CRYSTALSAE: INTERPRETING CRYSTAL DIFFUSION MODELS THROUGH SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Crystal diffusion models perform well in crystal structure prediction, but their internal representations remain poorly understood. To improve model interpretability and inform controllable generation, we propose CRYSTALSAE, a sparse autoencoder (SAE) framework for identifying and probing crystallographic features in crystal diffusion models. Specifically, we apply SAEs to the layer-wise residual updates of DiffCSP, a representative crystal diffusion model, and relate the resulting feature activations to crystallographic concepts. We then intervene along selected feature directions to examine their influence on generation. Extensive experiments on the MP-20 dataset confirm that sparse features selected on training data retain statistically significant associations with 17 crystallographic concepts on held-out data. Crystal visualizations further illustrate associations between selected features and 68 space groups. Moreover, our intervention experiments show that stronger perturbations of selected features produce larger average changes in the lattice parameters and density of generated crystals. These findings reveal crystallographically aligned features within DiffCSP and suggest a path toward more interpretable and controllable crystal generation.

1 INTRODUCTION

Crystal structure prediction (CSP) plays an important role in materials discovery, as lattice geometry and atomic arrangements strongly influence material properties. Diffusion models have advanced crystal structure prediction and generation by learning structural distributions from known crystals (Xie et al., 2022; Jiao et al., 2023; 2024; Zeni et al., 2025; Lin et al., 2024; 2026). Despite their predictive success, the internal representations supporting crystal generation remain poorly understood.

Studies of language and image models show that interpreting internal representations can support model diagnosis and controllable generation (Marks et al., 2025; Surkov et al., 2025). Crystallographic concepts also guide the search for functional inorganic materials. For example, structural similarities to cuprates, including layered structures and square-planar coordination, motivated the search for nickelate superconductors (Li et al., 2019). Prior work incorporating explicit space-group and Wyckoff-position constraints has improved the generation of structures that match known crystals, highlighting the value of crystallographic priors (Jiao et al., 2024). These results motivate examining whether crystal diffusion models learn recognizable crystallographic concepts within their internal representations. However, standard CSP benchmarks evaluate generated structures without revealing the internal computations that produce them. Moreover, lattice geometry, density, coordination, and symmetry are coupled, so individual features may reflect several properties. Interpreting these representations therefore requires both associating internal features with crystallographic concepts and testing their influence on generation.

Sparse autoencoders (SAEs) decompose dense activations into sparse combinations of learned directions, revealing interpretable features in language, protein, and image diffusion models (Bricken et al., 2023; Cunningham et al., 2024; Gao et al., 2025; Simon & Zou, 2025; Adams et al., 2025; Surkov et al., 2025). For crystal diffusion models, it remains unclear whether SAE features exhibit crystallographic associations that generalize to unseen crystals and whether intervening on these features affects generated structures.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce CRYSTALSAE, an SAE framework for identifying and probing crystallographic features in DiffCSP, a typical crystal diffusion model. We train Top-K SAEs on layer-wise residual updates and identify concept-associated features using training data, fixing feature identities and orientations before held-out validation. We examine feature locations across layers, track crystal-system activations during denoising, and visualize associations with selected space groups. Finally, we intervene along selected feature directions in the frozen model to test their influence on generated structures.

Our main contributions are summarized as follows:

- **SAE-based interpretation of crystal generative models.** To our knowledge, we are the first to apply sparse autoencoders to interpret crystal generative models, combining feature discovery with crystallographic analysis and functional intervention.
- **Crystallographic feature discovery and validation.** We identify features associated with seventeen crystallographic concepts on MP-20. All seventeen associations remain statistically significant on held-out data after multiple-testing correction.
- **Functional effects of feature interventions.** We show that feature interventions produce statistically significant changes in seven continuous properties: the three lattice lengths, three lattice angles, and density. Larger perturbations produce larger average effects within the tested range.

2 RELATED WORK

Crystal structure prediction. Classical CSP explores potential-energy landscapes with global methods such as evolutionary optimization, random structure search, particle-swarm methods, and basin-search procedures (Oganov & Glass, 2006; Glass et al., 2006; Pickard & Needs, 2011; Wang et al., 2012; Goedecker, 2004). Data-driven CSP increasingly replaces or augments explicit search with learned generative models over periodic structures.CDVAE introduced latent-variable generation for crystals (Xie et al., 2022); DiffCSP jointly diffuses fractional coordinates and lattice variables under periodic equivariance (Jiao et al., 2023); and DiffCSP++ incorporates explicit space-group and Wyckoff constraints (Jiao et al., 2024). Subsequent work has expanded crystal generation through Riemannian flow matching, autoregressive modeling, and large-scale conditional generation (Miller et al., 2024; Sriram et al., 2024; Antunes et al., 2024; Zeni et al., 2025). More recent methods increasingly emphasize crystallographic symmetry: SymmCD models asymmetric units together with symmetry transformations (Levy et al., 2025), WyckoffDiff performs discrete generation in a symmetry-adapted Wyckoff representation (Ekström Kelvinius et al., 2025), and SGEquiDiff incorporates space-group-equivariant diffusion dynamics (Chang et al., 2025). Recent alternatives further explore unified diffusion transformers, symmetry-aware Bayesian flows, and template-guided generation with energy-based refinement (Yi et al., 2026; Ruple et al., 2026; Yang et al., 2026). These methods are primarily evaluated by structural validity, stability, distribution matching, novelty, and target recovery. CRYSTALSAE instead examines which internal directions align with crystallographic concepts and how the frozen model responds to controlled edits.

Sparse autoencoders. SAEs build on sparse coding and k-sparse representation learning (Olshausen & Field, 1996; Makhzani & Frey, 2013) and are now used to separate features superposed in dense neural representations (Elhage et al., 2022; Bricken et al., 2023). Recent work develops Top-K, gated, and JumpReLU variants, studies scaling laws, releases broad open SAE suites, and benchmarks whether proxy reconstruction metrics translate into interpretability or causal utility (Cunningham et al., 2024; Gao et al., 2025; Rajamanoharan et al., 2024a;b; Lieberum et al., 2024; Karvonen et al., 2025). Sparse features have also been used to construct and edit causal circuits (Marks et al., 2025). Scientific applications now analyze protein language models with SAEs (Villegas Garcia & Ansuini, 2025; Adams et al., 2025; Simon & Zou, 2025; Liu et al., 2026), while diffusion-model work shows that factorizing residual updates can yield temporally localized and steerable concepts (Surkov et al., 2025). CRYSTALSAE adapts this residual-update view to periodic crystal diffusion and combines held-out crystallographic associations with direct feature interventions.

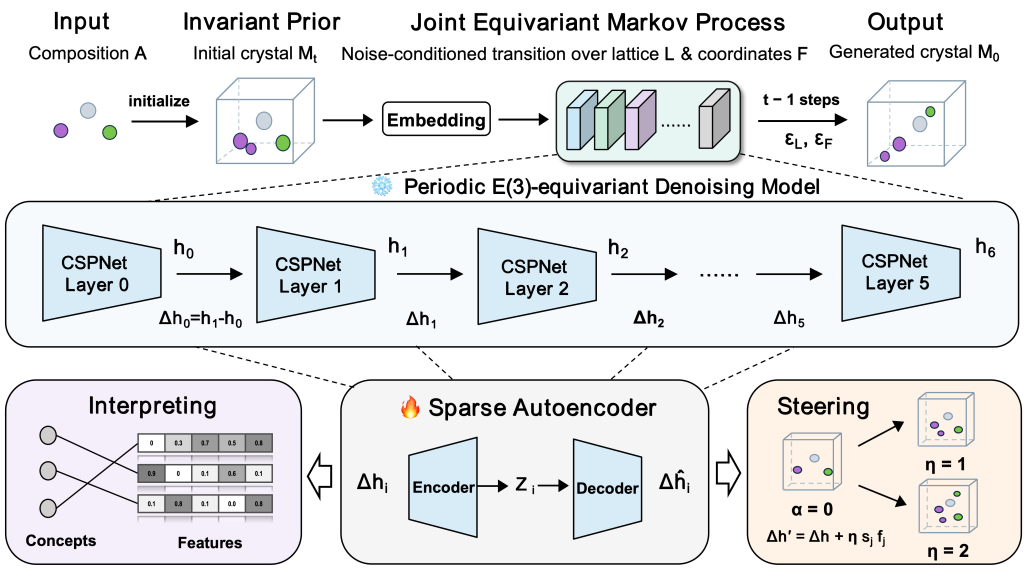


Figure 1: Overview of CrystalSAE. DiffCSP generates a crystal through a joint equivariant diffusion process over lattice variables and fractional coordinates. CrystalSAE factorizes the residual updates of the six CSPNet layers with layer-wise sparse autoencoders. The learned sparse features are associated with crystallographic concepts and analyzed across network depth and diffusion time, while selected decoder directions are perturbed to functionally probe the frozen generator.

3 Proposed Method

3.1 CrystalSAE

Interpreting crystal structure prediction models is important because accurate generated structures alone do not reveal which crystallographic information guides generation or how that information is transformed inside the denoiser. Following DiffCSP (Jiao et al., 2023), a crystal is represented as $\mathcal{M} = (\mathbf{A}, \mathbf{F}, \mathbf{L})$, where $\mathbf{A}$ is the composition, $\mathbf{F}$ is the fractional-coordinate matrix, and $\mathbf{L}$ is the lattice. Composition-conditioned CSP models $p(\mathbf{L}, \mathbf{F} \mid \mathbf{A})$ through joint diffusion over lattice variables and fractional coordinates. A periodic E(3)-equivariant denoiser progressively refines the noisy crystal state. As illustrated in Figure 1, CrystalSAE keeps this pretrained generator frozen and analyzes the residual transformations produced by its six CSPNet message-passing layers.

For layer $i \in \{0, \ldots, 5\}$, the residual update is

$$\Delta \boldsymbol{h}_i = \boldsymbol{h}_{i+1} - \boldsymbol{h}_i. \tag{1}$$

Rather than factorizing the accumulated hidden state, CrystalSAE decomposes each $\Delta \boldsymbol{h}_i$ so that the learned features describe the transformation introduced by a single CSPNet block. An independent Top-K sparse autoencoder is trained for each layer,

$$\boldsymbol{z}_i = \mathrm{ReLU}\Big[\mathrm{TopK}\Big(W_{\mathrm{enc}}^{(i)} \Delta \boldsymbol{h}_i + \boldsymbol{b}_{\mathrm{act}}^{(i)}; K\Big)\Big], \qquad \widehat{\Delta \boldsymbol{h}}_i = W_{\mathrm{dec}}^{(i)} \boldsymbol{z}_i, \tag{2}$$

with $\|\boldsymbol{z}_i\|_0 \le K$. Training minimizes residual reconstruction error together with the AuxK objective,

$$\mathcal{L}_{\mathrm{SAE}}^{(i)} = \mathbb{E}\left[\left\|\Delta \boldsymbol{h}_i - \widehat{\Delta \boldsymbol{h}}_i\right\|_2^2\right] + \lambda_{\mathrm{aux}} \mathcal{L}_{\mathrm{AuxK}}^{(i)}. \tag{3}$$

Layer-wise preprocessing and optimization details are provided in Appendix B. The resulting sparse codes $\boldsymbol{z}_i$ define the feature space used below to identify crystallographic concepts and study their organization.

3.2 Concept–Feature Discovery

The sparse codes learned above become interpretable only if individual features can be linked to meaningful crystallographic variables, so we first identify concept-associated directions and then keep

them fixed when analyzing where and when they are expressed. The analysis considers seventeen pre-specified concepts: eight continuous descriptors $(a, b, c, \alpha, \beta, \gamma, V/N, \rho)$, seven crystal systems, and CN$=4$ and CN$=6$. A separate exploratory atlas covers sixty-eight space-group labels. For crystal-level concepts, feature activations are averaged across residual rows belonging to the same crystal, while coordination concepts retain site-level activations and use grouped resampling by crystal.

For concept q, CSPNet layer i, and SAE feature j, a training-only score M_{qij} measures the association between feature activation and the crystallographic target. Continuous concepts are ranked by squared Pearson correlation and evaluated using one-dimensional R^2, while binary crystal-system and coordination concepts use AUPRC as the primary metric and AUROC as a secondary diagnostic. For each concept, the strongest layer–feature pair $(i_q^*, j_q^*) = \arg\max_{i,j} M_{qij}$ is selected using training data only. Its feature identity and orientation are then frozen before evaluation on the 9,047-crystal validation split. Held-out confirmation uses permutation tests followed by Benjamini–Hochberg correction across the seventeen pre-specified concepts. With 256 permutations, the smallest attainable Monte Carlo value is $1/257 \approx 0.0039$, so floor-reaching results are reported as $p_{\mathrm{perm}} \leq 0.0039$. Validation and test data never reselect a feature, and the exploratory space-group atlas follows the same training-only search.

The locked pair (i_q^*, j_q^*) also identifies the layer at which concept q has its strongest sparse association. Shared assignments are retained because geometrically coupled crystallographic variables may align with the same direction. To characterize temporal organization, the same frozen feature j_q^* is tracked across the saved reverse-diffusion steps without re-selection. Changes in activation therefore reflect changes in feature expression rather than changes in feature identity. Temporal peaks indicate when a selected direction is most active, but are not interpreted as the exact time at which a crystallographic concept is created or becomes uniquely decodable. These frozen concept-associated directions are then used for the intervention analysis below.

3.3 Feature Intervention

Concept–feature association shows that a sparse direction carries information about a crystallographic variable, but it does not show whether that direction can influence the generated structure, so we directly perturb the frozen directions inside DiffCSP. For feature j in layer i, the corresponding decoder column is mapped back to native residual units as $\boldsymbol{f}_j^{(i)} = \sigma_i W_{\mathrm{dec}}^{(i)}[:, j]$, where σ_i is the layer-specific scale used in SAE preprocessing. The residual update is then modified as

$$\Delta \boldsymbol{h}_i' = \Delta \boldsymbol{h}_i + \eta \, s_j^{\mathrm{train}} \boldsymbol{f}_j^{(i)}, \qquad \eta \in \{0, 1, 2\}, \tag{4}$$

where s_j^{train} is the training-set standard deviation of feature j. The edited residual is propagated through the remaining frozen CSPNet layers to obtain the final hidden representation $\boldsymbol{h}_6'$. The original lattice and coordinate denoising heads are then evaluated as

$$\hat{\epsilon}_L' = \mathbf{L}_t \, \varphi_L\left(N^{-1} \sum_{n=1}^{N} \boldsymbol{h}_{6,n}'\right), \qquad \hat{\epsilon}_F'[:, n] = \varphi_F\left(\boldsymbol{h}_{6,n}'\right), \tag{5}$$

which define the modified reverse transition

$$(\hat{\epsilon}_L', \hat{\epsilon}_F') \Longrightarrow p'(\mathcal{M}_{t-1} \mid \mathcal{M}_t). \tag{6}$$

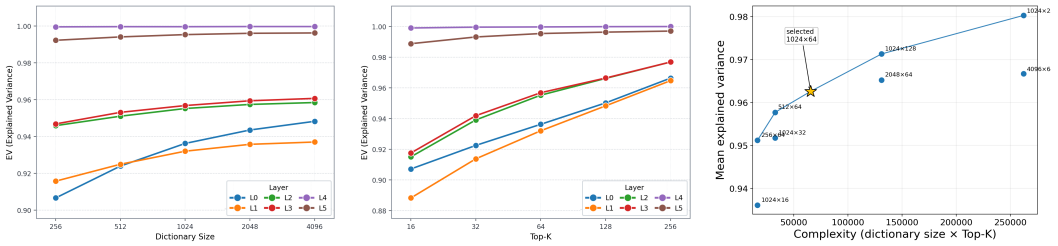
The intervention therefore acts only through the frozen DiffCSP computation. Crystallographic descriptors are measured from the final generated crystal and are never overwritten directly, and baseline and steered runs use matched initial states.

For continuous concepts, the endpoint is the corresponding descriptor of the generated crystal. For crystal system s, we use the tolerance set $\Theta = \{0.5^\circ, 1^\circ, 2^\circ, 3^\circ, 5^\circ\}$ and $\tau_0 = 0.05$ Å to define the continuous symmetry-robustness score

$$O_s = |\Theta|^{-1} \sum_{\theta \in \Theta} \exp[-\tau_{\mathrm{crit}}(s, \theta)/\tau_0], \tag{7}$$

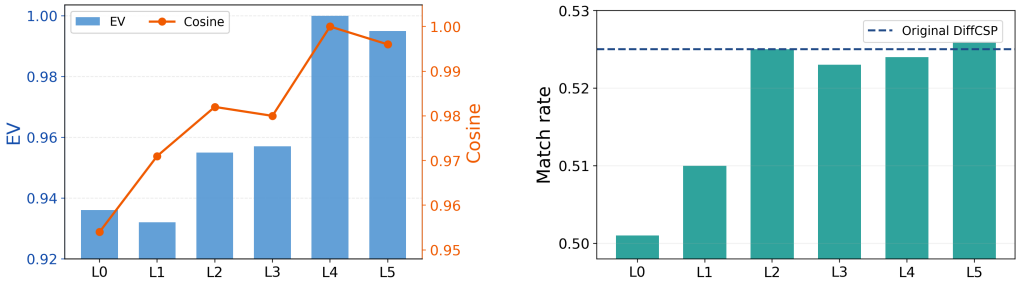
where $\tau_{\mathrm{crit}}(s, \theta)$ is the smallest positional symmetry tolerance at which the generated crystal is recovered as system s. For coordination, CrystalNN site weights $w_n(k)$ define

$$O_{\mathrm{CN}=k} = N^{-1} \sum_{n=1}^{N} w_n(k). \tag{8}$$



(a) **(b)** **(c)**

Figure 2: SAE hyperparameter sweep. (a) Layer-wise explained variance as dictionary size varies. (b) Layer-wise explained variance as Top-K varies. (c) Mean explained variance across layers versus the simple complexity proxy dictionary size$\times$Top-K; the star marks the selected 1024×64 operating point.



(a) **(b)**

Figure 3: Representation- and generator-level reconstruction fidelity. (a) Residual-reconstruction explained variance (bars) and mean row-wise cosine similarity (line) for L0–L5. (b) Structure match rate after reinjecting the SAE reconstruction at each layer; the dashed horizontal line is the unmodified DiffCSP match rate.

A response is considered dose-consistent when the mean changes at $\eta = 1$ and $\eta = 2$ have the same nonzero sign and the response at the larger dose has at least the same absolute magnitude. Paired two-sided tests are applied at $\eta = 2$, followed by BH correction across the seventeen pre-specified endpoints. Statistical significance is used to assess the reproducibility of paired effects rather than as a measure of steering strength. These interventions provide a functional test of the discovered sparse directions without assuming that they are necessary, sufficient, or causally independent.

4 Experiments

4.1 Experimental Setup

Dataset and Split. Experiments use the released composition-conditioned DiffCSP checkpoint and the official MP-20 split (Jiao et al., 2023; Jain et al., 2013), containing 27,136 training crystals, 9,047 validation crystals, and 9,046 test crystals. DiffCSP contains six CSPNet residual streams (L0–L5), each with hidden dimension $d = 512$, and one SAE is trained independently for each layer. The training split is used for SAE fitting, normalization, and concept–feature discovery. The validation split is used for held-out confirmation and representation analysis, while the test split is reserved for steering experiments.

Hyperparameter Selection. We first select the SAE operating point by sweeping dictionary size and Top-K sparsity. Figure 2(a) shows that explained variance generally improves as the dictionary becomes larger, with the largest gains occurring in the earlier layers and diminishing improvements at larger dictionary sizes. Figure 2(b) shows a similar trend as K increases: reconstruction improves across layers, but the gains become smaller beyond moderate sparsity levels. The later layers, especially L4 and L5, are already reconstructed with very high explained variance and are less sensitive to either hyperparameter. Figure 2(c) summarizes this trade-off using dictionary size$\times$Top-K as a simple complexity proxy. The 1024×64 configuration lies near the knee of the performance–

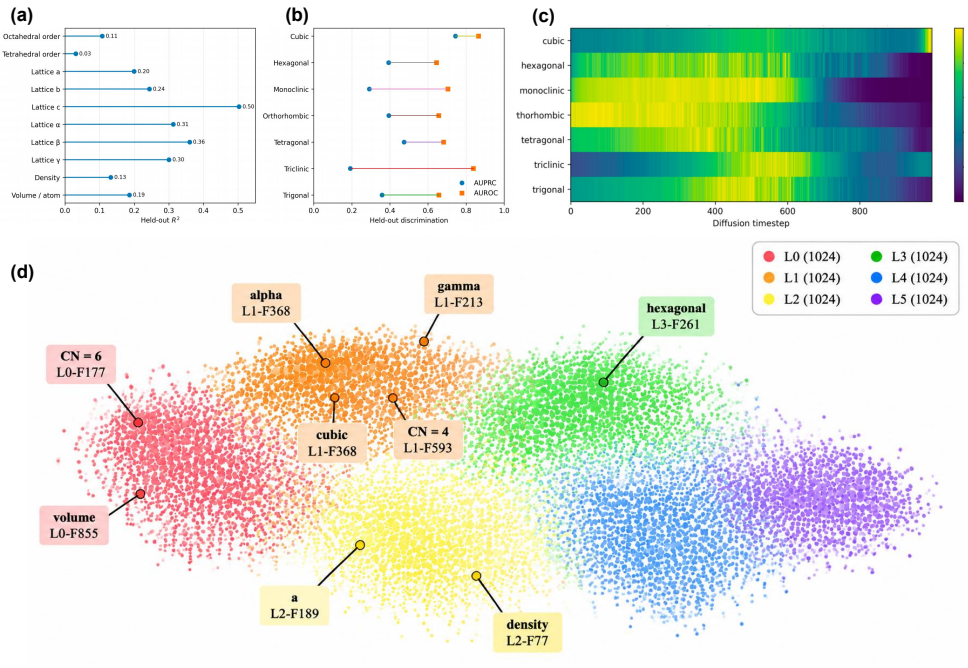


Figure 4: Concept–feature organization. (a) Held-out one-feature R^2 for continuous crystallographic descriptors and coordination-order scores. (b) Held-out AUPRC and AUROC for the seven crystal systems. (c) Activation profiles of the seven frozen crystal-system features across reverse-diffusion steps. (d) Two-dimensional visualization of the six SAE dictionaries with selected concept-associated features annotated; distances in the projection are not treated as quantitative evidence of semantic similarity.

complexity curve, providing strong reconstruction while avoiding the substantially higher cost of larger settings. We therefore use $m = 1024$ and $K = 64$ in all main experiments.

Reconstruction Fidelity. Before interpreting the learned sparse features, we examine whether the selected SAEs faithfully reconstruct the CSPNet residual streams. Figure 3(a) shows a clear layer-dependent trend: reconstruction quality generally improves toward the later CSPNet layers. L0–L3 obtain slightly lower explained variance and cosine similarity, whereas L4–L5 are reconstructed almost perfectly. One possible explanation is that later residual blocks operate on more refined representations and may therefore contain more regular structure that is easier to capture with a sparse dictionary. Figure 3(b) provides a complementary generator-level check by reinjecting the reconstructed residuals into DiffCSP. The same depth-dependent pattern is visible: L2–L5 remain close to the unmodified DiffCSP match rate, while L0–L1 show a larger decrease. This suggests that early-layer computation is more sensitive to reconstruction error, whereas later-layer approximations are better tolerated by the downstream generator. Overall, the representation- and generator-level results show that the selected SAEs provide high-fidelity reconstructions across all six layers.

4.2 Interpreting Response

Held-out Validation. Interpreting Response tests whether sparse directions discovered on the training split remain aligned with crystallographic variables on held-out data, and then examines where and when these directions are expressed in the frozen denoiser. Feature identities and orientations are selected from 27,136 training crystals and fixed before evaluation on the 9,047-crystal validation split. All seventeen locked concept–feature pairs remain statistically detectable under the pre-specified permutation audit and survive BH correction across the seventeen concepts ($q_{\mathrm{BH17}} \leq 0.0039$). Because all permutation tests reach the 256-permutation resolution floor, we use held-out effect sizes rather than p-value differences to assess association strength. As shown in Figure 4(a,b), the continuous descriptors achieve one-feature R^2 values from 0.1325 for density to 0.5025 for lattice c, while the seven crystal systems obtain AUPRC values from 0.1922 to 0.7442. Coordination features

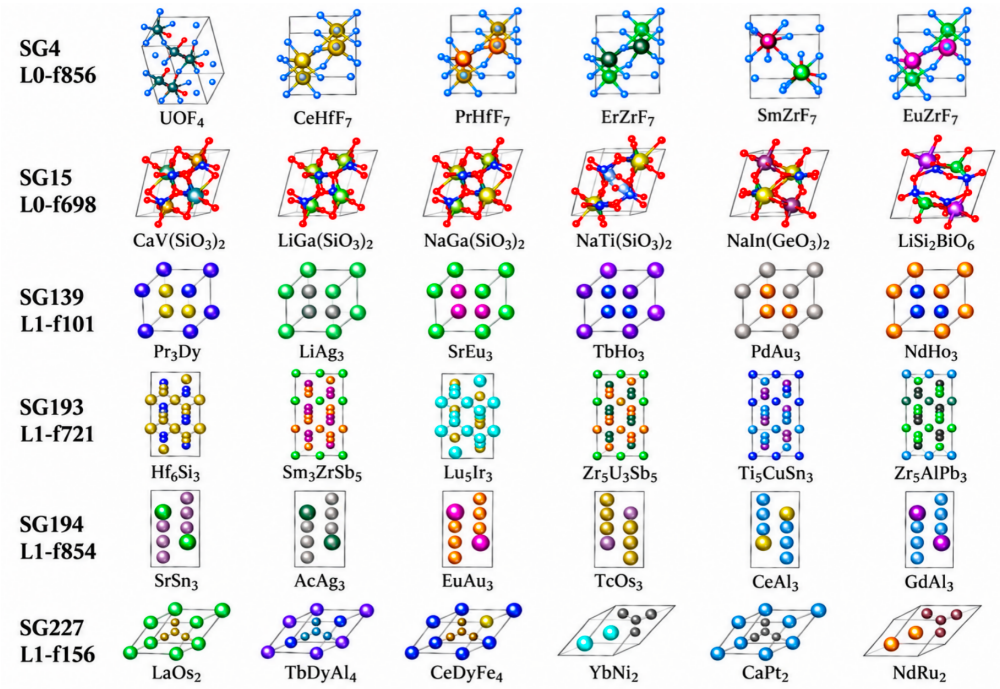


Figure 5: Exploratory high-activation structures for six space-group candidate features (SG4, SG15, SG139, SG193, SG194, and SG227). Each row shows structures with high activation of the corresponding training-discovered feature. The examples illustrate fine-grained symmetry associations but are not used as held-out confirmation of space-group semantics.

are weaker, with AUPRC values of 0.2224 and 0.3079 for CN$=4$ and CN$=6$, respectively. These results show that the train-selected sparse directions retain crystallographic information under an explicit train–validation transfer test.

Layer-wise Organization. The learned representation is neither fully separated nor concentrated in a single network block. Most concepts select distinct sparse directions, but some geometrically related variables share features. For example, β and γ both select L1-F213, while cubic and α select L1-F368. This is consistent with the geometric coupling among lattice variables. At the same time, separate directions are recovered for lattice lengths, volume per atom, density, coordination, and most crystal systems. Figure 4(d) shows that the selected features are distributed across several SAE dictionaries rather than concentrated in one terminal layer. Within the locked atlas, the strongest associations occur mainly in L0–L3, indicating that crystallographic information is already present in intermediate residual computations. The two-dimensional projection is used only for visualization; distances in this space are not interpreted as quantitative semantic similarity.

Temporal Dynamics. Figure 4(c) provides the complementary diffusion-time view. The same train-selected feature is tracked at every saved reverse-diffusion step, so changes in activation reflect changes in feature expression rather than changes in feature identity. The seven frozen crystal-system features show different temporal profiles: some remain active over broad intervals, while others are concentrated in narrower parts of the trajectory. Several also weaken toward one end of denoising. These curves are therefore interpreted as expression profiles rather than as evidence for the exact time at which a physical concept is created. Since lattice geometry, density, coordination, and symmetry constrain one another, information about a final crystal class can be present before the generated structure fully satisfies that class. Together, the layer and time analyses indicate that crystallographic information is distributed across intermediate computation and expressed non-uniformly over the reverse-diffusion trajectory.

Fine-grained Symmetry. The same training-only search is extended to sixty-eight observed space-group labels to obtain a finer-grained view of symmetry. Figure 5 shows representative high-activation structures for six candidate features. Structures associated with the same feature share the

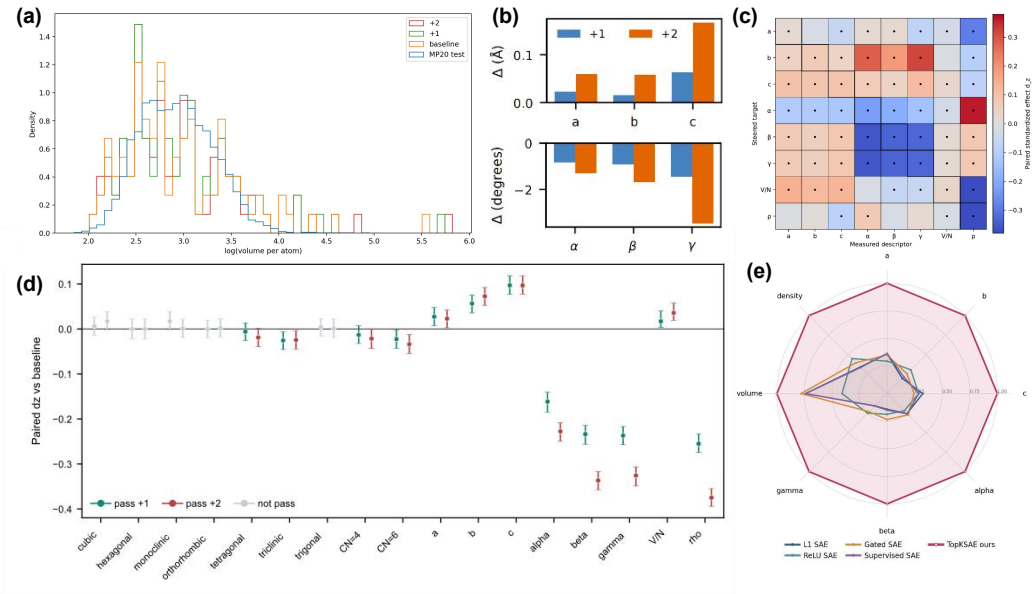


Figure 6: Full-test steering analysis on 9,046 MP-20 compositions. (a) Log-volume-per-atom distributions for the baseline and $\eta = 1, 2$ interventions, together with the MP-20 test distribution. (b) Mean changes in the six lattice parameters: a, b, c increase, while α, β, γ decrease, with larger responses at $\eta = 2$. (c) Paired standardized-effect matrix at $\eta = 2$. Structured off-diagonal responses reflect coupling among lattice lengths, angles, volume, and density. (d) Standardized endpoint changes for the seventeen locked directions at $\eta = 1$ and $\eta = 2$, highlighting same-sign, dose-consistent responses. (e) Descriptor-wise steering profiles for alternative SAE variants.

corresponding space-group label despite differences in composition, providing qualitative evidence that the sparse dictionaries can capture finer symmetry regularities beyond the seven crystal systems. However, these examples are selected conditionally on high activation and class support is uneven. The space-group atlas is therefore treated as exploratory rather than as held-out confirmation. Confirmatory claims remain restricted to the seventeen pre-specified concepts.

4.3 Steering Responses

Dose-dependent Steering. Association with a crystallographic variable does not by itself imply that the corresponding sparse direction can influence the generated structure. To test this, each frozen concept-associated direction is perturbed while the generator, composition, and initial sampling state remain fixed, and the resulting crystallographic descriptors are measured from the final crystal. Figure 6(a) shows that the log-volume-per-atom distributions under $\eta = 1$ and $\eta = 2$ still substantially overlap with both the unsteered baseline and the MP-20 test distribution, suggesting that steering changes crystal geometry without causing an obvious distributional collapse. Clear dose-dependent changes are observed in Figure 6(b,d): positive steering increases a, b, and c, while decreasing α, β, and γ, and all six lattice parameters show larger responses at $\eta = 2$ than at $\eta = 1$. Among the descriptor-audited continuous endpoints, $a, b, c, \alpha, \beta, \gamma$, and ρ exhibit same-sign, dose-consistent responses and pass the pre-specified BH-controlled test at $\eta = 2$. Figure 6(c) further shows structured off-diagonal responses, with lattice lengths, angles, volume, and density changing in coupled patterns. This is consistent with the geometric constraints of crystal structures and suggests that the sparse directions represent coordinated deformation modes rather than independent one-feature–one-property controls. Figure 6(e) further compares different SAE variants. Among the evaluated variants, the Top-K SAE exhibits the strongest overall steering responses while maintaining consistent effects across crystallographic descriptors, supporting its use as the primary SAE architecture in our experiments.

Baseline Comparison. A further comparison with neuron and random baselines is shown in Figure 7. Across the six lattice parameters a, b, c, α, β, and γ, SAE directions show smoother and more consistent dose–response behavior as intervention strength increases. Individual-neuron responses are

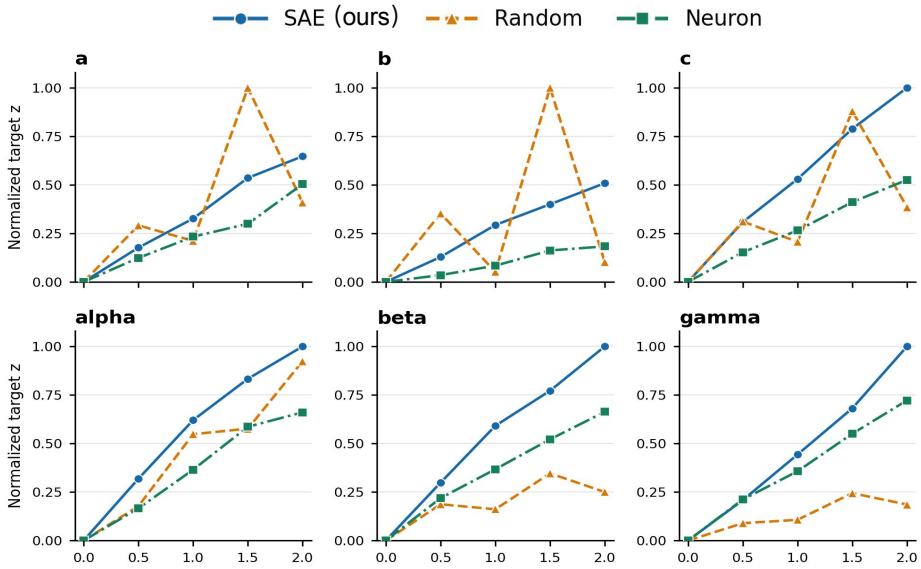


Figure 7: Steering comparison for the six lattice parameters a, b, c, α, β, and γ. Responses are normalized separately for each descriptor to facilitate comparison of dose–response patterns across intervention types. SAE directions generally exhibit smoother and more consistent responses across intervention strengths, while the neuron-based baselines are weaker or more irregular in several cases.

generally weaker, while random perturbations can occasionally produce large effects but often vary non-monotonically across doses. The contrast is especially clear for β and γ, where SAE responses increase steadily and remain stronger than both baselines. These results suggest that SAE features provide more reliable and dose-consistent steering directions than individual neurons, while avoiding the irregular responses observed for random directions.

Scope and future directions. As an initial study of sparse-autoencoder-based interpretability for crystal structure prediction, we focus on a representative CSP backbone, DiffCSP, and the widely used MP-20 benchmark. This setting allows us to establish a controlled framework for identifying, validating, and intervening on crystallographically aligned sparse features within a well-studied crystal diffusion model. A natural next step is to extend CrystalSAE to a broader range of crystal generative models and datasets, and to study whether similar sparse crystallographic structures emerge consistently across settings. Such extensions would also make it possible to examine the stability and transferability of discovered features across models and data distributions. Beyond interpretation, we view these internal feature directions as a promising basis for controllable crystal generation, where model representations could be steered toward desired structural characteristics and, ultimately, application-relevant material properties.

5 Conclusion

We introduced CrystalSAE, a framework for interpreting crystal diffusion models through sparse decompositions of DiffCSP residual representations. With training-only feature discovery and frozen held-out evaluation, CrystalSAE identifies sparse directions aligned with seventeen crystallographic concepts spanning lattice geometry, density, coordination, and crystal systems. These directions are distributed across network depth, and crystal-system features exhibit nonstationary expression over reverse diffusion. Interventions provide complementary functional evidence: seven descriptor-audited continuous endpoints show dose-consistent responses with multiplicity-controlled statistical support, together with structured cross-descriptor effects consistent with coupled crystal geometry. Overall, our results support a sparse, crystallographically aligned organization of DiffCSP representations and show how such internal structure can be discovered, localized, and functionally probed. This provides a step toward mechanistic understanding of scientific generative models beyond their output-level performance.

AI Use Statement. Generative AI tools were used to assist with drafting, organization, code review, and presentation of this manuscript. The authors reviewed all scientific claims, experimental results, code changes, figures, and tables against the corresponding experiment logs and source code. The authors take responsibility for the final content.

Ethics Statement. This work studies the interpretability and controllability of crystal generative models and does not involve human subjects or private personal data. To avoid overinterpretation, we distinguish held-out feature–concept associations from functional intervention results, report nonresponding endpoints, and do not treat feature steering as evidence of causal sufficiency or complete disentanglement.

Reproducibility Statement. The anonymous supplementary material provides the core CrystalSAE implementation, including SAE training, residual preprocessing, held-out evaluation, feature intervention, baseline analysis, and statistical evaluation. The release documents the main hyperparameters, random seeds, intervention strengths, normalization procedures, and crystallographic evaluation settings, and includes lightweight automated tests for the core implementation. Large pretrained checkpoints, intermediate residual caches, and generated crystal artifacts are not included in the current anonymous archive because of artifact-size and external-dependency constraints; the required external inputs are documented in the repository instructions.

REFERENCES

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 460–476, 2025.

Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15:10570, 2024. doi: 10.1038/s41467-024-54639-7.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Rees Chang, Angela Pak, Alex Guerra, Ni Zhan, Nick Richardson, Elif Ertekin, and Ryan P. Adams. Space group equivariant crystal diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2439.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Filip Ekström Kelvinius, Oskar B. Andersson, Abhijith S. Parackal, Dong Qian, Rickard Armiento, and Fredrik Lindsten. Wyckoffdiff—a generative diffusion model for crystal symmetry. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 15130–15147. PMLR, 2025.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025.

Colin W. Glass, Artem R. Oganov, and Nikolaus Hansen. Uspex—evolutionary crystal structure prediction. *Computer Physics Communications*, 175:713–720, 2006. doi: 10.1016/j.cpc.2006.07.020.

Stefan Goedecker. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *The Journal of Chemical Physics*, 120: 9911–9917, 2004. doi: 10.1063/1.1724816.

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

Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *Advances in Neural Information Processing Systems*, 2023.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *International Conference on Learning Representations*, 2024.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, et al. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability. *arXiv preprint arXiv:2503.09532*, 2025.

Daniel Levy, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Qiang Zhu, Kin Long Kelvin Lee, Mikhail Galkin, Santiago Miret, and Siamak Ravanbakhsh. Symmcd: Symmetry-preserving crystal generation with diffusion models. In *International Conference on Learning Representations*, 2025.

Danfeng Li, Kyuho Lee, Bai Yang Wang, Motoki Osada, Samuel Crossley, Hye Ryoung Lee, Yi Cui, Yasuyuki Hikita, and Harold Y. Hwang. Superconductivity in an infinite-layer nickelate. *Nature*, 572(7771):624–627, 2019. doi: 10.1038/s41586-019-1496-5.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *BlackboxNLP*, 2024.

Peijia Lin, Pin Chen, Rui Jiao, Qing Mo, Cen Jianhuan, Wenbing Huang, Yang Liu, Dan Huang, and Yutong Lu. Equivariant diffusion for crystal structure prediction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29890–29913. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/lin24b.html`.

Peijia Lin, Zihan Zhang, Zhangrui Zhao, Shaohao Rui, Junyi An, Yun-Fei Shi, Fenglei Cao, Weijie Ma, Yutong Lu, et al. Periodic bayesian flow networks with additive accuracy. In *Forty-third International Conference on Machine Learning*, 2026.

Xiangyu Liu, Haodi Lei, Yi Liu, Yang Liu, and Wei Hu. Protsae: Disentangling and interpreting protein language models via semantically-guided sparse autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 773–781, 2026. doi: 10.1609/aaai.v40i1.37044.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, 2025.

Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M. Wood. Flowmm: Generating materials with riemannian flow matching. In *International Conference on Machine Learning*, pp. 35664–35686, 2024.

Artem R. Oganov and Colin W. Glass. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. *The Journal of Chemical Physics*, 124:244704, 2006. doi: 10.1063/1.2210932.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996. doi: 10.1038/381607a0.

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

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024a.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024b.

Laura Ruple, Luca Torresi, Henrik Schopmans, and Pascal Friederich. Symmetry-aware bayesian flow networks for crystal generation. *npj Computational Materials*, 12:182, 2026. doi: 10.1038/s41524-026-02140-8.

Elana Simon and James Zou. Interplm: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22:2107–2117, 2025. doi: 10.1038/s41592-025-02836-7.

Anuroop Sriram, Benjamin Kurt Miller, Ricky T. Q. Chen, and Brandon M. Wood. Flowllm: Flow matching for material generation with large language models as base distributions. In *Advances in Neural Information Processing Systems*, 2024.

Viacheslav Surkov, Chris Wendler, Antonio Mari, Mikhail Terekhov, Justin Deschenaux, Robert West, Caglar Gulcehre, and David Bau. One-step is enough: Sparse autoencoders for text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Edith Natalia Villegas Garcia and Alessio Ansuini. Interpreting and steering protein language models through sparse autoencoders. In *International Conference on Learning Representations*, 2025.

Yanchao Wang, Jian Lv, Li Zhu, and Yanming Ma. Calypso: A method for crystal structure prediction. *Computer Physics Communications*, 183:2063–2070, 2012. doi: 10.1016/j.cpc.2012.05.008.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022.

Lu Yang, Tiantian Xu, Xiufeng Liu, Xu Cheng, Fan Shi, and Shengyong Chen. Tgd-csp: Reliable crystal structure prediction with template-guided diffusion and energy-based optimization. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 12633–12642, 2026. doi: 10.1145/3770855.3818992.

Xiaohan Yi, Guikun Xu, Zhong Zhang, Liu Liu, Yatao Bian, Xi Xiao, and Peilin Zhao. Crystaldit: Simple diffusion transformers for crystal generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(2):1462–1470, 2026. doi: 10.1609/aaai.v40i2.37121.

Claudio Zeni, Robert Pinsler, Daniel Z"ugner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. A generative model for inorganic materials design. *Nature*, 639:624–632, 2025. doi: 10.1038/s41586-025-08628-5.

A Interpreting Details

This appendix provides additional details for the crystallographic concepts and held-out interpretation analyses used in CrystalSAE. We first define the seventeen pre-specified concepts and summarize their distribution across MP-20, then report the frozen concept–feature assignments and held-out confirmation results. We finally provide the exploratory fine-grained space-group atlas used for qualitative symmetry analysis.

All concept–feature identities are selected exclusively on the 27,136-crystal training split and frozen before evaluation. The 9,047-crystal validation split is used for held-out confirmation and representation analysis, while the 9,046-crystal test split is reserved for feature intervention.

Crystallographic concepts. We consider seventeen pre-specified concepts spanning complementary levels of crystallographic organization. Eight are continuous crystal-level descriptors: the six lattice parameters $(a, b, c, \alpha, \beta, \gamma)$, volume per atom V/N, and density ρ. Seven correspond to the conventional crystal systems, and two describe local coordination environments, CN$= 4$ and CN$= 6$, using CrystalNN-derived coordination weights.

Figure S1 summarizes the geometric basis of these concepts. Panel (a) illustrates the six lattice parameters, while panel (b) shows the seven crystal systems and the coupled metric constraints that distinguish them. The latter are inherently multivariate: a crystal-system label depends on relationships among several lattice lengths and angles rather than on a single scalar descriptor.

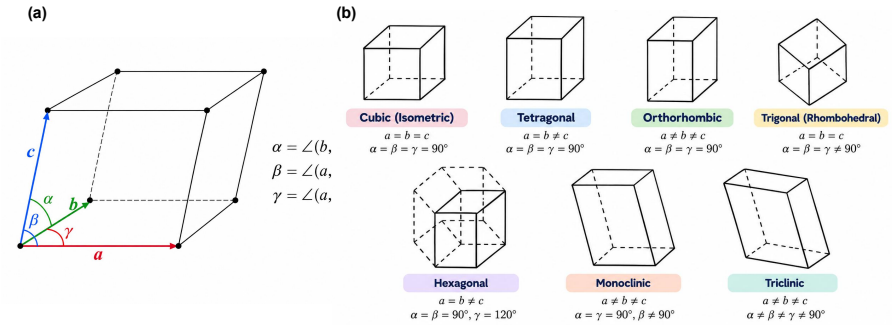


Figure S1: **Crystallographic concepts used in the interpretation analysis.** (a) Lattice geometry parameterized by edge lengths a, b, and c and interaxial angles $\alpha = \angle(b, c)$, $\beta = \angle(a, c)$, and $\gamma = \angle(a, b)$. (b) The seven crystal systems considered in the pre-specified concept set, illustrating their characteristic constraints on lattice lengths and angles. Unlike individual lattice descriptors, crystal-system identities are determined by coupled relationships among multiple geometric variables.

For held-out feature association, continuous descriptors are evaluated using one-feature R^2, whereas coordination and crystal-system concepts use AUPRC as the primary metric and AUROC as a secondary diagnostic. For intervention analysis, continuous concepts are evaluated using the corresponding descriptors measured from the generated crystal. Crystal systems use the symmetry-robustness score O_s defined in Eq. (7), and coordination concepts use the crystal-averaged CrystalNN weights defined in Eq. (8).

Dataset statistics and concept support. Table S1 reports the support of the seventeen concepts across the official MP-20 training, validation, and test splits. Figure S2 provides a complementary distributional view of the same data. Continuous descriptors are summarized after median-centering and IQR scaling to facilitate comparison across quantities with different physical units, while the crystal-system panel shows the relative class composition of each split.

Held-out concept–feature confirmation. Feature discovery is performed on the training split only. For each concept, the strongest layer–feature pair is selected according to the training-only association score, after which both feature identity and orientation are frozen. Table S2 and Table S3 report the resulting assignments and their held-out performance on the validation split. The seventeen locked pairs all satisfy the pre-specified held-out permutation test after BH correction. The reported effect sizes therefore characterize the strength of the association rather than being used for post-hoc feature selection. Continuous concepts are summarized by one-feature R^2, whereas binary coordination and crystal-system concepts are summarized primarily by AUPRC.

Exploratory space-group atlas. Beyond the seventeen pre-specified concepts, we apply the same training-only feature search to observed space-group labels to obtain a finer-grained view of symmetry-related structure in the learned representation. This analysis is exploratory: class support is highly non-uniform, and the resulting space-group candidates are not included in the seventeen-concept confirmatory family.

Table S1: Sample support for the 17 common concepts in the official MP-20 splits. Continuous concepts report the number of crystals with a valid descriptor. Crystal-system rows report positive crystals and their fraction of the split. Coordination rows report positive/valid crystals and the positive fraction among valid crystals. For CN$=4$/CN$=6$, the corresponding positive/valid site counts are 49,978/282,133 and 57,187/282,133 (train), 16,426/93,286 and 18,610/93,286 (validation), and 16,204/92,918 and 18,494/92,918 (test).

Concept	**Train**	**Validation**	**Test**
a	27,136	9,047	9,046
b	27,136	9,047	9,046
c	27,136	9,047	9,046
α	27,136	9,047	9,046
β	27,136	9,047	9,046
γ	27,136	9,047	9,046
V/N	27,136	9,047	9,046
ρ	27,136	9,047	9,046
Cubic	6,202 (22.9%)	2,020 (22.3%)	2,157 (23.8%)
Hexagonal	2,859 (10.5%)	947 (10.5%)	975 (10.8%)
Monoclinic	4,057 (15.0%)	1,282 (14.2%)	1,288 (14.2%)
Orthorhombic	5,272 (19.4%)	1,842 (20.4%)	1,725 (19.1%)
Tetragonal	4,538 (16.7%)	1,561 (17.3%)	1,561 (17.3%)
Triclinic	1,146 (4.2%)	358 (4.0%)	376 (4.2%)
Trigonal	3,062 (11.3%)	1,037 (11.5%)	964 (10.7%)
CN$=4$	10,452/27,115 (38.5%)	3,464/9,039 (38.3%)	3,379/9,041 (37.4%)
CN$=6$	13,249/27,115 (48.9%)	4,355/9,039 (48.2%)	4,377/9,041 (48.4%)

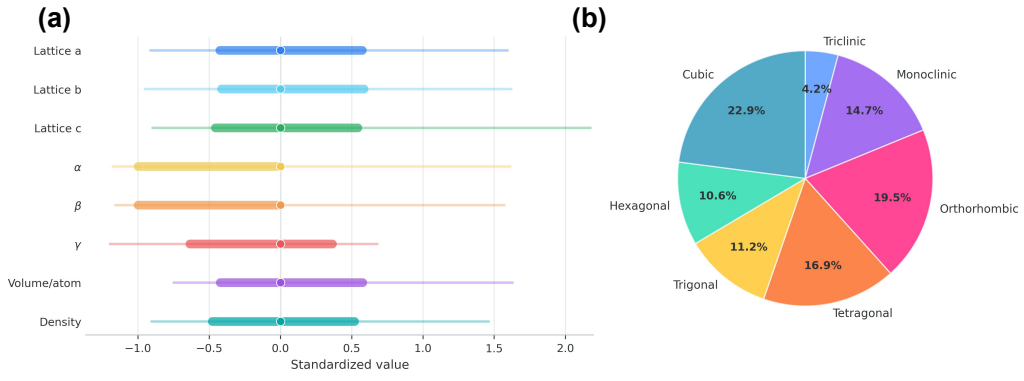


Figure S2: **MP-20 concept statistics across the official train, validation, and test splits.** (a) Split-level summaries of the eight continuous descriptors after median centering and IQR scaling. Thin segments denote the 5th–95th percentile range, thick segments denote the interquartile range, and circles mark the median. (b) Aggregate composition of the seven crystal systems. Split-specific support and class fractions are reported in Table S1.

Table S2: Held-out confirmation of the eight continuous concept–feature pairs. Feature identities are selected on the 27,136-crystal training split and frozen before evaluation on the 9,047-crystal validation split. All eight pairs satisfy the held-out BH-FDR criterion, with $p_{\mathrm{perm}} \leq 0.0039$ and $q_{\mathrm{BH17}} \leq 0.0039$. Association strength is reported using one-feature R^2.

Concept	**Feature**	**Layer**	R^2
Lattice a	F189	L2	0.1999
Lattice b	F46	L2	0.2441
Lattice c	F646	L0	0.5025
Lattice α	F368	L1	0.3129
Lattice β	F213	L1	0.3601
Lattice γ	F213	L1	0.3004
Density	F77	L2	0.1325
Volume / atom	F855	L0	0.1863

Table S3: Held-out confirmation of the nine binary coordination and crystal-system concept–feature pairs. Feature identities are selected on the 27,136-crystal training split and frozen before evaluation on the 9,047-crystal validation split. All nine pairs satisfy the held-out BH-FDR criterion, with $p_{\mathrm{perm}} \leq 0.0039$ and $q_{\mathrm{BH17}} \leq 0.0039$. AUPRC is the primary association metric and AUROC is reported as a secondary diagnostic.

Concept	**Feature**	**Layer**	**AUPRC**	**AUROC**
CN=4	F593	L1	0.2224	0.5329
CN=6	F177	L0	0.3079	0.5753
Cubic	F368	L1	0.7442	0.8653
Hexagonal	F261	L3	0.3932	0.6445
Monoclinic	F626	L2	0.2915	0.7048
Orthorhombic	F191	L1	0.3940	0.6563
Tetragonal	F101	L1	0.4747	0.6817
Triclinic	F71	L1	0.1922	0.8385
Trigonal	F158	L1	0.3585	0.6570

Table S4 lists representative space-group candidates from the frozen discovery output. These assignments are used only to support the qualitative structural examples shown in the main paper and should not be interpreted as held-out confirmation of space-group semantics.

Table S4: **Representative fine-grained space-group candidates from the frozen discovery output.** Counts refer to the analyzed MP-20 population. The space-group atlas is exploratory and is not part of the seventeen-concept confirmatory set.

SG	**Count**	**Layer**	**Feature**	**SG**	**Count**	**Layer**	**Feature**
1	164	L0	F258	44	46	L1	F311
2	212	L3	F614	62	371	L1	F227
4	47	L0	F856	63	307	L3	F574
5	76	L1	F140	65	89	L3	F574
8	119	L1	F1007	71	136	L1	F743
11	84	L3	F175	74	102	L1	F589
12	430	L1	F597	87	47	L3	F921
14	185	L1	F430	107	50	L2	F959
15	166	L0	F698	122	51	L0	F689
36	55	L3	F961	123	210	L1	F320

B SAE Training Details

This section provides implementation details for the layer-wise Top-K SAEs used throughout the interpretation analysis. An independent SAE is trained for each of the six DiffCSP residual streams. The main architecture and optimization settings are summarized in Table S5.

Table S5: **Main CrystalSAE training configuration.** The same SAE architecture and optimization settings are used independently for all six CSPNet residual streams.

SAE / training setting	Value	Optimization setting	Value
Residual dimension d	512	Aux loss weight	$1/32$
Dictionary size m	1024	Batch size	4096
Top-K	64	Learning rate	10^{-4}
Aux-K	256	Gradient clipping	1.0
Epochs	40	Weight decay	0

Residual normalization. For each DiffCSP layer, residual updates are centered and rescaled using statistics estimated from the SAE training corpus. The implementation uses a layer-specific scalar RMS scale, giving the normalized residual

$$\boldsymbol{x}_i^{(\ell)} = \frac{\Delta \boldsymbol{h}_i^{(\ell)} - \boldsymbol{b}_{\mathrm{pre}}^{(\ell)}}{\sigma_\ell}. \tag{S1}$$

The decoder reconstructs the normalized residual coordinates as $\widehat{\boldsymbol{x}} = W_{\mathrm{dec}}\boldsymbol{z}$ and maps them back to the native DiffCSP residual units according to $\widehat{\Delta \boldsymbol{h}} = \boldsymbol{b}_{\mathrm{pre}} + \sigma_\ell \widehat{\boldsymbol{x}}$. The same inverse scaling is used when decoder directions are mapped back into the frozen DiffCSP computation for intervention.

AuxK objective. To provide training signal to inactive dictionary features, we use the AuxK objective together with the main Top-K reconstruction objective. Let $\mathcal{D}_{\mathrm{dead}}$ denote the currently inactive feature set and

$$K_{\mathrm{aux}} = \min(K_{\mathrm{Aux}}, |\mathcal{D}_{\mathrm{dead}}|).$$

For encoder preactivation $\boldsymbol{u} = W_{\mathrm{enc}}\boldsymbol{x} + \boldsymbol{b}_{\mathrm{act}}$ and stop-gradient reconstruction residual $\boldsymbol{r} = \mathrm{sg}(\boldsymbol{x} - \widehat{\boldsymbol{x}})$, the auxiliary loss is

$$\mathcal{L}_{\mathrm{AuxK}} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_\ell}\left[\left\| W_{\mathrm{dec}}[:, \mathcal{D}_{\mathrm{dead}}]\ \mathrm{ReLU}\left(\mathrm{TopK}(\boldsymbol{u}_{\mathcal{D}_{\mathrm{dead}}}; K_{\mathrm{aux}})\right) - \boldsymbol{r}\right\|_2^2\right]. \tag{S2}$$

The auxiliary branch preserves the Top-K sparsity structure while supplying gradient signal to features that would otherwise remain inactive.

C Reconstruction Metrics

For an original residual update $\Delta \boldsymbol{h}$ and SAE reconstruction $\widehat{\Delta \boldsymbol{h}}$, explained variance is

$$\mathrm{EV} = 1 - \frac{\mathrm{Var}(\Delta \boldsymbol{h} - \widehat{\Delta \boldsymbol{h}})}{\mathrm{Var}(\Delta \boldsymbol{h})}. \tag{9}$$

The mean row-wise cosine similarity over N_{eval} evaluated residual rows is

$$\mathrm{Cosine} = \frac{1}{N_{\mathrm{eval}}} \sum_{i=1}^{N_{\mathrm{eval}}} \frac{\langle \Delta \boldsymbol{h}_i, \widehat{\Delta \boldsymbol{h}}_i \rangle}{\|\Delta \boldsymbol{h}_i\|_2 \, \|\widehat{\Delta \boldsymbol{h}}_i\|_2}. \tag{10}$$

Generator-level structural fidelity after reinjection is measured by

$$\mathrm{MatchRate} = \frac{1}{N_{\mathrm{eval}}} \sum_{i=1}^{N_{\mathrm{eval}}} \mathbf{1}\left[m(\widehat{\mathcal{M}}_i, \mathcal{M}_i^{\mathrm{ref}}) = 1\right], \tag{11}$$

where $\widehat{\mathcal{M}}_i$ is the reinjected generation, $\mathcal{M}_i^{\mathrm{ref}}$ is its reference, and $m(\cdot,\cdot)$ is the DiffCSP structure-matching rule.

Steering details. The deterministic protocol uses matched initial lattice and fractional-coordinate states for baseline and steering doses on the official MP-20 test split. For descriptor D, $\delta_i(\eta) = D_i(\eta) - D_i(0)$; dose consistency requires same-sign mean responses at $\eta = 1, 2$ and $|\bar{\delta}(2)| \geq |\bar{\delta}(1)|$. Paired tests at $\eta = 2$ are BH-corrected across the seventeen endpoints.

D FAILURE ANALYSIS

Among the seventeen pre-specified concepts, seven continuous endpoints ($a, b, c, \alpha, \beta, \gamma$, and ρ) exhibit same-sign, dose-consistent responses and pass the BH-controlled paired test at $\eta = 2$, while the remaining ten concepts do not satisfy the full intervention criterion. Importantly, all seventeen concept–feature pairs remain statistically detectable on held-out validation data, so failure in the intervention screen should not be interpreted as failure of feature discovery. Rather, the ten non-passing concepts reveal different limitations of one-dimensional feature steering.

Volume per atom. Volume per atom is the only continuous descriptor that does not pass the final intervention screen. Unlike an individual lattice length or angle, unit-cell volume is a nonlinear function of all six lattice parameters,

$$V = abc\sqrt{1 + 2\cos\alpha\cos\beta\cos\gamma - \cos^2\alpha - \cos^2\beta - \cos^2\gamma}.$$

Our steering experiments show coordinated changes in these primitive variables: a, b, c increase while α, β, γ decrease. Because these changes contribute jointly and nonlinearly to the determinant of the lattice matrix, their effects on V/N can partially offset one another. Thus, a feature direction can produce clear and reproducible changes in the constituent lattice parameters without inducing a correspondingly strong, monotonic, or statistically robust change in their aggregate volume. This failure therefore highlights the difference between controlling primitive geometric degrees of freedom and controlling a derived quantity that couples them nonlinearly.

Crystal systems. None of the seven crystal-system endpoints passes the full intervention criterion. This is consistent with the fact that crystal systems are discrete labels defined by joint constraints over multiple lattice variables rather than by a single scalar degree of freedom. For example, cubic symmetry requires $a = b = c$ together with $\alpha = \beta = \gamma = 90°$, while tetragonal and hexagonal systems impose different coupled equalities and angular constraints. Perturbing one sparse feature may therefore shift several geometric variables without moving the generated structure across the discrete boundary required to change its crystal-system assignment. Conversely, structures close to a symmetry boundary can undergo abrupt label changes under small perturbations, making the response inherently less smooth and less dose-consistent than for continuous descriptors. In addition, the same crystal-system label can correspond to many distinct atomic configurations, and symmetry can be broken along multiple geometric pathways. These observations suggest that crystal-system information can be reliably represented by a sparse direction without that direction being sufficient, by itself, to impose or remove the full set of constraints defining the discrete class.

Coordination classes. The CN=4 and CN=6 endpoints also fail the full intervention criterion and differ from the crystal-system case in an important way: coordination is a local, site-dependent structural property. Their held-out associations are already weaker than those of most lattice and symmetry concepts, with AUPRC values of 0.2224 and 0.3079, respectively. Changing coordination typically requires coordinated motion of several neighboring atoms and can depend on local bond geometry, competing neighbor assignments, and the CrystalNN weighting rule. A single globally applied sparse direction can therefore carry information about whether a local environment resembles fourfold or sixfold coordination without being sufficient to reorganize the relevant neighborhood into the target coordination state. This provides another example in which representational decodability does not imply a one-dimensional causal control axis.

Taken together, these ten failure cases indicate that the strongest steering responses occur for relatively direct continuous geometric variables, whereas derived, discrete, or locally defined concepts require coordinated changes across multiple degrees of freedom. We therefore interpret successful steering as evidence of functional influence rather than causal sufficiency or complete disentanglement. Improving control over the non-passing concepts may require multi-feature intervention, subspace-level steering, or intervention objectives that explicitly preserve the coupled geometric, local, and symmetry constraints defining the target concept.