Stress-Testing Multimodal Foundation Models for Crystallographic Reasoning

Can Polat[1], Hasan Kurban[2*], Erchin Serpedin[1], Mustafa Kurban[3,4*]

[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA

[2]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

[3]Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha, Qatar

[4]Department of Prosthetics and Orthotics, Ankara University, Ankara, Turkey

[*]Corresponding authors: hkurban@hbku.edu.qa, kurbanm@ankara.edu.tr

Abstract

Evaluating foundation models for crystallographic reasoning requires benchmarks that isolate generalization behavior while enforcing physical constraints. This work introduces, *xCrysAlloys*, a multiscale multicrystal dataset with two physically grounded evaluation protocols to stress-test multimodal generative models. The *Spatial-Exclusion* benchmark withholds all supercells of a given radius from a diverse dataset, enabling controlled assessments of spatial interpolation and extrapolation. The *Compositional-Exclusion* benchmark omits all samples of a specific chemical composition, probing generalization across stoichiometries. Nine vision–language foundation models are prompted with crystallographic images and textual context to generate structural annotations. Responses are evaluated via (i) relative errors in lattice parameters and density, (ii) a physics-consistency index penalizing volumetric violations, and (iii) a hallucination score capturing geometric outliers and invalid space-group predictions. These benchmarks establish a reproducible, physically informed framework for assessing generalization, consistency, and reliability in large-scale multimodal models. Dataset and implementation are available at `https://github.com/KurbanIntelligenceLab/StressTestingMMFMinCR`.

1 Introduction

Crystalline solids underpin a wide range of modern technologies. Their periodic atomic arrangements determine the band gaps of semiconductors, the ion-transport channels in battery electrodes, and the phonon spectra that govern thermal conductivity in microelectronics (Wyckoff, 1963a; Bhadeshia, 2001). Even a single misassigned lattice parameter can cascade through simulation pipelines, distorting derived physical models and impeding materials discovery (Levi and Kotrla, 1997; Lubarda, 2003). Structural resolution has traditionally relied on labor-intensive diffraction techniques or exhaustive structure enumeration followed by density functional theory (DFT) relaxation (Kohn and Sham, 1965). Synthesis methods such as hydrothermal growth (Baruah and Dutta, 2009), chemical vapor deposition (Carlsson and Martin, 2010), and high-pressure processing (Bertucco and Vetter, 2001) further introduce domain-specific variability by accessing distinct thermodynamic regimes and defect topologies.

Recent progress in generative modeling, particularly autoregressive language models capable of emitting crystallographic information files (Hall et al., 1991), enables rapid lattice generation with chemically plausible compositions. However, existing materials databases—such as AFLOW (Curtarolo et al., 2012), the Materials Project (Jain et al., 2013), and OQMD (Saal et al., 2013)—remain predominantly unimodal and typically lack expert-written, human-interpretable descriptions of crystal chemistry. This absence of multimodality impedes systematic evaluation of large vision–language models and language models in crystallographic reasoning. Current scientific multimodal benchmarks are limited in scale, visually simplistic, and textually sparse, constraining analysis of factual accuracy, hallucination patterns, and compliance with physical laws.

To overcome these limitations, *xCrysAlloys*, a new multimodal dataset of crystalline alloy materials is presented, accompanied by two physically grounded benchmarking protocols. The *spatial-exclusion* (SE) benchmark withholds supercells of a specific radius from the set $\{R_k\}_{k=7}^{10}$, enabling controlled evaluation of spatial interpolation (interior radii) and extrapolation (boundary radii). In parallel, the *compositional-exclusion* (CE) benchmark withholds all samples corresponding to a target chemical composition, assessing generalization across compositional space. State-of-the-art foundation models are evaluated under both

benchmarks by generating structural annotations from crystallographic images and textual prompts. Model outputs are parsed into a structured MATERIAL PROPERTIES schema and assessed for geometric accuracy, consistency with physical constraints, and hallucination risk. These benchmarks provide a reproducible, domain-informed framework for measuring generalization and reliability in large-scale generative models, and contribute to emerging efforts to probe, refine, and safely deploy scientific knowledge at scale.

The remainder of the manuscript is structured as follows. Section 2 surveys the theoretical foundations and related literature. Section 3 details the methodological framework. Section 4 describes the dataset construction, evaluation metrics, and experimental procedures. Section 5 presents the empirical findings. Section 6 discusses the study's limitations, and Section 7 concludes with final observations.

2 Background

2.1 Materials Modeling: From First-Principles to Data-Driven Representations

Accurate modeling of crystal structures has long relied on first-principles approaches such as DFT, which provides access to ground-state electronic properties, total energies, and atomic forces in periodic solids (Jensen and Wasserman, 2018). DFT remains the cornerstone of computational materials science, particularly for predicting band structures, charge distributions, and structural relaxations. However, its cubic scaling with respect to system size poses significant limitations for large supercell or high-throughput investigations (Hourahine et al., 2007).

To mitigate this computational burden, semi-empirical methods such as density functional tight binding (DFTB) (Gaus et al., 2011) offer an efficient approximation by expanding the Kohn–Sham energy around a reference density. Modern enhancements, including Slater–Koster parameterizations and self-consistent charge corrections (Papaconstantopoulos and Mehl, 2003), have extended DFTB's usability to heavier elements and time-dependent simulations. Nevertheless, both DFT and DFTB still require significant computational resources, especially when scaling across diverse compositions and large atomic configurations.

This work adopts an alternative route grounded in experimental crystallographic data. Rather than performing relaxation via electronic structure theory, all unit cell parameters are sourced from peer-reviewed literature. These serve as the foundation for constructing supercells and nanocluster models at varying spatial scales, enabling physically consistent benchmarking without reliance on simulation-based optimization.

2.2 Machine Learning and Multimodal Foundation Models in Materials Science

In parallel to physics-based approaches, machine learning has emerged as a powerful tool in materials discovery pipelines. Graph neural networks, such as SchNet (Schütt et al., 2017), DimeNet (Gasteiger et al., 2020), and FAENet (Duval et al., 2023), operate directly on atomic graphs to predict structural and functional properties with increasing fidelity (Zheng et al., 2018; Rane, 2023; Liao et al., 2023; Kurban et al., 2024). Despite their promise, these models often suffer from limitations related to data sparsity, distribution shift, and lack of interpretability.

Recent efforts focus on unifying visual, textual, and structural modalities via large multimodal models. Such systems—exemplified by ChemVLM (Li et al., 2025), MatterChat (Tang et al., 2025), and xChemAgents (Polat et al., 2025b)—are designed to capture complex structure–property relationships while supporting interactive reasoning tasks. Supporting benchmarks such as ScienceQA (Lu et al., 2022), MoleculeNet (Wu et al., 2018), and ChemLit-QA (Wellawatte et al., 2024) provide curated evaluation settings across physics, chemistry, and biology. In materials science specifically, TDCM25 (Polat et al., 2025a) and LAB-Bench (Laurent et al., 2024) advance this trend by offering multimodal, multi-property datasets.

While these efforts signal progress, current multimodal systems still exhibit limited capability in physical reasoning, compositional generalization, and geometric consistency (Miret and Krishnan, 2024). This motivates the development of targeted benchmarks—such as the Spatial-Exclusion and Compositional-Exclusion protocols introduced in this study—to systematically probe the crystallographic reasoning capabilities of foundation models at multiple scales.

3 Methods

3.1 Crystal Structure Generation

This study utilizes experimental lattice parameters from peer-reviewed literature to reconstruct unit cell geometries for ten crystalline materials: Ag, Au, $CH_3NH_3PbI_3$, Fe_2O_3, MoS_2, PbS, SnO_2, $SrTiO_3$, TiO_2, and ZnO. The reported crystallographic space groups and cell constants for each compound are listed in Appendix A.1.

For each material, a large periodic supercell of dimensions $30 \times 30 \times 30$ unit cells was constructed to approximate a bulk crystalline environment. This bulk structure served as the foundational source for subsequent nanoscale structure generation. Spherical nanoclusters were then carved from the center of this supercell using a radial cutoff criterion: atoms located within a prescribed distance from the geometric center were retained, while atoms beyond the cutoff were excluded.

To ensure systematic evaluation across multiple spatial scales, four target radii $R \in \{0.7, 0.8, 0.9, 1.0\}$ nm—labeled R7–R10—were selected. For each material, spherical nanoclusters of increasing size were carved out based on these radii. The resulting atom counts varied depending on the underlying crystal structure and unit cell complexity, typically yielding configurations with tens to hundreds of atoms. This procedure preserves the lattice symmetry and local coordination environments while introducing surface-dominated features relevant to nanoscale crystallographic reasoning.

3.2 Orientation Sampling and Rendering

To evaluate rotational invariance and visual robustness, each supercell was rendered under ten unique orientations. These include one canonical pose and nine additional orientations sampled using the Fibonacci-sphere algorithm (Stanley, 1975) to approximate uniform $\mathrm{SO}(3)$ coverage.

For each orientation, atomic configurations were orthographically projected onto the xy-plane. Visualization was performed by mapping atoms to Gaussian-blurred disks, scaled by covalent radius and colored using a CPK-inspired palette. This consistent rendering pipeline generated standardized 2D crystallographic images (64×64 px) that serve as visual input to the foundation models.

3.3 Structured Text Annotation

Each atomic structure is paired with a textual annotation formatted under a standardized MATERIAL PROPERTIES schema. Annotations include scalar properties—such as atom count, lattice parameters, supercell volume, and bulk density—as well as categorical attributes like space group and crystal system.

To support robust evaluation, each annotation also includes primitive-cell parameters, average nearest-neighbor distance, and a descriptive paragraph summarizing the crystal's physical characteristics. This structured multimodal representation enables the computation of multiple evaluation metrics—including geometric error, physical-law consistency, and hallucination rate—described in Section 4.

4 Experiments

Dataset. *xCrysAlloys*, comprises ten crystalline compounds of technological relevance: Ag, Au, $CH_3NH_3PbI_3$, Fe_2O_3, MoS_2, PbS, SnO_2, $SrTiO_3$, TiO_2, and ZnO. For each material, spherical nanoclusters were extracted at four target radii $R \in \{0.7, 0.8, 0.9, 1.0\}$ nm (R7–R10), yielding a multiscale corpus of 3D atomic structures.

Each nanocluster was rendered in ten orientations—one canonical and nine using Fibonacci-sphere rotations—to ensure quasi-uniform coverage over $\mathrm{SO}(3)$. This process generated over 400 crystallographic images per material—derived from 4 radius levels and 10 orientations per structure (i.e., $4 \times 10 = 40$ images per material–radius combination)—paired with expert-curated annotations conforming to the MATERIAL PROPERTIES schema. Full details on structure generation are provided in Section 3.1. An overview is shown in Figure 1.

Evaluation Metrics. PERCENT ERROR for each numerical property $p \in \{N_{\text{atoms}}, V_{\text{cell}}, a, b, c, \rho, a_{\text{p}}, b_{\text{p}}, c_{\text{p}}\}$ is computed as:

$$\Delta_p\,[\%] = 100 \cdot \frac{|p^{\text{gen}} - p^{\text{ref}}|}{|p^{\text{ref}}|}.$$

SPACE-GROUP MATCH is defined as:

$$I_{\text{SG}} = \mathbf{1}\big(\text{SG}^{\text{gen}} = \text{SG}^{\text{ref}}\big).$$

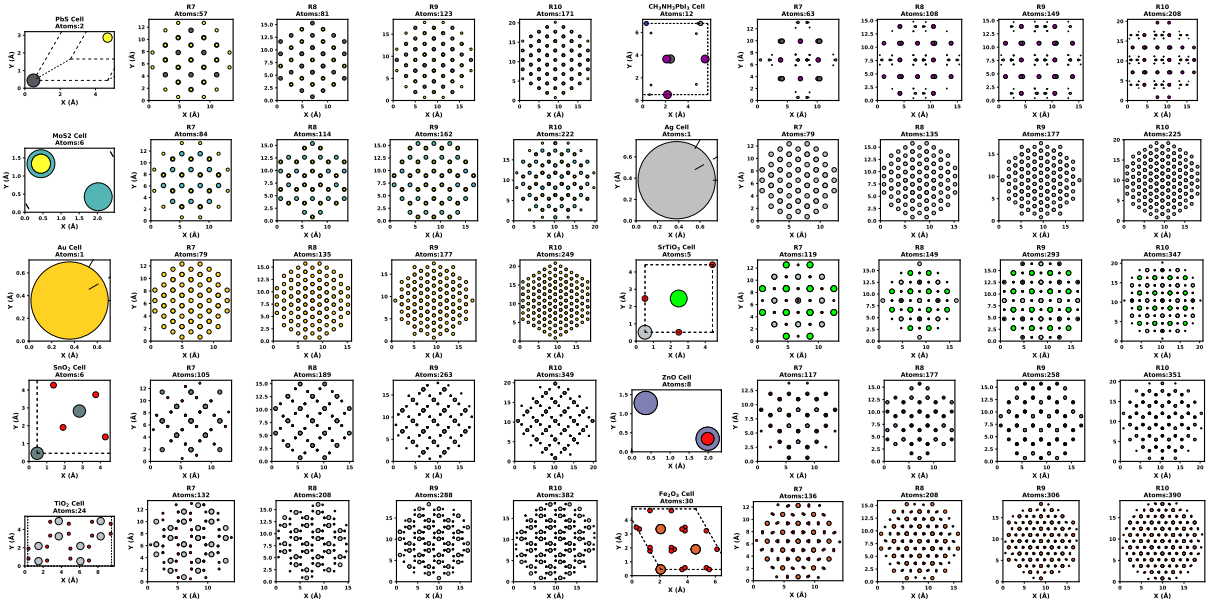

Figure 1: Gallery of atomic structures for each material in *xCrysAlloys*. The first column shows the primitive unit cell for each material, while the subsequent columns display nanocluster structures with increasing radii ($R7$, $R8$, $R9$, $R10$). Each structure is visualized in a canonical orientation, with the number of atoms indicated in each panel. Materials are sorted by the atom count of their largest ($R10$) nanocluster.

Group statistics over n examples are:

$$\mu_p = \frac{1}{n}\sum_{i=1}^{n} \%\Delta_p^{(i)},$$

$$\sigma_p = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\%\Delta_p^{(i)} - \mu_p\right)^2},$$

$$\mathrm{CI}_{95} = \mu_p \pm 1.96 \cdot \frac{\sigma_p}{\sqrt{n}}.$$

PREDICTION CONSISTENCY (ROTATIONS) is computed by:

$$C_{\text{pred}} = 1 - \min\left(\frac{\sigma_r}{\mu_r}, 1\right),$$

where μ_r and σ_r are the mean and standard deviation of a rotation-specific error set.

PHYSICAL-LAW COMPLIANCE is evaluated for:

$$p \in \left\{\rho, \frac{b}{a}, \frac{c}{a}, \left(\frac{b}{a}\right)_{\text{prim}}, \left(\frac{c}{a}\right)_{\text{prim}}\right\},$$

using:

$$\delta_p = \frac{|p^{\text{gen}} - p^{\text{ref}}|}{p^{\text{ref}}},$$

$$s_p = \begin{cases} 1.0 & \delta_p \le 0.10, \\ 0.5 & 0.10 < \delta_p \le 0.25, \\ 0.0 & \delta_p > 0.25 \text{ or on error.} \end{cases}$$

Aggregate score:

$$S_{\text{phys}} = \begin{cases} \frac{1}{N}\sum_p s_p & N > 0, \\ 0.0 & N = 0 \text{ or missing.} \end{cases}$$

HALLUCINATION SCORE is defined for all the percent error properties p. Let $g = p^{\text{gen}}$ and $r = p^{\text{ref}}$, then:

$$h_p = \begin{cases} 1.0 & g \le 0 \quad \text{(non-physical)}, \\ 1.0 & \frac{|g-r|}{|r|} > 0.25, \\ 0.5 & 0.10 < \frac{|g-r|}{|r|} \le 0.25, \\ 0.0 & \frac{|g-r|}{|r|} \le 0.10. \end{cases}$$

Let M be the number of valid checks:

$$S_{\text{hall}} = \begin{cases} \frac{1}{M}\sum_p h_p & M > 0, \\ 0.0 & M = 0, \\ 1.0 & \text{if input is } \texttt{None}. \end{cases}$$

Additional metric definitions are provided in Appendix A.2.

Spatial-Exclusion Protocol. SE protocol measures extrapolation across length scales. For each material m_i with radius set $\mathcal{R}_{m_i}$, one radius $R_* \in \mathcal{R}_{m_i}$ is held out. The model context includes:

$$\left|\mathcal{R}_{m_i} \setminus \{R_*\}\right| \times 5$$

examples (5 rotations for each of the remaining radii). Each test instance uses only the Cartesian

coordinates of (m_i, R_*, k), and the model must generate predictions without seeing any data at R_*. The overall SE error is:

$$E_{\mathrm{SE}} = \frac{1}{|\mathcal{M}| \ \sum_i |\mathcal{R}_{m_i}| \ \times 5}$$
$$\times \sum_i \sum_{R_* \in \mathcal{R}_{m_i}} \sum_{k=0}^{4} \ell\big(\hat{y}_{i,R_*,k},\, y_{i,R_*,k}\big),$$

where ℓ is the percent error loss.

Compositional-Exclusion Protocol. CE protocol assesses generalization across compositions. For each material m_i, all of its data are excluded from the context. The context size becomes:

$$\Big(\sum_{m_j \neq m_i} |\mathcal{R}_{m_j}| \Big) \times 5$$

At test time, only the Cartesian coordinates of (m_i, R_*, k) are given. The transfer error is:

$$E_{\mathrm{CE}} = \frac{1}{|\mathcal{M}| \ \sum_i |\mathcal{R}_{m_i}| \ \times 5}$$
$$\times \sum_i \sum_{R_* \in \mathcal{R}_{m_i}} \sum_{k=0}^{4} \ell\big(\tilde{y}_{i,R_*,k},\, y_{i,R_*,k}\big),$$

which captures model performance when required to infer from disjoint compositions. Comparing E_{CE} and E_{SE} helps isolate failure modes in spatial vs. chemical generalization.

5 Results

SE Evaluation. In SE evaluation, each language model was assigned the task of predicting a held-out radius value (R_7–R_{10}) for a given crystalline material, and its outputs for atom count (N_A), cell volume (V), lattice constants (a, b, c), and density (ρ) were compared against reference structures. Percent errors ($\%\Delta$) were averaged across all models and five random 3D orientations per configuration. As shown in Table 1(a), the resulting error rates remain consistently high, particularly for key physical properties—exceeding thresholds that render predictions scientifically unreliable.

These discrepancies reveal a fundamental limitation: the models fail to internalize core geometric and physical constraints that govern crystal structures. The inability to extrapolate structural properties across radii highlights the need for architectural enhancements, including explicit domain constraints, physical priors, and robust error-correction strategies to prevent hallucinated outputs and enforce consistency in atomic-scale reasoning.

CE Evaluation. In the CE evaluation, each language model received structural data from nine materials at a fixed radius R and was tasked with predicting N_A, primitive cell lengths (a_p, b_p, c_p), and angles (α_p, β_p, γ_p) for a held-out material. To ensure robustness, predictions were averaged over five random 3D orientations and multiple model variants. As reported in Table 1(b), percent errors in cell lengths frequently exceed 15%, and atom count errors surpass 30% for complex compounds at smaller radii—suggesting a failure to generalize geometric patterns across novel chemistries.

Additionally, absolute deviations in primitive angles often exceed 5° and reach beyond 20° in certain cases, reflecting substantial geometric inconsistencies and a tendency to hallucinate physical details. These results reinforce that purely data-driven training is insufficient for capturing atomic-scale regularities. Embedding explicit domain constraints, structured knowledge priors, and uncertainty-aware mechanisms is essential for enforcing physical plausibility and mitigating hallucination in generative crystallography.

Knowledge Transfer. CE evaluation reveals that current multimodal LLMs rely heavily on memorized numeric templates rather than internalized crystallographic principles. In the control setting (SE), all eight models achieve low mean percent errors ($0.04 \leq \mathrm{SE} \leq 0.18$). However, when evaluated on withheld compounds, performance collapses: the average error increases by several orders of magnitude, and the transfer ratio $T = \mathrm{CE}/\mathrm{SE}$ surges from 2.2×10^3 to 2.3×10^4, with one model diverging entirely ($T = \infty$).

A consistent failure pattern emerges across systems: six models record their largest relative error on the primitive-cell b-axis ($\%\Delta b_p$), while the remainder fail on $\%\Delta a_p$. PbS is the most challenging composition, ranked worst by all models except one, which instead fails on Fe_2O_3. The rock-salt symmetry of PbS demands reconciliation between cubic crystal geometry and its serialized representation; instead, most models generate inconsistent or arbitrary lattice parameters. These findings underscore that in-distribution performance does not imply genuine crystallographic reasoning. Even modest compositional perturbations destabilize the geometric priors learned by large-scale vision–language models, revealing a brittle foundation for generalization.

(a) Spatial-Exclusion (SE)

Material	R7						R8						R9						R10					
	$\%\Delta N_A$	$\%\Delta V$	$\%\Delta a$	$\%\Delta b$	$\%\Delta c$	$\%\Delta \rho$	$\%\Delta N_A$	$\%\Delta V$	$\%\Delta a$	$\%\Delta b$	$\%\Delta c$	$\%\Delta \rho$	$\%\Delta N_A$	$\%\Delta V$	$\%\Delta a$	$\%\Delta b$	$\%\Delta c$	$\%\Delta \rho$	$\%\Delta N_A$	$\%\Delta V$	$\%\Delta a$	$\%\Delta b$	$\%\Delta c$	$\%\Delta \rho$
Ag	26.53	46.74	9.21	13.00	21.32	14.00	10.21	14.59	5.11	5.98	8.12	13.63	7.31	15.00	7.88	8.52	10.05	7.96	7.48	9.64	5.65	5.07	9.81	8.36
Au	28.21	49.44	10.18	13.44	22.48	15.47	11.26	14.20	5.43	6.55	6.42	11.38	9.19	12.26	6.26	7.51	9.37	8.58	15.40	583.53	40.45	40.73	44.04	17.61
$CH_3NH_3PbI_3$	47.34	44.81	16.10	10.85	12.31	34.38	16.83	20.28	7.45	7.82	7.08	20.85	17.58	27.85	8.52	9.34	8.51	22.72	13.45	19.11	8.94	8.44	9.11	128.49
Fe_2O_3	26.21	31.41	8.92	10.46	12.99	13.37	13.42	20.18	6.80	4.00	7.31	11.34	12.23	15.81	5.84	6.60	5.18	10.46	11.92	12.86	4.45	5.23	4.97	6.93
MoS_2	15.48	27.46	9.21	9.17	22.55	10.29	16.80	14.31	5.84	5.53	11.78	16.10	9.59	17.76	7.18	7.67	7.22	9.63	5.69	19.28	5.63	7.79	10.97	19.16
PbS	17.54	29.71	9.07	10.78	11.75	39.28	18.66	23.90	6.16	7.38	11.53	19.91	12.90	22.60	9.69	8.78	9.62	28.27	12.27	14.45	7.25	5.99	8.00	13.85
SnO_2	29.48	19.31	8.25	7.02	10.03	26.84	9.78	18.99	4.31	4.04	9.33	7.56	8.24	12.57	5.32	5.25	6.90	7.21	6.80	10.86	4.42	4.15	8.57	8.34
$SrTiO_3$	30.59	55.64	15.59	16.05	15.82	17.99	37.42	20.10	7.12	7.77	8.25	38.31	20.30	22.56	7.87	7.19	7.56	17.26	21.58	21.49	6.84	6.80	7.36	16.79
TiO_2	23.54	22.99	6.71	6.27	12.77	6.42	8.08	9.54	4.48	4.09	4.35	5.09	6.39	9.32	4.88	5.92	4.75	6.35	5.76	6.95	4.84	4.12	3.44	5.48
ZnO	13.11	16.66	10.47	9.28	12.22	21.97	12.74	12.42	4.96	5.23	6.98	11.04	5.66	9.63	5.05	6.01	4.59	8.50	8.91	19.57	6.21	7.49	8.89	20.74

(b) Compositional-Exclusion (CE)

Material	R7							R8							R9							R10						
	$\%\Delta N_A$	$\%\Delta a_p$	$\%\Delta b_p$	$\%\Delta c_p$	$\lvert\Delta\alpha_p\rvert$	$\lvert\Delta\beta_p\rvert$	$\lvert\Delta\gamma_p\rvert$	$\%\Delta N_A$	$\%\Delta a_p$	$\%\Delta b_p$	$\%\Delta c_p$	$\lvert\Delta\alpha_p\rvert$	$\lvert\Delta\beta_p\rvert$	$\lvert\Delta\gamma_p\rvert$	$\%\Delta N_A$	$\%\Delta a_p$	$\%\Delta b_p$	$\%\Delta c_p$	$\lvert\Delta\alpha_p\rvert$	$\lvert\Delta\beta_p\rvert$	$\lvert\Delta\gamma_p\rvert$	$\%\Delta N_A$	$\%\Delta a_p$	$\%\Delta b_p$	$\%\Delta c_p$	$\lvert\Delta\alpha_p\rvert$	$\lvert\Delta\beta_p\rvert$	$\lvert\Delta\gamma_p\rvert$
Ag	6.39	10.49	10.49	10.49	7.50	7.50	7.50	3.39	10.48	10.48	10.48	7.50	7.50	7.50	4.28	10.47	10.47	10.47	7.50	7.50	7.50	14.09	13.76	13.76	13.76	6.75	6.75	6.75
Au	3.58	17.79	17.79	17.79	6.00	6.00	6.00	4.89	15.63	15.63	15.63	4.50	4.50	4.50	3.95	16.76	16.76	16.76	4.50	4.50	4.50	12.64	16.65	16.65	16.65	4.50	4.50	4.50
$CH_3NH_3PbI_3$	32.14	10.48	10.46	23.36	1.50	1.86	3.75	37.87	5.04	9.27	9.52	4.68	5.04	4.68	46.59	11.51	16.15	21.65	3.07	3.43	3.07	47.69	12.27	12.39	10.45	1.55	1.91	6.80
Fe_2O_3	18.42	2.95	1.10	8.65	1.50	1.50	5.25	27.82	2.79	2.77	9.70	0.75	0.78	5.25	26.00	3.46	3.46	11.38	1.50	1.50	6.00	19.60	5.01	3.36	10.34	1.74	1.74	6.24
MoS_2	13.57	0.04	0.04	0.02	0.00	0.00	0.00	23.42	7.45	7.42	3.72	0.00	0.03	2.25	18.61	0.01	0.01	0.02	0.00	0.00	0.00	26.55	0.04	0.04	0.03	0.00	0.00	0.00
PbS	30.95	40.57	40.55	40.57	22.31	22.33	22.31	56.05	40.13	40.13	39.16	24.00	24.00	24.00	58.68	41.16	41.16	41.16	24.75	24.75	24.75	44.59	40.11	40.11	40.11	24.75	24.75	24.75
SnO_2	19.17	4.71	0.78	3.08	0.00	0.00	0.00	19.79	5.67	1.73	13.80	0.00	0.01	0.75	31.33	2.42	0.40	1.59	0.00	0.00	0.00	16.14	2.36	0.39	1.55	0.00	0.00	0.75
$SrTiO_3$	27.72	8.54	4.44	5.79	0.43	0.43	0.43	22.48	11.82	11.74	13.02	0.00	0.06	0.00	27.52	9.53	3.81	7.68	0.00	0.00	0.60	19.49	1.53	1.52	1.54	0.00	0.01	0.00
TiO_2	21.02	51.78	18.81	55.18	0.00	0.00	1.50	19.81	51.80	19.62	47.97	0.00	0.01	2.25	32.06	51.96	19.12	53.58	0.00	0.00	1.50	20.75	53.15	21.12	49.80	0.00	0.00	3.75
ZnO	24.53	1.16	2.66	1.03	0.00	0.00	1.50	23.26	5.86	7.34	2.07	0.00	0.02	2.25	31.50	0.01	0.01	0.01	0.00	0.00	0.00	24.59	0.01	3.02	0.01	0.00	0.00	0.75

Table 1: Mean percent errors ($\%\Delta$) for (a) the spatial-extension (SE) protocol—evaluating extrapolation to unseen supercell radii—and (b) the compositional-exclusion (CE) protocol—evaluating cross-material transfer. Part (a) reports errors on atom count N_A, cell volume V, lattice parameters a, b, c, and density ρ; part (b) reports errors on N_A, primitive cell edges a_p, b_p, c_p, and absolute angular deviations $|\Delta\alpha_p|$, $|\Delta\beta_p|$, $|\Delta\gamma_p|$. Results are shown for each material and radius value (R7–R10), averaged over five random rotations per configuration and across all models. Lower values indicate better agreement with reference structures. These complementary metrics illustrate the model's capacity to capture atomic-scale patterns across variations in supercell size and material composition. Contrasting SE and CE errors highlights whether performance limitations stem from radius extrapolation or cross-material generalization. Colours indicate predictive difficulty: green marks the material with the lowest prediction error (easiest to predict), while red marks the material with the highest prediction error (hardest to predict).

Model	SE	CE	$T \times 10^3$	$G_{\max} \times 10$	t_{SE}	t_{CE}
Claude Opus 4 (Anthropic)	0.06	0.91	**2.17**	3.04	12.86	13.91
Claude Sonnet 4 (Anthropic)	**0.04**	0.68	3.93	3.04	6.43	8.23
DeepSeek-Chat (DeepSeek)	0.09	1.79	14.16	6.47	24.97	13.71
GPT-4.1 Mini (OpenAI)	0.18	**0.53**	2.63	6.00	8.08	7.26
Gemini 2.5 Flash (Google)	0.05	1.32	21.38	3.04	**3.06**	**5.00**
Grok 2 (X.ai)	0.07	2.34	15.55	3.04	6.37	8.99
Grok 2 Vision (X.ai)	0.06	2.02	22.54	6.47	7.32	9.50
Llama-4 Maverick (Meta)	0.09	0.89	3.70	**3.00**	4.33	6.72
Mistral Medium 3 (Mistral AI)	0.05	0.92	11.24	**3.00**	14.78	15.45

Table 2: Transfer degradation analysis with mean percent errors ($\%\,\Delta$) for the SE and CE splits. $T = \text{CE}/\text{SE}$; $G_{\max}$ is the largest absolute error observed in any single prediction. t_{SE} and t_{CE} represents the each models latency in seconds for SE and CE task, respectively. **Bold** indicates the top-performing model, while underlining denotes the runner-up.

Correlation Shift. Table 3 reports the average error–error correlation coefficients for fourteen property pairs under the SE and CE protocols, along with their differences. Notably, the transition from SE to CE increases the correlation between projected lattice constants a_p and b_p by 0.59, suggesting that prediction errors for these geometric features become more aligned when the model is exposed to entirely novel compositions. In contrast, the correlation between volume V and average formation energy $\bar{\varepsilon}$ drops by -0.64, indicating a breakdown in the learned volume–energy coupling under compositional generalization.

These shifts reverse when comparing CE to SE, confirming that the observed effects stem from the validation regime rather than intrinsic data asymmetries. This bidirectional sensitivity highlights a critical weakness: current foundation models preserve certain geometric relationships under run-wise exclusion but fail to maintain deeper physical dependencies—such as energetic coherence—when facing unfamiliar chemistries. The instability of error correlations under different evaluation settings undermines the robustness of model generalization and emphasizes the need for embedding invariant physical priors into model architecture and training.

Compliance and Hallucination. The models consistently struggle to enforce fundamental physical constraints and frequently fabricate ungrounded details, as quantified in Table 4. Physical-law compliance scores fall below acceptable thresholds for most materials, with particularly poor performance on TiO_2, where nearly half the predictions violate basic geometric or density-based relationships. Concurrently, hallucination scores indicate that a significant fraction of predicted properties—often over 40%—deviate substantially from reference values or represent nonphysical outputs. The co-occurrence of constraint violations and fictitious property generation highlights systemic limitations in current architectures. These results reinforce the need for models that integrate structural priors, conservation rules, and uncertainty-aware mechanisms

(a) SE ⇒ CE

	$N_{\text{atoms}} \leftrightarrow V$	$V \leftrightarrow \bar{\varepsilon}$	$V \leftrightarrow \rho$	$\gamma_p \leftrightarrow \bar{\varepsilon}$	$a \leftrightarrow \bar{\varepsilon}$	$a \leftrightarrow \rho$	$a \leftrightarrow b$	$a_p \leftrightarrow b_p$	$a_p \leftrightarrow c_p$	$b \leftrightarrow \bar{\varepsilon}$	$b \leftrightarrow \rho$	$b_p \leftrightarrow c_p$	$c \leftrightarrow \bar{\varepsilon}$	$c \leftrightarrow \rho$
ϵ_{SE}	+0.28	+0.81	+0.34	-0.00	+0.52	+0.13	+0.32	+0.09	+0.09	+0.50	+0.17	+0.09	+0.57	+0.21
ϵ_{CE}	+0.02	+0.17	-0.06	+0.22	+0.09	-0.10	-0.00	+0.69	+0.44	+0.07	-0.05	+0.44	+0.10	-0.14
Δ	-0.27	-0.64	-0.40	+0.22	-0.44	-0.22	-0.32	+0.59	+0.35	-0.43	-0.22	+0.35	-0.47	-0.35

(b) CE ⇒ SE

	$N_{\text{atoms}} \leftrightarrow V$	$V \leftrightarrow \bar{\varepsilon}$	$V \leftrightarrow \rho$	$\gamma_p \leftrightarrow \bar{\varepsilon}$	$a \leftrightarrow \bar{\varepsilon}$	$a \leftrightarrow \rho$	$a \leftrightarrow b$	$a_p \leftrightarrow b_p$	$a_p \leftrightarrow c_p$	$b \leftrightarrow \bar{\varepsilon}$	$b \leftrightarrow \rho$	$b_p \leftrightarrow c_p$	$c \leftrightarrow \bar{\varepsilon}$	$c \leftrightarrow \rho$
ϵ_{CE}	+0.02	+0.17	-0.06	+0.22	+0.09	-0.10	-0.00	+0.69	+0.44	+0.07	-0.05	+0.44	+0.10	-0.14
ϵ_{SE}	+0.28	+0.81	+0.34	-0.00	+0.52	+0.13	+0.32	+0.09	+0.09	+0.50	+0.17	+0.09	+0.57	+0.21
Δ	+0.27	+0.64	+0.40	-0.22	+0.44	+0.22	+0.32	-0.59	-0.35	+0.43	+0.22	-0.35	+0.47	+0.35

Table 3: Largest shifts in *error–error* correlation coefficients when transferring between SE and CE annotation protocols. Each sub-table displays the top 14 property pairs (ordered alphabetically) exhibiting the largest absolute changes in pairwise correlation, averaged over all models, materials, and $R7$–$R10$. Panel (a) shows the shift from SE to CE ($\Delta = \rho_{\text{CE}} - \rho_{\text{SE}}$), while panel (b) shows the reverse (CE to SE, $\Delta = \rho_{\text{SE}} - \rho_{\text{CE}}$). For each property pair, the table reports the correlation coefficients under each protocol and their difference Δ. Cells are color-coded: green for positive Δ (stronger coupling under the target protocol) and red for negative Δ (weaker coupling), highlighting which structural or physical property relationships are most sensitive to the choice of annotation protocol.

Material	Physical Law Compliance	Hallucination Score
Ag	<u>0.82</u> ± 0.03	0.21 ± 0.04
Au	**0.84** ± 0.03	0.24 ± 0.02
$CH_3NH_3PbI_3$	0.72 ± 0.03	0.42 ± 0.05
Fe_2O_3	0.74 ± 0.03	0.23 ± 0.02
MoS_2	0.78 ± 0.03	**0.18** ± 0.01
PbS	0.77 ± 0.03	0.53 ± 0.02
SnO_2	0.74 ± 0.03	0.24 ± 0.04
$SrTiO_3$	0.77 ± 0.02	0.28 ± 0.03
TiO_2	0.46 ± 0.02	0.43 ± 0.03
ZnO	0.77 ± 0.02	<u>0.21</u> ± 0.02

Table 4: Mean ± std physical-law compliance and hallucination scores for each material, averaged over all models and five runs per material–radius under both SE and CE protocols. Physical-law compliance measures adherence to fundamental structural constraints (e.g., density and lattice-parameter ratios), while the hallucination score quantifies the frequency of non-physical or highly erroneous predictions across a set of key properties. **Bold** denotes the material with the highest prediction accuracy, while <u>underlining</u> denotes the material with the second highest accuracy.

to produce physically plausible and trustworthy predictions at the atomic scale.

Model Latency. Table 2 presents the average inference latencies per sample across the SE and CE protocols. Gemini 2.5 Flash exhibits the lowest latency, requiring only 3.06 s under SE and 5.00 s under CE, making it well-suited for time-sensitive applications such as high-throughput materials screening. Llama-4 Maverick and GPT-4.1 Mini follow in the next performance tier with moderate latency (4 s to 8 s), while most other models cluster between 6 s to 15 s. DeepSeek-Chat is the slowest model in the SE evaluation (25 s), and Mistral Medium 3 exhibits the highest latency in CE (15.5 s). These trends broadly correlate with model size and architecture, where larger context windows and multimodal inputs tend to incur higher computational overhead. Although latency is not the primary evaluation criterion in this study, the results offer practical insights for downstream deployment scenarios, especially when balancing predictive accuracy against throughput constraints.

6 Limitations

This study isolates two complementary generalization regimes—geometric interpolation/extrapolation and chemical extrapolation—using a curated dataset of ten crystalline materials across four radii. While representative, this selection captures only a limited region of compositional and structural diversity present in real-world materials. All models are evaluated in a zero-shot setting with default decoding configurations, without fine-tuning, retrieval augmentation, or domain adaptation, which may underrepresent their full capabilities.

Evaluation emphasizes first-order structural properties such as lattice constants, density, and stoichiometry, along with a single volumetric consistency index. Higher-order descriptors—including phonon spectra, band topology, or symmetry-preserving deformations—are not considered. The analysis focuses on static prediction quality and does not measure model responsiveness to feedback, learning curves under domain supervision, or variance across decoding seeds.

7 Conclusion

This work introduces *xCrysAlloys* and its two complementary benchmarks—SE and CE—that isolate geometric interpolation and chemical extrapolation in crystallographic prediction. The evaluations reveal that current vision–language foundation models struggle to internalize core physical principles, as evidenced by high relative errors, substantial degradation in transfer settings, and disrupted inter-property correlations. The prevalence of hallucinated outputs and violations of basic physical laws further underscores the limitations of purely data-driven training in scientific domains.

To advance reliability and generalization, future models must incorporate explicit physical constraints, symmetry priors, and uncertainty-aware reasoning. The proposed benchmarks provide a reproducible and physically grounded testbed for evaluating model robustness in structured scientific settings. By bridging multimodal language understanding with domain-specific inductive biases, this work aims to foster the development of more trustworthy foundation models for materials science and beyond.

References

Sunandan Baruah and Joydeep Dutta. 2009. Hydrothermal growth of zno nanostructures. *Science and technology of advanced materials*, 10(1):013001.

W. H. Baur, R. A. Sass, et al. 1971. The rutile structure of sno_2. *Acta Crystallographica Section B*, 27:2133.

Alberto Bertucco and Gerhard Vetter. 2001. *High pressure process technology: fundamentals and applications*, volume 9. Elsevier.

HKDH Bhadeshia. 2001. *Geometry of crystals*, volume 8. Institute of Materials London.

Jan-Otto Carlsson and Peter M Martin. 2010. Chemical vapor deposition. In *Handbook of Deposition Technologies for films and coatings*, pages 314–363. Elsevier.

Stefano Curtarolo, Wahyu Setyawan, Gus LW Hart, Michal Jahnatek, Roman V Chepulskii, Richard H Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, et al. 2012. Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226.

Alexandre Agm Duval, Victor Schmidt, Alex Hernández-Garcıa, Santiago Miret, Fragkiskos D Malliaros, Yoshua Bengio, and David Rolnick. 2023. Faenet: Frame averaging equivariant gnn for materials modeling. In *International Conference on Machine Learning*, pages 9013–9033. PMLR.

L. W. Finger and R. M. Hazen. 1980. Crystal structure and isothermal compression of fe_2o_3, cr_2o_3, and v_2o_3 to 50 kbars. *Journal of Applied Physics*, 51:5362–5367.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. 2020. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*.

Michael Gaus, Qiang Cui, and Marcus Elstner. 2011. Dftb3: Extension of the self-consistent-charge density-functional tight-binding method (scc-dftb). *Journal of Chemical Theory and Computation*, 7(4):931–948.

R. Grau-Crespo and R. Lopez-Cordero. 2002. Mos_2 structural properties. *Phys. Chem. Chem. Phys.*, 4:4078.

Sydney R Hall, Frank H Allen, and I David Brown. 1991. The crystallographic information file (cif): a new standard archive file for crystallography. *Foundations of Crystallography*, 47(6):655–685.

M. Horn, C. R. Meagher, et al. 1972. Structure of anatase tio_2. *Zeitschrift für Kristallographie*, 136:273.

B Hourahine, S Sanna, B Aradi, C Köhler, Th Niehaus, and Th Frauenheim. 2007. Self-interaction and strong correlation in dftb. *The Journal of Physical Chemistry A*, 111(26):5671–5677.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1).

Daniel S Jensen and Adam Wasserman. 2018. Numerical methods for the inverse problem of density functional theory. *International Journal of Quantum Chemistry*, 118(1):e25425.

H. W. King. 2002a. *CRC Handbook of Chemistry and Physics*, 83 edition. CRC Press. Standard phase data for silver (Ag).

H. W. King. 2002b. *CRC Handbook of Chemistry and Physics*, 83 edition. CRC Press. Standard phase data for gold (Au).

Walter Kohn and Lu Jeu Sham. 1965. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133.

Mustafa Kurban, Can Polat, Erchin Serpedin, and Hasan Kurban. 2024. Enhancing the electronic properties of tio2 nanoparticles through carbon doping: An integrated dftb and computer vision approach. *Computational Materials Science*, 244:113248.

Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D

White, and Samuel G Rodriques. 2024. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*.

Andrea C Levi and Miroslav Kotrla. 1997. Theory and simulation of crystal growth. *Journal of Physics: Condensed Matter*, 9(2):299.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. 2025. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1, pages 415–423.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. 2023. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

VA Lubarda. 2003. On the effective lattice parameter of binary alloys. *Mechanics of materials*, 35(1-2):53–68.

Santiago Miret and Nandan M Krishnan. 2024. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*.

R. H. Mitchell and M. A. Carpenter. 2000. Physics and chemistry of $srtio_3$ perovskites. *Physics and Chemistry of Minerals*, 27:583.

DA Papaconstantopoulos and MJ Mehl. 2003. The slater–koster tight-binding method: a computationally efficient and accurateapproach. *Journal of Physics: Condensed Matter*, 15(10):R413.

Can Polat, Hasan Kurban, Erchin Serpedin, and Mustafa Kurban. 2025a. Tdcm25: A multi-modal multi-task benchmark for temperature-dependent crystalline materials. In *AI for Accelerated Materials Design-ICLR 2025*.

Can Polat, Mehmet Tuncel, Hasan Kurban, Erchin Serpedin, and Mustafa Kurban. 2025b. xchemagents: Agentic ai for explainable quantum chemistry. *arXiv preprint arXiv:2505.20574*.

Nitin Rane. 2023. Transformers in material science: roles, challenges, and future scope. *Challenges and Future Scope (March 26, 2023)*.

James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. 2013. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65:1501–1509.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. 2017. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems*, 30.

Richard P Stanley. 1975. The fibonacci lattice. *The Fibonacci Quarterly*, 13(3):215–232.

Yingheng Tang, Wenbin Xu, Jie Cao, Weilu Gao, Steve Farrell, Benjamin Erichson, Michael W Mahoney, Andy Nonaka, and Zhi Yao. 2025. Matterchat: A multi-modal llm for material science. *arXiv preprint arXiv:2502.13107*.

Aron Walsh, elds22, Federico Brivio, and Jarvist Moore Frost. 2019. Wmd-group/hybrid-perovskites: Collection 1 (v1.0). https://doi.org/10.5281/zenodo.2641358. Hybrid perovskite $CH_3NH_3PbI_3$ structural data.

Geemi Wellawatte, Huixuan Guo, Magdalena Lederbauer, Anna Borisova, Matthew Hart, Marta Brucka, and Philippe Schwaller. 2024. Chemlit-qa: A human evaluated dataset for chemistry rag tasks. In *AI for Accelerated Materials Design-NeurIPS 2024*.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530.

R. W. G. Wyckoff. 1963a. *Crystal Structures Volume 1*. Interscience Publishers.

R. W. G. Wyckoff. 1963b. *Crystal Structures Volume 1*. Interscience Publishers.

R. W. G. Wyckoff. 1963c. *Crystal Structures Volume 1*. Interscience Publishers.

Xiaolong Zheng, Peng Zheng, and Rui-Zhi Zhang. 2018. Machine learning material properties from the periodic table using convolutional neural networks. *Chemical Science*, 9(44):8426–8432.

A Appendix

A.1 Crystal Parameters

Silver (Ag). Silver adopts an FCC lattice with lattice constant $a = 4.0857$ Å. The cubic crystal belongs to space group Fm$\overline{3}$m (No. 225), Pearson symbol cF4, and Schoenflies notation O_h^5. A single Ag atom occupies the origin of the primitive cell (King, 2002a).

Gold (Au). Gold similarly adopts an FCC arrangement with lattice constant $a = 4.0782$ Å. It crystallizes in space group Fm$\overline{3}$m (No. 225), reflecting equivalent high symmetry. One Au atom resides at the (0,0,0) position within the unit cell (King, 2002b).

Methylammonium Lead Iodide ($CH_3NH_3PbI_3$). The hybrid perovskite $CH_3NH_3PbI_3$ forms a pseudo-cubic lattice with parameters $a = 6.290$ Å, $b = 6.274$ Å, $c = 6.297$ Å and angles close to 90°. It crystallizes in space group $P1$ (No. 1), accommodating slight distortions and dynamic disorder typical of organic–inorganic frameworks (Walsh et al., 2019).

Hematite (Fe_2O_3). Hematite (Fe_2O_3) exhibits a rhombohedral structure with lattice constants $a = b = 5.0346$ Å, $c = 13.7473$ Å, and angles $\alpha = \beta = 90°$, $\gamma = 120°$. It belongs to space group $R\overline{3}c$ (No. 167), underpinning its antiferromagnetic and catalytic properties (Finger and Hazen, 1980).

Molybdenum Disulfide (MoS_2). Molybdenum disulfide (MoS_2) adopts a layered hexagonal lattice with parameters $a = 3.1604$ Å, $c = 12.295$ Å, and angles $\alpha = \beta = 90°$, $\gamma = 120°$. It crystallizes in space group $P6_3/mmc$ (No. 194), reflecting its van der Waals–bonded layers (Wyckoff, 1963b; Grau-Crespo and Lopez-Cordero, 2002).

Galena (PbS). Galena (PbS) forms a rock-salt–type FCC structure with lattice constant $a = 5.9362$ Å. The cubic crystal belongs to space group Fm$\overline{3}$m (No. 225), with Pb and S atoms occupying alternating FCC sites (Wyckoff, 1963c).

Cassiterite (SnO_2). Cassiterite (SnO_2) displays a tetragonal rutile–type lattice with constants $a = 4.738$ Å, $c = 3.1865$ Å. It crystallizes in space group $P4_2/mnm$ (No. 136) and features an oxygen sublattice coordinating the Sn atoms (Baur et al., 1971).

Strontium Titanate ($SrTiO_3$). Strontium titanate ($SrTiO_3$) crystallizes in a cubic perovskite structure with lattice constant $a = 3.9053$ Å and space group $Pm\overline{3}m$ (No. 221). Its ideal symmetry underlies its prototypical ferroelectric and quantum paraelectric behavior (Mitchell and Carpenter, 2000).

Titanium Dioxide (TiO_2—Anatase). Anatase TiO_2 exhibits a body-centered tetragonal structure with $a = 3.7842$ Å, $c = 9.5146$ Å. It belongs to space group $I4_1/amd$ (No. 141), characteristic of the anatase polymorph's photocatalytic activity (Horn et al., 1972).

Zinc Oxide (ZnO—Zincite). Zinc oxide (ZnO) in the zincite phase adopts a hexagonal wurtzite lattice with parameters $a = 3.2495$ Å, $c = 5.2069$ Å and space group $P6_3mc$ (No. 186). This polar structure underpins its piezoelectric and optoelectronic applications (Wyckoff, 1963a).

A.2 Additional Metric Definitions

Absolute-error (angles). For each primitive-cell angle $\theta_p \in \{\alpha_p, \beta_p, \gamma_p\}$,

$$|\Delta\theta_p| = \left|\theta_p^{\text{gen}} - \theta_p^{\text{ref}}\right|.$$

Per-example mean error. If an example contains the set of properties P, then

$$\overline{\%\Delta} = \frac{1}{|P|}\sum_{p\in P} \%\Delta_p.$$

Format faithfulness. Let $\mathcal{F}_{\text{ref}}$ and $\mathcal{F}_{\text{gen}}$ be the non-null field sets, and $\mathcal{F}_\cap = \mathcal{F}_{\text{ref}} \cap \mathcal{F}_{\text{gen}}$. The following definitions are considered:

$$S_{\text{presence}} = \frac{|\mathcal{F}_\cap|}{|\mathcal{F}_{\text{ref}}|}$$

$$S_{\text{type}} = \frac{1}{|\mathcal{F}_\cap|}\sum_{f\in\mathcal{F}_\cap} \mathbf{1}\big(\text{type}_{\text{gen}}(f) = \text{type}_{\text{ref}}(f)\big),$$

and

$$S_{\text{format}} = 0.7\, S_{\text{presence}} + 0.3\, S_{\text{type}}.$$