Prophet: Scaling Atomistic Foundation Models Across Composition, Configuration, and Spin

Kairos Materials

research@kairosmaterials.com

Scaling remains a central driver of progress in foundation models, enabling AI systems to acquire increasingly broad and general capabilities. In materials science, atomistic foundation models turn large-scale first-principles data into transferable models of matter across broad chemical, structural and thermodynamic spaces. Scaling so far, however, has mostly enlarged the experience available to models whose inputs remain atomic species and geometry. Such models still inherit strong priors from the materials used to train them, have limited resolution over closely competing regions of the potential-energy landscape, and leave important physical degrees of freedom such as spin outside the representation. These limits affect more than numerical accuracy. They determine which material states and physical phenomena a model can distinguish at all. Here we introduce Prophet, a scalable atomistic foundation model spanning composition, configuration and spin. Prophet uses a spectrally decomposed equivariant graph neural network to resolve atomic interactions across multiple spatial scales, while Prophet-Spin introduces magnetic moments explicitly and uses seed denoising to recover self-consistent magnetic states. The Prophet family is trained on more than 300 million first-principles configurations, centered on the composition-complete, prototype-independent ELEMENTA corpus and its Vib and Spin extensions, together with other public datasets. This combines broad and low-prior sampling of chemical space with dense information on local energy landscapes and competing magnetic states. On nearly two million broadly sampled polymorphs spanning 84 elements, Prophet reduces energy error by 56.6% relative to the strongest evaluated baseline and more reliably resolves competing structures. Prophet achieves the highest Combined Performance Score on Matbench Discovery and the lowest lattice-thermal-conductivity error, with a symmetric relative mean error about 30% below the strongest baseline. Explicit spin opens a qualitatively new regime. Prophet-Spin can distinguish and search magnetic states that are inaccessible to structure-only models, while reducing energy error on 188 unseen magnetic materials from 63–72 to 19.2 meV/atom with DFT-converged magnetic moments (30.7 meV/atom from spin-denoised magnetic moments). Without system-specific fine-tuning, it searches magnetic orderings and reproduces phenomena including the antiferromagnetic ground state of LiMnAs, the Curie transition of bcc Fe and the pressure-induced b.c.c.–h.c.p. transition of iron. Together, these results point to a broader direction for atomistic foundation models in which scale is expressed not only through more data, but through broader physical coverage, finer configurational resolution and additional physical degrees of freedom.

KAIROS MATERIALS

1 Introduction

Scaling data, computation and model capacity has transformed the capabilities of artificial intelligence. Large-scale pre-training has established foundation models that transfer across tasks, while reinforcement learning and inference-time search have created further ways to convert scale into capability (Kaplan et al., 2020; Hoffmann et al., 2022; OpenAI, 2024; Snell et al., 2025; DeepSeek-AI, 2025). Materials science provides a physical setting for the same idea. First-principles calculations generate transferable supervision from the interactions that govern matter, including for materials that have never been synthesized. Successive advances in datasets and equivariant architectures have therefore enabled atomistic foundation models to describe broad regions of chemical and structural space, providing a common basis for predicting stability, dynamics and thermodynamic response across many material systems (Chen and Ong, 2022; Batatia et al., 2022; Deng

et al., 2023; Merchant et al., 2023; Yang et al., 2024; Schmidt et al., 2024; Batatia et al., 2025; Yang et al., 2026; Liao et al., 2024, 2026; Batzner et al., 2022; Lee et al., 2025).

The next limitation is no longer coverage alone. Existing datasets still reflect strong priors from known structures, prototype substitutions and stability filters, so increasing data volume can reinforce familiar regions of materials space rather than reveal new ones (Jain et al., 2013; Kirklin et al., 2015; Wang et al., 2021; Schmidt et al., 2022; Deng et al., 2023; Merchant et al., 2023; Horton et al., 2025; Barroso-Luque et al., 2026; Cavignac et al., 2026). Even where compositions are sampled, models must resolve small energy differences between competing structures and accurately reproduce the local shape of the potential-energy surface that controls vibrations and thermal transport (Togo and Tanaka, 2015; Deng et al., 2025; Póta et al., 2024, 2026). Low average energy errors do not guarantee either capability. A more fundamental boundary arises from degrees of freedom that are absent from the model itself. Previous pioneering atomistic foundation models describe energy as a function of atomic species and geometry. However, these variables do not uniquely specify the physical state. The same atomic configuration can support different local magnetic moments and spin orderings, with distinct energies, forces and thermodynamic behaviour. Spin is therefore not merely another property to predict, but an additional degree of freedom over which the energy landscape is defined (Xu et al., 2025; Ho et al., 2026; Yu et al., 2024b,a). No increase in geometric resolution can recover distinctions along a coordinate that the model does not represent. CHGNet provided an important early step toward incorporating electronic degrees of freedom, using site-resolved magnetic moments to enrich atomistic representations and infer charge states (Deng et al., 2023). Its energy model, however, remains structure-conditioned rather than explicitly conditioned on an independently specified spin configuration. A more complete description of magnetic materials requires the model to treat spin as an explicit variable, so that different magnetic states of the same structure can be evaluated, searched and evolved directly.

Here we introduce Prophet, a family of atomistic foundation models spanning composition, configuration and spin (Fig. 1a). The base model uses a spectrally decomposed equivariant graph neural network that represents atomic interactions over nested radial ranges and combines scale-specific features linearly. This gives short-range interactions finer spatial resolution while retaining longer-range context within the model cutoff, with forces and stresses derived from the same smooth and conservative energy. Prophet-Spin adds magnetic moments as explicit inputs and uses seed denoising to infer self-consistent magnetic moments from inexpensive initial guesses. An explicit exchange readout provides environment-dependent magnetic interactions, and at fixed geometry and magnetic moment magnitudes the model reduces exactly to a classical Heisenberg Hamiltonian. Magnetic-state search and large-scale spin simulation therefore become native inference capabilities of the same foundation-model framework. Fig. 1b summarizes these two design elements.

The Prophet family is trained on hundreds of millions of first-principles configurations centred on ELEMENTA and its Vib and Spin extensions (Kairos Materials, 2026), together with complementary public datasets. ELEMENTA systematically samples compositions and structures across 84 elements without relying on known crystal prototypes. ELEMENTA Vib probes local responses to controlled atomic and lattice perturbations, while ELEMENTA Spin resolves multiple magnetic states of the same material. Public datasets add extensive off-equilibrium configurations and reference energetics. Training proceeds from broad pre-training to energy–force–stress refinement and targeted supervision of local energy-surface curvature, followed by magnetic-state learning and separate training of the denoising head. In this way, the data and training stages progressively add chemical breadth, configurational resolution and magnetic degrees of freedom.

These capabilities are reflected across both statistical benchmarks and physical simulations. On nearly two million broadly sampled polymorphs spanning 84 elements, with no compositional overlap with training, Prophet reduces energy error by 56.6% relative to the strongest evaluated baseline and more reliably identifies and ranks competing structures. It also achieves the highest overall performance score on Matbench Discovery and the lowest lattice-thermal-conductivity error (Fig. 1c). For magnetic materials, explicit spin changes what the model can represent rather than merely improving an existing prediction. Prophet-Spin reduces the energy error on 188 composition-unseen magnetic materials from the 63–72 meV/atom range of spin-agnostic models to 19.2 meV/atom, while enabling magnetic-state search directly from crystal structures. Without material-specific fine-tuning, it recovers the antiferromagnetic ground state and convex-hull stability of LiMnAs, predicts the Curie transition of bcc Fe, and reproduces the pressure-induced bcc–hcp transition of iron. These results point to a broader view of atomistic scaling in which progress comes not only from more data, but from expanding the material space that models can explore, increasing the distinctions they can resolve, and

Figure 1 Prophet extends atomistic foundation models along composition, configuration and spin. **a**, Scope. Structure-only OAM potentials (grey) cover known compositions and structures at one implicit spin state; Prophet (purple) extends composition to enumerated chemistries (ELEMENTA), configuration to random structure sampling (ELEMENTA Vib) and spin to sampled magnetic states of the same structure (ELEMENTA Spin, Prophet-Spin). **b**, Design concepts. Nested cutoffs $c_1 < c_2 < c_3$ replace one broad radial kernel (top); the spin-aware energy takes the magnetic state as an explicit input, so that an initial state $\tilde{\mathbf{m}}_0$ is relaxed to the collinear state $\mathbf{m}^\star$ instead of being averaged into a single spin-agnostic value (bottom). **c**, Matbench Discovery, 15 September 2026 snapshot: CPS for the 32 highest-ranked entries (top), F_1 and κ_{SRME} for the best entry of each of the 16 leading model families down to MatterSim (bottom). Prophet has the highest CPS (0.912) and the lowest κ_{SRME} (0.065); the full table is Table 3.

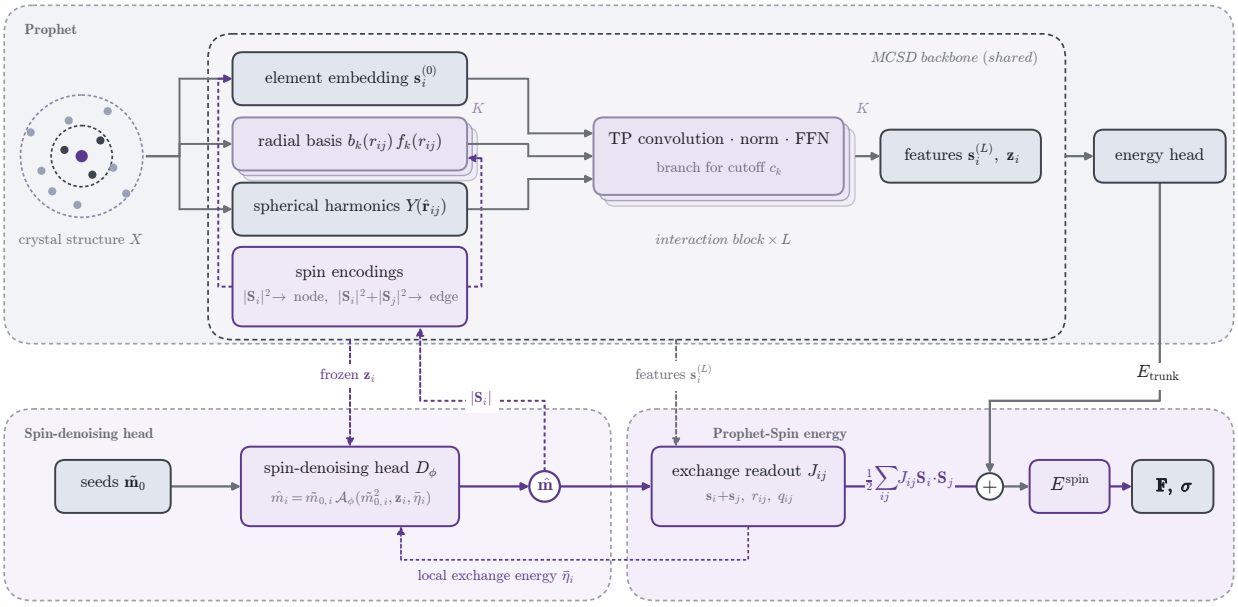

Figure 2 Architecture of the Prophet family. Top (Prophet): the crystal structure X is encoded by element embeddings $\mathbf{s}_i^{(0)}$, one radial basis $\mathbf{b}_k(r_{ij})\, f_k(r_{ij})$ with a smooth envelope per nested cutoff c_k, and spherical harmonics $Y(\hat{\mathbf{r}}_{ij})$. These feed L interaction blocks, each with K per-cutoff branches of tensor-product convolution, normalization and feed-forward update whose outputs are mixed with residual connections. The final scalar features $\mathbf{s}_i^{(L)}$ enter the energy head, which gives E_{trunk}; for Prophet this is the energy, and forces and stresses follow by differentiation. Prophet-Spin adds spin encodings inside the same backbone: Gaussian encodings of $|\mathbf{S}_i|^2$ are added at the node level and of $|\mathbf{S}_i|^2 + |\mathbf{S}_j|^2$ at the edge level, so the trunk depends on magnetic moment magnitudes only. Bottom left (spin-denoising head): seeds $\tilde{\mathbf{m}}_0$ are mapped to self-consistent spins $\hat{\mathbf{m}}$ by D_ϕ, a learned multiplier of the seed that acts on the frozen even-scalar features $\mathbf{z}_i$ of the backbone and on the local exchange energy $\bar{\eta}_i$. Bottom right (Prophet-Spin energy): an exchange readout builds J_{ij} from $\mathbf{s}_i^{(L)} + \mathbf{s}_j^{(L)}$, r_{ij} and q_{ij}, and the orientations enter only through $\frac{1}{2}\sum_{ij} J_{ij}\, \mathbf{S}_i \cdot \mathbf{S}_j$; its sum with E_{trunk} is E^{spin}, from which forces and stresses are obtained at fixed spins. Dashed arrows mark the couplings between the three parts; purple marks the components added by Prophet-Spin.

incorporating the physical degrees of freedom required to describe new classes of behaviour.

2 Model Architecture

Prophet learns a potential-energy function from which forces and stresses are obtained by differentiation. Its backbone represents atomic environments at several radial scales. Prophet-Spin additionally conditions the energy on spin and uses a separate spin-denoising head to estimate converged magnetic moments from initial seeds. The distinction between energy evaluation and spin recovery is central to the architecture. The first describes a specified state; the second supplies candidate states for evaluation. Figure 2 summarizes the resulting computational paths.

2.1 Spectrally Decomposed Equivariant Graph Neural Networks

The backbone must retain the surrounding chemical environment while resolving small changes in local bonding. We formulate this task as learning an energy $E_\theta(X)$ for a structure $X = (\mathbf{Z}, \mathbf{R}, \mathbf{A})$, where $\mathbf{Z} = \{Z_i\}_{i=1}^N$ gives the atomic elements, $\mathbf{R} = \{\mathbf{r}_i\}_{i=1}^N$ gives Cartesian positions, and the columns of $\mathbf{A}$ are the cell vectors. The parameters θ are shared across atoms and compositions. Energy is invariant under translations, joint rotations or reflections of positions and cell, and joint relabelling of elements and positions. Forces and stresses are

defined by

$$\begin{aligned}\mathbf{F}_i(X) &= -\frac{\partial E_\theta(X)}{\partial \mathbf{r}_i},\\ \boldsymbol{\sigma}(X) &= \frac{1}{V}\frac{\partial E_\theta(X_{\boldsymbol{\epsilon}})}{\partial \boldsymbol{\epsilon}}\bigg|_{\boldsymbol{\epsilon}=0},\end{aligned} \tag{1}$$

where $X_{\boldsymbol{\epsilon}}$ applies a symmetric strain $\boldsymbol{\epsilon}$ to positions and cell, and $V = |\det \mathbf{A}|$. Consequently, an orthogonal spatial transformation Q transforms the forces as $Q\mathbf{F}_i$ and the stress as $Q\boldsymbol{\sigma}Q^{\mathsf{T}}$. Learning one invariant energy therefore imposes a consistent relation between the energetic and mechanical outputs.

The network realizes this energy through an encoder, a sequence of interaction blocks, and an atomic readout. Each atom carries an invariant scalar stream $\mathbf{s}_i^{(t)}$ and an equivariant feature stream $\mathbf{h}_i^{(t)}$ after block t. The latter includes scalar, vector, and higher-order channels with prescribed transformation laws. Let $H^{(t)}$ collect both streams over all atoms. The complete computation is

$$\begin{aligned}(H^{(0)}, \mathcal{G}_\theta(X)) &= \mathcal{I}_\theta(X),\\ H^{(t+1)} &= \mathcal{B}_\theta^{(t)}(H^{(t)}; \mathcal{G}_\theta(X)), \qquad t = 0, \dots, L-1,\\ E_\theta(X) &= \sum_{i=1}^{N}\left[a_E\, \mathbf{w}_E^{\mathsf{T}}\mathbf{s}_i^{(L)} + b_E + E_{Z_i}^{\mathrm{ref}}\right].\end{aligned} \tag{2}$$

Here $\mathcal{I}_\theta$ returns initial node features and the fixed edge context $\mathcal{G}_\theta(X)$, which contains the graph, encoded edge features, cutoff masks and angular functions. L is the number of interaction blocks. The scalar readout has weights $\mathbf{w}_E$, an energy scale a_E, a shift b_E and elemental offsets E_Z^{ref}. Initially, the equivariant stream contains only the encoded scalars. Directional channels are generated through interactions with neighbouring atoms, and the final block returns scalar channels. Atomic summation ensures extensivity without restricting the learned interactions to a pair potential.

The radial design follows Multi-Cutoff Spectral Decomposition (MCSD) introduced in our companion work (Tan et al.). Its starting point is that neighbour aggregation acts as a spatial filter. In a reduced scalar description, a feature-weighted atomic density ϱ is aggregated through a radial kernel g, giving $u_{\mathrm{out}} \simeq \varrho * g$, where $*$ denotes convolution. The corresponding Fourier relation is $\widehat{u}_{\mathrm{out}} \simeq \widehat{\varrho}\,\widehat{g}$. A filter with a rapidly decaying high-frequency response suppresses fine radial differences in its input. For kernels related by spatial rescaling, a shorter length scale produces a broader frequency response. MCSD uses this relationship to motivate a family of radial filters,

$$g_{\mathrm{eff}}(r) = \sum_{k=1}^{K}\gamma_k g_k(r), \qquad \widehat{g}_{\mathrm{eff}}(\omega) = \sum_{k=1}^{K}\gamma_k \widehat{g}_k(\omega), \tag{3}$$

where g_k has its own radial support, γ_k is a learned mixing coefficient, and ω denotes spatial frequency. Shorter-scale filters provide access to finer variation while outer filters retain the wider environment. The filters overlap rather than forming disjoint frequency bands. This is a spectral interpretation of radial aggregation, not a Fourier decomposition of the entire nonlinear network. Detailed analysis is referred to the companion work.

Prophet implements this principle using nested cutoffs $c_1 < \cdots < c_K$. At scale k, $\mathcal{N}_k(i)$ contains neighbours within c_k, including periodic images. We write r_{ij} and $\widehat{\mathbf{r}}_{ij}$ for their distance and direction from atom i, suppressing image indices. The graph is constructed at the outer cutoff and masked for each scale. The encoder combines two elements embeddings with scale-specific sums over neighbours to obtain $\mathbf{s}_i^{(0)}$. It then constructs invariant edge features

$$\mathbf{e}_{ij}^{(k)} = \left(\mathbf{s}_i^{(0)} + \mathbf{s}_j^{(0)}\right) \odot \mathbf{b}_k(r_{ij})\, f_k(r_{ij}), \tag{4}$$

where $\odot$ denotes elementwise multiplication, $\mathbf{b}_k$ is a learned projection of the radial basis, and f_k is a smooth cutoff envelope. The Gaussian implementation places the basis functions separately over each cutoff interval; Bessel expansions are also supported by the code. These edge features and the spherical harmonics $Y(\widehat{\mathbf{r}}_{ij})$ are computed once per structure evaluation and reused across the interaction blocks.

Each block has an independent equivariant branch for each cutoff. In the base backbone, its neighbour aggregation has the form

$$\begin{aligned}\mathbf{u}_i^{(t,k)} &= \frac{1}{\sqrt{\bar{n}}} \sum_{j \in \mathcal{N}_k(i)} \left[\mathcal{L}_k^{(t)} \mathbf{h}_j^{(t)} \otimes_{\boldsymbol{\omega}_{ij}^{(t,k)}} Y(\widehat{\mathbf{r}}_{ij}) \right], \\ \boldsymbol{\omega}_{ij}^{(t,k)} &= W_k^{(t)}(\mathbf{e}_{ij}^{(k)}).\end{aligned} \tag{5}$$

The map $\mathcal{L}_k^{(t)}$ mixes matching equivariant channels, $\bar{n}$ is the reference neighbour count, and $W_k^{(t)}$ is a radial multilayer perceptron that generates tensor-product weights. The Clebsch–Gordan tensor product $\otimes$ combines the neighbour features and edge direction with the required transformation law. Each branch then applies a linear skip connection, gating, normalization and an equivariant feed-forward update. The feed-forward network uses invariant squared channel amplitudes to update scalar features and rescale equivariant channels. Branch outputs are concatenated and mixed separately in the scalar and equivariant streams, with residual connections where dimensions permit.

The cutoff envelopes vanish together with their first two derivatives. Importantly, the base implementation places the envelope inside the edge features of Equation (4). Its radial networks have no biases and satisfy $W_k^{(t)}(0) = 0$, so the interaction also vanishes smoothly as an edge leaves a cutoff. Smooth normalization and gating preserve this regularity. The resulting energy is twice continuously differentiable across cutoff boundaries for noncoincident atoms, and Equation (1) gives conservative forces.

2.2 Spin-Aware Network with Seed Denoising

For magnetic materials, the target state includes local spin in addition to atomic positions. A user normally supplies initial magnetic moments rather than a known self-consistent magnetic state. Prophet-Spin addresses this task by learning to recover the magnetic moments of a candidate state and then evaluating that state with a spin-conditioned energy model. At fixed ionic geometry, ordinary collinear spin-polarized DFT allows the initial magnetic moments to change during electronic self-consistency, where different initializations can reach different local magnetic solutions. ELEMENTA Spin samples such solutions and records their converged magnetic moments with the corresponding energies, forces and stresses (Kairos Materials, 2026). Electronic convergence here is distinct from relaxation of the atomic structure.

Learning magnetic relaxation by energy minimization would require reliable supervision away from these self-consistent endpoints. Constrained DFT can provide such information by holding local magnetic moments at prescribed values while optimizing the remaining electronic variables (Ma and Dudarev, 2015; VASP Software GmbH, 2026). Applying this procedure throughout the joint structural and magnetic configuration space would require an additional large-scale data-generation effort. Prophet-Spin instead learns a recovery map from the ordinary DFT endpoints. Seed denoising constructs nonconverged inputs from converged magnetic moment labels, without assigning the endpoint energy to those corrupted inputs.

Let $M = \{\mathbf{S}_i\}_{i=1}^N$ denote local magnetic moments. A collinear state is represented by signed amplitudes $\boldsymbol{m} \in \mathbb{R}^N$ and embedded as $M(\boldsymbol{m}) = \{m_i \widehat{\mathbf{z}}\}$ along a common spin axis. The seed-to-state prediction is

$$\begin{aligned}\widetilde{\boldsymbol{m}}_0 &= P_{\mathbf{Z}}(\boldsymbol{m}_0), \\ \widehat{\boldsymbol{m}} &= D_\phi(X, \widetilde{\boldsymbol{m}}_0; \bar{\theta}_{\mathrm{s}}), \\ \widehat{E} &= E_{\theta_{\mathrm{s}}}^{\mathrm{spin}}(X, M(\widehat{\boldsymbol{m}})).\end{aligned} \tag{6}$$

The user seed is $\boldsymbol{m}_0$, $P_{\mathbf{Z}}$ is the selected seed-preprocessing rule, and D_ϕ is the learned recovery head. The released predictor can use the seed unchanged or replace its magnetic moment magnitudes with element-dependent lookup values while retaining the signs on selected sites. The head uses a frozen feature backbone with parameters $\bar{\theta}_{\mathrm{s}}$ carried in its checkpoint bundle. A separate energy checkpoint with parameters θ_{s} scores the recovered magnetic moments. Hats denote estimates of DFT-converged quantities, not the result of another electronic self-consistency calculation.

The energy network accepts both collinear and vector magnetic moments. Its magnetic symmetries follow the no-spin–orbit reference setting. A common rotation of all magnetic moments leaves the energy unchanged, as

does reversal of every moment. These transformations are independent of spatial rotations and translations,

$$E^{\mathrm{spin}}_{\theta_s}(\mathcal{T}X, UM) = E^{\mathrm{spin}}_{\theta_s}(X, M), \qquad \mathcal{T} \in E(3), \quad U \in O(3). \tag{7}$$

Here $UM = \{U\mathbf{S}_i\}$, and the $O(3)$ spin action combines proper rotations with time reversal. Atom relabelling also relabels the magnetic moments. The collinear denoiser respects the corresponding reversal relation $D_\phi(X, -\widetilde{\boldsymbol{m}}_0; \bar{\theta}_s) = -D_\phi(X, \widetilde{\boldsymbol{m}}_0; \bar{\theta}_s)$. These constraints allow different relative magnetic orderings to have different energies without making the predictions depend on an arbitrary reference frame.

In the Heisenberg-mode energy network, magnetic moment magnitudes and relative orientations enter through different paths. Define $a_i = |\mathbf{S}_i|$, $\mathbf{a} = (a_1, \ldots, a_N)$, $q_{ij} = a_i^2 + a_j^2$ and $p_{ij} = \mathbf{S}_i \cdot \mathbf{S}_j$. Gaussian encodings of a_i^2 are added to the initial scalar nodes, and encodings of q_{ij} are added to the enveloped edge features. The backbone therefore depends on magnetic moment magnitudes but not directions. Its final scalar features feed both the atomic energy readout and a separate exchange readout,

$$\begin{aligned} E^{\mathrm{spin}}_{\theta_s}(X, M) &= E_{\mathrm{trunk}}(X, \mathbf{a}) + \frac{1}{2} \sum_{(j \to i) \in \mathcal{E}_{\mathrm{ex}}} J_{ji}(X, \mathbf{a})\, p_{ji}\, f_{\mathrm{ex}}(r_{ji}), \\ J_{ji}(X, \mathbf{a}) &= \mathcal{J}_{\theta_s}\left(\mathbf{s}_j^{(L)} + \mathbf{s}_i^{(L)}, \boldsymbol{\rho}_r(r_{ji}), \boldsymbol{\rho}_q(q_{ji})\right). \end{aligned} \tag{8}$$

E_{trunk} uses the atomic summation in Equation (2); $\mathcal{J}_{\theta_s}$ is a small coupling network, and $\boldsymbol{\rho}_r$ and $\boldsymbol{\rho}_q$ are Gaussian encodings. The exchange envelope f_{ex} imposes its radial support. The directed graph $\mathcal{E}_{\mathrm{ex}}$ contains both directions of each pair, accounting for the factor $1/2$. Keeping p_{ij} out of the nonlinear backbone makes the directional dependence exactly bilinear at fixed geometry and magnetic moment magnitudes.

The recovery head uses a different representation from the energy readout. It extracts the even-scalar channels $\mathbf{z}_i$ from the penultimate equivariant features of the frozen backbone. The exchange-conditioned variant also receives the local exchange energy $\bar{\eta}_i$, accumulated on edges entering site i in the same forward pass. This supplies information about the seed ordering that is absent from the magnitude-conditioned features. With all features evaluated at $M(\widetilde{\boldsymbol{m}}_0)$, the head can be written as

$$\widehat{m}_i = \widetilde{m}_{0,i}\, \mathcal{A}_\phi\left(\widetilde{m}_{0,i}^2, \mathbf{z}_i, \bar{\eta}_i\right). \tag{9}$$

$\mathcal{A}_\phi$ is a learned scalar multiplier built from Gaussian functions of squared seed amplitude and a separate neural network. Under global reversal of the seed, its arguments are unchanged, giving the required time-reversal relation. This parameterization also maps an exactly zero local seed to zero.

For training, let $\boldsymbol{m}^\star$ contain the DFT-converged collinear magnetic moments at structure X. The seed distribution $q_{\mathrm{seed}}(\widetilde{\boldsymbol{m}}_0 \mid X, \boldsymbol{m}^\star)$ combines element-based initializations, additive noise, amplitude rescaling and unchanged targets. With the feature backbone frozen, the head minimizes

$$\mathcal{L}_{\mathrm{denoise}}(\phi) = \mathbb{E}_{(X, \boldsymbol{m}^\star)} \mathbb{E}_{\widetilde{\boldsymbol{m}}_0 \sim q_{\mathrm{seed}}}\left[\frac{1}{N} \sum_{i=1}^{N} \rho_\delta\left([D_\phi(X, \widetilde{\boldsymbol{m}}_0; \bar{\theta}_s)]_i - m_i^\star\right)\right], \tag{10}$$

where ρ_δ is the Huber loss with threshold δ. This objective learns endpoint recovery without reconstructing the electronic iteration history or a fully supervised constrained magnetic energy surface. One head evaluation produces the predicted magnetic moments, which are then passed to a fresh energy evaluation. The prediction is neither an enforced fixed point of the denoiser nor an enforced stationary point of the learned energy. The corruption mixture and training settings are specified in Appendix B and Section 4.

The implementation keeps prescribed-state evaluation and seed-to-state recovery separate. The former returns energy, forces and stress at supplied magnetic moments; the latter first predicts magnetic moments and then evaluates these quantities. In both cases, mechanical derivatives hold the supplied or predicted magnetic moments fixed and do not differentiate through the recovery head. Multiple seeded evaluations can be combined in an external search over candidate magnetic orderings, as used in Section 8. Such a search selects the lowest-energy candidate explored in the chosen cell rather than guaranteeing a global magnetic ground state.

The same exchange representation connects these local predictions to collective spin simulations. At fixed X and $\mathbf{a}$, the trunk energy and all J_{ij} remain constant when the directions of the magnetic moments change. Equation (8) then reduces exactly to a classical Heisenberg Hamiltonian. The runtime extracts its coefficients once and reuses them for subsequent spin updates. This is an identity of the chosen model, not a claim that collinear endpoint data uniquely determine all noncollinear magnetic physics. Appendix B gives the extraction procedure, the released inference interfaces and the distinction between conditioned and composite energy derivatives.

3 Training Data

Prophet is trained on six first-principles datasets: OMat24, ELEMENTA, MPtrj, sAlex, ELEMENTA Vib, and ELEMENTA Spin. Together they contain more than 300 million labelled configurations. OMat24 and ELEMENTA provide the main pre-training data; MPtrj and sAlex are used to align the energy reference; ELEMENTA Vib adds near-equilibrium force and stress information; and ELEMENTA Spin provides magnetic-state supervision.

OMat24 contains approximately 10^8 inorganic configurations from structural relaxations, atomic rattling, and ab initio molecular dynamics. It provides broad sampling of both equilibrium and off-equilibrium geometries. ELEMENTA extends this coverage systematically across composition and structure. Across 84 elements, compositions are enumerated within defined stoichiometric limits and explored by random structure search without using known crystal prototypes as starting templates. Candidate structures are pre-relaxed, deduplicated, and refined with DFT. The resulting corpus contains approximately 1.9 million reduced formulas, more than 9.7 million relaxed polymorphs, and about 210 million labelled frames. Several relaxed minima are retained for each composition, so the dataset contains direct information on the relative energies of competing structures in addition to absolute energies and relaxation trajectories.

ELEMENTA Vib targets a different part of the potential-energy surface. Relaxation trajectories provide extensive force information along paths toward equilibrium, but they sample the immediate neighbourhood of a minimum less densely once the forces become small. ELEMENTA Vib starts from approximately 550,000 ELEMENTA polymorphs, expands small structures to compact supercells, samples volumes of $0.95V_0$, V_0, and $1.05V_0$, and applies random atomic displacements of 0.005–0.03 Å. More than eight million DFT single-point calculations provide energies, forces, and stresses around these local minima. In Prophet training, the force and stress labels are used to improve the local response of the potential, which is particularly relevant to phonons and lattice thermal transport.

MPtrj and **sAlex** are included primarily for energy-reference consistency. MPtrj contains approximately 1.58 million Materials Project relaxation frames, while sAlex contains subsampled Alexandria relaxation trajectories computed with Materials-Project-compatible settings. These datasets define the energy convention used in Matbench Discovery and are therefore used for reference alignment and replay during the final refinement stage.

ELEMENTA Spin extends the training data beyond atomic structure. Approximately 50,000 magnetic materials are sampled with multiple collinear magnetic initializations and relaxed using spin-polarized DFT. The resulting 4.7 million frames contain energies, forces, stresses, and converged site-resolved magnetic moments. Multiple self-consistent magnetic states can therefore be associated with the same atomic structure, with energy differences reaching tens to hundreds of meV/atom. These data are used to train the spin-conditioned energy model and provide the converged-moment targets for the denoising head. The current dataset is restricted to collinear calculations without spin–orbit coupling.

Across these datasets, the added information is complementary. OMat24 supplies broad off-equilibrium sampling; ELEMENTA increases chemical coverage and the density of competing polymorphs; ELEMENTA Vib resolves the neighbourhood of local minima; MPtrj and sAlex fix the external energy reference; and ELEMENTA Spin introduces magnetic state as an explicit degree of freedom. Section 4 describes how these sources are combined during training.

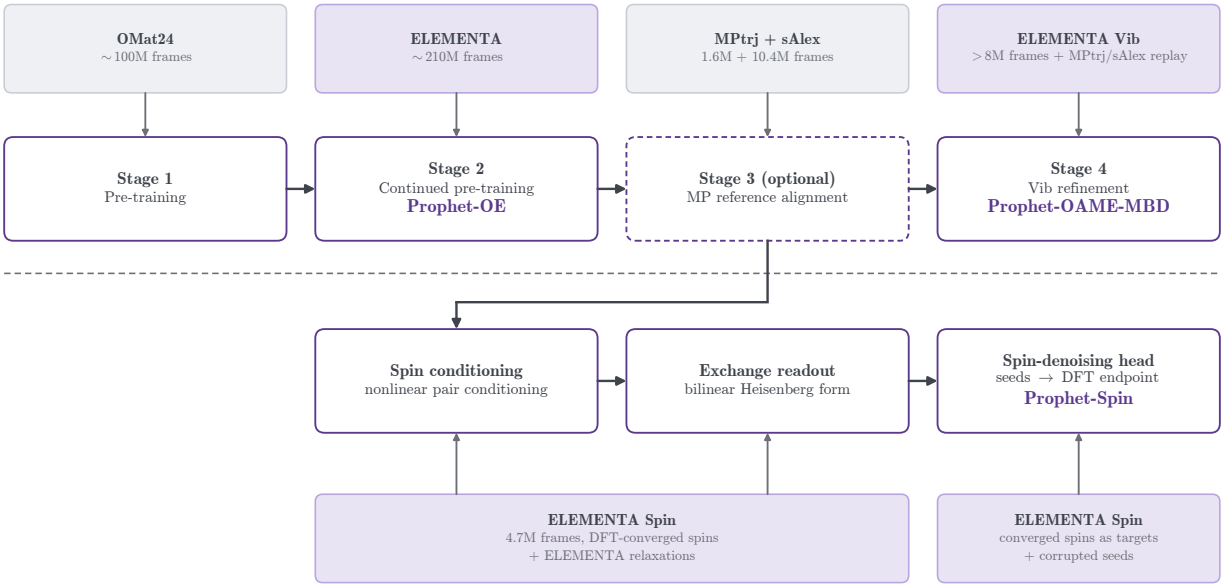

Figure 3 Training stages of the Prophet family. Top: the structural line pre-trains on OMat24 and ELEMENTA (Prophet-OE), optionally aligns to the Materials Project energy reference on MPtrj and sAlex, and refines the local force response on ELEMENTA Vib with replay (Prophet-OAME-MBD). Bottom: the magnetic line branches from the reference-aligned model, adapts the backbone to ELEMENTA Spin through nonlinear pair conditioning and the bilinear exchange readout, and finally trains the spin-denoising head with the energy model frozen (Prophet-Spin). Purple blocks are ELEMENTA datasets, grey blocks public datasets.

4 Training Strategy and Model Family

Prophet is trained progressively rather than on all data sources at once. Pretraining has two consecutive stages: initial training on OMat24 and continued pretraining on ELEMENTA. The model obtained after these two stages is the *pretrained Prophet*. ELEMENTA Vib is then used to refine the local mechanical response around relaxed structures. When direct comparison with the Materials Project energy reference is required, an optional MPtrj/sAlex alignment stage is inserted before the ELEMENTA Vib refinement. This alignment is useful for benchmarks such as Matbench Discovery, but is not required when all energies in a calculation are produced by Prophet itself. Prophet-Spin is trained as a separate magnetic extension of the pretrained model.

4.1 Pretraining on OMat24 and ELEMENTA

Pretraining begins on OMat24, which contains more than one hundred million first-principles configurations drawn from structural relaxation, atomic perturbations and molecular dynamics Barroso-Luque et al. (2026). These data expose the model to a broad range of off-equilibrium environments and provide the first large-scale training of its representation of atomic interactions. The physical targets are energy, forces, and stress. Their mean-absolute-error terms are weighted 20, 20 and 5, respectively, with the energy error normalized per atom. Because Prophet predicts a scalar energy and obtains forces and stresses by differentiation, these labels constrain different aspects of the same potential-energy surface rather than independent prediction heads.

OMat24 pretraining uses the energy, force, and stress objective only; no auxiliary structural denoising task is applied at this stage, and the only denoising objective in the Prophet family is the magnetic-moment denoising used later in Prophet-Spin. The model produced at this point is an intermediate pretrained state rather than the final pretrained Prophet.

The training then continues on ELEMENTA. OMat24 samples many distortions around a large collection of existing or generated structures, whereas ELEMENTA starts from systematic composition enumeration and prototype-independent structure search across 84 elements (Kairos Materials, 2026). For a given composition,

several relaxed polymorphs and their relaxation trajectories are retained. This adds direct supervision for the relative energies of competing structures, as well as for the forces leading into different local minima. The learning problem remains the same—a single energy surface constrained by energies, forces, and stresses—but the distribution of structures becomes substantially broader and less tied to known material prototypes.

The model obtained after continued pretraining on ELEMENTA is the *Prophet-OE*. This is a common starting point for the later structural and magnetic models. Unless stated otherwise, we use *Prophet native reference* to denote the energy convention learned through the OMat24 and ELEMENTA pretraining stages.

4.2 Optional Alignment to the Materials Project Energy Reference

Energies produced by different DFT workflows are not automatically interchangeable. Differences in pseudopotentials, Hubbard-U treatments and numerical settings can change elemental reference energies and can also introduce smaller structure-dependent shifts. A model can therefore be internally accurate on one DFT convention while showing a systematic discrepancy when its energies are mixed with values from another.

This matters in Matbench Discovery, where candidate structures are evaluated against a Materials-Project-based reference. For that use, pretrained Prophet is further trained on MPtrj and sAlex, sampled in an 8:1 ratio. Both datasets contribute energy, force and stress supervision. The purpose of this stage is not to relearn the underlying atomic interactions, but to adapt an already pretrained model to the energy convention used by the benchmark.

The model obtained after this stage is a *reference-aligned Prophet*. It has the same physical inputs and the same architecture as pretrained Prophet, but its energy scale has been adapted for comparison with Materials Project data. This step can be omitted when candidate phases, competing phases and elemental references are all evaluated with Prophet itself, because those energies then share a common model-defined reference.

4.3 Final Refinement with ELEMENTA Vib

The final structural-training stage uses ELEMENTA Vib to improve the local shape of the potential-energy surface around relaxed structures. Relaxation data tell the model how structures move toward local minima, but they do not sample the neighbourhood of those minima uniformly. Vibrational properties depend on a more local quantity: how the forces change under small atomic displacements and lattice strains. ELEMENTA Vib was constructed specifically to provide this information through controlled atomic and volume perturbations.

During this stage, ELEMENTA Vib contributes force and stress supervision but not an absolute-energy loss. The model nevertheless remains an energy-based potential: its forces and stresses are still derivatives of the predicted energy. The displaced-force labels constrain local force constants along the sampled directions, while the volume perturbations constrain the response to strain. Neither force constants nor thermal conductivity are fitted as direct targets.

This refinement can be applied to either energy reference. Applied after MPtrj/sAlex alignment, it produces *Prophet-OAME-MBD*, the structural model used for the Matbench Discovery results reported here. In the aligned training used in this work, the final stage samples MPtrj, sAlex and ELEMENTA Vib in proportions of 35, 35 and 30 optimization steps per 100. MPtrj and sAlex retain energy, force and stress supervision, while ELEMENTA Vib contributes only force and stress. Replaying the reference data keeps the aligned energetics stable while the local mechanical response is refined.

4.4 Training Prophet-Spin

Prophet-Spin is trained from the pretrained Prophet backbone, but its learning problem differs from the structural refinements above because the magnetic state becomes part of the input. The reference data come from ELEMENTA Spin together with ELEMENTA structural relaxation trajectories. ELEMENTA Spin uses ordinary collinear spin-polarized DFT with different magnetic initializations. The local moments are allowed to change during electronic self-consistency, so the training target is the DFT-converged magnetic state, together with the energy, forces and stress of that same state. The initial moments used to start the DFT calculation are not treated as if they were the state associated with the converged energy.

Training is divided into two parts. The first trains the magnetic energy model. DFT-converged local moments are supplied as additional inputs, and the model is fitted to the corresponding energy, forces and stress. The magnetic conditioning layers are initialized so that the inherited structural backbone is initially unchanged. In the reported training procedure, the conditioned network is first adapted with nonlinear pair conditioning. The final magnetic energy model is then initialized from this trained network, with relative spin orientation represented through the explicit bilinear exchange readout described in Section 2.2. This second step gives the model its exact Heisenberg form at fixed geometry and moment magnitudes.

At the end of the energy-model training, the result is a spin-conditioned Prophet energy model. It can evaluate the energy and mechanical response of a prescribed magnetic state, but using it on an unfamiliar crystal would still require the corresponding local moments to be supplied.

The second part trains moment recovery. The spin-conditioned energy model is frozen, and a separate denoising head learns to map inexpensive initial magnetic seeds to DFT-converged moments. Training seeds include the element-based guesses used during magnetic-state search, as well as noisy and amplitude-perturbed versions of converged moments. The head is trained against the converged moments themselves; no DFT energy is assigned to an arbitrary corrupted seed. Keeping this training separate prevents the moment-recovery objective from changing the already fitted energy, force and stress surface.

Combining the trained energy model with the moment-recovery head gives *Prophet-Spin*. Starting from a crystal structure, candidate magnetic seeds can be generated, mapped to candidate converged moments and then ranked by the spin-conditioned energy model. The model therefore supports both evaluation of prescribed magnetic states and magnetic-state search without requiring DFT moments as inputs.

4.5 Model Family

The Prophet family used in this report consists of three models with different intended uses.

Model	Training	Intended use
Prophet-OE	OMat24 pretraining, continued pretraining on ELEMENTA	General structural model using the Prophet native energy reference
Prophet-OAME-MBD	OMat24 pretraining, continued pretraining on ELEMENTA, MPtrj/sAlex reference alignment, followed by ELEMENTA Vib refinement	Materials-Project-compatible structural model used for the reported Matbench Discovery evaluation
Prophet-Spin	Magnetic adaptation of pretrained Prophet on ELEMENTA Spin, followed by separate training of the moment-recovery head	Explicit magnetic-state evaluation, magnetic-state search and spin-Hamiltonian extraction

Table 1 The Prophet model family used in this report. Prophet-Spin follows a separate magnetic adaptation from pretrained Prophet.

The structural model used for the Matbench Discovery results contains ten Multi-Cutoff Spectral Decomposition (MCSD) interaction blocks, with cutoffs of 3 and 7 Å, and 62.3 million parameters. Prophet-Spin uses a six-block backbone with 3 and 6 Å cutoffs, together with magnetic conditioning, an exchange readout and the moment-recovery model. Numerical results in later sections are associated with the corresponding member of the family and its energy reference.

Publicly released models are preview versions intended for research use only, without guarantees of accuracy or suitability for a particular application.

5 Matbench Discovery

We first evaluate Prophet on Matbench Discovery. Matbench Discovery Riebesell et al. (2025) is a community benchmark for universal interatomic potentials in a setting close to high-throughput materials discovery. Its

central tasks ask a model to determine the thermodynamic stability of approximately 257,000 candidate crystals after structural relaxation, reproduce their DFT-relaxed geometries, and predict lattice thermal conductivity from the model's own force constants. Additional tasks probe molecular-dynamics trajectories and the smoothness of diatomic potential-energy curves. Together, these evaluations test complementary aspects of the same potential-energy surface: the relative energies of competing phases, the location of equilibrium structures, the local curvature governing lattice vibrations, and the numerical stability of dynamics. The benchmark's Combined Performance Score (CPS) combines stability classification, thermal conductivity, and geometry optimization, with default weights of 50%, 40%, and 10%, respectively.

All results in this section use the released 62.3-million-parameter Prophet-OAME-MBD model. Following broad pre-training on OMat24 and continued pre-training on ELEMENTA, the model is aligned on MPtrj and sAlex to the Materials Project energy reference used by Matbench Discovery, and is then refined with ELEMENTA Vib while replaying the aligned reference data. No WBM labels or benchmark-derived quantities are used for training or model selection.

On the unique-prototype discovery split (Table 3), Prophet achieves an F_1 score of 0.928, with precision of 0.927, recall of 0.928, accuracy of 0.978, and a discovery acceleration factor of 6.065. Its formation-energy MAE is 18 meV/atom, with $R^2 = 0.869$. On the full WBM test set, the corresponding F_1 score is 0.910. These values are close to the best discovery results on the leaderboard: the leading models cluster within $F_1 = 0.924$–0.931 and MAE = 18–19 meV/atom. Prophet therefore retains competitive phase-stability classification after the additional curvature-focused training, rather than improving one physical task by sacrificing the benchmark's central discovery objective.

The geometry task asks whether the learned forces carry an initially proposed crystal into the same structural basin as DFT. Prophet obtains an RMSD of 0.059 between model- and DFT-relaxed WBM structures, again within the narrow range of the leading entries. Symmetry provides a complementary view of the same relaxations: at a strict tolerance of 10^{-5}, 60.7% of Prophet-relaxed structures retain the DFT space group, increasing to 81.5% at 10^{-2}. Thus the stability and geometry results together show that Prophet generally reaches the correct region of configuration space while preserving the energetic distinctions required for phase screening.

The clearest separation appears in lattice dynamics. Thermal conductivity is a stringent test of the local shape of the potential-energy surface because it depends on second- and third-order force constants: harmonic curvature sets phonon frequencies and group velocities, while anharmonic terms govern phonon scattering. On the 103-material phononDB-PBE set at 300 K, Prophet reaches $\boldsymbol{\kappa_{\mathrm{SRME}} = 0.065}$, compared with 0.093 for the next entry, a reduction of about 30%. The error in total conductivity is also small, with $\kappa_{\mathrm{SRE}} = 0.028$, while the signed error $\kappa_{\mathrm{SRD}} = -0.001$ indicates essentially no systematic bias. The Wasserstein-1 distance between predicted and DFT phonon-frequency spectra is 0.019 THz, close to the approximately 0.02 THz mesh-comparison noise floor, and all 103 calculations complete without a failed case or an imaginary mode.

The material-level results support the aggregate metrics. For MgO, GaN, and BAs, Prophet reproduces both acoustic and optical phonon branches closely, with the highest frequencies within approximately 0.03–0.26 THz of DFT. PbTe is more difficult: its acoustic branches are well reproduced, whereas parts of the optical spectrum are 0.2–0.3 THz too stiff and its thermal conductivity is overestimated by 19%. Across the full 103-material set, however, every predicted conductivity lies within a factor of 1.25 of DFT and the median scalar conductivity error is 1.8%. The largest mode-resolved errors occur mainly in soft lead chalcogenides and copper hydrides, where the total conductivity can remain accurate even when its distribution among phonon modes is less precise. This distinction is useful because it shows that the remaining errors are localized in difficult regions of the vibrational spectrum rather than appearing as a broad bias in thermal transport.

Combining discovery, geometry, and thermal transport gives Prophet a CPS of 0.912, the highest value in the 15 September 2026 leaderboard snapshot(Table 3). The gain is not driven by an exceptional stability or geometry score in isolation; those metrics are already tightly clustered among the leading models. Instead, Prophet maintains the same level of discovery and structural accuracy while substantially improving the local force response relevant to lattice dynamics. Additional benchmark tasks are consistent with this picture. Prophet completes all 17 DynaMat molecular-dynamics trajectories, spanning 293–1500 K and systems of 48–500 atoms, without a failed step or diverging trajectory, with RDF, angular-distribution, and vibrational-density-of-states errors of 3.4%, 1.4%, and 21.9%, respectively. The diatomic tests identify a narrower weakness: a small

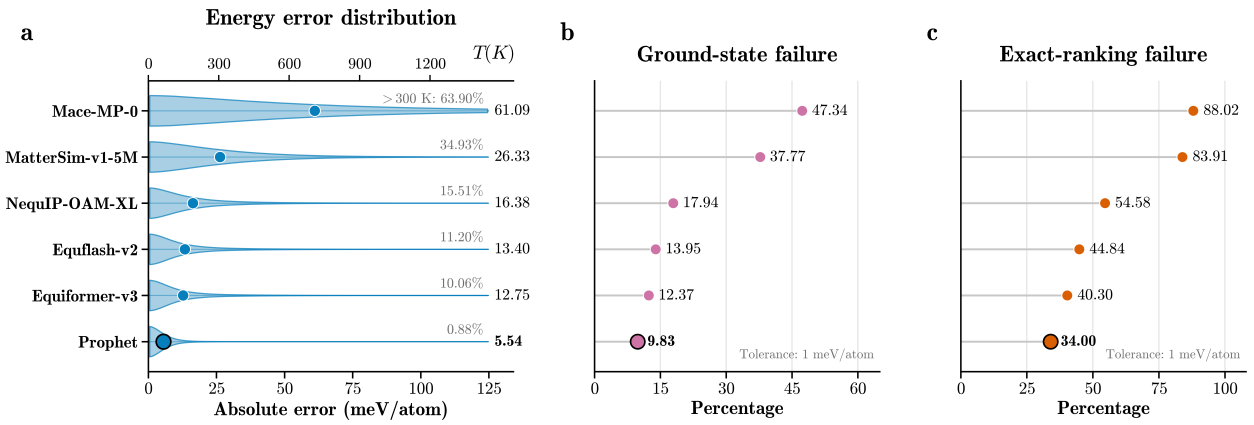

Figure 4 Benchmarking the energetic accuracy and polymorph ranking of Prophet against five representative atomistic foundation models: MACE-MP-0 (Batatia et al., 2025), MatterSim-v1-5M (Yang et al., 2024), NequIP-OAM-XL (Batzner et al., 2022), EquFlash-v2 (Lee et al., 2025), and Equiformer-v3 (Liao et al., 2024, 2026). (a) Distribution of absolute energy errors relative to DFT. Circles indicate the mean absolute error (MAE), with numerical values reported in meV/atom. The upper axis converts the energy-error scale to the corresponding thermal-energy scale, $T = \Delta E/k_{\mathrm{B}}$, and percentages denote the fraction of structures with errors exceeding $k_{\mathrm{B}}T$ at 300 K (≈ 25.9 meV/atom). (b) Ground-state identification failure rate, defined as the fraction of composition groups in which the model-selected ground state lies more than 1 meV/atom above the DFT ground state according to the DFT energy ordering. (c) Exact-ranking failure rate, defined as the fraction of composition groups for which the model fails to reproduce the complete DFT ordering of competing structures, using an energy tolerance of 1 meV/atom. Lower values indicate better performance in all three panels.

number of light-element and closed-shell dimers develop large force jumps deep inside the repulsive wall, outside the bonding range reached in the condensed-phase simulations. These cases provide a specific target for extending short-range training coverage without changing the overall conclusion of the benchmark: Prophet combines competitive materials-discovery accuracy with a markedly more faithful description of the local energy landscape that controls lattice dynamics.

6 Energetics and Phase Ranking

Accurate prediction of material energetics requires not only low errors for individual structures but also consistent resolution of the relative energies that distinguish competing polymorphs and determine phase stability. We therefore evaluate Prophet-OE at increasingly demanding levels of energetic consistency: energy accuracy across broad chemical space, ground-state identification and polymorph ranking at fixed composition, and phase stability across compositions through convex-hull reconstruction. The evaluation uses a held-out subset of ELEMENTA containing nearly 2 million structures grouped into approximately 0.4 million complete composition sets. These structures span binary, ternary, and quaternary systems across an 84-element chemical space, with no compositional overlap with the training data.

6.1 Energetic Accuracy Across Chemical Space

We first evaluate the energetic accuracy of Prophet on the held-out ELEMENTA subset. Figure 4a shows the distribution of absolute energy errors over nearly 2M structures with compositions not encountered during training. Prophet achieves an energy MAE of 5.54 meV/atom, substantially lower than all reference models and representing a 56.6% reduction relative to Equiformer-v3 (Liao et al., 2026), the strongest baseline with an MAE of 12.75 meV/atom. The improvement is particularly pronounced in the high-error tail: only 0.88% of Prophet predictions have errors exceeding the room-temperature thermal energy scale ($k_{\mathrm{B}}T \approx 25.9$ meV/atom), compared with 10.06% for Equiformer-v3 and 11.20% for EquFlash-v2 (Lee et al., 2025). Thus, the improvement of Prophet is reflected not only in a lower average error, but also in a substantially narrower error distribution with far fewer large deviations on unseen compositions.

To assess whether this accuracy is maintained uniformly across chemical space, we further resolve the prediction errors by element. For each element, the MAE is evaluated over all DFT-relaxed structures containing that element, such that multicomponent structures contribute to multiple elemental subsets. As shown in Fig. 5, Prophet maintains consistently low errors across most of the 84-element chemical space. For 79 of the 84 elements, the MAE remains below 7 meV/atom, with most values concentrated between approximately 4 and 6 meV/atom. The largest deviations occur for Gd, Pu, and Eu, with MAEs of 25.01, 15.72, and 10.81 meV/atom, respectively, reflecting the greater complexity of localized f-electron systems. Overall, the narrow error distribution across the elemental space indicates broadly uniform energetic accuracy across diverse chemical environments.

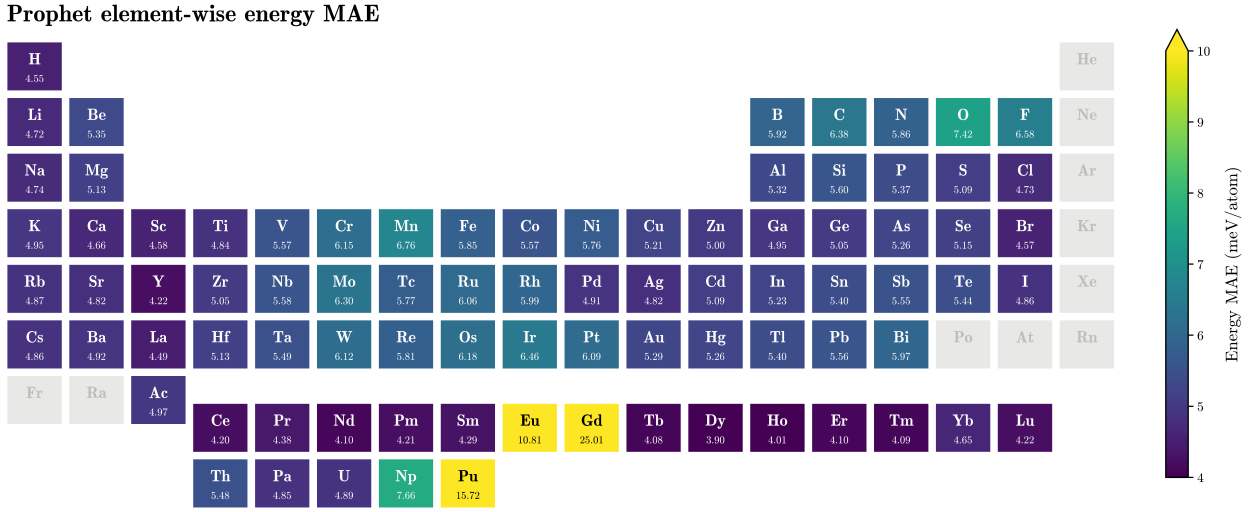

Figure 5 Element-resolved energy MAE of Prophet across the 84-element chemical space represented in the held-out ELEMENTA subset. For each element, the MAE is evaluated over all structures containing that element, such that multicomponent structures contribute to multiple elemental subsets. The colour scale is capped at 10 meV/atom to preserve contrast across the majority of elements, while numerical labels report the exact MAE for each element.

6.2 Ground-State Identification and Polymorph Ranking

Low errors in individual energies do not necessarily guarantee that a model preserves the relative stability of competing structures. At fixed composition, a material can exhibit multiple metastable structures corresponding to distinct local minima on the potential-energy landscape. These minima may differ in coordination geometry, bonding topology, atomic packing, or local structural distortions while being separated by relatively small energy differences. Accurately resolving their energetic hierarchy therefore provides a more stringent test of model fidelity, as even modest energy errors can change the identity of the ground state or invert the ordering of competing minima.

We first assess this capability statistically over approximately 0.4M complete composition groups in the held-out ELEMENTA subset. Resolving the energetic ordering of nearly degenerate structures at the meV/atom scale poses an exceptionally stringent requirement for current interatomic models. To avoid counting energetically negligible inversions as failures, we adopt an energy tolerance of 1 meV/atom for both ground-state identification and exact ranking, treating ordering differences below this threshold as equivalent. Under this criterion, Fig. 4b shows that Prophet achieves the lowest ground-state identification failure rate of 9.83%, compared with 12.37% for Equiformer-v3 and 13.95% for EquFlash-v2. We further consider the more stringent requirement of reproducing the complete energetic ordering of all competing structures within each composition. As shown in Fig. 4c, Prophet achieves an exact-ranking failure rate of 34.00%, compared with 40.30% for Equiformer-v3 and 44.84% for EquFlash-v2. The higher failure rates under exact ranking across all models reflect the greater difficulty of reproducing the complete hierarchy of competing minima: a model may correctly identify the ground state while still introducing inversions among higher-energy structures. Overall, Prophet achieves the lowest failure rate under both criteria, demonstrating improved resolution of relative

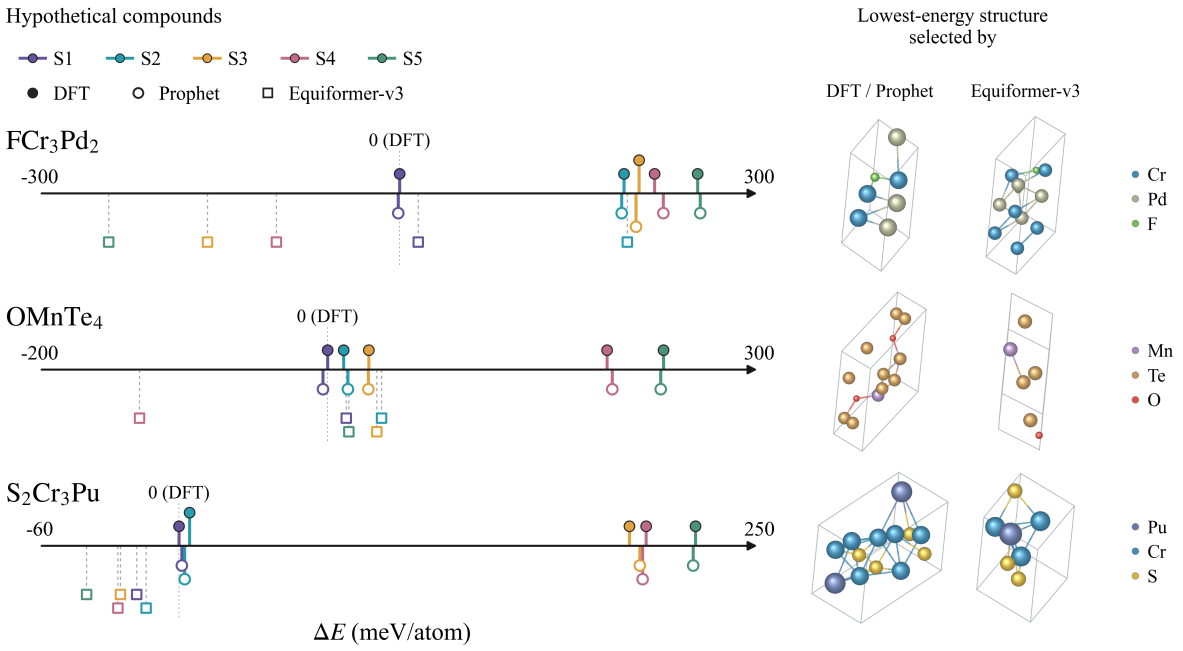

Figure 6 Representative cases in which Prophet recovers the DFT energetic ordering and ground-state structure where Equiformer-v3 fails. Five competing structures (S1–S5) are compared for three hypothetical compounds, FCr_3Pd_2, $OMnTe_4$, and S_2Cr_3Pu, with energies referenced to the DFT ground state for each composition. Filled circles, open circles, and open squares denote DFT, Prophet, and Equiformer-v3 predictions, respectively, with colors identifying the same structural candidate across methods. Prophet reproduces the complete DFT ordering and selects the DFT ground-state structure in all three cases, whereas Equiformer-v3 favors structurally distinct, higher-energy minima. The corresponding structures selected by DFT/Prophet and Equiformer-v3 are shown on the right.

energetics within previously unseen compositions.

To complement the statistical ground-state identification and polymorph-ranking benchmarks above, we examine three representative hypothetical compounds in which Prophet recovers the DFT energetic ordering and ground-state structure where Equiformer-v3 fails. These systems are not intended to represent experimentally established materials; rather, they provide controlled tests of model accuracy and robustness across unseen compositions and competing structural candidates, without prior knowledge of the underlying energy landscape. Figure 6 shows five competing structures for each of FCr_3Pd_2, $OMnTe_4$, and S_2Cr_3Pu, with energies referenced to the DFT minimum for each composition. Prophet reproduces the complete DFT ordering in all three cases, with energy MAEs of 3.02, 2.37, and 2.06 meV/atom, respectively. In contrast, Equiformer-v3 yields substantially larger MAEs of 236.98, 117.68, and 149.25 meV/atom and incorrectly selects structures lying 248.62, 194.28, and 222.49 meV/atom above the corresponding DFT ground states.

The structures shown in Fig. 6 illustrate how this recovery extends from energetic ordering to structural selection. In each case, Prophet selects the same ground-state structure as DFT, whereas Equiformer-v3 favors a structurally distinct, higher-energy minimum. These differences involve substantial rearrangements of local bonding environments, including changes in Cr–Pd coordination in FCr_3Pd_2, loss of short Mn–O contacts accompanied by altered O–Te coordination in $OMnTe_4$, and reduced S–Cr coordination in S_2Cr_3Pu. These cases therefore provide concrete examples of how improved resolution of relative energies enables Prophet to recover the DFT-preferred structure across distinct local bonding environments.

6.3 Thermodynamic Phase Stability

The preceding analysis establishes that Prophet more reliably identifies ground-state structures and preserves the energetic ordering of competing polymorphs at fixed composition. Thermodynamic phase stability imposes a more global requirement: relative energies must remain consistent not only among structures of

the same composition, but also across different compositions. Convex-hull topology depends on relative energy differences across chemically and structurally distinct compounds, making it particularly sensitive to systematic or composition-dependent errors (Bartel et al., 2020). Even modest shifts in formation energy can stabilize an otherwise metastable compound, remove a genuine ground-state phase, or alter its predicted decomposition products.

Figure 7 examines the thermodynamic stability of the binary Mo–O and ternary Ni–Pb–O systems. For each system, convex hulls are constructed independently from formation energies obtained with DFT, Prophet, and Equiformer-v3 using the same set of candidate structures. The comparison therefore probes whether each model preserves the relative energies required to recover the DFT-stable phases and the corresponding hull topology. For the binary Mo–O system [Fig. 7(a,c,e)], DFT identifies MoO_2 and MoO_3 as the stable intermediate compounds between the elemental endpoints. Prophet reproduces both phases and preserves the DFT hull topology across the full composition range. Equiformer-v3, in contrast, spuriously stabilizes the Mo-rich compound Mo_7O, which lies far above the DFT hull. This error introduces an artificial tie line and qualitatively changes the predicted phase diagram. The corresponding Mo_7O structure spuriously stabilized by Equiformer-v3 is shown in Fig. 7(g).

The ternary Ni–Pb–O system [Fig. 7(b,d,f)] provides a more demanding test because thermodynamic consistency must be maintained over a two-dimensional composition space containing both binary and ternary competing phases. The DFT hull contains NiO and NiO_2 along the Ni–O edge; PbO, Pb_3O_4, and PbO_2 along the Pb–O edge; and the ternary $NiPbO_3$ phase. Prophet recovers all of these phases and closely reproduces their connectivity within the ternary hull. Equiformer-v3 instead misses the DFT-stable NiO_2 and $NiPbO_3$ phases while spuriously stabilizing Ni_3O_4 and $NiPb_2O$. Together, these errors alter the tie-line connectivity and the resulting decomposition pathways. Representative structures of the missed $NiPbO_3$ phase and the spuriously stabilized $NiPb_2O$ phase are shown in Fig. 7(h,i).

These results illustrate why low average energy errors alone do not guarantee accurate phase stability: even localized or composition-dependent errors can change the predicted stable-phase set and hull topology. The close agreement between Prophet and the DFT hulls indicates that its improved energetic accuracy extends across compositions, enabling more reliable prediction of stable phases and decomposition relationships.

7 Defects, Surfaces, and Interfaces

The preceding results focus primarily on bulk crystals and their relative phase stability. We next examine representative defects, surfaces, and interfaces, which introduce local or extended atomic environments distinct from equilibrium bulk structures through changes in composition, coordination, bonding, and symmetry. These examples provide an initial assessment of model transferability to non-bulk environments (Freysoldt et al., 2014). Details of the relaxation protocols and energy definitions are provided in Appendix C.

7.1 Defect Formation Energies

We examine six neutral point defects in CdTe and SiC, comprising two antisites, two interstitials, and two vacancies by Prophet-OE. These defects probe energetic responses to local chemical and structural perturbations, ranging from reduced coordination at vacancies to strongly distorted local environments around interstitials. Starting from identical initial geometries, the defective and matched pristine structures are independently relaxed using DFT, Prophet, and Equiformer-v3, with formation energies evaluated using consistent elemental references (Appendix C).

Figure 8a compares the absolute errors in defect formation energies. Prophet gives lower errors for five of the six defects, with the largest improvement for the Si interstitial, whose error decreases from approximately 0.58eV/defect with Equiformer-v3 to 0.11eV/defect with Prophet. Prophet also gives appreciably lower errors for both SiC vacancies, whereas the C interstitial and TeCd antisite show similar accuracy between the two models. Cd_{Te} is the only case for which Equiformer-v3 is more accurate. Thus, within this limited set, the improvement is most pronounced for SiC interstitial and vacancy energetics rather than being uniform across defect types.

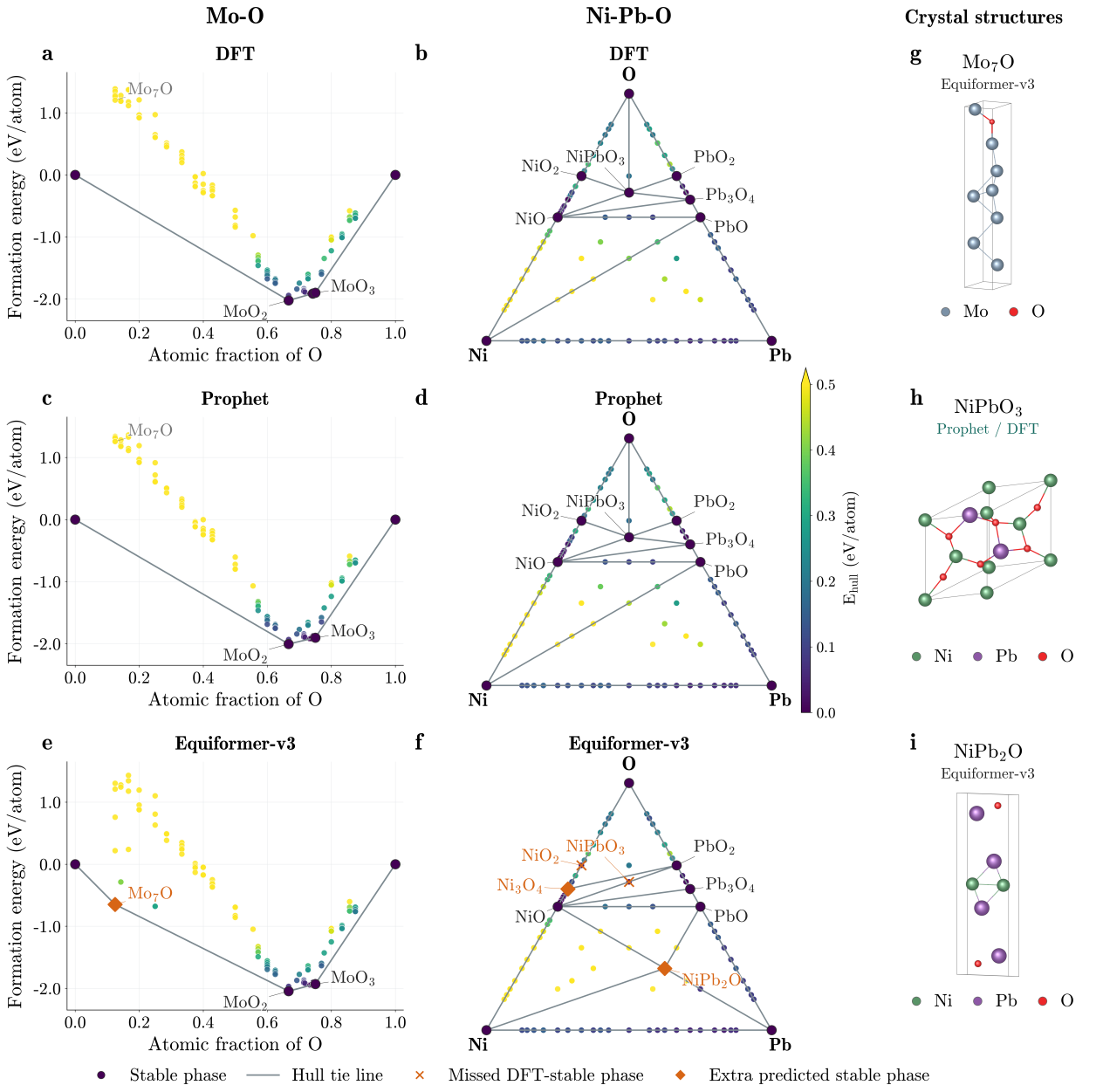

Figure 7 Thermodynamic phase stability and convex-hull reconstruction. Convex hulls obtained from (a,b) DFT, (c,d) Prophet, and (e,f) Equiformer-v3 for the binary Mo–O system (a,c,e) and ternary Ni–Pb–O system (b,d,f). All methods are evaluated on the same set of candidate structures. For each method, points are colored by E_{hull} evaluated with respect to its independently constructed convex hull, and lines connect the predicted stable phases. Relative to the DFT reference, crosses indicate missed stable phases, whereas diamonds indicate spuriously stabilized phases. (g) Mo_7O, spuriously stabilized by Equiformer-v3. (h) $NiPbO_3$, stable according to both DFT and Prophet but missed by Equiformer-v3. (i) $NiPb_2O$, spuriously stabilized by Equiformer-v3.

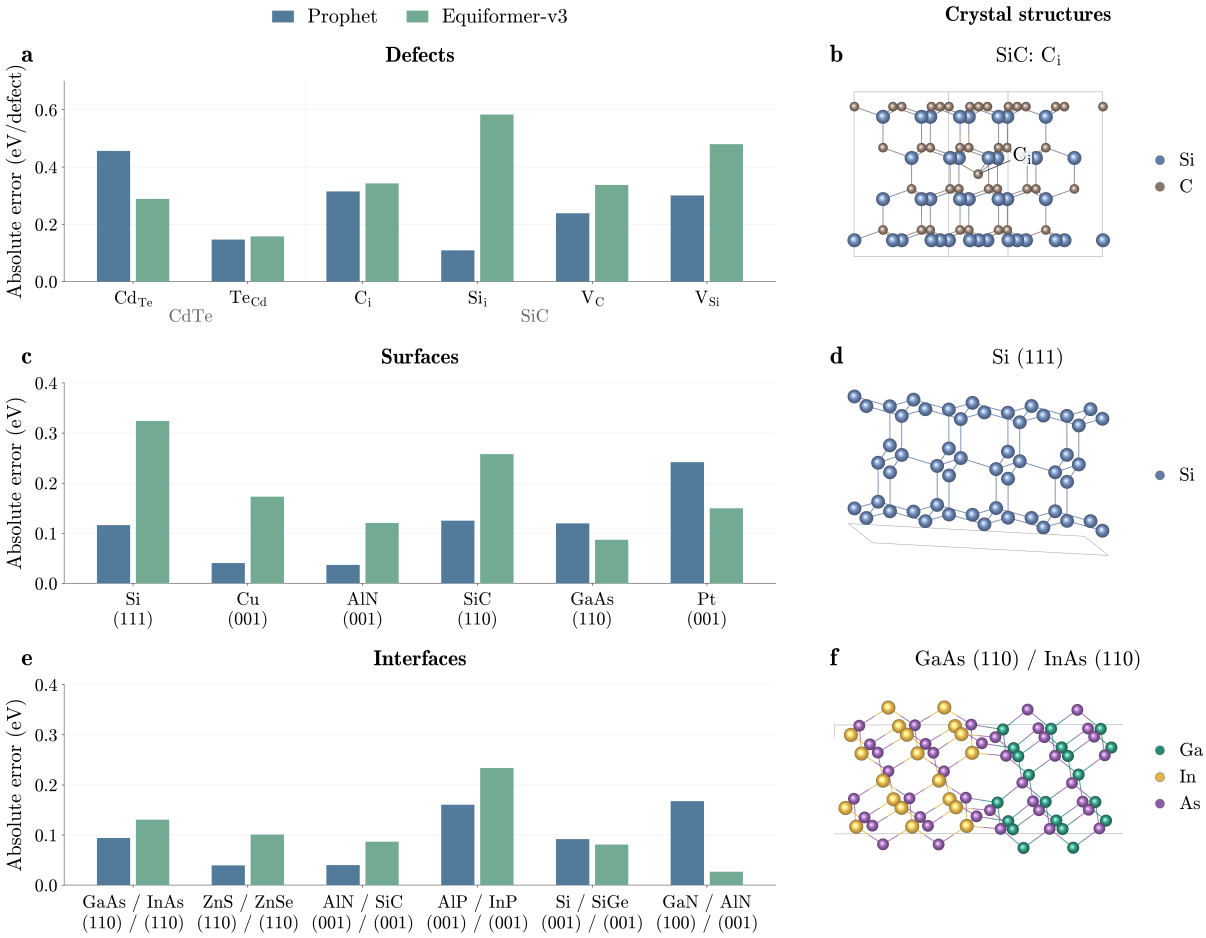

Figure 8 Representative energetics of defects, surfaces, and interfaces. Absolute deviations from DFT are compared for Prophet (blue) and Equiformer-v3 (green) across (a) six neutral defect formation energies in CdTe and SiC, (c) six surface excess energies, and (e) six heterogeneous-interface excess energies. Panels (b), (d), and (f) show representative atomic configurations of the C interstitial in SiC, the Si(111) surface, and the GaAs(110)/InAs(110) interface, respectively.

7.2 Surface and Interface Energetics

We further evaluate six surfaces and six heterogeneous interfaces spanning different compositions and crystallographic orientations. All calculations start from identical geometries, with the non-bulk structures and corresponding bulk references relaxed independently by each method before evaluating the excess energies (Appendix C).

Among the six surfaces in Fig.8c, Prophet yields lower errors in four cases, with the clearest reductions for Si(111), Cu(001), and SiC(110), while Equiformer-v3 is more accurate for GaAs(110) and Pt(001). A representative Si(111) structure is shown in Fig.8d. For the interfaces (Fig.8e), Prophet gives lower errors for four of the six examples, including GaAs/InAs, ZnS/ZnSe, AlN/SiC, and AlP/InP. The two models perform comparably for Si/SiGe, whereas GaN/AlN is a clear exception for which Equiformer-v3 is substantially more accurate. A representative GaAs(110)/InAs(110) interface is shown in Fig.8f.

Overall, Prophet gives lower errors for 13 of the 18 defect, surface, and interface configurations considered here. Although the benchmark is intentionally limited in scope, the results indicate improved transferability across several non-bulk environments while also highlighting specific cases where further improvement is needed.

8 Magnetic States and Phase Behavior

The preceding sections established the accuracy and transferability of Prophet across structural energetics, phase stability, and non-bulk atomic environments. These capabilities, however, remain within a structure-based description in which atomic species and geometry uniquely determine the predicted energy. Magnetic materials introduce an additional dimension: the same atomic structure can support multiple magnetic states with distinct energies, forces, and thermodynamic behavior. We now turn to Prophet-Spin, a spin-aware foundation model that treats local magnetic moments as explicit degrees of freedom and generalizes across materials without material-specific fine-tuning. We examine its capabilities from competing magnetic states and phase stability to finite-temperature magnetic ordering and pressure-induced magnetostructural transitions.

8.1 Magnetic Energetics, Ordering, and State Search

We evaluate Prophet-Spin on 188 experimentally characterized collinear magnetic materials from MAGNDATA Gallego et al. (2016), recomputed under the **ELEMENTA** protocol with converged site-resolved magnetic moments. All reduced formulas are excluded from training. We construct two complementary benchmarks. MAG188-EXP contains the DFT-relaxed experimentally reported magnetic state of each material (188 endpoints), whereas MAG188-SAMPLE contains 1,195 DFT-relaxed endpoints spanning multiple self-consistent collinear magnetic states of the same materials. The former evaluates accuracy on experimentally established magnetic states, while the latter probes the resolution of competing magnetic states within the same material.

On MAG188-EXP, spin-agnostic foundation models give energy MAEs of 63–72 meV/atom, whereas Prophet-Spin conditioned on DFT-converged local magnetic moments reaches 19.2 meV/atom, a 3.3× reduction relative to Prophet-OAME-MBD, while reducing the force MAE from 33–45 to 18 meV/Å (Table 2). The distinction becomes more fundamental on MAG188-SAMPLE: spin-agnostic models have no explicit variable specifying the magnetic state, whereas Prophet-Spin explicitly conditions the energy on the local magnetic moments. Across the sampled magnetic states, Prophet-Spin achieves energy and force MAEs of 16.3 meV/atom and 20 meV/Å, respectively, compared with energy MAEs of 36–46 meV/atom for the spin-agnostic models. The comparable accuracy on MAG188-EXP and MAG188-SAMPLE shows that Prophet-Spin retains its accuracy across distinct self-consistent magnetic states rather than only the experimentally reported ordering.

Table 2 Energy and force errors on MAG188-EXP and MAG188-SAMPLE. E and F denote the energy MAE per atom and force-component MAE, respectively. Raw energies share the **ELEMENTA** reference for in-house models; public models use their own isolated-atom references (*). For Prophet-Spin, "DFT-converged" denotes evaluation with DFT-converged local magnetic moments supplied directly to the spin-conditioned energy model, whereas "spin-denoised" denotes evaluation with magnetic moments predicted from simple initial magnetic-moment guesses by the spin-denoising head.

Model	Magnetic-moment source	MAG188-EXP		MAG188-SAMPLE	
		E (meV/at)	F (meV/Å)	E (meV/at)	F (meV/Å)
Prophet-OAME-MBD	–	62.84	33.4	35.63	31.0
MACE-MP-0	–	71.97*	40.1	45.69*	38.3
MACE-MPA-0	–	65.08*	37.5	39.94*	34.0
MatterSim v1	–	66.03*	45.3	36.90*	43.6
EquiformerV3	–	64.95*	40.1	37.72*	37.0
ORB v3	–	66.51*	42.8	40.96*	38.1
Nequix-MP-1	–	65.48*	37.3	38.91*	34.0
Nequix-OAM-1	–	65.70*	36.1	40.83*	32.5
Prophet-Spin	DFT-converged	**19.23**	**18.3**	**16.34**	**20.1**
Prophet-Spin	spin-denoised	30.73	26.8	20.75	30.9

The Prophet-Spin results with DFT-converged magnetic moments evaluate the spin-conditioned energy model independently of errors in magnetic-moment prediction. In practical inference, however, the converged

magnetic moments are not known in advance. We therefore use the spin-denoising head introduced in Section 2.2 to predict them from simple initial magnetic-moment guesses. For each candidate collinear ordering, the sign of the initial magnetic moment at each magnetic site specifies its spin orientation, while its magnitude is initialized to a fixed element-specific value. The spin-denoising head maps these initial magnetic moments to the predicted site-resolved magnetic moments used by the spin-conditioned energy model.

Among the 188 MAG188-EXP materials, 177 contain magnetic sites with DFT-converged magnetic moments above $0.5\,\mu_B$ and are used to evaluate spin denoising, comprising 721 magnetic sites. The spin-denoising head achieves a site-wise magnetic-moment MAE of $0.21\,\mu_B$, and in 75% of these materials all magnetic sites are recovered within $0.2\,\mu_B$ of their DFT-converged magnetic moments. Detailed error definitions and element-resolved results are provided in the Appendix D.

We next evaluate whether these predicted magnetic moments are sufficiently accurate for magnetic-state energetics and ranking. Among the 188 MAG188-SAMPLE materials, 178 have multiple DFT-enumerated magnetic configurations and are used for this comparison. With DFT-converged magnetic moments, Prophet-Spin selects the same lowest-energy magnetic state as DFT in 63% of these materials and gives a Spearman rank correlation of 0.63 with the DFT energy ordering. Using spin-denoised magnetic moments, these values decrease to 57% and 0.58, respectively. On MAG188-SAMPLE, replacing DFT-converged magnetic moments with spin-denoised magnetic moments increases the energy MAE from 16.3 to 20.8 meV/atom. Thus, spin denoising introduces only a modest loss in accuracy, enabling magnetic-state evaluation without DFT magnetic moments or material-specific fine-tuning.

We finally apply the complete magnetic-state search workflow to $BaCr_2As_2$, $BaMn_2Bi_2$, and $RbFeSe_2$, whose DFT magnetic states span more than 50 meV/atom (Figure 9). Starting from the crystal structure alone, we enumerate candidate collinear magnetic orderings, predict their site-resolved magnetic moments using the spin-denoising head, and alternate magnetic-moment prediction with structural relaxation until convergence. A zero-moment initialization is additionally included as the non-magnetic (NM) candidate.

Without material-specific fine-tuning or prior knowledge of the magnetic ground state, Prophet-Spin identifies the DFT AFM ground state in all three materials and closely reproduces the relative energetics of the competing magnetic states. The ground-state energy errors are only 2–8 meV/atom, while the higher-energy FiM and FM states are reproduced within 9 meV/atom for $BaCr_2As_2$ and $RbFeSe_2$, and within 13 meV/atom for $BaMn_2Bi_2$. In contrast, the spin-agnostic Prophet-OAME-MBD yields errors of 56, 92, and 100 meV/atom for the corresponding ground-state structures, respectively. These errors are comparable to the energy splittings among competing magnetic states, highlighting the importance of explicitly resolving magnetic degrees of freedom for predicting magnetic-state energetics.

8.2 Magnetism-Dependent Phase Stability

The thermodynamic stability of a magnetic material is determined jointly by its atomic structure and magnetic state. This poses a fundamental challenge for conventional spin-agnostic foundation models: even for a fixed crystal structure, different magnetic configurations can differ substantially in energy, and the magnetic ground state may require a periodicity larger than the crystallographic unit cell. Neglecting magnetic degrees of freedom can therefore lead not only to inaccurate energies, but also to an incorrect ordering of competing phases and an altered thermodynamic phase diagram.

We demonstrate the effect for the Li–Mn–As ternary system by constructing convex hulls with the spin-agnostic model and Prophet-Spin using the same set of structural candidates (Fig. 10). In the spin-aware calculation, competing collinear magnetic configurations are explored on the Mn sublattice using multiple moment initializations and spin patterns. Supercells are introduced when necessary to capture AFM order beyond the primitive-cell periodicity. The low-energy magnetic configurations are then relaxed together with the atomic structure, allowing the preferred magnetic state to emerge from the search rather than being prescribed in advance.

The impact is particularly clear for LiMnAs. The spin-agnostic model places LiMnAs above the convex hull and therefore predicts it to be metastable [Fig. 10(a)]. In contrast, the spin-aware search identifies an AFM ground state and stabilizes LiMnAs on the convex hull [Fig. 10(b)]. LiMnAs has been experimentally synthesized, and its AFM ground state is experimentally established (Beleanu et al., 2013). Without system-specific fine-tuning

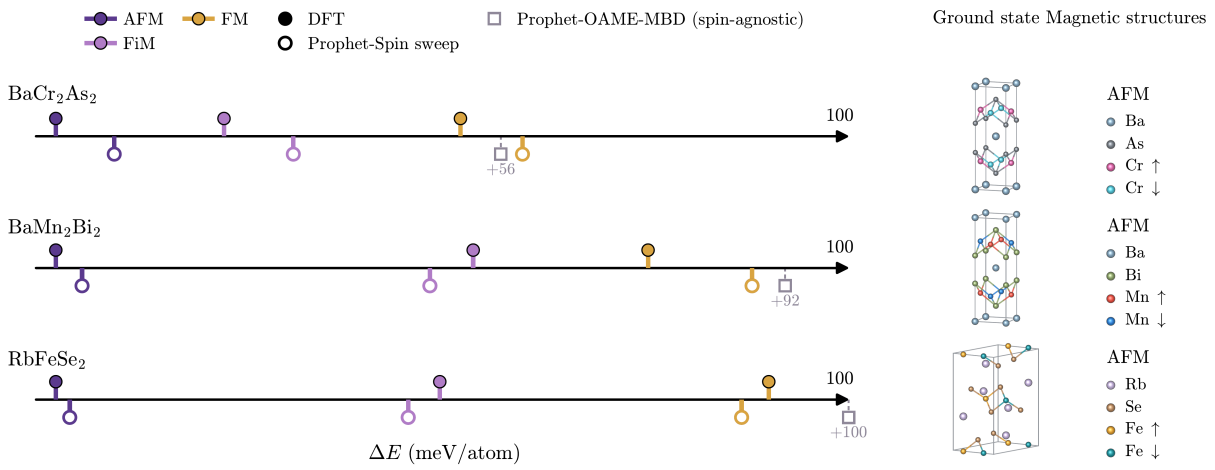

Figure 9 Magnetic-state energetics of $BaCr_2As_2$, $BaMn_2Bi_2$, and $RbFeSe_2$ from MAG188-SAMPLE. Energies are referenced to the DFT magnetic ground state of each material. Filled and open circles denote DFT and Prophet-Spin energies, respectively, for competing collinear AFM, FiM, and FM states. Grey squares show the energies assigned by the spin-agnostic Prophet-OAME-MBD to the DFT ground-state structures; annotated values give the corresponding energy errors in meV/atom. Prophet-Spin identifies the DFT AFM ground state in all three materials without material-specific fine-tuning. The corresponding DFT ground-state magnetic structures are shown on the right, with opposite local-moment orientations indicated by distinct colors.

or prior knowledge of its magnetic order, Prophet-Spin therefore recovers both its observed magnetic state and thermodynamic stability. The agreement captures the dominant in-plane AFM ordering shown in Fig. 10(c). Prophet-Spin does not distinguish between the two interlayer AFM stackings considered here; the corresponding independently relaxed DFT states differ by approximately 0.7 meV/atom, with the alternating stacking lower in energy (Appendix D). The effect extends beyond LiMnAs. Explicit magnetic ordering also stabilizes AFM Mn_2As on the convex hull, while MnAs remains stable with FM ordering (Austin et al., 1962; Ishikawa et al., 2004). These changes in the stable-phase set reconstruct the hull tie-line topology and, consequently, the predicted decomposition and phase-coexistence relationships. The results demonstrate that explicitly treating magnetic configurations—including orders that require supercell representations—allows magnetic ground states to participate directly in thermodynamic phase-stability prediction.

8.3 Finite-Temperature Magnetic Transitions

Finite-temperature magnetic ordering is one of the defining collective phenomena of magnetic materials, linking microscopic exchange interactions to macroscopic magnetic functionality. Predicting ordering temperatures such as T_C and T_N requires not only accurate energetics of individual magnetic configurations, but also a faithful description of the collective spin interactions that govern the emergence and thermal destruction of long-range magnetic order. This regime remains inaccessible to conventional spin-agnostic MLFFs: although they can describe thermally driven atomic motion, the absence of explicit magnetic degrees of freedom prevents them from capturing thermal spin fluctuations and magnetic phase transitions. These properties are therefore conventionally studied using material-specific spin Hamiltonians, with exchange parameters obtained by fitting DFT energies of multiple magnetic configurations to a small set of shell-resolved interactions such as J_1, J_2, and J_3 (Xiang et al., 2013). In contrast, Prophet-Spin provides a native exchange readout that directly yields environment-dependent magnetic interactions, enabling large-scale statistical spin simulations without additional DFT fitting or material-specific parameterization. To demonstrate this capability across distinct magnetic regimes, we consider bcc Fe as a prototypical three-dimensional FM, together with layered FeSn and KV_2Se_2O exhibiting intralayer FM and AFM order, respectively.

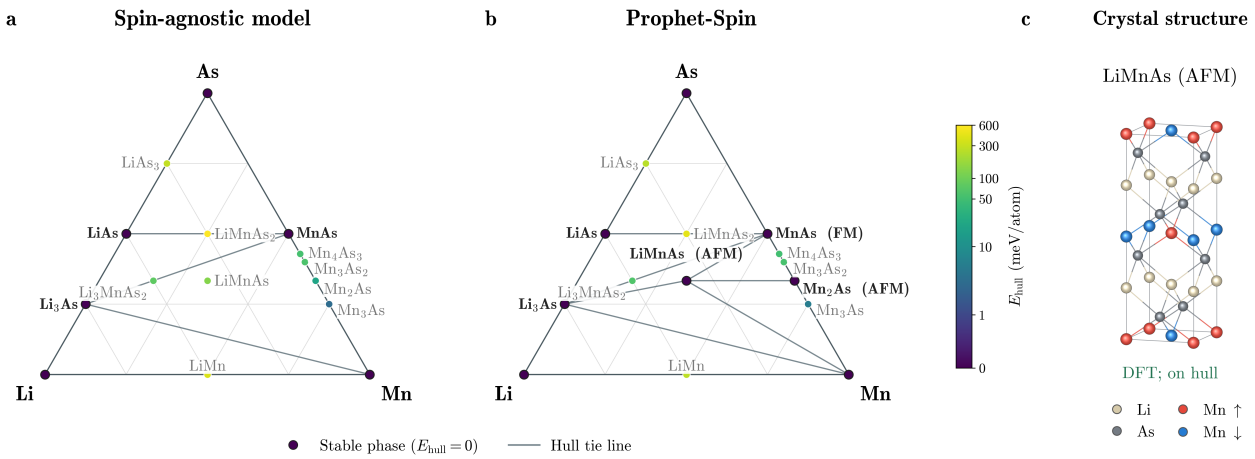

Figure 10 Effect of magnetic degrees of freedom on the predicted phase stability of the Li–Mn–As system. Convex hulls predicted by (a) the spin-agnostic model and (b) Prophet-Spin. In the spin-agnostic description, LiMnAs and Mn_2As lie above the convex hull. Explicit exploration of magnetic configurations stabilizes AFM LiMnAs and AFM Mn_2As on the hull, while MnAs remains stable with FM ordering. Points are colored by their predicted energy above the hull, E_{hull}, with dark points denoting stable phases. (c) Representative DFT-stable structures of LiMnAs. Red and blue Mn atoms indicate opposite spin orientations.

For a fixed atomic structure, the magnetic Hamiltonian used in the Monte Carlo simulations is

$$\mathcal{H} = \frac{1}{2} \sum_{e:i\to j} \widetilde{J}_e \, \boldsymbol{\mu}_i \cdot \boldsymbol{\mu}_j + E_{\mathrm{anis}},$$

where $\boldsymbol{\mu}_i$ denotes the local magnetic moment at site i, including both its magnitude and orientation, and the factor of $1/2$ avoids double counting because each interacting pair is represented by two directed edges, $i \to j$ and $j \to i$. The exchange term is provided directly by the native exchange readout of the Prophet-Spin. The bond-specific coefficient $\widetilde{J}_e$ depends on the local chemical and structural environment, retaining the environment dependence learned by the foundation model rather than reducing the magnetic interactions to a small number of coordination-shell parameters such as J_1, J_2, and J_3. The anisotropy contribution E_{anis} is introduced separately to account for the leading symmetry-breaking interactions, with the corresponding anisotropy parameters obtained from independent DFT calculations including spin–orbit coupling (SOC). For bcc Fe, cubic magnetocrystalline anisotropy is described by $E_{\mathrm{anis}} = K_1 \sum_i (\hat{\mu}_{ix}^2 \hat{\mu}_{iy}^2 + \hat{\mu}_{iy}^2 \hat{\mu}_{iz}^2 + \hat{\mu}_{iz}^2 \hat{\mu}_{ix}^2)$, with $K_1 = 0.004$ meV/Fe (Graham, 1958). For the layered systems, uniaxial single-ion anisotropy is described by $E_{\mathrm{anis}} = -A \sum_i (\hat{\boldsymbol{\mu}}_i \cdot \hat{\mathbf{n}})^2$, where $\hat{\mathbf{n}}$ is the out-of-plane unit vector. FeSn exhibits easy-plane anisotropy with $A = -0.012$ meV/Fe, while KV_2Se_2O exhibits easy-axis anisotropy with $A = 0.03$ meV/V. Here, $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i / |\boldsymbol{\mu}_i|$ denotes the local-moment direction.

We perform fixed-lattice classical Monte Carlo simulations in which the atomic structures and local-moment magnitudes $|\boldsymbol{\mu}_i|$ are fixed at their relaxed values, while the orientations of $\boldsymbol{\mu}_i$ are thermally sampled. Under this approximation, the local chemical and structural environments, and consequently the exchange coefficients $\widetilde{J}_e$, remain unchanged throughout the simulation. The native exchange readout therefore needs to be evaluated only once, enabling efficient large-scale Monte Carlo sampling. Structural and moment-magnitude fluctuations are in principle accessible to the Prophet-Spin, but incorporating these degrees of freedom would require repeated full model evaluations and substantially increase the computational cost. Simulations are performed using a $15 \times 15 \times 15$ supercell containing 6750 Fe spins for bcc Fe, and $15 \times 15 \times 5$ supercells containing 6750 Fe and 4500 V spins for FeSn and KV_2Se_2O, respectively.

As shown in Fig. 11(a–c), the magnetic order parameters decrease with temperature, while the magnetic heat capacities exhibit maxima that define the corresponding ordering temperatures. For bcc Fe, the model predicts a Curie temperature of $T_{\mathrm{C}} = 1150$ K, close to the experimental value of 1043 K (Hrubiak et al., 2017). For the layered systems, it reproduces the intralayer FM order of FeSn and AFM order of KV_2Se_2O, yielding ordering temperatures of 480 and 210 K, respectively, compared with experimental values of 365 and approximately

400 K (Sales et al., 2019; Sun et al., 2025). Their ground-state spin arrangements and representative exchange pathways are illustrated in Fig. 11(g–i). The larger deviations for the layered systems may reflect the difficulty of resolving weak interlayer exchange interactions, which establish the three-dimensional A-type AFM order in FeSn and G-type AFM order in KV_2Se_2O. Because these interactions can lie at the sub-meV scale, we focus here on the more robust intralayer order parameters. Monte Carlo simulations based on independently DFT-derived, shell-resolved exchange parameters are provided in Appendix D for comparison. Overall, the results show that the native exchange readout can be transferred directly to collective simulations involving thousands of spins without material-specific fine-tuning or auxiliary fitting of exchange parameters.

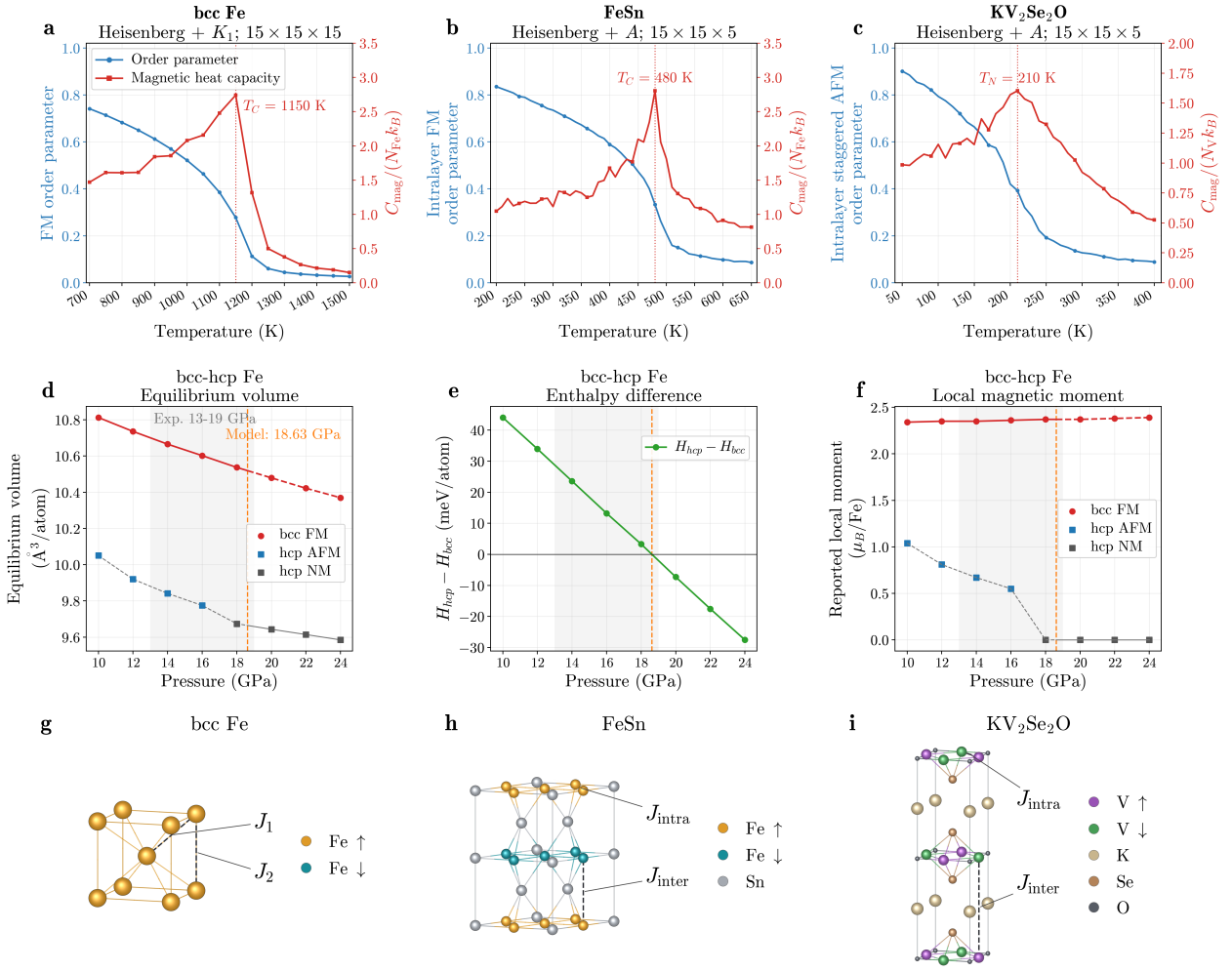

Figure 11 Finite-temperature magnetic ordering and the pressure-induced magnetostructural transition in Fe-based magnetic systems. (a–c) Temperature-dependent magnetic order parameters and magnetic heat capacities for bcc Fe, FeSn, and KV_2Se_2O, respectively. Vertical dotted lines mark the heat-capacity maxima used to determine the ordering temperatures. (d) Equilibrium volumes of FM bcc, AFM hcp, and nonmagnetic hcp Fe under pressure. The shaded region indicates the experimental transformation range of 13–19 GPa (Merkel et al., 2020), and the orange dashed line marks the predicted transition pressure of 18.63 GPa. (e) Enthalpy difference $H_{\mathrm{hcp}} - H_{\mathrm{bcc}}$, whose zero crossing determines the transition pressure. (f) Pressure evolution of the local magnetic moments in the bcc and hcp phases. (g–i) Corresponding magnetic structures and representative exchange pathways. Opposite colors denote opposite local-moment orientations; the indicated neighbor shells and intra- and interlayer pathways are illustrative, while the simulations retain the full bond-specific exchange readout.

8.4 Pressure-Induced Magnetostructural Transitions

Pressure provides a direct thermodynamic route for tuning the interplay between lattice and magnetic degrees of freedom, and pressure-induced magnetostructural transitions are a fundamental manifestation of this coupling in magnetic materials. Changes in interatomic spacing under compression can strongly modify magnetic moments and exchange interactions, which in turn affect the relative stability of competing crystal structures. Capturing such transitions therefore requires a simultaneous description of structural energetics and the evolution of magnetic states. Conventional spin-agnostic MLFFs can accurately describe pressure-dependent structural energetics, but lack explicit magnetic degrees of freedom and cannot resolve the accompanying evolution of magnetic order and local moments. Elemental Fe provides a canonical example, undergoing a pressure-driven transformation from FM bcc to the denser hcp phase, accompanied by a pronounced volume reduction and suppression of magnetism (Zarkevich and Johnson, 2015). We therefore use compressed Fe to demonstrate how Prophet-Spin captures structural stability, volume evolution, and magnetic-state changes within a unified pressure-dependent framework.

At zero pressure, Prophet-Spin correctly identifies FM bcc Fe as the stable phase. We then follow the bcc and hcp branches independently from 10 to 24 GPa, allowing the lattice parameters, atomic positions, and local-moment magnitudes to relax at each pressure. Their relative stability is determined from the static enthalpy, $H = E + PV$. As shown in Fig. 11(d,e), compression progressively stabilizes the denser hcp phase, and $H_{\mathrm{hcp}} - H_{\mathrm{bcc}}$ crosses zero at 18.63 GPa. This transition pressure lies within the experimental transformation range of approximately 13–19 GPa (Merkel et al., 2020) and is accompanied by a predicted volume collapse of approximately 8.1%. Independent DFT calculations using the same first-principles protocol give a lower enthalpy crossing of approximately 11.7 GPa (Appendix D), while retaining the same pressure-driven stabilization of the hcp phase.

The calculated local moments reveal the accompanying magnetic evolution [Fig. 11(f)]. The bcc branch retains a robust FM moment of approximately $2.4\,\mu_{\mathrm{B}}$/Fe throughout the pressure range. In contrast, hcp Fe exhibits an intermediate AFM state whose local moment decreases from $1.04\,\mu_{\mathrm{B}}$/Fe at 10 GPa to $0.55\,\mu_{\mathrm{B}}$/Fe at 16 GPa before the nonmagnetic hcp state becomes energetically favored near the predicted structural transition. Independent DFT calculations likewise show progressive suppression of the hcp AFM moment and an eventual transition to a nonmagnetic state, although the magnetic crossover occurs at substantially higher pressure (Appendix D). Prophet-Spin therefore captures the coupled evolution of structural stability, equilibrium volume, and magnetic order under compression, with its predicted structural transition pressure lying closer to the experimental transformation range than the corresponding DFT reference.

Taken together, these results show that explicitly incorporating magnetic degrees of freedom extends foundation MLFFs beyond structure-based energetics to magnetic phase behavior across composition, temperature, and pressure. Without system-specific fine-tuning, Prophet-Spin connects magnetic-state energetics to thermodynamic phase stability, collective finite-temperature ordering, and pressure-induced magnetostructural transitions. This extension illustrates how expanding the physical state space represented by an atomistic foundation model can enable phenomena that are inaccessible to structure-only descriptions.

9 Future Directions

Prophet provides evidence that scaling an atomistic foundation model along three axes—composition, configuration, and physical state—can lead to broader and more reliable capabilities. The present results are nevertheless only a starting point. There are many directions in which this work can be extended; three are particularly important. First, the level of theory used to generate the training data needs to improve. The Materials Project setting combines system-dependent Hubbard corrections with other choices that are not always consistent across elements, structures, and magnetic states, which can produce an internally inconsistent potential-energy surface. DFT+U also has known accuracy limitations, especially for localized and strongly correlated electrons. A larger dataset built on a more accurate and consistently defined level of theory would therefore be as important as increasing its size. Second, the current inference cost remains too high for workflows requiring millions or billions of evaluations. Atomistic models need a dedicated inference stack, with the same separation between training and deployment that has become standard for large language models. Compiled kernels, efficient graph construction, batching, and hardware-specific execution could make

large-scale screening and simulation substantially more practical.

A longer-term challenge is to connect microscopic predictions to macroscopic material behaviour. Properties measured in practice depend not only on atomic energetics, but also on defects, interfaces, microstructure, processing history, transport across length scales, and device geometry. Addressing this gap will require combinations of atomistic simulation, coarse-grained and continuum models, process-aware modelling, experimental data, and methods for tracking uncertainty as information passes between scales. The present results on defects, surfaces, and interfaces are encouraging, but their accuracy remains noticeably below that obtained for bulk crystals. These environments are also much less thoroughly represented in existing datasets. Large, deliberately designed calculation campaigns covering surfaces, interfaces, defects, reconstructions, amorphous structures, and realistic processing conditions could therefore produce substantial gains. The directions listed here are only a subset of the available opportunities; together they point toward atomistic foundation models that are not only more accurate on benchmark structures, but useful across the full chain from electronic structure to materials performance.

Data, Code, and Model Availability

Code. The Prophet source code, including the model definition and inference stack and the ASE calculators used in this report (`KairosCalculator` for the structural models, `SpinCalculator` and `MagmomPredictor` for Prophet-Spin), is available at https://github.com/kairosmaterial/prophet under the MIT license and can be installed with `pip install "prophet-mlip @ git+https://github.com/kairosmaterial/prophet.git"`.

Model weights. The checked points are distributed through the Hugging Face model repository https://huggingface.co/kairosmaterial/prophet under the CC-BY-4.0 license: `prophet-oame-mbd.pt` (Prophet-OAME-MBD, the structural model of the Matbench Discovery evaluation), `prophet-v1-mag.pt` (the Prophet-Spin energy model with exchange readout), and `prophet-v1-mag-head.pt` (the same model bundled with the spin-denoising head).

Training data. ELEMENTA, ELEMENTA-Vib and ELEMENTA-Spin were generated by Kairos Materials with the single DFT protocol described in Kairos Materials (2026). A subset ELEMENTA set plus the ELEMENTA-Spin set are released in extended-XYZ format under the CC-BY-NC-4.0 license at https://huggingface.co/datasets/kairosmaterial/ELEMENTA; the ELEMENTA technical report is archived at https://doi.org/10.5281/zenodo.22638068.

Benchmarks and evaluation data. Matbench Discovery Riebesell et al. (2025) predictions of Prophet-OAME-MBD for the discovery, geometry-optimization, thermal-conductivity, diatomic and molecular-dynamics tasks, together with the model card, are deposited in the Matbench Discovery repository. The MAG188-EXP benchmark of Section 8, derived from the MAGNDATA database Gallego et al. (2016), and its MAG188-SAMPLE extension, including the recomputed relaxation trajectories and reference magnetic states, will be released on Hugging Face at https://huggingface.co/datasets/kairosmaterial/MAG188.

Correspondence. Requests for materials should be sent to research@kairosmaterials.com.

Appendix

A Matbench Discovery Leaderboard Snapshot

Table 3 reproduces the public Matbench Discovery leaderboard as of 15 September 2026, restricted to the 42 models that have a Combined Performance Score (CPS); entries marked as superseded on the site, and models without thermal-conductivity or geometry-optimization results, are omitted. All metrics are taken from the model cards in the Matbench Discovery repository, and CPS is recomputed from them with the site's weights (50% F_1, 40% κ_{SRME}, 10% RMSD). Figure 1c and the discussion in Section 5 draw on this snapshot.

B Spin Symmetry, Exchange, and Denoising

This appendix specifies the magnetic components of Prophet-Spin in the form used by the released checkpoints: the symmetry of the spin-conditioned energy, the spin encodings and exchange readout, the corruption mixture and training of the spin-denoising (recovery) head D_ϕ, the inference interfaces, the derivatives returned at a predicted magnetic state, and the extraction of the Heisenberg Hamiltonian. Energies are in eV and magnetic moments in μ_{B} throughout.

B.1 Spin symmetry and completeness of the invariant basis

The magnetic moments enter the energy network only through three invariants: the squared magnitudes $a_i^2 = |\mathbf{S}_i|^2$ at the nodes, the pair sums $q_{ij} = a_i^2 + a_j^2$ on the trunk edges, and the products $p_{ij} = \mathbf{S}_i \cdot \mathbf{S}_j$ in the exchange readout. All three are unchanged by a common $U \in O(3)$ acting on every moment, so Equation (7) holds exactly for collinear and vector inputs, independently of the $E(3)$ invariance of the spatial backbone. Time reversal is the element $U = -I$ and needs no separate treatment. The following statement shows that restricting the inputs to inner products loses no admissible dependence.

Proposition 1 (Completeness of the invariant basis without spin–orbit coupling)**.** *Let the energy $E(X; \mathbf{S}_1, \ldots, \mathbf{S}_N)$ be invariant under $\mathbf{S}_i \to U\mathbf{S}_i$ for all $U \in O(3)$ at fixed X. Then E is a function of the Gram entries $\{\mathbf{S}_i \cdot \mathbf{S}_j\}_{i \le j}$ alone.*

Proof. For polynomials this is the first fundamental theorem of invariant theory for $O(3)$ acting diagonally on tuples of vectors: every invariant polynomial is a polynomial in the pairwise inner products. For general functions it suffices that the Gram matrix separates orbits. Two moment configurations with equal Gram matrices span subspaces of equal dimension, and an orthogonal map carries one onto the other. They therefore lie on the same $O(3)$ orbit, and every invariant function takes the same value on both. □

Under the smaller group $SO(3)$ the triple products $\mathbf{S}_i \cdot (\mathbf{S}_j \times \mathbf{S}_k)$ would also be admissible. They are odd under $-I$ and are excluded by time reversal at linear order, while their squares are Gram determinants and hence already functions of the inner products. The inner-product basis is thus complete for the magnetic symmetry group without spin–orbit coupling. Prophet-Spin uses one particular split of this basis: the diagonal entries enter the trunk nonlinearly through a_i^2 and q_{ij}, and the off-diagonal entries enter linearly through the exchange readout. The linearity in p_{ij} is a modelling choice; the basis itself omits no invariant.

The collinear head inherits an exact reversal relation from the same construction. In Heisenberg mode the trunk receives no p_{ij} encoding, so its features $\mathbf{z}_i$ are even in the seed, and the local exchange energy $\bar{\eta}_i$ depends on the seed only through products $\widetilde{m}_{0,i}\widetilde{m}_{0,j}$. Both are unchanged under $\widetilde{\boldsymbol{m}}_0 \to -\widetilde{\boldsymbol{m}}_0$, and so is $\widetilde{m}_{0,i}^2$. The multiplier $\mathcal{A}_\phi$ in Equation (9) is therefore invariant, and $D_\phi(X, -\widetilde{\boldsymbol{m}}_0) = -D_\phi(X, \widetilde{\boldsymbol{m}}_0)$ holds for every structure and every seed.

B.2 Spin encodings and exchange readout

All magnetic encodings use Gaussian bases $\exp[-((x - c_k)/w)^2]$ with 16 equally spaced centres c_k and width w equal to 1.5 centre spacings. The node encoding expands a_i^2 on $[0, 32]$ μ_{B}^2 and is added to the initial scalar node features through a linear map. The edge encoding expands q_{ij} on $[0, 64]$ μ_{B}^2, uses a separate linear map

Table 3 Matbench Discovery leaderboard, 15 September 2026 snapshot: all 42 models on the public leaderboard that have a Combined Performance Score (entries the site marks as superseded, and models without thermal-conductivity or geometry-optimization results, are omitted), sorted by CPS. Discovery metrics are reported on the unique-prototype split: F_1, discovery acceleration factor (DAF), precision, accuracy, and formation-energy MAE in eV/atom. RMSD is the normalized structure-matcher RMSD between model- and DFT-relaxed WBM structures; κ_{SRME} is evaluated on the 103-material phononDB-PBE set. Higher is better for CPS, F_1, DAF, precision, and accuracy; lower is better for MAE, RMSD, and κ_{SRME}. Params in millions; OAM denotes MPtrj+OMat24+sAlex.

Model	CPS ↑	F_1 ↑	DAF ↑	Prec ↑	Acc ↑	MAE ↓	κ_{SRME} ↓	RMSD ↓	Params	Training data
Prophet-OAME-MBD	**0.912**	0.928	6.07	0.927	**0.978**	**0.018**	**0.065**	0.0591	62.3	OAM+ELEMENTA
TECE-OAM-RRA-1.0	0.908	0.929	6.07	0.928	**0.978**	**0.018**	0.093	**0.0575**	221.9	OAM
EquFlashV2	0.907	0.929	6.07	0.928	**0.978**	**0.018**	0.094	0.0577	44.9	OAM
EquiformerV3+DeNS-OAM	0.902	**0.931**	6.07	0.928	**0.978**	**0.018**	0.118	0.0595	30.3	OAM
GRACE-3L-OAM-L	0.900	0.925	6.04	0.923	0.977	**0.018**	0.121	**0.0575**	42.1	OAM
PET-OAM-XL	0.898	0.924	**6.08**	**0.929**	0.977	0.019	0.119	0.0596	730.0	OAM
TACE-OAM-L	0.889	0.910	5.90	0.902	0.972	0.020	0.126	0.0606	82.9	OAM
eSEN-30M-OAM	0.888	0.925	6.07	0.928	0.977	**0.018**	0.170	0.0608	30.2	OAM
EquFlash	0.888	0.919	5.98	0.915	0.975	0.019	0.158	0.0602	28.7	OAM
Nequip-OAM-XL	0.886	0.906	5.87	0.897	0.971	0.020	0.125	0.0630	32.1	OAM
MatRIS-10M-OAM	0.877	0.921	6.04	0.923	0.976	0.019	0.218	0.0601	10.4	OAM
SevenNet-Omni-i12	0.873	0.906	5.95	0.910	0.971	0.021	0.192	0.0618	54.9	COSMOS
Nequip-OAM-L	0.870	0.893	5.82	0.890	0.967	0.022	0.166	0.0647	9.6	OAM
GRACE-2L-OAM-L	0.865	0.883	5.84	0.893	0.964	0.022	0.169	0.0639	26.4	OAM
ORB v3	0.860	0.905	5.91	0.904	0.971	0.024	0.210	0.0750	25.5	MPtrj+Alex+OMat24
DPA-4.0.1-Pro-MPtrj	0.840	0.857	5.61	0.857	0.956	0.029	0.211	0.0687	22.8	MPtrj
Allegro-OAM-L	0.840	0.895	5.67	0.867	0.966	0.022	0.319	0.0651	9.7	OAM
GRACE-2L-OAM	0.837	0.880	5.77	0.883	0.963	0.023	0.294	0.0666	12.6	OAM
EquiformerV3+DeNS-MP	0.830	0.863	5.48	0.838	0.956	0.029	0.275	0.0701	30.3	MPtrj
DPA-3.1-3M-FT	0.802	0.884	5.67	0.866	0.963	0.023	0.469	0.0693	3.3	OpenLAM
eSEN-30M-MP	0.797	0.831	5.26	0.804	0.946	0.033	0.340	0.0752	30.1	MPtrj
MACE-MPA-0	0.795	0.852	5.58	0.853	0.954	0.028	0.412	0.0731	9.1	MPtrj+sAlex
MatRIS-10M-MP	0.778	0.847	5.42	0.829	0.951	0.031	0.489	0.0717	10.4	MPtrj
AlphaNet-v1-OAM	0.769	0.901	5.75	0.879	0.968	0.024	0.643	0.0793	4.7	OAM
MatterSim v1 5M	0.767	0.862	5.85	0.895	0.959	0.024	0.575	0.0733	4.5	MatterSim
GRACE-1L-OAM	0.761	0.824	5.25	0.803	0.944	0.031	0.517	0.0721	3.4	OAM
Eqnorm MPtrj	0.756	0.786	4.84	0.741	0.929	0.040	0.408	0.0837	1.3	MPtrj
Nequix MP PFT	0.755	0.748	4.48	0.685	0.914	0.044	0.307	0.0866	0.7	MPtrj+MDR ω_q
Nequip-MP-L	0.730	0.761	4.70	0.719	0.921	0.043	0.466	0.0856	9.6	MPtrj
Nequix MP	0.729	0.751	4.46	0.681	0.914	0.044	0.446	0.0853	0.7	MPtrj
Allegro-MP-L	0.720	0.751	4.52	0.690	0.915	0.044	0.504	0.0816	18.7	MPtrj
SevenNet-l3i5	0.714	0.760	4.63	0.708	0.920	0.044	0.550	0.0847	1.2	MPtrj
HIENet	0.707	0.777	4.93	0.754	0.929	0.041	0.642	0.0795	7.5	MPtrj
GRACE-2L-MPtrj	0.681	0.691	4.16	0.636	0.895	0.052	0.525	0.0897	15.3	MPtrj
MACE-MP-0	0.637	0.669	3.78	0.577	0.878	0.057	0.682	0.0915	4.7	MPtrj
BAM-MP-core	0.584	0.623	3.68	0.563	0.869	0.060	0.848	0.0863	17.0	MPtrj
eqV2 M	0.558	0.917	6.05	0.924	0.975	0.020	1.771	0.0691	86.6	MPtrj+OMat24
ORB v2 MPA	0.528	0.880	6.04	0.924	0.965	0.028	1.734	0.0973	25.2	MPtrj+Alex
eqV2 S DeNS	0.522	0.815	5.04	0.771	0.939	0.036	1.676	0.0757	31.2	MPtrj
ORB v2 MPtrj	0.470	0.765	4.70	0.719	0.922	0.045	1.726	0.1007	25.2	MPtrj
M3GNet	0.428	0.569	2.88	0.441	0.812	0.075	1.409	0.1117	0.2	MPF
CHGNet	0.400	0.613	3.36	0.514	0.851	0.063	1.717	0.0949	0.4	MPtrj

for each cutoff branch, and is multiplied by the envelope of that branch before it is added to the edge features. All of these maps are initialized to zero, so the spin-conditioned network initially reproduces the inherited structural backbone.

The coupling network $\mathcal{J}_{\theta_s}$ of Equation (8) receives the sum $\mathbf{s}_j^{(L)} + \mathbf{s}_i^{(L)}$ of the 256 final scalar features of the two sites, a 16-function expansion of r_{ji} on $[0, r_{\mathrm{ex}}]$ and a 16-function expansion of q_{ji} on $[0, 64]\ \mu_{\mathrm{B}}^2$. It has two hidden layers of width 128 with SiLU activations and a zero-initialized scalar output. The exchange envelope f_{ex} is the polynomial envelope with $p = 6$ at the exchange radius r_{ex}. In the released model r_{ex} equals the largest trunk cutoff, 6 Å, so exchange and trunk share one neighbour graph. When a larger exchange radius is configured, the trunk still uses only the edges within its own cutoffs. The spin encodings contain 1.3×10^4 parameters and the exchange readout 5.4×10^4, which together form the 6.7×10^4 magnetic-pathway parameters.

The exchange radius bounds the orderings the model can distinguish. Two collinear states with equal magnitudes whose products $m_i m_j$ differ only for site pairs farther apart than r_{ex} give identical inputs to the trunk and to the exchange readout, and hence identical energies. Section 8 discusses the consequences for layered and chain compounds.

B.3 Seed-path corruption and head training

The released head is the exchange-conditioned variant of Equation (9). The site features $\mathbf{z}_i$ are the 256 even-scalar channels of the output of the penultimate interaction block, evaluated at the seed state. The local exchange energy $\bar{\eta}_i$ is expanded in 12 Gaussian functions on $[-0.15, 0.15]$ eV. A network with two hidden layers of width 128 maps $(\mathbf{z}_i, \bar{\eta}_i)$ to 17 coefficients $g_{i,0}, \ldots, g_{i,16}$, and the multiplier is

$$\mathcal{A}_\phi\left(\widetilde{m}_{0,i}^2, \mathbf{z}_i, \bar{\eta}_i\right) = g_{i,0} + \sum_{k=1}^{16} g_{i,k}\, \exp\left[-\left(\frac{\widetilde{m}_{0,i}^2 - c_k}{w}\right)^2\right], \tag{11}$$

with 16 centres c_k on $[0, 64]\ \mu_{\mathrm{B}}^2$. The head has 5.3×10^4 parameters.

Element-based seeds are drawn from a table of three magnitude levels per element, set to the 30th, 60th and 90th percentiles of the converged $|m|$ of that element in the magnetic training data. The table covers 24 elements; every other element receives 0.1 μ_{B} at all three levels. For a training structure with converged moments $\boldsymbol{m}^\star$, one corruption mode is drawn per structure and applied to the magnetic sites, defined as those with $|m_i^\star| > 0.5\ \mu_{\mathrm{B}}$:

$$\widetilde{m}_{0,i} = \begin{cases} \operatorname{sgn}(m_i^\star)\, \ell_{Z_i,k_i}, & k_i \sim \mathcal{U}\{1,2,3\}, & \text{probability } 0.4, \\ m_i^\star + \sigma\, \varepsilon_i, & \sigma \sim \mathcal{U}(0.1, 0.6)\ \mu_{\mathrm{B}},\ \varepsilon_i \sim \mathcal{N}(0,1), & \text{probability } 0.3, \\ u_i\, m_i^\star, & u_i \sim \mathcal{U}(0.5, 1.5), & \text{probability } 0.2, \\ m_i^\star, & & \text{probability } 0.1. \end{cases} \tag{12}$$

Here $\ell_{Z,k}$ is the k-th level of element Z, the level index k_i and the factor u_i are drawn per site, and σ is drawn per structure. Nonmagnetic sites receive $\widetilde{m}_{0,i} \sim \mathcal{N}(0, 0.1^2)$ in every mode. The level and rescaling modes keep the DFT sign pattern and change only the magnitudes; the noise mode can also change the sign of small moments; the identity mode anchors the converged state as a fixed point of the map.

The feature backbone is the frozen exponential-moving-average model of the spin-conditioned energy network. The head is trained on the magnetic training data of Section 4.4 with the Huber loss of Equation (10), $\delta = 0.05\ \mu_{\mathrm{B}}$, averaged over all sites of a mini-batch. Optimization uses AdamW with learning rate 10^{-3} and weight decay 10^{-5}, a linear warm-up of 500 steps followed by cosine decay to zero, 32 structures per GPU with data-parallel training, and two epochs. After each epoch the head is scored by the magnitude error on the magnetic sites of 4,992 validation frames of ELEMENTA Spin, seeded with the middle element level and the DFT signs. The released head is the second-epoch model, with a validation error of 0.147 μ_{B}.

B.4 Inference interfaces

The released code exposes the two tasks of Equation (6) as separate ASE calculators. `SpinCalculator` loads `prophet-v1-mag.pt` and performs prescribed-state evaluation. It reads the magnetic state from the initial

magnetic moments of the structure, either collinear with shape (N) or vector-valued with shape $(N, 3)$, and returns energy, forces and stress.

`MagmomPredictor` performs seed-to-state recovery. It loads the energy checkpoint together with `prophet-v1-mag-head.pt`, a bundle of the frozen feature backbone, the head and the element seed table, and accepts collinear seeds only. With the default preprocessing rule, sites whose seed magnitude exceeds 0.5 μ_{B} keep the sign of the seed and take the middle level of their element, and all other sites are set to zero; the alternative rule passes the seed unchanged. One head evaluation returns $\widehat{\boldsymbol{m}}$, and the energy checkpoint then evaluates energy, forces and stress at $M(\widehat{\boldsymbol{m}})$. The predicted moments are returned through the magnetic-moment channel of the calculator.

The underlying runtime additionally evaluates many candidate magnetic states of one structure in a single batch and extracts the Heisenberg Hamiltonian of Section B.6. The ordering search of Section 8 is an external procedure that combines head predictions, batched energies and fixed-moment relaxations; it is not part of the released code.

B.5 Conditioned and composite derivatives

At a predicted state two derivatives of the energy with respect to the atomic positions $\mathbf{R}_a$ can be distinguished. The conditioned derivative holds the magnetic moments fixed,

$$\mathbf{F}_a^{\mathrm{cond}} = -\left.\frac{\partial E_{\theta_{\mathrm{s}}}^{\mathrm{spin}}(X, M)}{\partial \mathbf{R}_a}\right|_{M=M(\widehat{\boldsymbol{m}})}, \tag{13}$$

whereas the composite derivative follows the head through its dependence on the geometry,

$$\mathbf{F}_a^{\mathrm{comp}} = \mathbf{F}_a^{\mathrm{cond}} - \sum_{i=1}^{N} \left.\frac{\partial E_{\theta_{\mathrm{s}}}^{\mathrm{spin}}}{\partial m_i}\right|_{M(\widehat{\boldsymbol{m}})} \frac{\partial \widehat{m}_i}{\partial \mathbf{R}_a}. \tag{14}$$

Both calculators return the conditioned force, and the stress is the corresponding derivative with respect to a homogeneous strain at fixed moments. The head is evaluated without gradient tracking, so the second term of Equation (14) is never formed.

The conditioned derivative is the quantity that was supervised. ELEMENTA Spin forces and stresses are computed at self-consistent DFT solutions, where the variational property of the density functional removes the implicit dependence on the electronic and magnetic degrees of freedom, and the energy network is fitted to these labels at the converged moments. The two derivatives coincide only if the predicted moments are a stationary point of the learned energy, $\partial E/\partial m_i = 0$ at $\widehat{\boldsymbol{m}}$, and neither training stage imposes this condition. The composite force would therefore add the geometric sensitivity of the head, weighted by an unsupervised energy gradient. A structural relaxation with Prophet-Spin consequently relaxes the atoms at a fixed magnetic state. Magnetic and structural degrees of freedom are brought to mutual consistency by alternating head predictions and fixed-moment relaxations, as in the search of Section 8.

B.6 Extraction of the Heisenberg Hamiltonian

The extraction runs one forward pass of the energy network at a reference state M_{ref} with magnitudes $\mathbf{a}$. For every directed exchange edge $e = (j \to i)$, including edges to periodic images, it records the coefficient

$$c_e = \tfrac{1}{2}\, J_{ji}(X, \mathbf{a})\, f_{\mathrm{ex}}(r_e), \qquad E_{\theta_{\mathrm{s}}}^{\mathrm{spin}}(X, M) = E_{\mathrm{trunk}}(X, \mathbf{a}) + \sum_e c_e\, \mathbf{S}_j \cdot \mathbf{S}_i. \tag{15}$$

Edges joining a site to one of its own periodic images contribute $c_e a_i^2$, which does not depend on the directions. Collecting all remaining edges between the same pair of sites, in both directions and over all images, gives the folded couplings and the offset

$$\widetilde{J}_{ij} = \sum_{e:\{j,i\}} c_e \quad (i < j), \qquad E_0 = E_{\mathrm{ref}} - \sum_{i<j} \widetilde{J}_{ij}\, \mathbf{S}_i^{\mathrm{ref}} \cdot \mathbf{S}_j^{\mathrm{ref}}, \tag{16}$$

so that

$$E^{\mathrm{spin}}_{\theta_s}(X, M) = E_0 + \sum_{i<j} \widetilde{J}_{ij}\, \mathbf{S}_i \cdot \mathbf{S}_j \tag{17}$$

holds exactly for every collinear or vector state M with the same geometry and magnitudes. E_0 contains the trunk energy and all self-image terms. In this convention $\widetilde{J}_{ij} < 0$ favours parallel and $\widetilde{J}_{ij} > 0$ antiparallel alignment, in eV μ_{B}^{-2}. The couplings depend on $\mathbf{a}$ through the trunk and the q_{ij} encoding, so a change of magnitudes or geometry requires a new extraction.

The folded form reflects the periodicity of the cell rather than an approximation. The energy of a cell-periodic state depends only on the sums over images in Equation (16), and a coupling between a site and its own image acts as a constant. Couplings at wavevectors not commensurate with the cell are resolved by extracting in a supercell. Because each site interacts with a fixed list of neighbours, the energy change of a single-site update costs $O(z_i)$ operations for z_i exchange neighbours. Spin Monte Carlo can therefore run on the extracted $\widetilde{J}_{ij}$ and E_0 without further network evaluations, as long as geometry and magnitudes stay fixed.

C Computational Details for Defects, Surfaces, and Interfaces

For a neutral point defect, the formation energy is calculated as (Freysoldt et al., 2014)

$$E_{\mathrm{f}} = E_{\mathrm{def}} - E_{\mathrm{bulk}} - \sum_i \Delta n_i \mu_i, \tag{18}$$

where E_{def} and E_{bulk} are the relaxed energies of the defective and pristine structures, respectively, Δn_i is the change in the number of atoms of element i, and μ_i is its elemental chemical potential. The same elemental reference phases are used for all methods, with their energies evaluated consistently using DFT, Prophet, or Equiformer-v3.

For a periodic slab, the surface excess energy of the simulation cell is defined as (Fiorentini and Methfessel, 1996)

$$E^f_{\mathrm{surf}} = E_{\mathrm{slab}} - nE_{\mathrm{bulk}}, \tag{19}$$

where E_{slab} and E_{bulk} are the relaxed energies of the slab and parent bulk cell, respectively, and n is the number of equivalent bulk cells represented in the slab.

For a heterogeneous interface, the excess energy is defined as

$$E^f_{\mathrm{int}} = E_{\mathrm{int}} - n_{\mathrm{film}} E^{\mathrm{bulk}}_{\mathrm{film}} - n_{\mathrm{sub}} E^{\mathrm{bulk}}_{\mathrm{sub}}, \tag{20}$$

where E_{int} is the relaxed interface energy, $E^{\mathrm{bulk}}_{\mathrm{film}}$ and $E^{\mathrm{bulk}}_{\mathrm{sub}}$ are the relaxed energies of the constituent bulk cells, and n_{film} and n_{sub} are their respective numbers in the interface structure.

D Supplementary results and DFT References for Prophet-Spin

This appendix provides additional validation and DFT reference results for the Prophet-Spin analyses in the main text, including site-resolved magnetic-moment prediction, Li–Mn–As phase stability, finite-temperature magnetic ordering, and pressure-dependent phase behavior of Fe.

The supplementary DFT calculations follow the first-principles protocol used for ELEMENTA (Kairos Materials, 2026). Calculations are performed using VASP with the PBE exchange–correlation functional (Perdew et al., 1996) and the projector-augmented-wave (PAW) method (Blöchl, 1994; Kresse and Joubert, 1999). Input settings are based on the Materials Project relaxation protocol implemented through `MPRelaxSet` in pymatgen (Ong et al., 2013). Spin-polarized calculations are initialized according to the magnetic configurations considered here. Unless otherwise noted, the remaining settings follow the ELEMENTA protocol.

For the spin-denoising benchmark, we retain 177 MAGNDATA-188 materials with DFT-converged magnetic moments above 0.5 μ_{B}, comprising 721 magnetic sites. As summarized in Table 4, the site-resolved magnetic-moment MAE is 0.211 μ_{B}, with a median absolute error of 0.115 μ_{B}; in 75% of the materials, all magnetic sites are recovered within 0.2 μ_{B} of their DFT values.

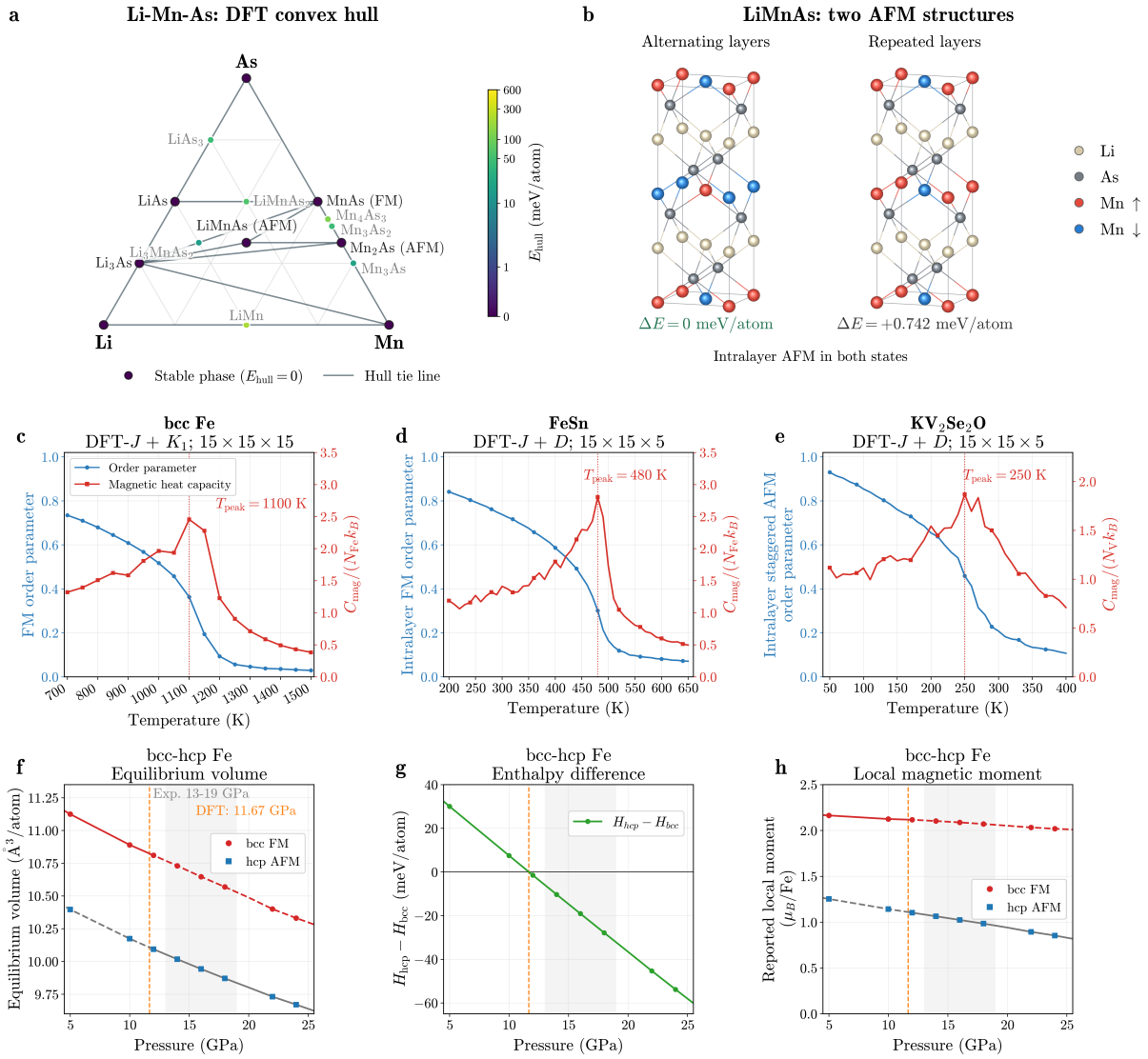

Figure 12 DFT references for the magnetic results in the main text. (a) DFT Li–Mn–As convex hull. (b) LiMnAs states with alternating and repeated interlayer AFM stacking; energies are referenced to the lower-energy alternating state. (c–e) Magnetic order parameters and heat capacities from Monte Carlo simulations using DFT-derived exchange for bcc Fe, FeSn, and KV_2Se_2O. Dotted lines mark the heat-capacity maxima. (f–h) DFT equilibrium volume, hcp–bcc enthalpy difference, and local magnetic moments of Fe under pressure.

As an independent reference for the Li–Mn–As phase-stability analysis, we construct a DFT convex hull from the same 15 compositions as the model hulls. DFT places Li_3As, LiAs, LiMnAs, Mn_2As, and MnAs on the convex hull [Fig. 12(a)], matching the five stable compound compositions obtained with Prophet-Spin. The detailed tie-line connectivity differs: DFT contains the Li_3As–Mn_2As tie line, whereas Prophet-Spin contains the LiMnAs–Mn tie line. The reaction energy separating these alternatives is +3.41 meV/atom in DFT and −3.77 meV/atom in Prophet-Spin. The two LiMnAs states in Fig. 12(b) share the same intralayer AFM order but differ in their interlayer stacking. Independently relaxed DFT structures place the repeated-layer state 0.742 meV/atom above the alternating-layer state. Prophet-Spin does not resolve this splitting; the nearest interlayer Mn–Mn separations, 6.08–6.14 Å, lie beyond its 6.0 Å exchange cutoff.

For comparison with the native Prophet-Spin exchange readout, we also perform Monte Carlo simulations using shell-resolved exchange parameters derived from collinear DFT energies. For bcc Fe, three exchange shells are fitted to nine magnetic configurations, giving an energy RMSE of 7.25 meV/atom. FeSn and

Table 4 Site-resolved magnetic-moment errors of the spin-denoising head on MAGNDATA-188, evaluated over 721 magnetic sites.

Metric	Spin-denoised
Site MAE (μ_B)	0.211
Site median absolute error (μ_B)	0.115
Materials with all sites within 0.2 μ_B	75%
Per-element site MAE (μ_B)	
Mn	0.31
Fe	0.15
Cr	0.09
Co	0.16
Ni	0.06
V	0.06
Os	0.18
Re	0.44
Ru	0.07
Cu	0.03

KV_2Se_2O use the corresponding DFT-derived shell mappings. Using the same Monte Carlo protocol as in the main text, the resulting heat-capacity maxima are 1100, 480, and 250 K for bcc Fe, FeSn, and KV_2Se_2O, respectively [Fig. 12(c–e)], compared with 1150, 480, and 210 K using the native Prophet-Spin exchange readout.

Finally, independent pressure-dependent DFT calculations provide a reference for the Fe magnetostructural transition. The hcp–bcc enthalpy crossing occurs at approximately 11.7 GPa [Fig. 12(f,g)], compared with 18.63 GPa predicted by Prophet-Spin. The lowest-energy hcp magnetic branch remains AFM over the displayed pressure range, with its local moment decreasing from 1.143 μ_B/Fe at 10 GPa to 0.854 μ_B/Fe at 24 GPa [Fig. 12(h)]. At higher pressure, the nonmagnetic hcp branch becomes energetically favored between 40 and 45 GPa. Thus, both DFT and Prophet-Spin show progressive suppression of magnetism in hcp Fe and an eventual AFM-to-NM transition, although the structural and magnetic crossover pressures differ quantitatively.

References

A. E. Austin, E. Adelson, and W. H. Cloud. Magnetic structures of Mn2As and Mn2Sb0.7As0.3. *Journal of Applied Physics*, 1962.

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.

Christopher J. Bartel, Amalie Trewartha, Qi Wang, Alexander Dunn, Anubhav Jain, and Gerbrand Ceder. A critical examination of compound stability predictions from machine-learned formation energies. *npj Computational Materials*, 6(1):97, 2020. doi: 10.1038/s41524-020-00362-y. URL https://doi.org/10.1038/s41524-020-00362-y.

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, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John

L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, 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.

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.

A. Beleanu, J. Kiss, G. Kreiner, C. Köhler, L. Müchler, W. Schnelle, U. Burkhardt, S. Chadov, S. Medvediev, D. Ebke, C. Felser, G. Cordier, B. Albert, A. Hoser, F. Bernardi, T. I. Larkin, D. Pröpper, A. V. Boris, and B. Keimer. Large resistivity change and phase transition in the antiferromagnetic semiconductors LiMnAs and LaOMnAs. *Physical Review B*, 88(18):184429, 2013. doi: 10.1103/PhysRevB.88.184429. URL https://doi.org/10.1103/PhysRevB.88.184429.

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.

Théo Cavignac, Jonathan Schmidt, Pierre-Paul De Breuck, Antoine Loew, Tiago F. T. Cerqueira, Hai-Chen Wang, Anton Bochkarev, Yury Lysogorskiy, Aldo H. Romero, Ralf Drautz, Silvana Botti, and Miguel A. L. Marques. AI-driven expansion and application of the alexandria database. *Journal of Physics: Materials*, 9(2):025014, 2026. doi: 10.1088/2515-7639/ae6620. URL https://doi.org/10.1088/2515-7639/ae6620.

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.

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.

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.

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.

Vincenzo Fiorentini and Michael Methfessel. Extracting convergent surface energies from slab calculations. *Journal of Physics: Condensed Matter*, 8(36):6525–6529, 1996. doi: 10.1088/0953-8984/8/36/005. URL https://doi.org/10.1088/0953-8984/8/36/005.

Christoph Freysoldt, Jörg Neugebauer, and Chris G. Van de Walle. First-principles calculations for point defects in solids. *Reviews of Modern Physics*, 86(1):253–305, 2014. doi: 10.1103/RevModPhys.86.253. URL https://doi.org/10.1103/RevModPhys.86.253.

Samuel V. Gallego, J. Manuel Perez-Mato, Luis Elcoro, Emre S. Tasci, Robert M. Hanson, Koichi Momma, Mois I. Aroyo, and Gotzon Madariaga. MAGNDATA: Towards a database of magnetic structures. I. the commensurate case. *Journal of Applied Crystallography*, 49(5):1750–1776, 2016. doi: 10.1107/S1600576716012863. URL https://doi.org/10.1107/S1600576716012863.

Jr. Graham, C. D. Magnetocrystalline anisotropy constants of iron at room temperature and below. *Physical Review*, 112(4):1117–1120, 1958. doi: 10.1103/PhysRev.112.1117. URL https://doi.org/10.1103/PhysRev.112.1117.

Cheuk Hin Ho, Cas van der Oord, James P. Darby, Theo Keane, Raz L. Benson, Cristian Rebolledo Espinoza, Rutvij Kulkarni, Elina Spinu, Michail Papanikolaou, Richard Tomsett, Robert M. Forrest, Jonathan J. Bean, Gábor Csányi, and Christoph Ortner. Equivariant many-body message passing interatomic potentials for magnetic materials. *arXiv preprint arXiv:2604.08143*, 2026. doi: 10.48550/arXiv.2604.08143. URL https://arxiv.org/abs/2604.08143.

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.

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.

Rostislav Hrubiak, Yue Meng, and Guoyin Shen. Thermodynamics and equations of state of iron to 350 gpa and 6000 k. *Scientific Reports*, 7:41863, 2017. doi: 10.1038/srep41863. URL https://doi.org/10.1038/srep41863.

Fumihiro Ishikawa, Keiichi Koyama, Kazuo Watanabe, and Hirofumi Wada. X-ray diffraction study of the structural phase transition in MnAs under high magnetic fields. *Physica B: Condensed Matter*, 346–347:408–412, 2004. doi: 10.1016/j.physb.2004.01.116. URL https://doi.org/10.1016/j.physb.2004.01.116.

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.

Kairos Materials. ELEMENTA: Toward unbiased scaling of materials data from first principles. Technical report, Kairos Materials, 2026. URL https://www.kairosmaterials.com/papers/ELEMENTA.pdf.

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.

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.

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.

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*, 2024. 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.

Pui-Wai Ma and S. L. Dudarev. Constrained density functional for noncollinear magnetism. *Physical Review B*, 91(5): 054420, 2015. doi: 10.1103/PhysRevB.91.054420.

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.

Sébastien Merkel, Ainhoa Lincot, and Sylvain Petitgirard. Microstructural effects and mechanism of bcc-hcp-bcc transformations in polycrystalline iron. *Physical Review B*, 102(10):104103, 2020. doi: 10.1103/PhysRevB.102.104103. URL https://doi.org/10.1103/PhysRevB.102.104103.

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.

OpenAI. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/, September 2024. Accessed 2026-07-19.

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.

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.

Balázs Póta, Paramvir Ahlawat, Gábor Csányi, and Michele Simoncelli. Thermal conductivity predictions with foundation atomistic models. *Nature Communications*, 2026. doi: 10.1038/s41467-026-76391-w. URL https://doi.org/10.1038/s41467-026-76391-w.

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

Brian C. Sales, Jiaqiang Yan, William R. Meier, Andrew D. Christianson, Satoshi Okamoto, and Michael A. McGuire. Electronic, magnetic, and thermodynamic properties of the kagome layer compound FeSn. *Physical Review Materials*, 3(11):114203, 2019. doi: 10.1103/PhysRevMaterials.3.114203. URL https://doi.org/10.1103/PhysRevMaterials.3.114203.

Jonathan Schmidt, Mário R. G. Marques, Silvana Botti, and Miguel A. L. Marques. A machine-learning-assisted materials discovery platform for the prediction of stable materials. *npj Computational Materials*, 8:134, 2022. doi: 10.1038/s41524-022-00820-5. URL https://doi.org/10.1038/s41524-022-00820-5.

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.

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.

Yili Sun, Yu Huang, Jingwen Cheng, Shanshan Zhang, Zezhong Li, Huiqian Luo, Xiaobai Ma, Wenyun Yang, Jinbo Yang, Dongfeng Chen, Kai Sun, Matthias Gutmann, Silvia C. Capelli, Feiran Shen, Jiazheng Hao, Lunhua He, Genfu Chen, and Shiliang Li. Antiferromagnetic structure of KV2Se2O: A neutron diffraction study. *Physical Review B*, 112(18):184416, 2025. doi: 10.1103/27n8-5q4l. URL https://doi.org/10.1103/27n8-5q4l.

Junfu Tan, Mengfan Wu, Yu Zhu, Yu Shi, Yifang Qin, Chang Liu, Bonan Zhu, and Ziheng Lu. Spectrally Decomposed Equivariant Graph Neural Networks for Interatomic Potentials. To be published.

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.

VASP Software GmbH. VASP wiki: Constraining local magnetic moments. https://www.vasp.at/wiki/index.php/Constraining_local_magnetic_moments, 2026. Accessed 2026-09-15.

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

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.

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.

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.

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.

Hongyu Yu, Boyu Liu, Yang Zhong, Liangliang Hong, Junyi Ji, Changsong Xu, Xingao Gong, and Hongjun Xiang. Physics-informed time-reversal equivariant neural network potential for magnetic materials. *Physical Review B*, 110 (10):104427, 2024a. doi: 10.1103/PhysRevB.110.104427. URL https://doi.org/10.1103/PhysRevB.110.104427.

Hongyu Yu, Yang Zhong, Liangliang Hong, Changsong Xu, Wei Ren, Xingao Gong, and Hongjun Xiang. Spin-dependent graph neural network potential for magnetic materials. *Physical Review B*, 109(14):144426, 2024b. doi: 10.1103/PhysRevB.109.144426. URL https://doi.org/10.1103/PhysRevB.109.144426.

Nikolai A. Zarkevich and Duane D. Johnson. Coexistence pressure for a martensitic transformation from theory and experiment: Revisiting the bcc–hcp transition of iron under pressure. *Physical Review B*, 91(17):174104, 2015. doi: 10.1103/PhysRevB.91.174104. URL https://doi.org/10.1103/PhysRevB.91.174104.