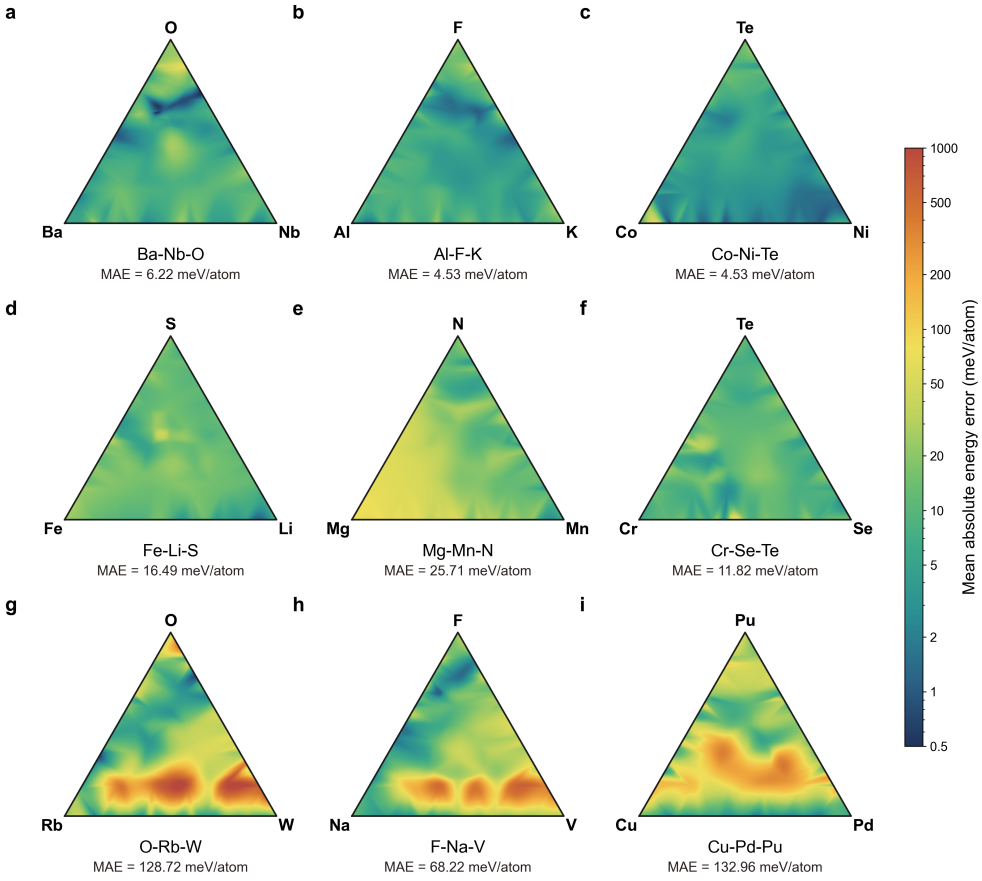

Figure 7 Inhomogeneous composition-dependent errors of Equiformer-v3 on the **ELEMENTA** benchmark. Ternary maps of composition-resolved energy prediction errors for representative chemical systems. Panels a-c, d-f, and g-i show systems with low, intermediate, and high overall errors, respectively. Colors indicate the absolute prediction error, and the corresponding MAE is reported below each map. The pronounced variations within individual ternary spaces reveal strongly inhomogeneous model accuracy across compositions.

Prediction errors are distributed unevenly across chemical space. Taking Equiformer-v3 as a representative example, Figure 7 shows substantial differences in prediction accuracy across ternary chemical systems and among compositions within the same ternary space. Some compositions exhibit relatively low errors, whereas others show markedly larger deviations, indicating a dependence of model accuracy on chemical composition. These heterogeneous error patterns motivate a closer examination of the conditions under which current atomistic foundation models exhibit reduced accuracy. We further examined representative cases in which the model incorrectly identifies the ground-state structure. Importantly, these are not merely near-degenerate ambiguities: the model-selected structures often lie tens to hundreds of meV/atom above the DFT ground state, with some penalties approaching or exceeding 1 eV/atom. This indicates that the observed failures correspond to substantial errors in ground-state ranking and relative stability, rather than small uncertainties among nearly equivalent structures.

As shown in Figure 8, failures in identifying the DFT ground state arise from multiple physically distinct mechanisms rather than a single universal error mode. One common failure mode is the preference for structures containing weakly coordinated or more isolated alkali atoms, suggesting that the model systematically over-

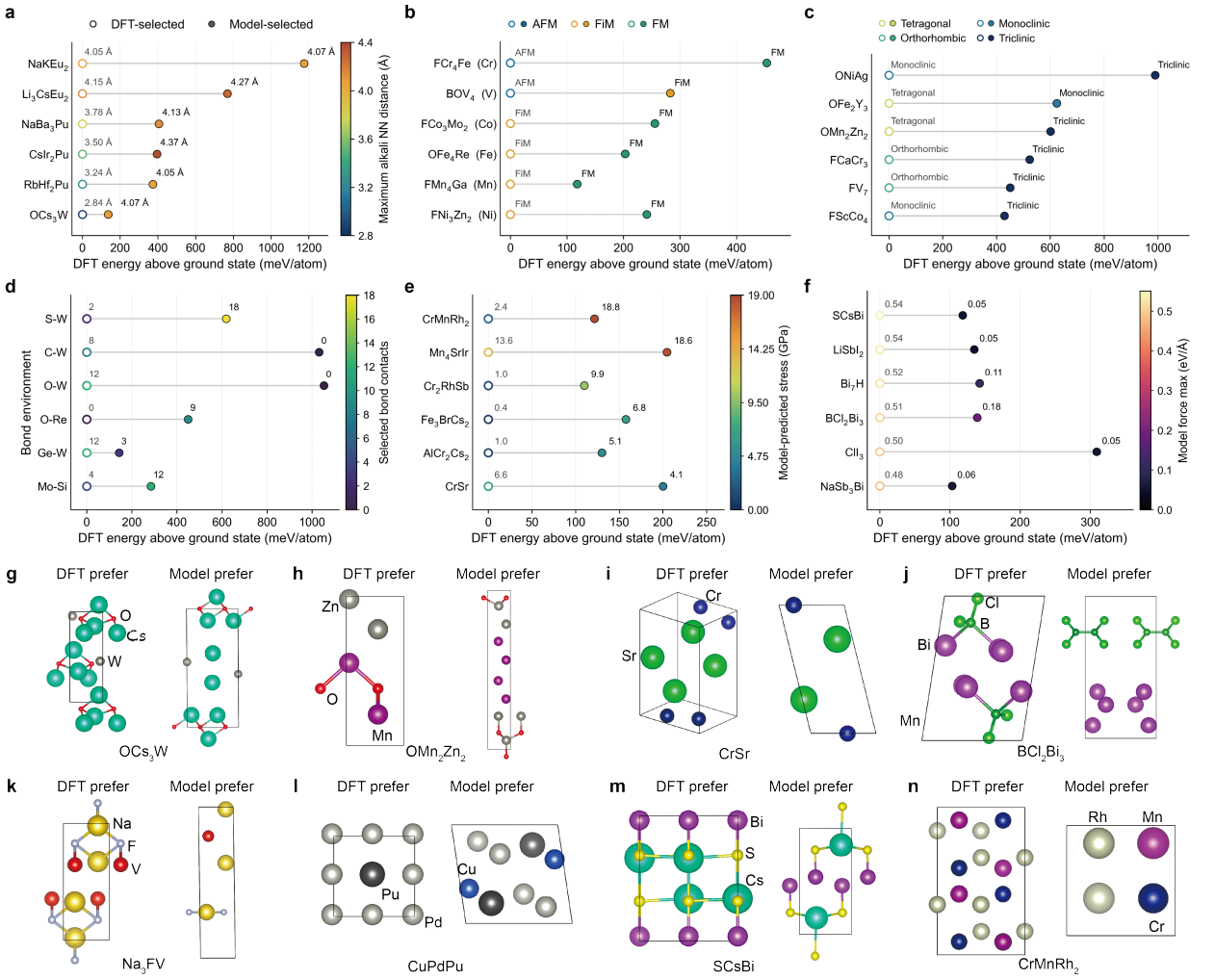

Figure 8 Representative failure modes in model ground-state selection. Open circles denote the DFT ground state among the five candidate structures, whereas filled circles denote the structure identified as the ground state by the model. The horizontal axis reports the DFT energy above the true ground state, corresponding to the DFT energy penalty of the model-selected structure. Colors and annotations indicate either the defining structural descriptor (a–d) or the diagnostic quantity (e,f) associated with each failure mode: (a) maximum alkali nearest-neighbor distance, revealing a tendency to favor structures with more isolated alkali ions; (b) magnetic ordering, showing discrepancies between the DFT and model-preferred magnetic states; (c) crystal system, illustrating symmetry changes between the DFT and model-selected structures; (d) number of selected bond contacts, highlighting differences in local bonding environments; (e) model-predicted stress (Frobenius norm); and (f) model-predicted maximum atomic force. (g–n) Representative structural examples corresponding to several of these failure modes. Compared with the DFT ground states, the model-selected structures exhibit more isolated alkali arrangements, altered local coordination environments, or different crystallographic motifs.

stabilizes under-constrained ionic environments. In magnetic materials, the model frequently favors FM configurations while DFT predicts AFM or FiM ground states, indicating an inaccurate description of the relative stability of competing spin orderings. Additional failures are associated with symmetry lowering and changes in local bonding environments. In these cases, the model incorrectly ranks lower-symmetry structures or structures with modified coordination motifs below the DFT ground state, implying systematic biases in its representation of local chemical interactions. For example, the model-selected structures may contain substantially more S–W contacts or substantially fewer O–W and C–W contacts than the corresponding DFT ground states, demonstrating that relatively small changes in specific bonding motifs can produce qualitatively different stability rankings.

Mechanical diagnostics provide further evidence that the learned potential-energy landscape differs systematically from that obtained by DFT. Although all structures shown are fully relaxed DFT local minima with negligible residual forces and stresses, the model predicts substantially larger mechanical responses for many of them. As shown in Figure 8e, several model-selected structures exhibit model-predicted stresses approaching 19 GPa despite corresponding to relaxed DFT configurations. Conversely, Figure 8f shows that the model predicts residual forces of approximately 0.5 eV/Å for several DFT ground-state structures, whereas its own preferred structures typically exhibit much smaller model-predicted forces, generally below 0.2 eV/Å. These observations indicate that DFT stationary points are not necessarily stationary points on the learned potential-energy surface, implying a systematic displacement between the learned and DFT energy landscapes.

Collectively, these analyses demonstrate that failures in ground-state identification originate from multiple coupled physical factors rather than a single dominant source of error. Misrepresentations of ionic coordination, magnetic interactions, crystal symmetry, local bonding, and mechanical equilibrium collectively distort the relative stability of competing structures, ultimately leading to incorrect energetic rankings despite relatively small absolute energy errors. These observations also explain why excellent aggregate regression metrics do not necessarily translate into reliable ground-state identification. More broadly, they motivate evaluating atomistic foundation models not only by prediction accuracy but also by their ability to faithfully preserve the topology of the underlying potential-energy landscape and the energetic ordering of competing metastable structures. Such physically informed evaluation provides a complementary perspective to conventional regression metrics and may better guide the development of next-generation foundation models for crystal structure prediction and materials discovery.

8 Limitations and Future Directions

ELEMENTA achieves broad coverage of local atomic environments by sampling across elements, stoichiometries, structural basins, magnetic states and local perturbations. Within its domain of unary, binary and ternary systems containing no more than twelve atoms per cell, it captures a wide range of coordination motifs, bonding regimes and short-range interactions that constitute much of the local vocabulary of inorganic matter. Its present scope remains less complete at longer length scales. Long-period polymorphs, dilute defects, extended stacking disorder, amorphous phases, surfaces and interfaces generally require larger cells and different boundary conditions. Long-range electrostatics, dispersion, collective magnetic order and long-wavelength lattice modes are also only partially represented. The dataset further inherits the finite search budget and the approximations of its first-principles protocol, which are most consequential for strongly correlated systems, unusual oxidation states and closely competing phases.

The computational investment required to construct **ELEMENTA** is already sufficiently large that another order-of-magnitude increase through brute-force first-principles calculation alone would be difficult to sustain. Future progress will therefore depend on new methods for scaling data generation, not only additional hardware. Promising directions include adaptive search that allocates computation according to newly discovered environments and model failure modes, multi-fidelity workflows that combine broad inexpensive exploration with selective high-accuracy calculations, and generative or active-learning procedures that increase physical coverage with less redundant sampling. These methods should extend **ELEMENTA** from broad local coverage towards larger cells, higher-order compositions and collective phenomena while improving the resolution of polymorph ordering, magnetic competition and chemically difficult regimes. The relevant measure of future scale is therefore not frame count alone, but the amount of new physical information obtained per first-principles calculation.

Data Availability

The first public release comprises nearly 39 million DFT-labeled frames from the core **ELEMENTA** dataset together with the complete 4.7 million-frame **ELEMENTA Spin** dataset. The core release covers the complete unary and binary chemical spaces across the periodic table, together with ternary compositions whose reduced-formula stoichiometric coefficients sum to no more than four. Although this represents a subset of the full 210-million-frame corpus, it is defined by explicit compositional and structural boundaries rather than random subsampling, thereby preserving complete and continuous coverage within the released chemical domain. To our knowledge, **ELEMENTA** is the first low-prior atomistic dataset to openly provide chemically complete coverage at this scale within clearly defined boundaries. The accompanying **ELEMENTA Spin** dataset further provides large-scale magnetic initializations together with DFT-resolved magnetic energy landscapes.

The released datasets establish a common benchmark domain for the development and evaluation of atomistic foundation models. Their continuous chemical-space coverage enables reproducible construction of training, validation, and test sets while facilitating fair comparison across different models. Beyond conventional energy and force prediction, they support studies of out-of-composition generalization, polymorph recovery, ground-state identification, exact local-minima ranking, crystal structure generation, and magnetic-state prediction, while enabling systematic analysis of the chemical compositions and structural environments in which current models fail. The datasets are publicly available at https://huggingface.co/datasets/kairosmaterial/ELEMENTA.

Code Availability

Code for dataset generation, data processing, and benchmark evaluation is publicly available at https://github.com/kairosmaterial/ELEMENTA. The repository includes reproducible pipelines, benchmark scripts, VASP input templates, and complete pseudopotential-selection metadata.

References

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. doi: 10.48550/arXiv.2001.08361. URL https://arxiv.org/abs/2001.08361.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. doi: 10.48550/arXiv.2501.12948. URL https://arxiv.org/abs/2501.12948.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. URL https://mlanthology.org/tmlr/2024/wang2024tmlr-voyager/.

P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864–B871, 1964. doi: 10.1103/PhysRev.136.B864. URL https://doi.org/10.1103/PhysRev.136.B864.

W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140 (4A):A1133–A1138, 1965. doi: 10.1103/PhysRev.140.A1133. URL https://doi.org/10.1103/PhysRev.140.A1133.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. doi: 10.1063/1.4812323. URL https://doi.org/10.1063/1.4812323.

Matthew K. Horton et al. Accelerated data-driven materials science with the materials project. *Nature Materials*, 24 (10):1522–1532, 2025. doi: 10.1038/s41563-025-02272-0. URL https://doi.org/10.1038/s41563-025-02272-0.

Scott Kirklin, James E. Saal, Bryce Meredig, Alex Thompson, Jeff W. Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1:15010, 2015. doi: 10.1038/npjcompumats.2015.10. URL https://doi.org/10.1038/npjcompumats.2015.10.

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

Jonathan Schmidt, Tiago F. T. Cerqueira, Aldo H. Romero, Antoine Loew, Fabian Jäger, Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, 48:101560, 2024. doi: 10.1016/j.mtphys.2024.101560. URL https://doi.org/10.1016/j.mtphys.2024.101560.

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

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022. doi: 10.1038/s43588-022-00349-3. URL https://doi.org/10.1038/s43588-022-00349-3.

Ilyes Batatia, Dávid P. Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, pages 11423–11436, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4a36c3c51af11ed9f34615b81edb5bbc-Abstract-Conference.html.

Ilyes Batatia et al. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18): 184110, 2025. doi: 10.1063/5.0297006. URL https://doi.org/10.1063/5.0297006.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. MatterSim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024. doi: 10.48550/arXiv.2405.04967. URL https://arxiv.org/abs/2405.04967.

Aaron D. Kaplan, Runze Liu, Ji Qi, Tsz Wai Ko, Bowen Deng, Janosh Riebesell, Gerbrand Ceder, Kristin A. Persson, and Shyue Ping Ong. A foundational potential energy surface dataset for materials. *arXiv preprint arXiv:2503.04070*, 2025. doi: 10.48550/arXiv.2503.04070. URL https://arxiv.org/abs/2503.04070.

Matthew C. Kuner, Aaron D. Kaplan, Kristin A. Persson, Mark Asta, and Daryl C. Chrzan. MP-ALOE: An r2SCAN dataset for universal machine learning interatomic potentials. *npj Computational Materials*, 11(1):352, 2025. doi: 10.1038/s41524-025-01834-9. URL https://doi.org/10.1038/s41524-025-01834-9.

Ryan Jacobs et al. A practical guide to machine learning interatomic potentials–status and future. *Current Opinion in Solid State and Materials Science*, 35:101214, 2025. doi: 10.1016/j.cossms.2025.101214. URL https://doi.org/10.1016/j.cossms.2025.101214.

Xiang Fu, Brandon M. Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 17875–17893. PMLR, 2025. URL https://proceedings.mlr.press/v267/fu25h.html.

Balázs Póta, Paramvir Ahlawat, Gábor Csányi, and Michele Simoncelli. Thermal conductivity predictions with foundation atomistic models. *arXiv preprint arXiv:2408.00755*, 2024. doi: 10.48550/arXiv.2408.00755. URL https://arxiv.org/abs/2408.00755.

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

Han Yang, Xixian Liu, Chenxi Hu, Yichi Zhou, Yu Shi, Chang Liu, Junfu Tan, Jielan Li, Guanzhi Li, Qian Wang, Yu Zhu, Zekun Chen, Shuizhou Chen, Fabian Thiemann, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Lixin Sun, Yicheng Chen, Lingyu Kong, Yeqi Bai, Deniz Gunceler, Frank Noé, Hongxia Hao, and Ziheng Lu. MatterSim-MT: A multi-task foundation model for *In Silico* materials characterization. *arXiv preprint arXiv:2605.07927*, 2026. doi: 10.48550/arXiv.2605.07927. URL https://arxiv.org/abs/2605.07927.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin D. Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023. doi: 10.1038/s41586-023-06735-9. URL https://doi.org/10.1038/s41586-023-06735-9.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025. doi: 10.1038/s41586-025-08628-5. URL https://doi.org/10.1038/s41586-025-08628-5.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, Matt Uyttendaele, C. Lawrence Zitnick, and Zachary Ulissi. The open materials 2024 (OMat24) inorganic materials dataset and models. *Nature Computational Science*, 6(6):642–652, 2026. doi: 10.1038/s43588-026-00996-w. URL https://doi.org/10.1038/s43588-026-00996-w.

Ryosuke Jinnouchi, Kazutoshi Miwa, Ferenc Karsai, Georg Kresse, and Ryoji Asahi. On-the-fly active learning of interatomic potentials for large-scale atomistic simulations. *The Journal of Physical Chemistry Letters*, 11(17):6946–6955, 2020. doi: 10.1021/acs.jpclett.0c01061. URL https://doi.org/10.1021/acs.jpclett.0c01061.

Daniel Schwalbe-Koda, Aik Rui Tan, and Rafael Gómez-Bombarelli. Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nature Communications*, 12(1):5104, 2021. doi: 10.1038/s41467-021-25342-8. URL https://doi.org/10.1038/s41467-021-25342-8.

Kangming Li, Daniel Persaud, Kamal Choudhary, Brian DeCost, Michael Greenwood, and Jason Hattrick-Simpers. Exploiting redundancy in large materials datasets for efficient machine learning with less data. *Nature Communications*, 14(1):7283, 2023. doi: 10.1038/s41467-023-42992-y. URL https://doi.org/10.1038/s41467-023-42992-y.

Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Predicting stable crystalline compounds using chemical similarity. *npj Computational Materials*, 7(1):12, 2021. doi: 10.1038/s41524-020-00481-6. URL https://doi.org/10.1038/s41524-020-00481-6.

Michael J. Mehl, David Hicks, Cormac Toher, Ohad Levy, Robert M. Hanson, Gus Hart, and Stefano Curtarolo. The AFLOW library of crystallographic prototypes: Part 1. *Computational Materials Science*, 136:S1–S828, 2017. doi: 10.1016/j.commatsci.2017.01.017. URL https://doi.org/10.1016/j.commatsci.2017.01.017.

Lowik Chanussot et al. Open catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021. doi: 10.1021/acscatal.0c04525. URL https://doi.org/10.1021/acscatal.0c04525.

Richard Tran et al. The open catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023. doi: 10.1021/acscatal.2c05426. URL https://doi.org/10.1021/acscatal.2c05426.

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

Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013. doi: 10.1103/PhysRevB.87.184115. URL https://doi.org/10.1103/PhysRevB.87.184115.

Chris J. Pickard and R. J. Needs. *Ab Initio* random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201, 2011. doi: 10.1088/0953-8984/23/5/053201. URL https://doi.org/10.1088/0953-8984/23/5/053201.

Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for *Ab Initio* total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169–11186, 1996. doi: 10.1103/PhysRevB.54.11169. URL https://doi.org/10.1103/PhysRevB.54.11169.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. doi: 10.1016/j.commatsci.2012.10.028. URL https://doi.org/10.1016/j.commatsci.2012.10.028.

John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865–3868, 1996. doi: 10.1103/PhysRevLett.77.3865. URL https://doi.org/10.1103/PhysRevLett.77.3865.

P. E. Blöchl. Projector augmented-wave method. *Physical Review B*, 50(24):17953–17979, 1994. doi: 10.1103/PhysRevB.50.17953. URL https://doi.org/10.1103/PhysRevB.50.17953.

Georg Kresse and Daniel Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B*, 59(3):1758–1775, 1999. doi: 10.1103/PhysRevB.59.1758. URL https://doi.org/10.1103/PhysRevB.59.1758.

Lei Wang, Thomas Maxisch, and Gerbrand Ceder. Oxidation energies of transition metal oxides within the GGA+U framework. *Physical Review B*, 73(19):195107, 2006. doi: 10.1103/PhysRevB.73.195107. URL https://doi.org/10.1103/PhysRevB.73.195107.

Wenhao Sun, Stephen T. Dacek, Shyue Ping Ong, Geoffroy Hautier, Anubhav Jain, William Davidson Richards, Anthony C. Gamst, Kristin A. Persson, and Gerbrand Ceder. The thermodynamic scale of inorganic crystalline metastability. *Science Advances*, 2(11):e1600225, 2016. doi: 10.1126/sciadv.1600225. URL https://doi.org/10.1126/sciadv.1600225.

Rohan Yuri Sanspeur and John R. Kitchin. Circumventing data imbalance in magnetic ground state data for magnetic moment predictions. *Machine Learning: Science and Technology*, 5(1):015023, 2024. doi: 10.1088/2632-2153/ad23fb. URL https://doi.org/10.1088/2632-2153/ad23fb.

Wenbin Xu, Rohan Yuri Sanspeur, Adeesh Kolluru, Bowen Deng, Peter Harrington, Steven Farrell, Karsten Reuter, and John R. Kitchin. Spin-informed universal graph neural networks for simulating magnetic ordering. *Proceedings of the National Academy of Sciences*, 122(27):e2422973122, 2025. doi: 10.1073/pnas.2422973122. URL https://doi.org/10.1073/pnas.2422973122.

Hongjun Xiang, Changhoon Lee, Hyun-Joo Koo, Xingao Gong, and Myung-Hwan Whangbo. Magnetic properties and energy-mapping analysis. *Dalton Transactions*, 42(4):823–853, 2013. doi: 10.1039/C2DT31662E. URL https://doi.org/10.1039/C2DT31662E.

Yi-Lun Liao, Tess Smidt, Muhammed Shuaibi, and Abhishek Das. Generalizing denoising to non-equilibrium structures improves equivariant force fields. *Transactions on Machine Learning Research*, 2024a. URL https://openreview.net/forum?id=whGzYUbIWA.

Bowen Deng, Yunyeong Choi, Peichen Zhong, Janosh Riebesell, Shashwat Anand, Zhuohan Li, KyuJung Jun, Kristin A. Persson, and Gerbrand Ceder. Systematic softening in universal machine learning interatomic potentials. *npj Computational Materials*, 11(1):9, 2025. doi: 10.1038/s41524-024-01500-6. URL https://doi.org/10.1038/s41524-024-01500-6.

Atsushi Togo and Isao Tanaka. First principles phonon calculations in materials science. *Scripta Materialia*, 108:1–5, 2015. doi: 10.1016/j.scriptamat.2015.07.021. URL https://doi.org/10.1016/j.scriptamat.2015.07.021.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022. doi: 10.1038/s41467-022-29939-5. URL https://doi.org/10.1038/s41467-022-29939-5.

Seung Yul Lee, Hojoon Kim, Yutack Park, Dawoon Jeong, Seungwu Han, Yeonhong Park, and Jae W. Lee. FlashTP: Fused, sparsity-aware tensor product for machine learning interatomic potentials. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 33143–33156. PMLR, 2025. URL https://proceedings.mlr.press/v267/lee25l.html.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. In *International Conference on Learning Representations*, 2024b. URL https:

//proceedings.iclr.cc/paper_files/paper/2024/hash/ab12e8f3443c1a789f595b18d8c597b4-Abstract-Conference.html.

Yi-Lun Liao, Alexander J. Hoffman, Sabrina C. Shen, Alexandre Duval, Sam Walton Norwood, and Tess Smidt. EquiformerV3: Scaling efficient, expressive, and general SE(3)-equivariant graph attention transformers. *arXiv preprint arXiv:2604.09130*, 2026. doi: 10.48550/arXiv.2604.09130. URL https://arxiv.org/abs/2604.09130.