What is Your Force Field *Really* Learning? Gaining Scientific Intuition with A Dual-Level Explainability Framework

Yi Cao[1], Peter Mastracco[1], Jieneng Chen[2], Alan Yuille[2], Paulette Clancy[1]

[1] Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD 21218
[2] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218
ycao73@jh.edu, pmastra2@jh.edu, jchen293@jhu.edu, ayuille1@jhu.edu, pclancy3@jhu.edu

Abstract

Machine learning force fields (MLFFs) are emerging as a key player in accurate simulations of materials systems. Despite this, their internal logic remains rather opaque—a critical barrier to both trust and scientific discovery. We introduce **DUAL-X**, a *Dual-Level Explainability Framework* that bridges model reasoning with human understanding. **DUAL-X** unites two complementary perspectives: a model-centric level identifying which atoms and local environments in the structure the model should focuses its attention on, and a **human-centric** level revealing *what* physically meaningful interactions it prioritizes. Implemented with Grad-CAM for spatial attribution and SHAP-on-SOAP for physical interpretation, **DUAL-X** provides a general, human-in-the-loop paradigm for interpretable scientific AI.

Applied to dopant migration (Miskin et al. 2025) in a crystalline 2D material—a challenging task for MLFFs—**DUAL-X** reveals that different training strategies lead to models that capture different aspects of the underlying chemical physics. Multi-temperature fine-tuned MACE models exhibit over $10^2\times$ stronger selectivity for Cr–Cr f-type ($l = 3$) angular correlations than ones trained from scratch, emphasizing that complex and contextually relevant 3D coordination motifs are essential for accuracy. Models with sharper Grad-CAM focus also display coherent SHAP importance for dopant clustering, revealing consistent internal reasoning across scales.

By aligning model logic with human physical knowledge, **DUAL-X** transforms opaque predictors into interpretable scientific partners—advancing trustworthy, explainable, and insight-driven AI for materials discovery.

Code — https://github.com/yicao-elina/X-FORCE.git

Introduction

Machine learning force fields (MLFFs) promise to revolutionize computational materials science by achieving quantum-mechanical accuracy at a fraction of the cost of traditional simulations (Behler and Parrinello 2007; Batatia et al. 2022). However, as these models transition into studies using large, complex neural networks (Chen and Ong 2022), their inherently *black-box* nature imposes a critical barrier to scientific progress. This gap between predictive power and physical reasoning creates a *crisis of trust and discovery*: without understanding the physical principles underlying a prediction, researchers cannot trust MLFFs in high-stakes applications, diagnose their failures, or extract the latent scientific insights they may have learned (Rudin 2019). Indeed, while all MLFFs are sophisticated curve-fitting engines, their learned representations often encode non-trivial physical relationships. Uncovering these representations transforms MLFFs from mere interpolators into partners for scientific discovery. Thus, interpretability is not an auxiliary goal but a prerequisite for turning predictive accuracy into genuine understanding.

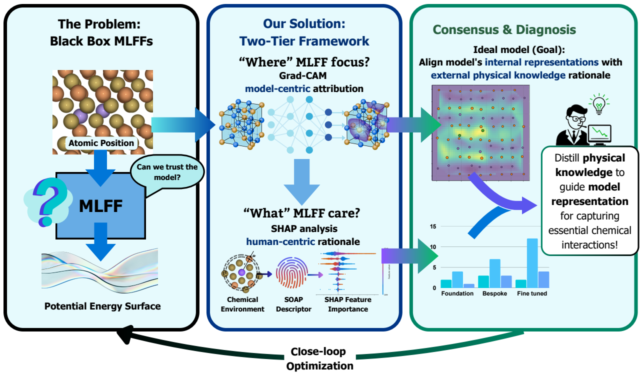

Figure 1: Our hierarchical XAI framework transforms opaque MLFFs into interpretable tools that reveal how training strategies shape chemical intuition.

The core challenge lies in reconciling two fundamentally distinct explanatory modes: **model-centric attribution** and **human-centric scientific rationale**. The former asks, *According to the model's learned representation, which atomic features drive a given prediction?* The latter asks, *Do these learned structure–property relationships align with known physical or chemical principles?* Existing explainable AI (XAI) approaches primarily address the first mode, offering *post-hoc* justifications without grounding them in scientific theory (Doshi-Velez and Kim 2017; Molnar 2020). Bridging this divide is essential for interpretable and trustworthy scientific ML.

To address this gap, we introduce **DUAL−X, a *Dual-Level Explainability Framework*** that unifies these two explanatory perspectives. **DUAL−X** provides a hierarchical, physically grounded understanding of MLFFs by sequentially answering two key questions:

1. **Model-Centric Locus:** Where in the atomic structure does the model focus its attention to make predictions? (The ”Where”)
2. **Human-Centric Mechanism:** What physically meaningful interactions within those regions does the model prioritize? (The ”What”)

In this work, we instantiate the **DUAL−X** hierarchy by synergistically combining two complementary techniques. To identify the *model-centric locus* (the “where”), we implement this level using a Gradient-weighted Class Activation Mapping (Grad-CAM) method, (Selvaraju et al. 2017) which back-propagates gradients to reveal which atoms learned the embeddings that most influence the final prediction. To uncover the *human-centric mechanism* (the “what”), we apply SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017) to a basis of Smooth Overlap of Atomic Positions (SOAP) descriptors. (Bartók, Kondor, and Csányi 2013) By analyzing SHAP values on these physical descriptors, we decompose the model's decision-making process into interpretable geometric and chemical components.

We demonstrate **DUAL−X** using a state-of-the-art MACE force field (Batatia et al. 2022) to study a common material problem, namely doping, which is technologically important in the semiconductor industry, for instance. Our analysis reveals how different training strategies instill distinct physical intuitions into the model, providing a route toward interpretable and scientifically grounded MLFF development.

Our main contributions are threefold:

- **Methodological:** We propose **DUAL−X**, the first two-tier XAI framework for MLFFs that explicitly bridges internal, model-centric attributions with external, human-centric scientific rationales.
- **Scientific:** Applied to a 2D material doping task (Cr-doped Sb_2Te_3), **DUAL−X** uncovers the physical mechanism learned by the MLFF, revealing that predictive accuracy arises from a spatial focus on Cr dopant clusters (the ”where”) and the prioritization of specific high-order angular correlations (the ”what”).
- **Practical:** We demonstrate that multi-temperature fine-tuning produces MACE-based MLFFs with more robust and physically selective representations than either a “vanilla” foundation model or training a model from scratch (completely user-defined), offering a principled approach to building reliable and interpretable models.

Related Work

The Accuracy–Interpretability Trade-off in MLFFs

So far, the evolution of MLFFs has been marked by a persistent tension between predictive accuracy and interpretability. Early architectures such as the Behler–Parrinello neural network used hand-crafted, physically motivated descriptors like Atom-Centered Symmetry Functions (ACSFs) and Smooth Overlap of Atomic Positions (SOAP) (Behler and Parrinello 2007; Bartók, Kondor, and Csányi 2013). While these fixed bases offered interpretability, they constrained model expressiveness, meaning the models were limited in their ability to represent complex, high-dimensional interactions beyond the predefined descriptor space. This restriction often prevented capturing subtle many-body or long-range effects that are crucial in chemically and structurally heterogeneous systems.

Subsequent end-to-end graph neural networks (GNNs), including SchNet and DimeNet++ (Schütt et al. 2017; Gasteiger et al. 2020), shifted this paradigm by learning representations directly from atomic coordinates. This trend culminated in equivariant architectures such as MACE (Batatia et al. 2022), which constructs high-order tensor features respecting physical symmetries and achieve unprecedented accuracy. However, these advances come at the cost of transparency: learned representations are no longer tied to human-defined physical quantities but emerge from optimization, making them difficult to interpret. Resolving this accuracy–interpretability trade-off is the central motivation of our work. To contextualize this challenge, Table 2 provides a comparative overview of existing explainable AI approaches for MLFFs, highlighting both their methodological advances and inherent limitations across interpretability paradigms.

Paradigms in Scientific Interpretability

Interpretability methods in scientific ML have evolved along three major paradigms, corresponding to model-centric, human-centric, and intrinsic interpretability.

Human-Centric Explanations via Feature Attribution. This approach explains model outputs in terms of predefined, physically meaningful features. Methods such as LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) assign importance scores to input descriptors, allowing mechanistic interpretation. In materials science, SHAP has been used to identify key bond lengths, angles, or solvent effects governing predictions (Wang and Chen 2024; Schütt et al. 2019). However, these approaches rely on fixed, human-defined bases and assume feature independence—limitations that make them ill-suited for explaining correlated, emergent features in modern end-to-end MLFFs.

Model-Centric Explanations via Spatial Attribution. This paradigm probes internal representations directly, identifying which input regions most influence predictions. Gradient-based methods such as Saliency Maps (Simonyan, Vedaldi, and Zisserman 2013), Grad-CAM (Selvaraju et al. 2017), and GNN-specific tools like GNNExplainer (Ying et al. 2019) reveal the model's spatial “attention.” More recent variants, such as EquiGX, adapt these techniques to tensor operations in equivariant GNNs. While such methods indicate *where* the model focuses, they cannot alone explain *what* physical principles drive those attentions.

Intrinsic Interpretability via Model Design. An alternative approach builds transparency into the model itself. Symbolic regression, for example, can distill GNN knowledge into compact analytical expressions. Architectures like SchNet4AIM directly predict physically rigorous quantities (*e.g.,* atomic charges, delocalization indices), yielding inherently interpretable outputs (Schütt et al. 2017). These methods, however, often require new architectures and may sacrifice peak accuracy.

Our Contribution in Context

Prior work has largely treated these paradigms as mutually exclusive: researchers were forced to choose between human-centric explanations (the "what") for simpler models, model-centric attributions (the "where") for complex ones, or intrinsically interpretable but less-performant architectures.

Our work, **DUAL-X**, resolves this false dichotomy, in terms of a hierarchical framework that systematically unifies these perspectives. It first uses a model-centric approach to identify the *locus of prediction*—the specific atoms or regions that the model deems critical (the "where"). It then employs a human-centric approach to decode the *physical mechanism*—the interpretable chemical or geometric features the model prioritizes within that high-attention locus (the "what").

In this paper, we implement **DUAL-X** by synthesizing Grad-CAM for spatial attribution and SHAP-on-SOAP for mechanistic decomposition. This synthesis leverages the strengths of both paradigms, transforming state-of-the-art MLFFs from opaque predictors into transparent and trustworthy tools for scientific discovery.

A Hierarchical Framework for MLFFs Interpretability

To bridge the gap between the predictive accuracy of modern MLFFs and their physical interpretability, we introduce **DUAL-X**—a dual-level hierarchical framework designed to translate model decisions into physically meaningful explanations. It first locates *where* in an atomic structure the model focuses its attention during prediction (Level 1), and then decomposes *what* physical interactions within those regions govern this focus (Level 2). Figure 1 schematically summarizes this "Where–then–What" interpretability pipeline.

Level 1: Spatial Attribution via Grad-CAM ("The Where")

The first level identifies which atoms are most influential to the MLFF's prediction of the total energy. We adapt Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al. 2017), originally developed for convolutional neural networks, to the setting of equivariant graph neural networks (GNNs) used in atomistic modeling.

Target Model and Embedding Extraction. Our analysis employs the MACE architecture (Batatia et al. 2022), a state-of-the-art equivariant GNN that encodes many-body interactions through tensorial message passing. After L interaction blocks, MACE produces a set of atomic embeddings

$$\{\mathbf{A}_i^{(L)} \in \mathbb{R}^K\}_{i=1}^{N},$$

for a system of N atoms, where K denotes the embedding dimensionality. These vectors encapsulate high-order geometric and chemical correlations within each atom's local environment.

To probe spatial importance, we attach `PyTorch` hooks to the final embedding layer—the most comprehensive environmental representation before aggregation into the total energy E_{total}. A forward pass yields both E_{total} and $\{\mathbf{A}_i^{(L)}\}$.

Grad-CAM Attribution. We compute the gradient of E_{total} with respect to each embedding component:

$$\frac{\partial E_{\text{total}}}{\partial A_{ik}^{(L)}},$$

where $A_{ik}^{(L)}$ denotes the k-th channel of atom i's embedding. The Grad-CAM importance α_i for atom i is then defined as

$$\alpha_i = \mathrm{ReLU}\left(\sum_{k=1}^{K} \frac{\partial E_{\text{total}}}{\partial A_{ik}^{(L)}} A_{ik}^{(L)}\right).$$

A larger α_i indicates that atom i's local environment exhibits both strong activation and high gradient sensitivity—implying a strong positive contribution to the predicted energy. The resulting set $\{\alpha_i\}$ forms a spatial importance map highlighting the structural regions that the model prioritizes when evaluating the total energy. This "attention map" constitutes the **model-centric perspective** of **DUAL-X**, addressing the question: *"Where in the atomic structure does the model focus to make its prediction?"*

Level 2: Mechanistic Decomposition via Surrogate Modeling and SHAP ("The What")

Having localized the atoms most influential to the model's decision, the second level of **DUAL-X** reveals *why* these regions matter by identifying the underlying physical interactions that govern prediction quality. Directly applying SHAP to large GNNs like MACE is computationally prohibitive; we therefore employ a **surrogate modeling** approach that maps human-defined physical descriptors to the model's predictive behavior.

Descriptor Construction. To capture global geometric signatures, each atomic configuration is represented by a rotationally invariant Smooth Overlap of Atomic Positions (SOAP) descriptor (Bartók, Kondor, and Csányi 2013). While Level 1 focuses on localized spatial saliency, Level 2 abstracts to global structural motifs, providing complementary insight into the physical origins of the model's focus. SOAP vectors are computed using the `dscribe` package (Himanen et al. 2020) with the `average="inner"` setting to obtain global power spectra:

$$\mathbf{s} = \{s_{nl}\}_{n,l},$$

where the radial index n and angular index l define the resolution of structural features, with larger l values capturing higher-order many-body correlations.

Surrogate Model Training. We treat the pre-trained MACE model as an oracle to generate labeled data. For each configuration, we compute its SOAP descriptor $\mathbf{s}$ and the absolute energy prediction error

$$|\Delta E| = |E_{\text{MACE}} - E_{\text{DFT}}|.$$

We then train a `GradientBoostingRegressor` (Friedman 2001) to learn the mapping

$$f_{\text{sur}} : \mathbf{s} \rightarrow |\Delta E|,$$

thereby modeling how structural motifs influence the reliability of MACE predictions. This reframes interpretability from "*What features predict energy?*" to the more actionable "*What structural motifs predict model failure?*"

SHAP-Based Mechanistic Analysis. Once trained, the surrogate model is analyzed using `TreeExplainer` (Lundberg and Lee 2017), which computes exact SHAP values for tree-based models. For each SOAP feature s_{nl}, the corresponding SHAP value ϕ_{nl} quantifies its marginal contribution to $f_{\text{sur}}(\mathbf{s})$, satisfying local accuracy and consistency. Aggregating ϕ_{nl} across structures yields a ranked spectrum of geometric and chemical importance, revealing which motifs are most associated with model unreliability.

This mechanistic decomposition of model *error* complements Level 1's focus analysis, providing a human-centric perspective on *what* physical interactions the model has implicitly learned. By jointly interpreting the Grad-CAM spatial maps and SHAP-based feature spectra, **DUAL-X** establishes a coherent bridge between model logic and physical intuition—turning black-box MLFFs into interpretable scientific instruments.

Synthetic Experimental Setup

System and *Ab Initio* Dataset

We investigate Cr-doped Sb_2Te_3, a prototypical layered phase-change material widely studied for topological and thermoelectric applications. Reference data were generated using Density Functional Theory (DFT) with the Perdew–Burke–Ernzerhof (PBE) functional and Grimme's D3 dispersion correction, as implemented in QUANTUM ESPRESSO (Giannozzi et al. 2009; Perdew, Burke, and Ernzerhof 1996). A total of ∼20,000 atomic configurations were sampled from *ab initio* molecular dynamics (AIMD) trajectories of a 120-atom, two-quintuple-layer supercell containing varying Cr dopant concentrations. Simulations were performed at 300 K, 600 K, and 1200 K to ensure broad thermal diversity across the potential energy surface. The dataset was randomly divided into training (80%), validation (10%), and test (10%) subsets. Detailed computational parameters and AIMD protocols are provided in Appendix .

MLFFs Training Protocols

MLFFs were trained using the MACE architecture (Batatia et al. 2022) with a 6.0 Å cutoff radius. To probe how pre-training and data diversity influence learned representations, we compared four models:

1. **Scratch:** A MACE-trained model from a random initialization using the full AIMD dataset.
2. **Foundation:** A pre-trained MACE-MP-0 model (Batatia et al. 2022) evaluated in a zero-shot setting without fine-tuning (constituting an as-is "vanilla" model).
3. **FT-600K:** A foundation model that has been fine-tuned on 600 K AIMD data to assess single-temperature transferability.
4. **FT-MultiT:** A foundation model that has been fine-tuned on combined data from three different temperatures, 300–1200 K, of AIMD data to encourage temperature-aware generalization.

All models employed a dual-branch output head for simultaneous energy and force prediction. Hyperparameters and training details are summarized in Appendix.

Hierarchical Interpretability Protocol

To evaluate model explainability, we applied the proposed **DUAL-X** (Dual-Level Explainability) framework to 200 representative test configurations spanning different dopant concentrations, structural motifs, and thermal conditions.

Level 1: "Where" — Spatial Attribution Grad-CAM was used to compute atom-wise importance scores (α_i), indicating which atomic regions most strongly influenced the energy predicted by the MLFF. We further analyzed the chemical identity and coordination environment of the top-10 most influential atoms across test samples to characterize each model's attention patterns.

Level 2: "What" — Mechanistic Decomposition Following spatial attribution, a SHAP analysis was performed to reveal which geometric motifs drive model prediction errors defined as $\Delta E = E_{\text{MACE}} - E_{\text{DFT}}$. For each model variant, a Gradient Boosting Regressor was trained to approximate ΔE using global SOAP descriptors ($n_{\max} = 4$, $l_{\max} = 4$). Feature-wise SHAP values quantified each descriptor's marginal contribution to the surrogate model's prediction, thereby linking Grad-CAM-derived spatial importance ("Where") to descriptor-level physical mechanisms ("What").

Together, these two interpretability levels establish a unified mechanistic understanding of MLFF behavior.

Results and Discussion

We structure our results to follow the **DUAL-X** framework: We first apply Level 1 to identify the *model-centric locus* ("where" the models focus) and then apply Level 2 to uncover the *human-centric mechanism* ("what" they have learned).

Level 1: Model-Centric Locus of Prediction and Architectural Paradigms

Applying the Level 1 (Grad-CAM) analysis to our test set immediately reveals that the different training protocols (Scratch, Foundation, FT-MultiT) instill fundamentally distinct spatial attention mechanisms for processing chemical

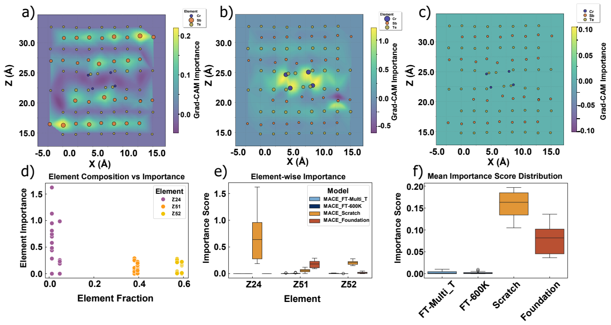

Figure 2: **Grad-CAM visualization reveals model-specific attention mechanisms.** (a–c) Spatial activation maps for DFT-relaxed Cr-doped Sb_2Te_3: (a) The foundation model exhibits diffuse Sb-centered attention; (b) The scratch model shows sharply localized Cr-centric responses; (c) The fine-tuned Multi-T model displays minimal activation, indicating internalized equilibrium knowledge. (d–f) Element-wise analyses demonstrate distinct chemical selectivity as follows: Sb-dominant behavior (Foundation), Cr-dominant behavior (Scratch), and a more balanced behavior (Fine-tuned)—reflecting progressive specialization across learning stages.

information (Figure 2). When analyzing DFT-relaxed equilibrium structures, these differences are particularly stark.

The vanilla, as parameterized, **Foundation** model exhibits diffuse, low-intensity activation (max score: 0.20) with a clear preference for giving its attention to Sb atoms in the host matrix (see Figure 2a). This pattern suggests that it predicts the system's energy by integrating collective matrix behavior, treating Cr dopants as minor perturbations. In contrast, the **Scratch** model, with its training focused on the dopant movement, displays intense, spatially-localized activation (max score: 1.00) that sharply targets Cr dopants and their immediate coordination environments (see Figure 2b). This dopant-centric focus indicates it has learned these specific sites are critical decision points for energy prediction. Most remarkably, the **FT-MultiT** model shows virtually zero activation on the same equilibrium structures (Figure 2c), implying it has successfully internalized these stable patterns and requires minimal feature extraction—an optimal state for well-characterized regimes.

This divergence in spatial attention is mirrored by element-specific chemical intelligence (Figure 2e). The Scratch model's pronounced focus on Cr-selectivity (mean importance: 0.985) and the Foundation model's strong bias towards Sb atoms (0.110) confirm their respective dopant-centric versus matrix-dominated learning dynamics. The overall importance score hierarchy—Scratch (0.185) > Foundation (0.045) > FT-MultiT (0.001) (Figure 2f)—directly correlates with the degree of model specialization.

While the fine-tuned models appear dormant on equilibrium structures, they correctly reactivate when processing high-energy, non-equilibrium configurations from AIMD trajectories (SI Figure 6). This demonstrates a sophisticated, selective processing ability to distinguish stable states from challenging dynamic regimes. However, this specialization is not without risk; we also observe concerning activations on seemingly random Sb atoms in distorted structures, suggesting possible overfitting to training-specific artifacts. This phenomenon aligns with previously reported "catastrophic forgetting" in naively fine-tuned MACE models, wherein excessive specialization can degrade generalizability. (Luo et al. 2023; Kirkpatrick et al. 2017)

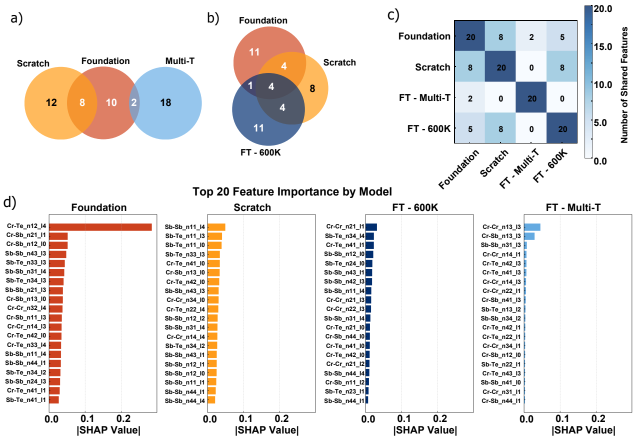

Figure 3: **SHAP analysis reveals distinct feature importance hierarchies across MLFFs architectures.** (a–b) Venn diagrams and (c) overlap matrix show limited shared top features among models. (d) Feature rankings highlight divergent chemical focus: Foundation (Cr–Te, Sb–Te), Scratch (Sb–Sb), FT-600K (Cr–Sb mix), and FT-Multi-T (Cr-dominant).

This Grad-CAM analysis successfully answers the first question of our framework: *where* the models focus. We have established three distinct processing paradigms: **Distributed** (Foundation), **Localized** (Scratch), and **Selective** (Fine-tuned). We know, for instance, that the specialized models learn that the Cr-dopant environment is the critical locus of prediction, while the generalist Foundation model relies on the Sb-Te matrix. However, this spatial map alone does not explain *what* physical or chemical information the models extract from these regions to make their predictions. To bridge the gap from model-centric attribution to scientific rationale, we must now apply Level 2 of our framework. We proceed with a SHAP-based analysis to decompose the models' decisions into physically meaningful geometric and chemical components, thereby uncovering *what* specific interactions the models have learned to prioritize.

Level 2: The Human-Centric Mechanism Reveals Divergent Physics

Having established the *model-centric locus* (the "where") with Level 1, we now apply Level 2 of the **DUAL-X** framework to uncover the *human-centric mechanism* (the "what"). This analysis immediately reveals a paradox: While the fine-tuned and scratch models achieve similar predictive accuracy (MAE: 0.5–1.4 eV) (Table 3), they rely on fundamentally different geometric feature hierarchies for decision-making. This exposes the profound limitations of using performance metrics, like MAE, alone to assess model quality.

This mechanistic divergence is illustrated by the minimal consensus among models regarding critical geometric features (Figure 3a,b). In a three-way comparison of the Foundation, Scratch, and FT-600K models, only four features are shared among all variants, representing just 20% of each model's top-20 features. The pairwise overlap matrix (Figure 3c) quantifies this divergence, revealing that the FT-Multi-T model shares zero top features with the Scratch model and only two with the Foundation model. Given their similar predictive accuracy, this lack of consensus is remarkable and unexpected. The FT-Multi-T model exhibits 18 unique features (90% of its top-20) not prioritized by other variants, proving that multi-temperature training fundamentally reshapes the model's geometric feature space rather than simply optimizing it.

Analysis of the top-ranked features (Figure 3d) reveals the distinct "chemical intelligence" patterns that explain this divergence:

- **Foundation Model:** This model exhibits mixed chemical priorities, emphasizing both dopant-host interactions (*e.g.*, `Cr-Te_n12_l4`) and matrix correlations (*e.g.,* `Sb-Te_n43_l3`). Its diverse angular complexities ($l =$ 1–4) reflect the broad structural sensitivity inherited from its generalist pre-training.
- **Scratch Model:** Demonstrates a pronounced matrix-centric bias, with 12 of its top 20 features involving Sb-Sb interactions. Its focus on high-order host-host motifs (e.g., `Sb-Sb_n11_l4`) means it achieves accuracy while largely missing the critical dopant-specific physics.
- **FT-Multi-T Model:** This model exhibits the most chemically sophisticated hierarchy, with 13 of its top 20 features involving Cr interactions. It establishes a systematic, dopant-centric framework absent in the others, prioritizing: (1) Cr-Cr clustering (*e.g.,* `Cr-Cr_n13_l3`, `Cr-Cr_n31_l1`) to capture direct dopant correlations; (2) Cr-matrix coupling (*e.g.,* `Cr-Sb_n44_l1`, `Cr-Te_n43_l3`) to encode dopant-host perturbations; and (3) Extended Cr networks (*e.g.*, `Cr-Cr_n34_l1`).

The emergent hierarchy in the FT-Multi-T model is the key scientific finding. Its highest-importance feature, `Cr-Cr_n13_l3`, represents f-type angular correlations corresponding to direct Cr-Cr bonding and next-nearest Cr interactions, capturing the extended dopant network formation critical for phase-switching behavior (Wang et al. 2016). (Here, "f-type" strictly refers to the geometric angular descriptor and does not imply the presence of f-electrons on Cr, which has no occupied f-orbitals.) This chemical narrative, which progresses from local dopant clustering to global Cr-matrix coupling, reflects a deep understanding of the hierarchical mechanism for the dopant interaction within 2D material interface. This entire physical picture is completely absent in the Scratch model.

This cross-model analysis identifies two distinct categories of features: Stable Universal Features (*e.g.*, fundamental `Sb-Te_n34_l2` matrix interactions) that capture essential bonding patterns, and Training-Specific Specialized Features (see Figure 8). The Multi-T model's unique emphasis on high-order `Cr-Cr` correlations represents the latter (namely, TSS features): It demonstrates a learned sophistication, enabled by diverse thermal sampling, that single-temperature models fail to develop.

This mechanistic divergence proves that **traditional accuracy metrics, such as RMSE, are insufficient for evaluating MLFF quality**. Our **`DUAL-X`** framework demonstrates that models can achieve correct predictions while learning fundamentally different physics. This has critical implications for reliability and transferability, as models reasoning differently will likely fail differently under out-of-distribution conditions. Multi-temperature training thus emerges as an essential strategy for developing chemically meaningful and physically robust representations in complex doped systems.

`DUAL-X` Deep Dive: Cr-Cr f-type Correlations as Emergent Chemical "Intelligence"

Our Level 2 (Human-Centric Mechanism) analysis established that the FT-Multi-T model develops a unique, dopant-centric feature hierarchy. We now perform a "deep dive" on its single most dominant descriptor: the **`Cr-Cr_n13_l3`** feature. This feature, which represents f-type angular correlations ($l = 3$) between chromium atoms, exemplifies how the **`DUAL-X`** framework can uncover sophisticated, emergent chemical intelligence that captures the essential physics of dopant-mediated phase transitions.

An analysis of this feature's physical basis confirms its chemical significance. SOAP field strength visualizations (Figure 3b–c) demonstrate pronounced spatial localization, with high-intensity regions across representative z-slices concentrated exclusively around the Cr dopant sites (marked by white circles). This spatial selectivity is not an artifact; 3D structural analysis (Figure 3d) reveals the feature's high importance is driven by a dominant Cr-Cr interaction at **4.34 Å**. This geometrically significant distance corresponds to dopants in adjacent octahedral sites, the critical separation wherein Cr atoms begin to form the extended networks that facilitate collective phase-switching behavior. The feature's f-type angular character ($l = 3$) is thus essential for encoding the complex directional correlations that distinguish random dopant distributions from these ordered, functional clustering patterns.

The emergence of this feature's dominance is entirely dependent on the training strategy. We observe a dramatic 170-fold enhancement in `Cr-Cr_n13_l3` importance in the FT-Multi-T model (0.0332) compared to the Scratch model (0.0002) (Figure 3e). This enhancement is not a simple byproduct of performance tuning; it signifies a qualitative shift in the model's internal representation. The Scratch model, lacking this focus, develops a "democratic" attention allocation, leading to a flat feature importance distribution (Figure 3a) that prioritizes common matrix (Sb-Sb) interactions but fails to isolate the critical dopant physics.

In contrast, multi-temperature fine-tuning fundamentally reshapes this attention landscape. By exposing the model to diverse thermal conditions, it learns to distinguish temperature-dependent matrix fluctuations (noise) from temperature-invariant dopant-dopant correlations (signal). This selective pressure drives the Cr−Cr_n13_l3 feature to hierarchical dominance—exceeding the second-ranked feature by a factor of 2.1—as the model correctly identifies this specific, complex correlation as the most robust predictor of system energy across all thermal regimes.

Ultimately, the identification of Cr−Cr_n13_l3 as the key chemical descriptor provides the actionable, **human-centric insight** that the **DUAL−X** framework is designed to find. The 4.34 Å interaction distance suggests that optimal material performance requires controlled dopant spacing within this critical range. The f-type angular dependence further indicates that not just distance but geometric arrangement is critical. This analysis bridges the gap between the black-box model and physical intuition, demonstrating how our Level 2 analysis extracts chemically meaningful design rules and provides a quantitative metric for assessing the "chemical sophistication" of a given training strategy.

Angular Momentum (l) Patterns Reveal Systematic Geometric Preferences

The dominance of the specific $l = 3$ Cr−Cr feature, detailed in the previous section, is not an isolated artifact. Rather, it indicates a systematic geometric preference learned by the FT-Multi-T model. Broadening our Level 2 (Human-Centric Mechanism) analysis to the full angular momentum (l) distribution across all top-ranked SHAP features provides crucial insights into each model's geometric sophistication (Figure 4).

The FT-Multi-T model exhibits a pronounced **bimodal angular preference**, with 30% of its key features utilizing $l = 1$ (p-type) correlations and another 30% utilizing $l = 3$ (f-type) correlations. This selective emphasis on directional ($l = 1$) and complex octahedral ($l = 3$) geometries shows the model has learned that the phase-change behavior depends on specific coordination environments, not merely on simple radial distances (which would correspond to $l = 0$).

Conversely, the Scratch model shows a more uniform angular distribution, with a significant reliance on $l = 4$ correlations (30% of features). This suggests a dependency on high-order mathematical correlations that may not map directly to physically meaningful bonding patterns. The Foundation model demonstrates intermediate behavior (25% $l = 1$, 20% $l = 3$, 25% $l = 4$), reflecting its generalist pre-training on diverse materials.

The scientific interpretation of these patterns, uncovered by our **DUAL−X** framework, is clear: The FT-Multi-T model's $l = 1/l = 3$ bimodal preference directly corresponds to the known crystallography of Cr-doped Sb_2Te_3. The $l = 1$ correlations capture directional Cr-Sb bonding at heterointerfaces, while the $l = 3$ correlations encode the octahedral coordination environments that Cr atoms adopt when substituting into the layered chalcogenide structure. This geometric selectivity represents a form of learned chemical knowledge, confirming that the model, guided by multi-temperature training, discovered the specific coordination symmetries most predictive of this system's functional properties.

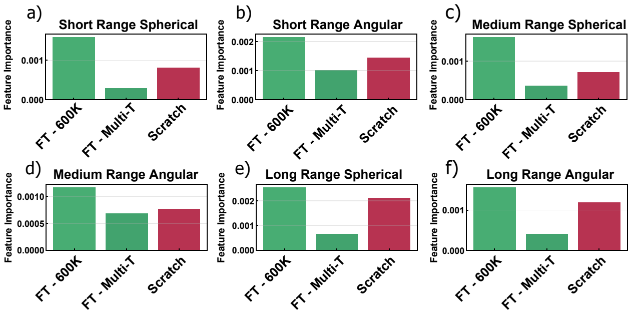

Figure 4: **Angular momentum feature distribution.** Multi-T model shows strong preference for f-type ($l = 3$) correlations, while Scratch model exhibits uniform distribution across all angular channels.

DUAL−X Synthesis and Implications for MLFF Development

Our **DUAL−X** framework has sequentially answered the *model-centric locus* (the "where") and the *human-centric mechanism* (the "what"). We now integrate these two perspectives to construct a unified mechanistic picture. This synthesis reveals that the *locus* and the *mechanism* are deeply intertwined, providing a powerful lens through which to evaluate model reliability, guide development, and understand the fundamental trade-offs between pre-training and task-specific fine-tuning.

A Unified Picture: Pre-training Bias vs. Task-Specific Specialization

The divergence in learned physics begins with the locus of prediction. The Foundation model's distributed attention on the Sb-Te matrix (the "where") is a direct consequence of its pre-training heritage on diverse chalcogenides where such bonds dominate energetics. This creates a systematic bias toward a generalist, matrix-dominated strategy. Consequently, its mechanistic priorities (the "what") are a diverse ensemble of SOAP features reflecting broad structural correlations. This approach is thermodynamically sound for predicting a collective property like total energy, but it is mechanistically incomplete for understanding dopant-driven phenomena, treating the critical Cr atoms as mere perturbations.

Conversely, the FT-Multi-T model, fine-tuned specifically on the Cr-doped system, learns a specialized, dopant-centric strategy. Its attention is sharply localized on Cr atoms (the "where"), correctly identifying them as the critical decision points. This spatial focus enables it to discover and prioritize the specific, high-order Cr−Cr_n13_l3 feature (the "what") that governs the material's phase-change behavior. The model moves beyond general thermodynamic intuition and learns the precise kinetic pathways driven by dopant clustering.

The Right Tool for the Scientific Question: Task-Dependent Reliability

This divergence in learned physics implies that no single model is universally optimal; its suitability depends critically on the scientific question being asked.

- For **Total Energy Prediction**, a property governed by collective behavior, the Foundation model's distributed processing is superior. By integrating contributions from the entire matrix, it captures the bulk energetics more robustly, even if it misses the specific dopant chemistry.
- For **Migration Energy Barriers**, a phenomenon driven by local atomic events, the scratch model's localized attention is essential. Migration kinetics are determined by subtle changes in the Cr coordination environment during transition states. The model's focused "where" and specific "what" allow it to accurately capture these local perturbations, while the Foundation model's distributed approach averages out the very signals that define the energy barrier. (See the benchmark of the migration energy of different models in our previous paper (Cao and Clancy 2025b,a)

This highlights a fundamental principle for scientific machine learning: the optimal model architecture depends not just on the chemical system but on the specific physical phenomenon under investigation.

Actionable Insights for Principled Model Development

This **`DUAL-X`** synthesis provides a clear, actionable framework for the development and assessment of MLFFs, moving beyond opaque accuracy metrics:

1. **Evaluating Model Convergence and Sophistication.** The evolution from a distributed "where" and mixed "what" (Foundation/Scratch) to a localized "where" and specific "what" (FT-Multi-T) serves as a diagnostic tool. The emergence of zero-importance scores on stable structures, combined with a sharp, physically meaningful SHAP hierarchy, signals that a model has achieved a state of efficient, specialized chemical intelligence.
2. **Guiding Principled Active Learning.** Our framework provides a blueprint for targeted data collection. Configurations with high Grad-CAM importance identify regions of the potential energy surface where the model is uncertain. SHAP analysis can then reveal which specific chemical motifs within those regions are under-represented, allowing for the targeted generation of new training data to improve model generalization and robustness.
3. **Predicting Generalizability and Failure Modes.** The attention profile is a strong predictor of transferability. Models showing extreme selectivity (like FT-Multi-T) are likely to excel at their specific task but may fail on novel configurations outside their learned dopant chemistry. Conversely, the Foundation model's distributed sensitivity suggests better (though less precise) generalization. This understanding allows for a more informed selection of models for deployment in new chemical spaces.

In conclusion, by developing the **`DUAL-X`** framework, we have moved beyond treating MLFFs as black boxes. Our framework establishes an interpretable methodology to validate that a model has not only found the right answer but has done so for the right scientific reasons, paving the way for the development of truly reliable and physically-grounded models for materials discovery.

Conclusions and Future Work

In this work, we introduced **`DUAL-X`**, a dual-level explainability framework, and demonstrated that predictive accuracy is a poor proxy for physical fidelity. By hierarchically bridging the *model-centric locus* ("where" models focus) with the *human-centric mechanism* ("what" they learn), we established that training strategy is the critical determinant of a model's learned chemical intuition.

Our work provides three key contributions:

1. **`DUAL-X` reveals how training strategy shapes intuition.** We demonstrate that multi-temperature fine-tuning on foundation models cultivates a specialized, dopant-centric focus (the "where"), which is absent in scratch-trained models.
2. **`DUAL-X` uncovers the learned physical mechanisms.** By decoding the human-centric "what," our framework shows that the model's predictive accuracy is governed by high-order f-type ($l = 3$) angular correlations, not simpler geometric descriptors.
3. **`DUAL-X` provides actionable scientific insights.** **`DUAL-X`** identifies the specific `Cr-Cr_n13_13` interaction as the dominant predictor, extracting actionable chemical knowledge that traditional accuracy metrics completely obscure.

While this work establishes the **`DUAL-X`** paradigm, its current implementation is limited to a single material system. Future work should therefore extend the framework to other complex material classes (*e.g.,* catalysts, battery electrodes) and integrate alternative XAI techniques for the locus and mechanism levels, building a more robust multi-method consensus. This reinforces the generalizability of **`DUAL-X`**, as its core philosophy is independent of the specific tools (like Grad-CAM or SHAP). Most importantly, the mechanistic insights **`DUAL-X`** provides can guide targeted synthesis and experimental validation, transforming MLFFs from "black-box" predictors into active tools for *in silico* knowledge discovery and establishing a principled path toward automated materials discovery.

Acknowledgments

This work was supported by the United States Department of Defense-funded Center of Excellence for Advanced Electrophotonics with 2D materials—Morgan State University, under Grant No. W911NF2120213. The authors thank Frank Gardea and Owen Vail, the cooperative agreement managers of the Center, for their interest and support. They also thank Dr. Ramesh Budhani for his guidance on the project. Yi thanks Johns Hopkins University for their support in AY 24–25. Computational resource support was provided by

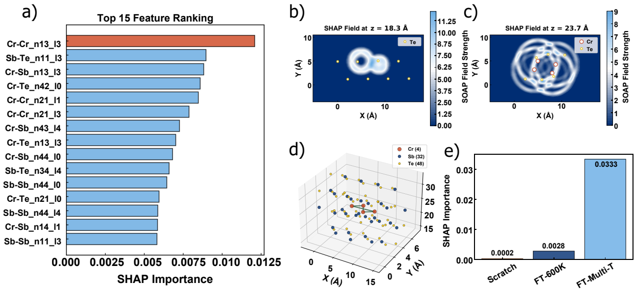

Figure 5: **Dominance of Cr–Cr f-type angular correlations revealed by SHAP analysis in Cr-doped SbTe$_3$.** **(a)** Feature ranking showing Cr–Cr_n13_l3 as the most important descriptor among 50 candidates. **(b–c)** SOAP field strength heatmaps at representative z-slices showing localized high-intensity regions around Cr dopants (white circles). **(d)** 3D structure with Cr dopants (red spheres) connected by the critical 4.34 Å Cr–Cr interaction. **(e)** Training strategy comparison, showing 170-fold enhancement in Cr–Cr_n13_l3 importance under multi-temperature fine-tuning (0.0332) vs. scratch training (0.0002).

the petascale Hopkins facility, Advanced Research Computing at Hopkins (ARCH) (rockfish.jhu.edu), supported by National Science Foundation award OAC 1920103. Partial funding for ARCH's infrastructure was originally provided by the State of Maryland.

We thank the anonymous reviewers for their constructive feedback.

References

Bartók, A. P.; Kondor, R.; and Csányi, G. 2013. On representing chemical environments. *Physical Review B—Condensed Matter and Materials Physics*, 87(18): 184115.

Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; and Csányi, G. 2022. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35: 11423–11436.

Behler, J.; and Parrinello, M. 2007. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14): 146401.

Cao, Y.; and Clancy, P. 2025a. Atomic Whispers to System Health Diagnosis and Prognosis: First-Principles-Based Degradation Modeling of 2D Materials in Next-Generation Bioelectronics. In *Annual Conference of the PHM Society*, volume 17.

Cao, Y.; and Clancy, P. 2025b. Migration as a Probe: A Generalizable Benchmark Framework for Specialist vs. Generalist Machine-Learned Force Fields. In *AI for Accelerated Materials Design - NeurIPS 2025*.

Chen, C.; and Ong, S. P. 2022. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11): 718–728.

Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Gasteiger, J.; Giri, S.; Margraf, J. T.; and Günnemann, S. 2020. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*.

Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; et al. 2009. QUANTUM ESPRESSO: a modular and open-source software project for quantumsimulations of materials. *Journal of physics: Condensed matter*, 21(39): 395502.

Grimme, S.; Antony, J.; Ehrlich, S.; and Krieg, H. 2010. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of chemical physics*, 132(15).

Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; and Foster, A. S. 2020. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247: 106949.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Luo, Y.; Yang, Z.; Meng, F.; Li, Y.; Zhou, J.; and Zhang, Y. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2023. *URL https://arxiv. org/abs/2308.08747*, 2308: 60.

Miskin, K.; Cao, Y.; Marland, M.; Shaikh, F.; Moore, D. T.; Marohn, J. A.; and Clancy, P. 2025. Low-energy pathways lead to self-healing defects in CsPbBr 3. *Physical Chemistry Chemical Physics*, 27(29): 15446–15459.

Molnar, C. 2020. *Interpretable machine learning*. Lulu. com.

Perdew, J. P.; Burke, K.; and Ernzerhof, M. 1996. Generalized gradient approximation made simple. *Physical review letters*, 77(18): 3865.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Schlipf, M.; and Gygi, F. 2015. Optimization algorithm for the generation of ONCV pseudopotentials. *Computer Physics Communications*, 196: 36–44.

Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; and Müller, K.-R. 2017. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30.

Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; and Maurer, R. J. 2019. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature communications*, 10(1): 5024.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Wang, Q.; Jiang, M.; Liu, B.; Wang, Y.; Zheng, Y.; Song, S.; Wu, Y.; Song, Z.; and Feng, S. 2016. Reversible phase change characteristics of Cr-Doped Sb2Te3 films with different initial states induced by femtosecond pulses. *ACS Applied Materials & Interfaces*, 8(32): 20885–20893.

Wang, S.; and Chen, Y. 2024. Improved yield prediction and failure analysis in semiconductor manufacturing with xgboost and shapley additive explanations models. In *2024 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*, 01–08. IEEE.

Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; and Leskovec, J. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32.

Computational and Methodological Details

Ab Initio Simulation Details

All reference data were generated using DFT as implemented in QUANTUM ESPRESSO (v6.8) (Giannozzi et al. 2009). The key computational settings are summarized below:

- **Exchange–Correlation Functional:** Perdew–Burke–Ernzerhof (PBE) (Perdew, Burke, and Ernzerhof 1996).
- **Dispersion Correction:** Grimme's DFT-D3 scheme (Grimme et al. 2010).
- **Pseudopotentials:** The SG15 collection of Optimized Norm-Conserving Vanderbilt (ONCV) pseudopotentials (Schlipf and Gygi 2015).
- **Plane-Wave Cutoff:** 400 eV for wavefunctions.
- ***k*-Point Mesh:** $4 \times 4 \times 1$ Monkhorst–Pack grid.
- **System Composition:** 120-atom supercell doped with 1–7 Cr atoms (1/8–7/8 monolayer coverage) at substitutional and interstitial van der Waals gap sites.
- **Data Partitioning:** Structures from the same AIMD trajectory were kept within the same data split to avoid leakage.

MACE Model Training Hyperparameters

All models were trained using the public MACE implementation (GitHub: https://github.com/ACEsuit/mace). The architectural and training parameters are listed in Table 1.

Table 1: MACE Architecture and Training Hyperparameters.

Parameter	**Value**
Architecture	
Interaction Layers (L)	1
Cutoff Radius ($r_{\max}$)	6.0 Å
Feature Channels	32
Irreducible Representations	128×0e + 128×1o
Correlation Order	3
Training	
Optimizer	Adam
Learning Rate	1×10^{-4}
Batch Size (Train / Val)	32 / 16
Max Epochs	1000
SWA Start Epoch	200
Early Stopping Patience	20 epochs
Energy / Force Loss Weight	1.0 / 10.0
Stress Loss Weight	0.0
Fine-tuning Strategy	Multi-head, all layers unfrozen
Foundation Model Checkpoint	`MACE-matpes-pbe-omat`
Random Seed	123

The **Scratch** model was trained for 500–1000 epochs or until convergence via early stopping. The **FT-600K** and **FT-MultiT** models were fine-tuned for approximately 100 epochs.

Interpretability Protocol Implementation

Grad-CAM Analysis The hierarchical interpretability framework was applied to a curated test set of 200 structures encompassing all Cr dopant concentrations, site types, and simulation temperatures. To probe model reliability, we included the 20 structures with the highest and lowest absolute prediction errors per model.

The surrogate Gradient Boosting Regressor used for SHAP analysis achieved cross-validated $R^2 > 0.85$ and a mean absolute error below 0.5 meV/atom in predicting ΔE, confirming that the surrogate accurately reproduced the MACE error distribution. This validates the subsequent SHAP-based mechanistic decomposition as a faithful representation of model uncertainty sources.

SHAP Analysis To understand the physical factors driving MACE model performance and identify failure modes, we employed SHAP (SHapley Additive exPlanations) analysis using a model-agnostic surrogate approach. Since MACE employs message-passing neural networks that are computationally expensive for direct gradient-based explanations, we developed a two-stage interpretability framework.

Feature Engineering for Physical Interpretability. We extracted physically meaningful descriptors of local atomic environments using the Smooth Overlap of Atomic Positions (SOAP) method(Bartók, Kondor, and Csányi 2013). SOAP descriptors were computed with a cut-off radius of 5.0 Å, maximum radial basis functions $n_{\max} = 4$, and maximum angular momentum $l_{\max} = 4$, yielding rotationally and translationally invariant representations of atomic neighborhoods. These parameters capture the typical range of chemical bonding and coordination environments in our Cr-$SbTe_3$ system. Each SOAP feature corresponds to specific physical interactions: species pairs (Cr-Cr, Cr-Sb, Cr-Te, Sb-Sb, Sb-Te, Te-Te), radial complexity (short-, medium-, and long-range interactions), and angular character (spherical $l = 0$, dipolar $l = 1$, quadrupolar $l = 2$, and higher-order multipolar contributions).

Surrogate Model Training. For each MACE variant, we trained gradient-boosting regressors (200 estimators, learning rate 0.05, maximum depth 8) to predict both absolute prediction errors and signed prediction errors from SOAP descriptors. This surrogate approach enables efficient SHAP value computation, while maintaining the physical interpretability of the original atomic descriptors. The surrogate models achieved $R^2 > 0.7$ for error prediction across all MACE variants, validating their utility for feature importance analysis.

SHAP Value Computation and Analysis. We computed TreeSHAP values(Lundberg and Lee 2017) for 500 randomly sampled structures per model to identify which atomic environment features most strongly correlate with prediction accuracy. SHAP values quantify each feature's contribution to the prediction error, with positive values indicating features that increase prediction uncertainty. We analyzed feature importance across multiple physical categories: interaction type (homo- vs. heteronuclear), distance scale (short-, medium-, long-range), and angular complexity (spherical vs. directional interactions).

Comparative Analysis Framework. To identify model-specific strengths and failure modes, we performed cross-model comparisons of feature importance patterns. We computed the top 50 most influential SOAP features across all models and analyzed their physical significance. Additionally, we identified features showing high importance variability across models, indicating potential areas where different training strategies or architectural choices lead to different learned representations of the atomic environment.

Model Performance Analysis

Table 3 presents a comprehensive evaluation of the three training strategies across train, validation, and test sets. While all models demonstrate acceptable RMSE values for both energy and force predictions, detailed analysis reveals significant differences in their learning characteristics and generalization capabilities.

The single temperature model (600 K) exhibits clear signs of overfitting, with energy RMSE increasing from 0.4 meV/atom on the training set to 0.5 meV/atom on validation and test sets (25% degradation). More critically, the RMSE in the force shows substantial deterioration from 13.3 meV/Å during training to 43.3 meV/Å and 37.9 meV/Å on validation and test sets, respectively, representing a 225% performance degradation. This pattern indicates that the model has memorized temperature-specific patterns rather than learning generalizable interatomic interactions.

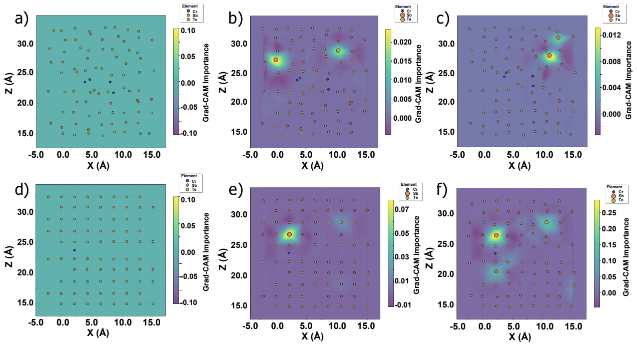

Figure 6: Multi-T model Grad-CAM activation patterns across different snapshots during the AIMD trajectory

In contrast, the multi-temperature model demonstrates superior generalization with consistent energy predictions across all data splits (0.5 meV/atom) and more stable force prediction performance (training: 20.3 meV/Å, test: 45.5 meV/Å, 124% increase). The "from-scratch" model, while showing the most consistent train-test performance gaps, achieves the poorest absolute accuracy with test set RMSE values of 1.4 meV/atom for energy and 82.1 meV/Å for forces.

Importantly, the observation that all models yield seemingly acceptable RMSE values despite exhibiting fundamentally different learning behaviors highlights a critical limitation of conventional evaluation metrics. RMSE alone cannot reveal whether models violate physical principles, maintain thermodynamic consistency, or capture the correct underlying physics. The apparent adequacy of these statistical measures may mask significant deficiencies in the models' understanding of interatomic interactions, force-energy relationships, and transferability to unseen chemical environments. This limitation necessitates the development of more sophisticated diagnostic approaches, including physics-informed validation metrics, explainable AI techniques for feature attribution analysis, and systematic uncertainty quantification to truly assess model reliability and physical fidelity.

Detailed Analysis

Effect of Dopant Spatial Arrangement on Scratch Model Attention

While the main text establishes that the scratch model develops a strong, localized attention on Cr dopants, a more granular analysis reveals that this attention is highly sensitive to the dopants' spatial arrangement. Figure 7 illustrates this phenomenon by comparing the scratch model's Grad-CAM activations across four distinct dopant configurations.

A clear dichotomy emerges between isolated and clustered dopants. In configurations where the Cr dopants are spatially isolated from one another (Figure 7a,b), the model's attention on the Cr atoms is present but relatively

Table 2: Comparison of explainable AI methods for machine learning force fields.

Method	Type	Achievement	Limitation
Linear B-splines	Intrinsic	Direct physical interpretation	Limited expressiveness
Symbolic Regression	Intrinsic	Analytical formulas	Requires distillation
SchNet4AIM	Intrinsic	Predicts physical descriptors	Architecture-specific
SHAP	Post-hoc	Feature importance	Assumes independence
GNNExplainer	Post-hoc	Identifies subgraphs	Computationally expensive
Grad-CAM	Post-hoc	Spatial attribution	No physical grounding

Table 3: Comprehensive performance comparison of MLFFs models across different training strategies.

Training Strategy	RMSE E (meV/atom)			RMSE F (meV/Å)			Relative F RMSE (%)		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
From Scratch	1.2	1.3	1.4	73.9	85.5	82.1	15.27	17.99	16.71
Fine-tuning (600 K)	0.4	0.5	0.5	13.3	43.3	37.9	2.67	8.56	7.59
Fine-tuning (Multi-T)	0.5	0.5	0.5	20.3	49.1	45.5	4.20	10.32	9.26

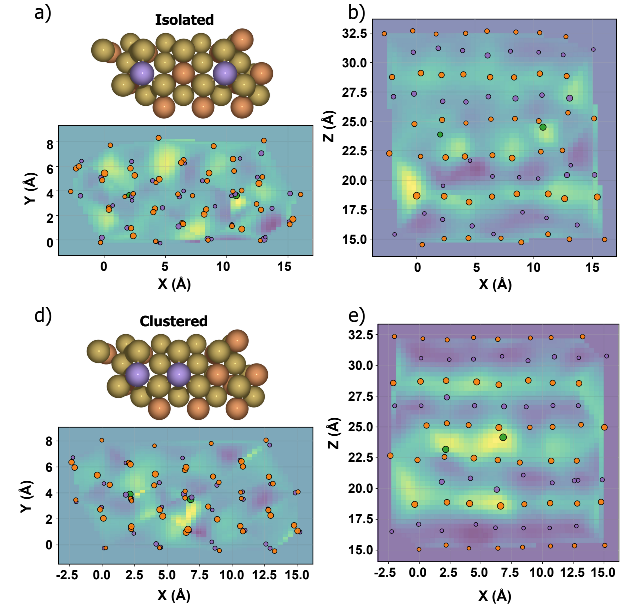

Figure 7: **Scratch Model's Attention is Modulated by Cr Dopant Clustering.** Grad-CAM visualization of atomic energy contributions from the scratch MACE model for four representative dopant configurations. **(a, b)** In configurations with spatially isolated Cr dopants, the model exhibits weak and diffuse activation on the dopant sites. **(c, d)** In contrast, for configurations with clustered Cr dopants, the model displays intense, localized activation on the Cr atoms and, critically, in the interstitial region between them. This emergent focus on the Cr-Cr interaction zone indicates the model has learned to identify dopant clustering as an energetically significant structural motif, consistent with its known physical effects on the material's electronic properties.

weak and diffuse. The activation intensity is low, suggesting the model perceives these atoms as minor perturbations to the host matrix.

In contrast, when the Cr dopants are clustered in close proximity (Figure 7c,d), the model's attention mechanism changes dramatically. The Grad-CAM activation becomes intense and sharply localized not only on the Cr atoms themselves but also in the interstitial region *between* them. This emergent attention on the Cr-Cr interaction zone signifies that the model has learned to recognize dopant clustering as a distinct, energetically significant chemical feature.

This finding is particularly compelling as it aligns with our independent *ab initio* investigations (Cao and Clancy 2025b), which show that Cr dopant clustering directly modulates the electronic properties of the Sb_2Te_3 host, including its band structure. The scratch model, despite its less refined strategy compared to the FT-Multi-T variant, has thus learned a physically correct and non-trivial structure-property relationship. This demonstrates the power of our XAI framework to validate that the model is learning scientifically sound principles directly from the training data.

Hierarchical Feature Learning in MACE Models

To elucidate how machine learning interatomic potentials encode structural information, we performed an attention flow analysis on the MACE model trained for the $Cr_4Sb_{32}Te_{48}$ supercell. This 84-atom system provides a representative framework for analyzing hierarchical information processing in layered chalcogenides.

Layer-specific information processing. The evolution of mean attention magnitudes (Fig. 9b) reveals a 30-fold reduction from node embedding (3.41 ± 0.18) to radial embedding (0.11 ± 0.11), signifying a transition from dense atomic representations to sparse, chemically selective encodings. The high standard deviation in the node-embedding stage reflects heterogeneous atomic environments inherent to the mixed Cr–Sb–Te lattice. This pronounced attention decay indicates that the model progressively distills structural complexity into compact, chemically relevant represen-

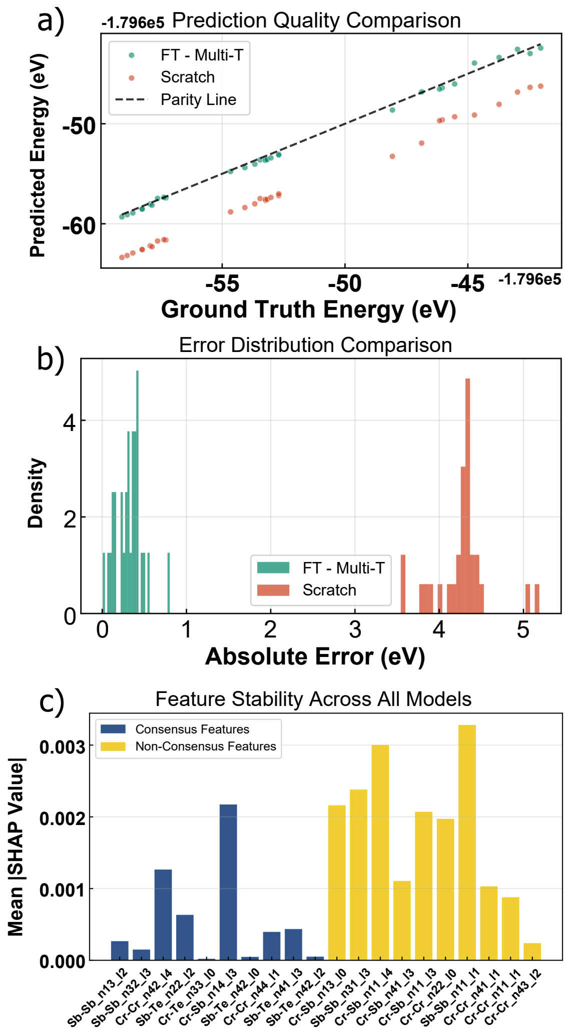

Figure 8: **Supplementary Analysis of Error Predictor Performance and Feature Stability. (a)** Parity plot comparing predicted vs. ground truth energies for best (FT–Multi-T) and worst (scratch) MACE models, showing significant fine-tuning improvement. **(b)** Absolute prediction error distributions. The fine-tuned model shows lower mean error and narrower distribution, indicating more reliable predictions. **(c)** Feature stability across models. Average importance of 10 most stable ('Consensus") vs. 10 most unstable ('Non-Consensus") features, determined by SHAP value coefficient of variation. Fine-tuning drives convergence on consistent physically important features.

tations.

Chemical environment recognition. The pairwise attention-weight heatmap (Fig. 9c) demonstrates that MACE learns chemically intuitive patterns. Distinct Te–Te and Sb–Sb blocks confirm element-specific bonding recognition, while Cr atoms (80–83) display broad attention across both Sb and Te species. This behavior aligns with Cr's role as an intercalant that perturbs local bonding symmetry and introduces diverse coordination environments—consistent with experimental observations of hybridization between Cr d-orbitals and the host p-states.

Information complexity and feature specialization. Entropy analysis (Fig. 9d) shows a non-monotonic trend, with information complexity increasing from node (7.72 bits) to radial embeddings (9.69 bits). This counter-intuitive rise suggests that radial embeddings capture higher-order geometric relationships that require richer representations. Correspondingly, effective-rank analysis (Fig. 9e) indicates that radial layers encode a broader diversity of spatial correlations despite reduced variance, highlighting the network's efficiency in compressing yet preserving chemical diversity.

Validation of chemical relevance. The distribution of edge attention (Fig. 9f) displays three distinct peaks corresponding to Te–Te (2.8 Å), Sb–Te (3.2 Å), and interlayer (4.5 Å) distances, confirming that the model's attention naturally aligns with chemically meaningful bonding motifs. Notably, these characteristic scales emerge without explicit prior knowledge, underscoring the model's ability to infer structural chemistry directly from energy–force data.

Implications for interpretability and transferability. Node embeddings exhibit high activation variance and negligible sparsity, suggesting efficient feature utilization and minimal redundancy. As information flows to deeper layers, variance decreases while effective rank increases, indicating enhanced feature specialization. This hierarchical compression mirrors chemical bonding hierarchies—transitioning from local atomic identity to extended structural correlations—thereby enhancing both interpretability and generalization across temperature and composition domains.

Comparison with handcrafted descriptors. Unlike conventional descriptors requiring manual definition, the MACE attention weights act as dynamically learned, system-specific structural fingerprints. Their emergent chemical selectivity provides superior transferability and insight into structure–property relationships. The hierarchical attention patterns observed here establish that MACE not only predicts energies with high accuracy but also encodes chemically interpretable features directly linked to the underlying bonding topology.

Overall, the attention flow analysis reveals that MACE models develop a chemically grounded hierarchy of features that capture both local coordination and long-range correlations, providing a transparent framework for understanding model decision-making in complex chalcogenide systems.

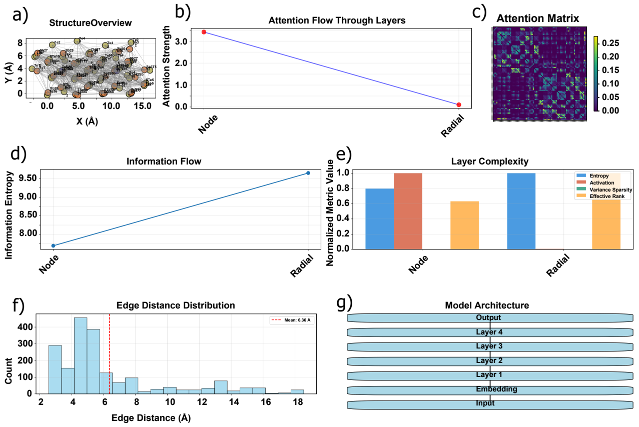

Figure 9: **MACE attention flow analysis reveals hierarchical information processing in Cr-doped $\mathbf{SbTe_3}$.** (a) 2D projection of the $Cr_4Sb_{32}Te_{48}$ supercell showing atomic species and connectivity within a 5.0 Å cutoff. (b) Layer-wise decay of mean attention magnitude, indicating progressive information compression. (c) Attention heatmap revealing element-specific correlations and Cr-mediated cross-interactions. (d–e) Entropy and activation metrics demonstrate transition from dense, high-variance node embeddings to compact, diverse radial representations. (f) Edge attention distribution aligns with characteristic Te–Te, Sb–Te, and interlayer distances. (g) Schematic of the two-layer MACE architecture. Together, these results highlight MACE's hierarchical encoding of both local coordination and long-range structural correlations critical for accurate energy prediction.