Latent Crystallographic Microscope: Probing the Emergent Crystallographic Knowledge in Large Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models have demonstrated their capabilities in materials science, generating thermodynamically stable crystal structures without explicit domain training. However, the internal mechanisms enabling this scientific reasoning remain unclear, limiting our ability to develop reliable and controllable AI systems for materials discovery. This work investigates how LLMs encode crystallographic knowledge, process multi-structure reasoning, and whether mechanistic insights can enable controlled crystal structure optimization. We introduce the Latent Crystallography Microscope (LCM), the first mechanistic interpretability framework designed to reverse-engineer crystallographic reasoning in large language models. Through systematic linear probing across Llama 3.1-70B's transformer layers, we identify a hierarchical knowledge architecture where crystallographic concepts emerge across distinct processing phases from early chemical composition through intermediate thermodynamic and geometric reasoning to final symmetry classification. Our attention flow analysis reveals strong position bias effects in computational resource allocation. We further expose the limitations of prompt-based control approaches through ablation experiments. Moving beyond prompt-level control, we demonstrate that mechanistic insights enable targeted manipulation of crystal structure generation through layer-specific neural interventions, achieving systematic improvements in thermodynamic stability while preserving structural diversity. This work investigates scientific reasoning mechanisms in large language models and demonstrates that mechanistic interpretability can enable practical control over materials discovery processes, providing critical foundations for developing interpretable and controllable AI systems that can serve as reliable tools in autonomous materials discovery.

1 Introduction

The emergence of Large Language Models (LLMs) as powerful tools for scientific discovery has revolutionized computational materials science, with recent breakthroughs demonstrating their ability to generate thermodynamically stable crystal structures [2, 10, 11]. These models can predict material properties with accuracy comparable to specialized graph neural networks [5, 30] and generate novel crystalline structures through autoregressive text modeling, functioning as "innate crystal structure generators" without explicit domain training [10]. However, despite these impressive capabilities, the internal mechanisms by which LLMs encode, process, and manipulate crystallographic knowledge remain fundamentally unclear, which represent a critical gap that limits both scientific understanding and practical control over materials discovery applications.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

This gap constitutes a fundamental barrier to trustworthy AI-driven materials discovery. When LLMs generate crystal structures or demonstrate sophisticated understanding of space group symmetries and thermodynamic properties, we lack insight into the computational pathways enabling these behaviors. The lack of mechanistic interpretability prevents systematic debugging when models generate thermodynamically unstable structures, limits targeted interventions to guide generation toward specific material properties, and hinders the development of more controllable and reliable materials discovery systems.

Modern materials discovery faces a persistent stability-novelty dilemma, as comprehensively documented in recent benchmarking studies [21]. Traditional techniques like data-driven ion exchange achieve impressive thermodynamic stability rates (9.2%) by performing substitutions on known stable compounds, but exhibit zero prototype novelty because they rely exclusively on existing structural templates. Conversely, advanced generative models like Crystal Diffusion Variational Autoencoder (CDVAE) [31] and MatterGen [33] excel at structural innovation with prototype novelty rates up to 8.2%, but suffer from poor stability rates (1.8-3.0%) as most novel prototypes are thermodynamically unstable [21]. This trade-off presents researchers with a limiting choice: pursue stability at the expense of innovation, or chase novelty at the risk of instability.

Recent advances in mechanistic interpretability offer a promising path forward by developing sophisticated techniques for reverse-engineering neural network computational mechanisms into human-understandable components [8, 18]. Linear probing methods reveal how different types of knowledge are encoded across transformer layers [1, 7], while causal intervention techniques like activation patching enable direct manipulation of model behaviors [15, 25]. Sparse autoencoders decompose neural representations into interpretable features [4, 22], providing unprecedented insight into complex concept representation.

However, these interpretability techniques have focused primarily on linguistic tasks, leaving scientific reasoning domains largely unexplored. Materials science presents unique interpretability challenges, requiring models to integrate discrete categorical variables (space groups), continuous geometric parameters (lattice constants), and emergent physical properties (formation energies). Understanding how LLMs process these diverse crystallographic concepts could unlock principled approaches to the stability-novelty dilemma through targeted neural interventions.

In this work, we introduce the **Latent Crystallography Microscope (LCM)** framework to reverse-engineer crystallographic reasoning in LLMs and translate mechanistic insights into practical crystal structure optimization strategies. We investigate three fundamental questions: How do LLMs internally represent crystallographic knowledge across computational layers? How does attention allocation reveal processing mechanisms during multi-structure reasoning? How can mechanistic understanding enable targeted control over materials discovery?

We map the hierarchical organization of crystallographic knowledge through linear probing across transformer layers of Llama 3.1-70B, identifying distinct processing phases for different crystallographic properties. Through attention flow analysis with position bias correction, we reveal strong recency effects that dominate content-based preferences, while prompt engineering provides only modest control. Most significantly, we demonstrate how mechanistic insights translate into practical improvements: targeted neural interventions at specific layers enhance thermodynamic stability in generated crystal structures.

This work represents the first mechanistic analysis of crystallographic knowledge in LLMs. Our findings establish that layer-specific interventions can achieve meaningful control over materials properties, suggesting a promising direction for crystal structure optimization through targeted neural modifications. The framework provides both theoretical understanding of how LLMs process scientific knowledge and practical tools for enhancing materials discovery applications.

2 Related Work

2.1 Mechanistic Interpretability of Large Language Models

Mechanistic interpretability seeks to reverse-engineer the computational mechanisms learned by neural networks into human-understandable algorithms and concepts [8, 18]. This field has developed sophisticated methodologies for understanding transformer architectures, moving beyond black-box analysis to provide granular, causal understanding of model behaviors. Linear probing has emerged as

a fundamental technique for assessing what information is encoded within neural representations [1], with Conneau et al. [7] demonstrating that sentence embeddings contain rich linguistic structure accessible through linear classifiers. Subsequent work revealed hierarchical organization of syntactic knowledge across transformer layers [23], while recent advances have extended these insights to reveal linear structure in how models represent truth [14] and sentiment [24], supporting the hypothesis that many concepts are encoded as linear directions in activation space.

Understanding correlation between representations and concepts requires establishing causal relationships through direct intervention. Vig et al. [26] pioneered causal mediation analysis for language models, while Meng et al. [15] developed techniques for locating and editing specific factual associations in GPT models. Activation patching has emerged as a particularly powerful intervention technique [25], with recent work establishing best practices for applying these methods systematically [34]. A major breakthrough has been the development of sparse autoencoders (SAEs) for decomposing neural representations into interpretable features [4]. Building on theoretical foundations in superposition [9], SAEs address the fundamental challenge of polysemanticity by extracting monosemantic features from entangled neural activations. Recent scaling work has demonstrated that these techniques can be applied to production-scale models like Claude 3 Sonnet [22], revealing highly abstract features that capture sophisticated reasoning patterns. Understanding how features interact requires mapping the computational circuits that connect inputs to outputs. Olsson et al. [19] identified induction heads as a key mechanism for in-context learning, while Conmy et al. [6] developed automated techniques for discovering computational circuits, enabling detailed understanding of specific reasoning tasks [27].

2.2 Large Language Models in Materials Science

The application of LLMs to materials science has emerged as a transformative approach that challenges traditional computational methods. Unlike specialized architectures that require explicit geometric encodings, LLMs process crystallographic information through unified textual representations, offering unprecedented flexibility and generality. Before the emergence of LLMs, materials property prediction was dominated by graph neural network approaches. Crystal Graph Convolutional Neural Networks (CGCNNs) [30] provided the first systematic framework for learning from crystal structures, while MEGNet [5] established graph networks as a universal framework for molecular and crystalline systems. Recent scaling efforts have pushed these approaches to handle massive datasets [16], setting performance benchmarks for materials discovery applications.

The breakthrough application of LLMs to crystal generation began with CrystaLLM [2], which demonstrated that autoregressive modeling of Crystallographic Information Files (CIF) could generate thermodynamically plausible structures. Gruver et al. [11] demonstrated that fine-tuning pre-trained LLMs could generate stable inorganic materials with higher success rates than specialized diffusion models, achieving 49% vs 28% metastable generation rates compared to CDVAE baselines. Recently MatLLMSearchGan et al. [10] has revealed remarkable capabilities of pre-trained LLMs functioning as "innate crystal structure generators" without additional training, achieving high metastable rates through evolutionary search. Beyond direct structure generation, sophisticated frameworks have emerged that combine multiple AI components. GenMS [32] integrates language models with diffusion models and graph neural networks to enable hierarchical materials search from natural language descriptions. Technical innovations include invariant tokenization approaches [28] that ensure SE(3) and periodic invariance in crystal representations, and text-conditioned generation methods [20] that enable targeted design through natural language prompts.

2.3 Interpretability Challenges in Scientific Domains

While mechanistic interpretability has made significant progress in linguistic tasks, its application to scientific reasoning presents unique challenges that remain largely unaddressed. Recent systematic investigations have revealed fundamental limitations when interpretability methods are applied to scientific domains. In biology-inspired deep learning, systematic analysis has demonstrated that interpretations lack robustness upon repeated training and are systematically influenced by biases in knowledge graphs, with interpretation variability increasing with network depth and knowledge incompleteness creating spurious feature attributions [3]. Comprehensive reviews of biologically-informed models have found that interpretation reliability decreases significantly when models

encounter data distributions outside their training domain, highlighting fundamental generalization challenges in scientific interpretability [29].

Existing analysis of explainable AI methods applied to regulatory genomics has revealed that commonly used attribution methods produce inconsistent explanations across model architectures and training procedures [17]. The evaluations show that gradient-based interpretations often fail to capture true causal relationships in biological systems, with explanation fidelity varying dramatically based on input representation format choices. Recent work in computational biology has identified specific failure modes where attention-based explanations highlight irrelevant regions due to models' reliance on long-range dependencies that attention mechanisms cannot reliably capture [3], while gradient-based interpretations systematically conflate input relevance with gradient magnitude [13].

The application of interpretability to scientific domains faces the fundamental challenge of validation against domain-specific ground truth [12]. Many interpretability techniques designed for natural data assume clear categorical distinctions that apply to continuous scientific phenomena. In materials science, this lead to the challenge of validating interpretations against quantum mechanical principles, crystallographic theory, and thermodynamic constraints—domains where "ground truth" itself may be computationally intractable or experimentally inaccessible. These challenges motivate the development of specialized interpretability frameworks that can handle the unique requirements of scientific reasoning domains while providing reliable insights into model behavior.

3 Experiment 1: Hierarchical Knowledge Architecture via Linear Probing

3.1 Motivation and Experimental Setup

Understanding how crystallographic knowledge is organized within large language models is fundamental to developing controllable materials discovery systems. Linear probing provides a systematic methodology to map the hierarchical organization of crystallographic concepts across transformer layers, revealing where and how scientific knowledge are encoded. We identified several key crystallographic concepts including pace groups, formation energies, bulk moduli, and lattice parameters. Then we investigate how different crystallographic concepts are distributed across Llama 3.1-70B's 80 layers to establish the mechanistic foundation for subsequent intervention experiments.

Our analysis uses 10,000 diverse crystal structures from the Materials Project database, covering 193 space groups and formation energies spanning -13.214 to -0.437 eV/atom. For each structure, we transform it into POSCAR format inputs and extract mean-pooled activation vectors $\mathbf{h}_\ell \in \mathbb{R}^{8192}$ from each transformer layer. We train linear probes for space group classification (logistic regression, F1 scores) and continuous property prediction (linear regression, R^2 scores), with control probes on shuffled labels ensuring genuine knowledge detection.

3.2 Results and Analysis

Our analysis reveals distinct layer-wise specializations for different crystallographic properties, providing the computational roadmap for targeted interventions.

Space group classification shows early emergence (F1 = 0.1 at layer 10) with progressive improvement to peak performance (F1 = 0.54) at layers 65-75. This extended development demonstrates that while basic symmetry patterns are detected early, complete crystallographic classification requires sustained processing across nearly the entire network depth. Formation energy prediction achieves sustained high performance ($R^2 \approx$ 0.95-1.0) across lower to middle layers, with peak in the middle layers. This sustained plateau indicates robust thermodynamic reasoning circuits that integrate chemical composition with energetic principles. Bulk modulus maintains consistent performance ($R^2 \approx$ 0.6-0.7) across middle layers, while lattice parameters show peak processing ($R^2 \approx 0.65$) around lower to middle layers, indicating geometric reasoning demands in intermediate layers.

Through the analysis, we identify four groups of transformer layers that intrinsically focus on specific knowledge: **Chemistry Foundation (0-15):** Initial chemical composition processing as the foundation for crystallographic reasoning; **Formation Energy Processing (8-45):** Sustained thermodynamic reasoning integrates chemical principles with energy considerations; **Lattice Processing (15-45):** Geometric reasoning circuits handle structural arrangement optimization; **Space Group Processing (45-75):** Final symmetry classification resolves complex crystallographic relationships.

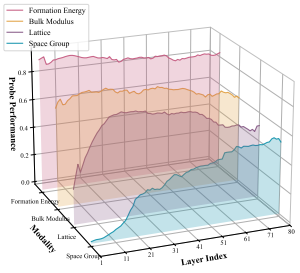

Figure 1: **Knowledge Emergence Pattern.** Temporal progression of crystallographic concept accessibility across transformer layers.

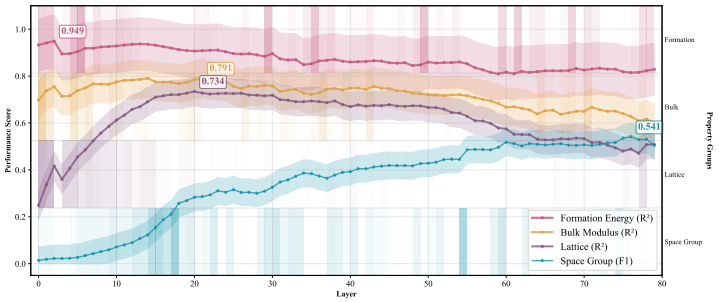

Figure 2: **Layer-wise Crystallographic Knowledge Architecture.** Performance curves showing probe accuracy across all 80 layers for four crystallographic concepts, revealing distinct processing specializations that inform subsequent intervention strategies.

3.3 Implications for Mechanistic Control

These layer specializations directly inform our experimental design. For Experiment 2, we focus on the formation energy layers (8-45) and space group layers (45-75) as these show the most distinct processing characteristics, enabling us to test whether attention allocation during multi-structure reasoning reflects these computational specializations. For Experiment 3, we employ the identified layer groups to perform targeted neural interventions: chemistry circuits (0-15) for compositional stability, formation energy circuits for thermodynamic optimization, lattice circuits for geometric reasoning, and space group circuits for symmetry processing. This mechanistic understanding enables specific interventions that enhance specific crystallographic reasoning capabilities while preserving the crystal structure generation ability.

4 Experiment 2: Attention Flow Analysis

4.1 Motivation and Research Framework

Building on the layer-wise knowledge architecture identified in Experiment 1, we investigate computational attention allocation during crystallographic reasoning involving multiple competing structural representations. We address the research question: *How does computational attention allocation reveal crystallographic processing mechanisms during multi-structure reasoning?* Our analysis examines attention distribution between competing structures, evaluates prompt influence on attention patterns, and distinguishes architectural constraints from task-dependent behaviors.

4.2 Experimental Design and Position Bias Correction

We analyze attention distribution using "later token attention" methodology, measuring attention from final processing tokens to each parent structure across the four processing phases. Our dataset consists of 1,000 crystal structure pairs, each containing a flawed prototype (A) generated by DiffCSP and a stable reference (B) retrieved from MatBench-bandgap through composition similarity.

The attention analysis extracts raw attention weights from all transformer layers, averages across attention heads, and calculates attention ratios by computing (Attention to Parent B) / (Attention to Parent A), where values above 1.0 indicate greater focus on the stable reference.

Initial analysis revealed substantial recency effects: when we tested both presentation orders ($A \rightarrow B$ and $B \rightarrow A$), attention patterns shifted systematically based on which structure appeared later in the prompt. To isolate genuine computational mechanisms from presentation order artifacts, we implemented a dual-ordering correction framework:

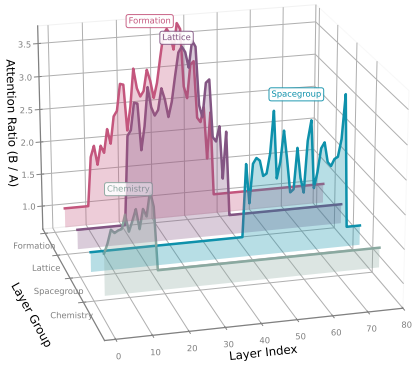

Figure 3: **Attention Flow Analysis.** Average attention ratios (B/A) across all 1,000 pairs from A→B ordered prompt before correction.

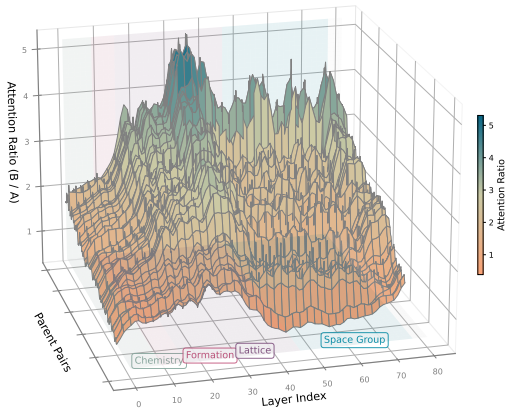

Figure 4: **Attention Landscape.** Attention ratios for 200 random pairs from A→B order, sorted by chemistry layers' attention. The attention ratio surface shows distinct patterns across layer groups.

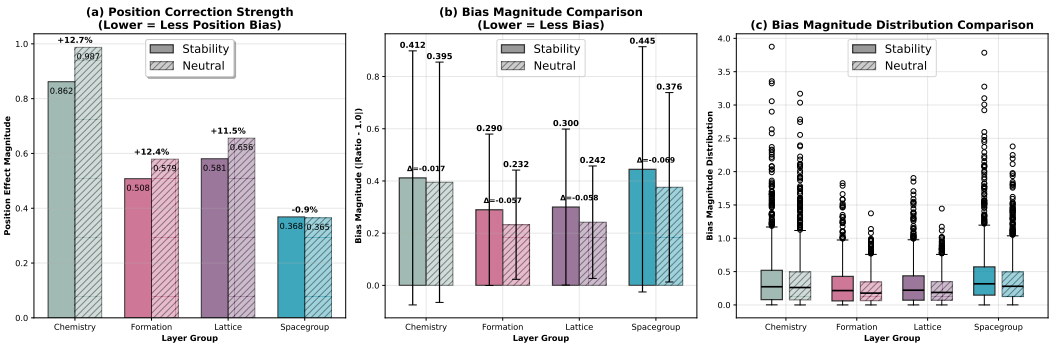

Figure 5: **Position Effect Comparison.** Stability-focused prompts reduce position dependency across layer groups. (a) Position effect magnitude measures position-dependent bias strength. (b) Bias magnitude shows deviation from balanced attention. (c) Distribution comparison reveals improved consistency.

$$R_{A\to B} = \frac{A^{A\to B}_{\text{to B}}}{A^{A\to B}_{\text{to A}}}, \quad R_{B\to A} = \frac{A^{B\to A}_{\text{to B}}}{A^{B\to A}_{\text{to A}}}, \quad R_{\text{corrected}} = \sqrt{R_{A\to B} \cdot R_{B\to A}} \tag{1}$$

where $R_{\text{corrected}}$ represents position-independent attention allocation using geometric mean correction.

We further quantify position effects using complementary measures presented in Figure 5. Position effect magnitude (a) measures how much attention depends on presentation order. Bias magnitude (b) measures deviation from balanced attention, calculated as the absolute difference from perfect balance (1.0 ratio).

4.3 Key Observations

Strong recency effects dominate attention allocation. Figure 4 demonstrates systematic position-dependent biases across all layer groups, with raw attention ratios varying dramatically based on presentation order rather than crystallographic content. The attention landscape reveals that positional encoding creates architectural constraints that override specialized reasoning, with later-presented

structures receiving disproportionate attention regardless of their stability properties. This finding establishes position correction as essential for isolating genuine computational mechanisms.

Prompt engineering provides measurable but limited control. Figure 5 quantifies the effects of instruction design on attention patterns. Stability-focused prompts reduce position dependency by 11-13% across layer groups (a), decrease bias magnitudes toward more balanced attention (b), and produce tighter distributions with improved consistency (c). However, these improvements remain modest, indicating that prompt-based interventions face fundamental limitations imposed by architectural constraints.

Layer groups exhibit minimal attention differentiation after correction. Following position correction, attention preferences show narrow variation: Chemistry layers (bias magnitude 0.41), Formation layers (0.29), Lattice layers (0.30), and Space Group layers (0.45) as shown in Figure 5(b). The uniform pattern across processing phases indicates that attention allocation may not simply follows a specialization organization.

4.4 Implications for Mechanistic Control

These findings reveal fundamental limitations of attention-based approaches for controlling crystallographic reasoning. The dominance of recency effects over content-based preferences, combined with modest prompt engineering benefits and minimal layer-specific differentiation, demonstrates that attention patterns provide insufficient leverage for precise mechanistic intervention. The disconnect between attention allocation and the layer-wise knowledge architecture established in Experiment 1 indicates that achieving meaningful control over specific crystallographic properties requires direct manipulation of computational pathways rather than attention-based approaches. These limitations directly motivate Experiment 3's causal intervention methodology.

5 Experiment 3: Layer-wise Causal Intervention Analysis

5.1 Motivation and Research Framework

The findings from Experiments 1 and 2 reveal fundamental constraints on controllable AI systems for scientific discovery. Experiment 2 demonstrates that prompt engineering provides only modest control over crystallographic reasoning, with attention allocation governed primarily by architectural properties rather than instruction-following mechanisms. This establishes that prompting are insufficient for precise control, necessitating direct intervention in internal computational pathways.

Building on the layer-wise knowledge architecture identified in Experiment 1, we investigate our central research question: *Can mechanistic insights enable precise causal control over crystallographic reasoning and materials optimization?* We systematically test which layer groups demonstrate causal control over specific properties, how interventions affect parent-child inheritance mechanisms, and the trade-offs between stability improvement and structural diversity.

5.2 Experimental Design and Intervention Framework

We implement targeted neural interventions during crystal structure optimization, modifying activations at specific layers while the model processes pairs of structures: a thermodynamically unstable prototype (Parent A) and a stable reference (Parent B). Six intervention conditions are tested: baseline control (1.0× strength), chemistry enhancement (layers 0–15, 1.5×), formation energy enhancement (layers 8–45, 1.5×), lattice enhancement (layers 15–45, 1.5×), spacegroup enhancement (layers 45–75, 1.5×), and integrated enhancement combining multiple pathways with conservative strengths.

The experiment analyzes 1,000 parent structure pairs across all intervention conditions, providing 6,000 total generation attempts. We evaluate structural similarity using RMS-based structural similarity metrics, defining novel structures as those with similarity < 0.5 to both parents simultaneously. This threshold ensures generated structures are genuinely innovative rather than simple copies or blends of parent structures.

Table 1: Intervention Performance Comparison

Intervention	Success Rate (%)	Avg E_d (eV/atom)	Stability Improvement (eV/atom)	Avg Tokens
Integrated Optimal	**70.3**	**0.627**	-0.259	416.5
Baseline	69.8	–	–	389.8
Chemistry Enhancement	68.4	0.872	-0.248	396.0
Spacegroup Enhancement	68.4	–	–	**386.3**
Formation Energy Enhancement	67.9	0.844	**-0.226**	399.9
Lattice Enhancement	67.7	1.007	-0.341	391.1

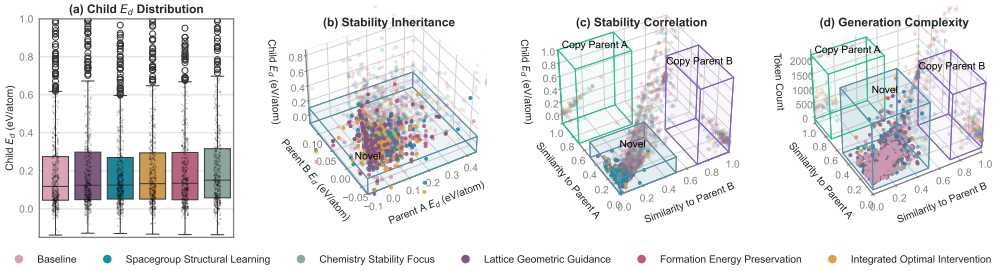

Figure 6: **Intervention Effects on Crystal Structure Generation.** Analysis showing (a) thermodynamic stability distributions across interventions, (b) stability inheritance patterns, (c) parent similarity correlations, and (d) generation complexity metrics.

5.3 Key Observations

Novel structures dominate generation and achieve superior stability. Figure 6(a,c) reveals that 45.4% of generated structures are novel (structurally distinct from both parents), achieving deformation energies of 0.13-0.16 eV/atom compared to 0.70-0.80 eV/atom for hybrid structures. This demonstrates that interventions preferentially generate innovative structures with enhanced thermodynamic properties rather than simple parent combinations.

Computational complexity correlates with structural innovation. Higher token generation requirements correlate with novel structure production, indicating that creating genuinely innovative crystal structures demands more sophisticated computational pathways. Figure 6(d) shows this relationship holds consistently across all interventions, suggesting an intrinsic connection between computational effort and crystallographic innovation.

Different interventions implement distinct inheritance strategies. Figure 6(b,c) reveals intervention-specific patterns in parent-child relationships. Formation energy interventions achieve the most balanced inheritance from both parents, chemistry enhancements show stronger prototype dependence, while spacegroup interventions enable structural reorganization with minimal parent similarity. These patterns demonstrate controllable inheritance mechanisms rather than random generation.

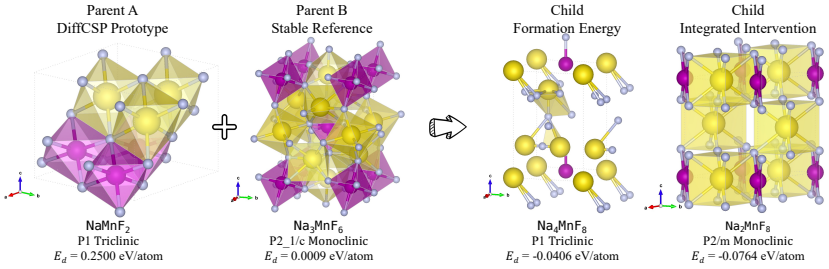

Figure 7: **Representative Parent-Child Inheritance Patterns.** Structural comparison showing Parent A (unstable prototype), Parent B (stable reference), and generated child structures across intervention conditions, demonstrating distinct inheritance mechanisms.

5.4 Implications for Mechanistic Control

The experimental results establish layer-wise causal intervention as a viable approach for controlling crystallographic reasoning in LLMs. The consistent preference for novel structure generation across all interventions, combined with superior thermodynamic properties and distinct inheritance patterns, demonstrates that mechanistic understanding translates into practical control over materials discovery. The correlation between computational complexity and structural innovation suggests that targeted interventions amplify existing architectural capabilities rather than creating entirely new optimization mechanisms.

6 Limitations and Future Directions

While our work establishes the first comprehensive mechanistic analysis of crystallographic reasoning in LLMs, several methodological limitations present valuable opportunities for future research. Our linear probing approach reveals correlational rather than definitively causal relationships between layer activations and crystallographic knowledge, and mean-pooled representations may obscure fine-grained activation patterns that could provide deeper mechanistic insights. Our intervention methods, while demonstrating systematic improvements, operate at coarse layer-group granularity rather than feature-level precision. The analysis demonstrates that certain regions are causally significant but cannot identify confounding factors or the specific computational circuits within those regions responsible for the observed effects.

Additionally, our evaluation framework relies primarily on CHGNet-based stability metrics, creating a "model-evaluates-model" paradigm that lacks grounding in first-principles physics and risks optimizing for evaluator biases rather than true physical stability. The analysis is currently limited to a single model architecture (Llama 3.1-70B), and the parent structure pairs sometimes involve simple compositions from DiffCSP prototypes, which may be less representative of complex materials discovery challenges. Future research should prioritize adoption of first-principles validation through Density Functional Theory calculations, systematic generalization studies across different LLM architectures, and incorporation of monosemantic feature discovery to enable more surgical interventions on complex crystal structures with higher compositional complexity.

7 Conclusion

This research provides the first mechanistic understanding of how large language models process crystallographic knowledge, revealing that LLMs organize scientific concepts through distinct hierarchical processing phases that mirror the conceptual structure of crystallographic theory. Our attention flow analysis further reveals computational resource allocation of LLMs when performing crystal structure generation as well as the limitations of prompt-based control approaches. Most significantly, our causal intervention experiments demonstrate that mechanistic insights can be translated into practical control over materials discovery processes, with targeted layer-specific interventions achieving improvements in thermodynamic stability while preserving structural innovation. Our Latent Crystallography Microscope framework validates mechanistic interpretability as a viable approach for controllable LLMs-based scientific discovery, which is generalizable for investigating domain-specific reasoning in transformers. This work provides practical insights on applying LLMs to materials science and broader scientific domains where precise control over model behavior is essential for reliable discovery.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR)*, 2017. 2, 3

[2] Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15:10570, 2024. 1, 3

[3] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley.

Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, 2021. doi: 10.1038/s41592-021-01252-x. 3, 4

[4] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yuxin Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features. 2, 3

[5] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31 (9):3564–3572, 2019. 1, 3

[6] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. 3

[7] Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, 2018. 2, 3

[8] Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html. 2

[9] Nelson Elhage, Tristan Hume, Catherine Olsson, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html. 3

[10] Jingru Gan, Peichen Zhong, Yuanqi Du, Yanqiao Zhu, Chenru Duan, Haorui Wang, Carla P. Gomes, Kristin A. Persson, Daniel Schwalbe-Koda, and Wei Wang. Large language models are innate crystal structure generators. *arXiv preprint arXiv:2502.20933*, 2025. 1, 3

[11] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary Ulissi. Fine-tuned language models generate stable inorganic materials as text. *arXiv preprint arXiv:2402.04379*, 2024. 1, 3

[12] Frederick Klauschen, Jonas Dippel, Philipp Keyl, Philipp Jurmeister, Michael Bockmayr, Oliver Buchstab, Maximilian Alber, Lynton Grabbed, Helena Kindermann, Eva Krieghoff-Henning, et al. Toward explainable artificial intelligence for precision pathology. *Annual Review of Pathology: Mechanisms of Disease*, 19:541–570, 2024. doi: 10.1146/annurev-pathmechdis-051222-014750. 4

[13] Antonio Majdandzic, Chandana Rajesh, and Peter K Koo. Correcting gradient-based interpretations of deep neural networks for genomics. *Genome Biology*, 24(1):109, 2023. doi: 10.1186/s13059-023-02956-3. 4

[14] Samuel Marks and Max Tegmark. The geometry of truth: emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023. 3

[15] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022. 2, 3

[16] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023. 3

[17] Gherman Novakovsky, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, 2023. doi: 10.1038/s41576-022-00532-2. 4

[18] Christopher Olah. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/mech-interp-essay. 2

[19] Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html. 3

[20] Md Rashidul Rahman, Akib Haque, Tamzidul Hassan, et al. Crystext: A generative ai approach for text-conditioned crystal structure generation using llm. *ChemRxiv preprint*, 2024. doi: 10.26434/chemrxiv-2024-qkxzm. 3

[21] Nathan J. Szymanski and Christopher J. Bartel. Establishing baselines for generative discovery of inorganic crystals. *Materials Horizons*, 2025. 2

[22] Adly Templeton, Trenton Bricken, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yuxin Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity. 2, 3

[23] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[24] Cathy Tigges, Oskar Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023. 3

[25] Alexandre Variengien and Eric Winsor. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2023. 2, 3

[26] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Shusen Qian, Daniel Nevo, Yoav Singer, and Emma Strubell. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020. 3

[27] Kevin Wang, Vatsal Varma, Alex Chris, Ryan Li, Aman BELLE, Jacob Meister, Jacob Andreas, Catherine Olsson, David Bau, et al. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10839–10854, 2022. 3

[28] Yuxin Wang, Hao Chen, Yang Liu, and Wei Zhang. Invariant tokenization of crystalline materials for language model enabled generation. *arXiv preprint arXiv:2405.09341*, 2024. 3

[29] Magdalena Wysocka, Oskar Wysocki, Marie Zufferey, Dario Landers, and Andre Freitas. A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data. *BMC Bioinformatics*, 24(1):198, 2023. doi: 10.1186/s12859-023-05262-8. 4

[30] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14):145301, 2018. 1, 3

[31] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, Tommi Jaakkola, et al. Crystal diffusion variational autoencoder for periodic material generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 21961–21973, 2021. 2

[32] Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Pierson, Ekin Dogus Cubuk, Stefano Ermon, and Doina Precup. Generative hierarchical materials search. *arXiv preprint arXiv:2409.06762*, 2024. 3

[33] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, et al. A generative model for inorganic materials design. *Nature*, 639:624–632, 2025. doi: 10.1038/s41586-025-08628-5. 2

[34] Jing Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. 3