HARNESS MEETS DIFFUSION: SINGLE-REACT-GUIDED CRYSTAL GENERATION AND KNOWLEDGE-GROUNDED SUPERCONDUCTIVITY SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific crystal discovery requires generators that can adapt their search to explicit property targets. We introduce a single-ReAct controller for a frozen LLM-diffusion crystal generator. The controller uses batch summaries and retrieved materials knowledge to guide subsequent LLM proposals. Selected candidates then undergo diffusion refinement. In open-ended generation, it raises the stable, unique, and novel fraction (SUN) from 0.2131 ± 0.0015 to 0.3436 ± 0.0749 across three matched seeds, with a gain in every seed and a decline in reference-distribution fidelity. We then apply the controller to superconductivity screening using a frozen T_c predictor and a separate structural predictor for evaluation. In comparisons with the same generator and candidate budget, the controller achieves SUN $= 0.455 \pm 0.040$ across four seeds, compared with 0.257 ± 0.007 for best-of-k and 0.241 ± 0.009 for verifier-only selection. Across six seeds, it generates $2.26\times$ as many candidates with predicted $T_c \geq 20$ K and nearest-training-composition Jaccard distance at least 0.20 as a knowledge-only policy. Adding a superconductivity-tuned generator further increases the joint stable, unique, novel, and predicted-$T_c \geq 20$ fraction from 0.1723 to 0.2843 in a matched run. These results show that feedback-driven control improves the computational screening yield of an existing crystal generator across two scientific objectives.

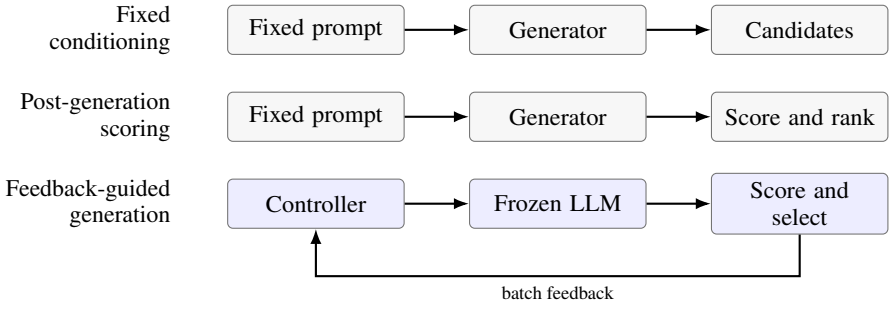

Figure 1: Crystal-generation control regimes. Fixed conditioning uses one prompt throughout generation; post-generation scoring ranks the resulting candidates. Our controller uses verifier feedback to update subsequent rounds. Superconductivity search adds a predicted-T_c objective and retrieved materials knowledge.

1 INTRODUCTION

Crystal generators support scientific discovery by proposing structures that satisfy both geometric constraints and target properties. Recent LLM-diffusion systems combine prompt-based scientific constraints with continuous refinement of lattice parameters and atomic coordinates (Khastagir et al., 2025; Jiao et al., 2023; Zeni et al., 2025). Their usefulness for a search task depends on how generated candidates are evaluated and how that feedback affects subsequent proposals. Reference-distribution fidelity and the yield of candidates meeting a property target measure different aspects of this process.

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 study how feedback from previous proposals can improve the yield of a frozen crystal generator. The controller tracks recurring compositions, gaps in the sampled distribution, and property scores. It uses these summaries and retrieved materials knowledge to revise the next prompt. Candidate selection takes place before diffusion refinement, allowing the controller to guide the LLM's proposals while preserving the backbone's geometry model. We evaluate this design on open-ended materials discovery and superconductivity screening.

In open-ended generation, the controller increases SUN by 61.2% and MSUN by 32.3% relative to the same frozen checkpoint, averaged over three matched seeds. For superconductivity screening, the controller uses a frozen local TCPROXY and a retrieved knowledge pack. A separate structural predictor assigns predicted $T_c \geq 20$ K to 859 of 3000 generated candidates, compared with 226 for the base generator. Matched candidate-budget comparisons with best-of-k, verifier-only, and random policies test the contribution of iterative control. A six-seed comparison with knowledge-only prompting measures whether the controller also increases yield away from training-set compositions.

The primary contributions are:

- A batch-level controller that combines search history and materials knowledge to guide proposals from a frozen LLM-diffusion generator.
- A matched open-ended evaluation showing higher CHGNet-based discovery yield alongside reduced reference-distribution fidelity.
- Controlled superconductivity-screening experiments showing higher joint discovery yield than global best-of-k selection and more high-scoring candidates with distinct element sets than fixed knowledge prompting.

2 RELATED WORK

Crystal generation with language models and diffusion. Recent crystal-generation systems combine discrete proposal with continuous refinement: DiffCSP studies joint equivariant diffusion (Jiao et al., 2023), MatterGen scales inorganic materials design (Zeni et al., 2025), and *LLM Meets Diffusion* motivates the hybrid decomposition used by our backbone (Khastagir et al., 2025). Text-native LLM generators and geometry-aware alternatives further broaden this landscape (Gruver et al., 2024; Antunes et al., 2024; Jiao et al., 2024; Miller et al., 2024; Levy et al., 2025). Our controller updates generation conditions from verifier feedback while keeping the backbone fixed.

Search with language models. ReAct and tool-augmented scientific planning use external feedback to update a language model's actions (Yao et al., 2022; Bran et al., 2024). More directly related, GenMS combines language and diffusion models with property-guided tree search (Yang et al., 2024), while MatLLMSearch evolves a population of structures through LLM proposals and interatomic-potential evaluation (Gan et al., 2025). Our controller revises a single prompt from batch summaries and retrieved knowledge, selects provisional LLM structures, and then refines the selected set with a frozen diffusion model. The experiments compare this procedure with fixed prompting, local selection, and global best-of-k selection on the same backbone.

Directed inverse generation. Inverse-design methods use conditioning, auxiliary property heads, and domain-specific fine-tuning to guide generation. Examples include guided superconductivity generation (Prakash et al., 2025), diffusion-based family generation, and learned T_c prediction (Yuan & Dordevic, 2024; Stanev et al., 2018). We study an additional source of adaptation: updating the search policy around a frozen generator using verifier feedback and retrieved knowledge.

3 FEEDBACK-GUIDED CRYSTAL GENERATION

3.1 BASE GENERATOR

We use CRYSLLMGEN (Khastagir et al., 2025), which first generates atom types, fractional coordinates, and lattice parameters with an LLM, then refines the coordinates and lattice with an equivariant diffusion model. Our controller operates on the provisional LLM structures. Once selection is complete, the retained candidates pass through the diffusion model with their atom types held fixed.

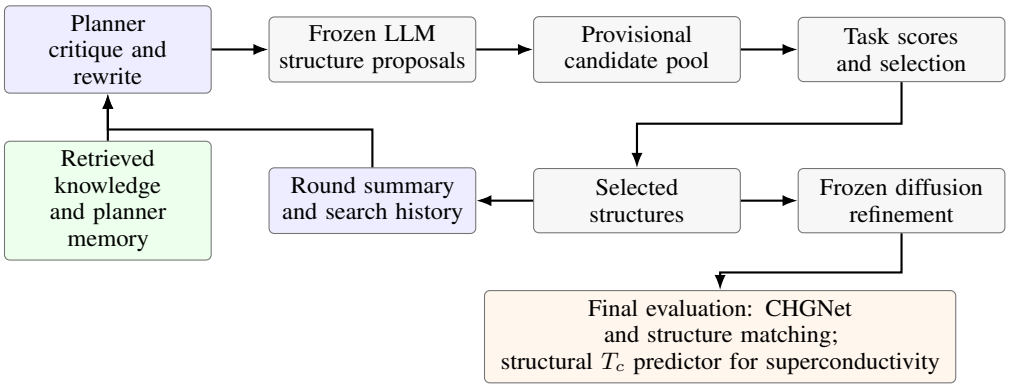

Figure 2: Feedback-guided generation. The planner uses summaries of selected LLM proposals to revise the next prompt. In superconductivity search, TCPROXY scores provisional structures and retrieved knowledge guides planning. After selection, the frozen diffusion model refines the retained structures. CHGNet screening and the structural T_c predictor evaluate the final structures.

Both models remain frozen in the control-policy comparisons; Section 4.3 also evaluates a tuned generator. Appendix L describes the diffusion model.

3.2 SINGLE-REACT CONTROL LOOP

The controller maintains one prompt and a memory of the search (Figure 2). After each round, it summarizes the selected proposals, retrieves relevant knowledge, and revises the prompt. Let $\mathcal{B}_t$ denote the selected provisional structures in round t, $\mathcal{V}$ the scoring and summarization procedure, $\mathcal{S}_t$ the resulting summary, $\mathcal{C}_t$ the prompt conditions, $\mathcal{P}_t$ the memory, and $\mathcal{K}_t$ the retrieved knowledge. The planner produces a decision d_t and updates the conditions and memory:

$$\begin{aligned} \mathcal{S}_t &= \mathcal{V}(\mathcal{B}_t), \\ d_t &= \mathcal{R}(\mathcal{C}_t, \mathcal{P}_t, \mathcal{S}_t, \mathcal{K}_t), \\ \mathcal{C}_{t+1} &= \mathcal{A}(\mathcal{C}_t, d_t), \quad \mathcal{P}_{t+1} = \mathcal{U}(\mathcal{P}_t, \mathcal{S}_t, \mathcal{K}_t, d_t). \end{aligned} \tag{1}$$

Single-ReAct refers to a single maintained prompt, updated through a critique followed by a rewrite. Updates occur every B selected proposals: $B = 200$ for open-ended generation and $B = 100$ for superconductivity search. Open-ended selection uses composition coverage, diversity, geometric plausibility, and a repetition penalty. Superconductivity selection ranks candidates first by validity and then by TCPROXY; the remaining criteria break ties. The planner changes the proposal conditions, while this ranking rule remains fixed. Algorithm 1 gives the sampling procedure, and Appendix K specifies the selection rules.

Search feedback and final evaluation. During open-ended search, candidates are scored for validity, composition coverage, repetition, and geometric plausibility. CHGNet relaxation and SUN/MSUN evaluation are performed after diffusion refinement. For superconductivity search, the controller also uses a frozen local TCPROXY that combines UMA (Wood et al., 2025), ORB (Neumann et al., 2024), and DPA (Zhang et al., 2024) structure features with an explicit composition vector. Retrieved family cards, motif cues, and element-role priors provide additional guidance. We evaluate predicted T_c on the final CIFs with a second predictor, referred to below as the *structural verifier*, that omits the explicit composition vector. Both predictors share the SuperCon–MP training domain and UMA/ORB/DPA representation family; their errors may therefore be correlated. CHGNet relaxation and reference-set matching determine SUN and MSUN separately. Appendix C reports agreement between the two T_c predictors.

We report reference-distribution fidelity alongside discovery yield to characterize how adaptive prompting changes the generated set.

Algorithm 1 Feedback-guided proposal and refinement

Input: frozen LLM and diffusion model, scoring procedure $\mathcal{V}$, prompt u_0, target count N, round size B, optional knowledge pack K
Initialize conditions $\mathcal{C}_0$ from u_0, memory $\mathcal{P}_0$, selected set $G \leftarrow [\,]$, history $H \leftarrow [\,]$, and round index $t = 0$
while $|G| < N$ **do**
  Generate provisional structures with the LLM under $\mathcal{C}_t$ (eight proposals per prompt in superconductivity search)
  Score the parseable proposals using $\mathcal{V}$
  Accept one candidate using the current objective-specific rule and append it to G
  if $|G|$ reaches the next multiple of B **then**
    Summarize the most recent B selected structures as $\mathcal{S}_t$, together with the cumulative search history
    Retrieve relevant knowledge $\mathcal{K}_t$ from K, if used
    Query planner in two stages (critique $\rightarrow$ rewrite) to obtain decision d_t
    Update prompt conditions $\mathcal{C}_{t+1}$ and memory $\mathcal{P}_{t+1}$ using Eq. 1
    Record $(\mathcal{S}_t, \mathcal{K}_t, d_t)$ in H and set $t \leftarrow t + 1$
  end if
end while
Refine each structure in G with the frozen diffusion model, preserving atom types
return refined structures and search history H

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate the harness in two settings. The first uses the same Materials Project-style inorganic crystal benchmark and protocol as the base generator, so standard metrics ask how closely generated CIFs match held-out MP-like crystals while novelty and uniqueness are judged against the MP2020correction reference set. The second is knowledge-grounded superconductivity search, where the same generator family is retargeted with a frozen local TCPROXY during search but evaluated with a separate final-CIF structural verifier without explicit composition features. Throughout Part I, *Base* means the frozen generator without round-level control and *Harness* means the same checkpoint family with the single-ReAct controller.

Metrics. For open-ended quality we report standard benchmark-referenced metrics together with CHGNet-based discovery metrics SUN and MSUN (Zeni et al., 2025; Deng et al., 2023): the fraction of evaluator-ready samples that are simultaneously novel, non-duplicate, and energetically plausible after relaxation (SUN: $E_{\text{hull}} \leq 0$; MSUN: $E_{\text{hull}} \leq 0.1$ eV/atom). In the superconductivity setting we report final-CIF structural-verifier summaries (predicted T_c in kelvin) plus joint metrics SUN $\cap$ $(T_c \geq 20)$ and MSUN $\cap$ $(T_c \geq 20)$; full reproducibility details are in Appendix A.1.

4.2 PART I: OPEN-ENDED MATERIALS DISCOVERY

The open-ended experiment compares the same frozen checkpoint with and without the controller. Appendix A lists the run identifiers and configurations. Table 1 reports the matched comparison alongside published methods for context.

Table 1: Open-ended generation. SUN and MSUN for Base and Harness are means $\pm$ standard deviations over three matched seeds; the other columns describe one representative matched run. External rows are from the re-evaluation in Khastagir et al. (2025), Tables 2–3, and use different protocols. Per-method citations identify the original methods. The planner is `deepseek-v4-pro`.

Method	comp.	struct.	valid	W_{dens}	W_{elem}	recall	prec.	SUN	MSUN
CDVAE (Xie et al., 2021)	86.70	100.00	-	0.687	1.432	99.15	99.49	0.0256	0.2199
DiffCSP (Jiao et al., 2023)	83.25	100.00	-	0.352	0.339	99.71	99.76	0.0561	0.3219
DiffCSP++ (Jiao et al., 2024)	85.12	99.94	-	0.235	0.375	99.73	99.59	0.0655	0.3056
FlowMM (Miller et al., 2024)	83.19	96.85	-	0.239	0.083	99.49	99.58	0.0306	0.2246
SymmCD (Levy et al., 2025)	87.13	92.30	-	0.531	0.210	97.33	98.78	0.0676	0.3169
Llama-2 (7B) (Gruver et al., 2024)	93.55	97.70	-	1.575	0.272	96.95	99.32	0.0484	0.2666
CrysLLMGen (7B) (Khastagir et al., 2025)	93.55	99.94	-	0.972	0.272	98.52	99.84	0.0921	0.3594
Base (`r64e2_A`)	89.20	99.97	89.17	0.177	0.090	93.57	99.83	0.2131 ± 0.0015	0.3153 ± 0.0033
Harness (`r64e2_A`)	83.70	99.87	83.57	1.382	0.258	91.57	99.67	$\mathbf{0.3436 \pm 0.0749}$	$\mathbf{0.4171 \pm 0.0960}$

The harness raises SUN from 0.2131 ± 0.0015 to 0.3436 ± 0.0749 ($+61.2\%$) and MSUN from 0.3153 ± 0.0033 to 0.4171 ± 0.0960 ($+32.3\%$), exceeding the base in all three matched seeds. Composition validity and recall decline, and the density and element Wasserstein distances increase. The improvement in discovery yield therefore comes with lower fidelity to the reference distribution. Appendix A.2 reports selected single-run results from nearby configurations.

4.3 Part II: Knowledge-Grounded Superconductivity Search

We compare three configurations of the same generator family. *Base* uses the frozen generator under the superconductivity task configuration. *Harness* adds the controller and superconductivity objective. *Harness + tuned* uses the same controller with a superconductivity-tuned generator. Each configuration returns 3000 samples. The harnessed runs also perform planner updates and in-loop TcProxy scoring; Section 4.5 compares control policies with matched candidate budgets.

Tables 2–3 report predicted T_c from the structural verifier and discovery metrics from CHGNet and structure matching. Table 4 uses TcProxy rescoring for its six-seed T_c summaries. The joint counts use the structural verifier for the $T_c \geq 20$ threshold, together with CHGNet stability and structure matching.

Table 2: Directed superconductivity generation within a shared generator family: final-CIF summaries from the structural verifier without explicit composition features (predicted T_c) for 3000 samples per method (*single matched run*).

Method	Mean T_c (K)	Median T_c (K)	$T_c \geq 5$ (count)	$T_c \geq 10$ (count)	$T_c \geq 20$ (count)	$T_c \geq 30$ (count)	SUN $\cap$ $T_c \geq 20$
Base	6.69	3.13	816	400	226	131	0.0453
Harness	17.05	11.78	2279	1719	859	520	0.1723
Harness + tuned generator	27.37	23.77	2376	2025	**1596**	**1340**	**0.2843**

The harness increases the number of candidates with predicted $T_c \geq 20$ K from 226 to 859 and with predicted $T_c \geq 30$ K from 131 to 520 (Table 2). Adding generator tuning raises the mean predicted T_c from 17.05 to 27.37 K and the $T_c \geq 20$ count to 1596. The joint SUN $\cap$ $(T_c \geq 20)$ fraction increases from 0.1723 to 0.2843, and MSUN $\cap$ $(T_c \geq 20)$ from 0.1743 to 0.3060 (Table 3). Tuning reduces the overall SUN fraction from 0.5683 to 0.5103. The corresponding mean CHGNet-relaxed hull energies are -0.43 and -0.19 eV/atom. These surrogate-based stability estimates require higher-fidelity validation; the joint metrics show how stability screening and the predicted-T_c target combine in candidate selection.

Table 3: Joint discovery and target fractions for one matched run of 3000 samples per configuration. CHGNet and structure matching determine SUN and MSUN; the structural verifier supplies predicted T_c. Joint fractions use all 3000 samples as the denominator.

Method	SUN	MSUN	SUN $\cap$ $T_c \geq 10$	SUN $\cap$ $T_c \geq 20$	MSUN $\cap$ $T_c \geq 10$	MSUN $\cap$ $T_c \geq 20$
Base	0.1417	0.2258	0.0737	0.0453	0.0790	0.0480
Harness	**0.5683**	**0.6073**	0.3407	0.1723	0.3530	0.1743
Harness + tuned generator	0.5103	0.5850	**0.3440**	**0.2843**	**0.3803**	**0.3060**

4.4 Agreement Between Property Predictors

On the three matched runs, Pearson correlations between TcProxy and the structural verifier are 0.34, 0.40, and 0.56 for Base, Harness, and Harness + tuned, respectively. Appendix C reports rank correlations and threshold-based agreement. The following experiments compare control policies across multiple seeds. Appendix D separately reports repeated component-removal experiments for an earlier four-prompt controller.

4.5 Control-Policy Comparisons, Novelty, and Sensitivity

We compare six policies using the same frozen generator and TcProxy, with a budget of 24,000 LLM proposals and 3000 candidates retained for diffusion refinement. **React** updates the prompt

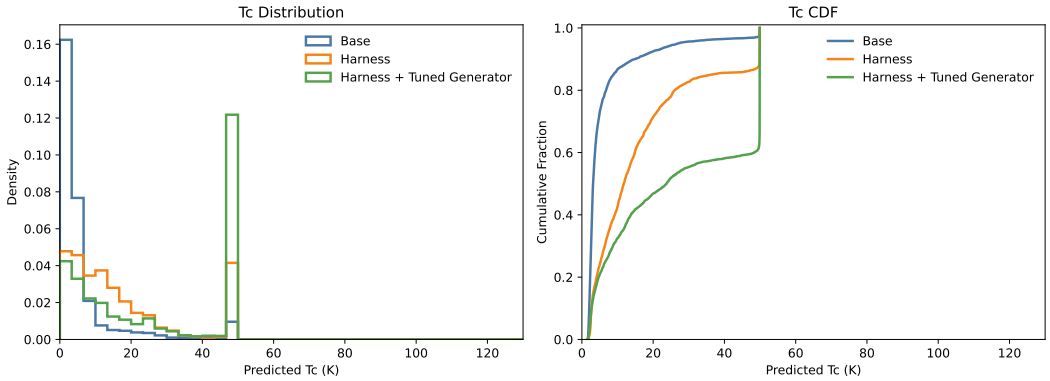

Figure 3: Targeting shift in the superconductivity setting. Both harnessed runs move the score distribution of the structural verifier (without explicit composition features) upward relative to the untargeted base.

using search summaries and retrieved knowledge. **Best-of-k** keeps the globally top-ranked 3000 parseable proposals from the full pool, prioritizing validity and TcProxy. **Verifier-only** uses fixed prompts and selects one candidate per local pool by the same ranking rule. **Random** uses seeded random prompt updates with local selection. **Knowledge-only** reuses a fixed knowledge-derived prompt, and **Control-only** uses the planner without knowledge retrieval. Table 11 reports their inference costs.

Table 4: Control policies with the same frozen generator and candidate budget. Entries are means ± standard deviations. TcProxy summaries use seeds 11, 22, 33, 44, 55, and 66. Discovery and joint metrics use seeds 22, 33, 55, and 66: CHGNet and structure matching determine SUN and MSUN, and the structural verifier supplies predicted T_c for the joint count. Each run retains 3000 candidates; discovery fractions use evaluator-ready samples. Dashes denote unavailable discovery evaluations.

	TcProxy (six seeds)		Discovery and joint metrics (four seeds)		
Control policy	Mean T_c (K)	# $T_c \geq 20$	SUN	MSUN	# SUN ∩ ($T_c \geq 20$)
React	27.83 ± 3.36	1701.5 ± 218.7	$\mathbf{0.455 \pm 0.040}$	$\mathbf{0.528 \pm 0.047}$	$\mathbf{545.5 \pm 95.7}$
Best-of-k	25.31 ± 0.53	1232.3 ± 45.6	0.257 ± 0.007	0.329 ± 0.005	274.8 ± 7.1
Verifier-only	21.24 ± 0.73	966.5 ± 32.0	0.241 ± 0.009	0.314 ± 0.011	256.5 ± 13.0
Random	6.70 ± 0.44	166.8 ± 30.3	0.149 ± 0.010	0.214 ± 0.012	134.8 ± 12.0
Knowledge-only	31.08 ± 8.79	1718.7 ± 560.2	—	—	—
Control-only	$\mathbf{35.68 \pm 8.73}$	$\mathbf{1763.8 \pm 339.2}$	—	—	—

React achieves SUN $= 0.455 \pm 0.040$, compared with 0.257 ± 0.007 for Best-of-k and 0.241 ± 0.009 for Verifier-only (Table 4). Across the four seeds, React's lowest SUN is 0.3940, above the highest baseline value of 0.2675 (Table 12). Its joint candidate count is 545.5, compared with 274.8 for Best-of-k. Best-of-k nevertheless has a higher extreme predicted-T_c tail: its maximum and top-10 mean are 120.8 and 113.3 K, compared with React's 106.9 and 95.9 K, and it returns 507 candidates above 50 K compared with 350. Knowledge-only and Control-only also have higher mean TcProxy scores than React. React's advantage over Best-of-k appears in joint discovery yield; the next comparison examines its yield at a specified distance from training compositions.

Compositional novelty of high predicted-T_c candidates. We measure nearest-neighbor Jaccard distance between the elements present in each candidate and those in MP-20 training structures, denoted nn_{Jac} (Appendix A.1). This measures differences in element sets; stoichiometry and geometry are evaluated separately. At $\text{nn}_{\text{Jac}} \geq 0.20$, React produces 749.3 ± 213.7 candidates with predicted $T_c \geq 20$ K, compared with 330.8 ± 56.0 for Knowledge-only and 531.7 ± 28.6 for Best-of-k (Table 5). React exceeds Knowledge-only in all six seeds, with a $2.26\times$ ratio of mean counts. Mean composition distance is 0.131 for React and 0.072 for Knowledge-only, indicating a broader search under this metric.

Table 5: Counts of candidates with predicted $T_c \geq 20$ K and nearest-training-composition Jaccard distance above each threshold, and mean distance (six-seed means $\pm$ standard deviations). Ratios compare React with Knowledge-only. React exceeds Knowledge-only in six of six seeds at thresholds 0.1 and 0.2, and five of six at 0.5.

Control policy	# ($T_c \geq 20 \wedge$ nn ≥ 0.1)	# ($T_c \geq 20 \wedge$ nn ≥ 0.2)	# ($T_c \geq 20 \wedge$ nn ≥ 0.5)	mean nn-Jac. distance
React	802.8 ± 195.4	$\mathbf{749.3 \pm 213.7}$	11.8 ± 7.9	0.131 ± 0.033
Knowledge-only	366.5 ± 67.8	330.8 ± 56.0	3.8 ± 2.5	0.072 ± 0.015
Best-of-k	681.0 ± 41.0	531.7 ± 28.6	4.2 ± 1.5	0.120 ± 0.004
React advantage	$2.19\times$	$\mathbf{2.26}\times$	$3.09\times$	$1.82\times$

Generator fine-tuning. Under six-seed TcProxy rescoring, the mean count above 20 K is 306.7 for the frozen generator, 289.2 for fine-tuning alone, and 1701.5 for React with the frozen generator (Table 13). The combined controller and tuned-generator experiment uses the structural verifier and is reported separately in Tables 2–3.

Sensitivity and compute. In a single-seed, one-at-a-time sweep, every tested controller configuration produces more than 1300 candidates with TcProxy ≥ 20 K, compared with the Best-of-k six-seed mean of 1232. Removing memory lowers mean predicted T_c by 5.78 K; increasing the retained fraction from 25% to 50% raises it by 5.39 K (Appendix E). All six policies use a 24,000-proposal budget and retain 3000 candidates for refinement. React uses 30 planning rounds and approximately 553K planner tokens, with a generation time of 4.47 ± 0.17 seconds per retained candidate versus 2.44 ± 0.04 for Best-of-k (Appendix F). Appendix I reports the two-seed family-card removal experiments.

5 Discussion and Conclusion

A feedback-driven controller improves the screening yield of a frozen LLM-diffusion crystal generator. In open-ended generation, the gain in CHGNet-based discovery metrics accompanies reduced reference-distribution fidelity. In superconductivity search, iterative control increases the number of candidates meeting predicted-T_c and joint discovery thresholds under matched candidate budgets. It also yields more high-scoring candidates distant from training compositions than knowledge-only prompting. The additional planner cost is measured explicitly, and generator tuning provides a further improvement in the joint screening objective.

Limitations. The experiments cover one generator family and surrogate-based evaluations. All T_c values are predictions; candidate superconductivity, phonon stability, and synthesizability have not been validated by DFT or experiment. The two T_c predictors share training data and structural representations, and their material-identity splits permit the same reduced formula to occur across splits. Calibration near the 20 K threshold and shared surrogate biases remain concerns. The family-card ablations use two seeds; open-ended generation uses three, with higher seed variation for the harness ($\sigma_{\text{SUN}} \approx 0.075$) than the base ($\sigma_{\text{SUN}} \approx 0.002$). Further evaluation should include formula-disjoint predictor splits, additional generator families, and higher-fidelity physics calculations.

Reproducibility statement

Sections 3–4 describe the controller and evaluation. Appendix A.1 defines the discovery metrics; Appendices B–C give predictor training details; Appendix K specifies planner outputs and update rules. Run configurations, sensitivity tests, and compute budgets appear in Appendices A, E, and F. Code, configurations, and knowledge-pack artifacts will be released upon publication.

AI use statement

Generative AI tools assisted with literature review, code scaffolding, manuscript editing, and review of the presentation and interpretation of results. The authors are responsible for checking AI-assisted code, claims, and reported values against the implementation and source results, and take full responsibility for the final manuscript.

REFERENCES

Luis M. Antunes, Tiarnan D. S. Crabb, Edward O. Pyzer-Knott, Joseph M. Cole, and Keith T. Butler. Crystal structure generation with large language modeling, 2024. URL https://arxiv.org/abs/2407.17791.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, September 2023. doi: 10.1038/s42256-023-00716-3. URL http://dx.doi.org/10.1038/s42256-023-00716-3.

Jingru Gan, Peichen Zhong, Yuanqi Du, Yanqiao Zhu, Chenru Duan, Haorui Wang, Daniel Schwalbe-Koda, Carla P. Gomes, Kristin A. Persson, and Wei Wang. MatLLMSearch: Crystal structure discovery with evolution-guided large language models, 2025. URL https://arxiv.org/abs/2502.20933v2.

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.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), July 2013. doi: 10.1063/1.4812323. URL http://dx.doi.org/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. *Advances in Neural Information Processing Systems*, 36:17464–17497, 2023.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. *arXiv preprint arXiv:2402.03992*, 2024.

Subhojyoti Khastagir, Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. LLM meets diffusion: A hybrid framework for crystal material generation, 2025. URL https://arxiv.org/abs/2510.23040.

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. *arXiv preprint arXiv:2502.03638*, 2025.

Benjamin Kurt Miller, Ricky TQ Chen, Anuroop Sriram, and Brandon M Wood. FlowMM: Generating materials with Riemannian flow matching. *arXiv preprint arXiv:2406.04713*, 2024.

Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. ORB: A fast, scalable neural network potential. *arXiv preprint arXiv:2410.22570*, 2024.

Pawan Prakash, Jason B. Gibson, Zhongwei Li, Gabriele Di Gianluca, Juan Esquivel, Eric Fuemmeler, Benjamin Geisler, Jung Soo Kim, Adrian Roitberg, Ellad B. Tadmor, Mingjie Liu, Stefano Martiniani, Gregory R. Stewart, James J. Hamlin, Peter J. Hirschfeld, and Richard G. Hennig. Guided diffusion for the discovery of new superconductors, 2025. URL https://arxiv.org/abs/2509.25186.

Valentin Stanev, Corey Oses, A Gilad Kusne, Efrain Rodriguez, Johnpierre Paglione, Stefano Curtarolo, and Ichiro Takeuchi. Machine learning modeling of superconducting critical temperature. *npj Computational Materials*, 4(1):29, 2018.

Brandon M Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R Kitchin, Daniel S Levine, et al. UMA: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*, 2025.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

Sherry Yang, Simon Batzner, Ruiqi Gao, Muratahan Aykol, Alexander L. Gaunt, Brendan McMorrow, Danilo J. Rezende, Dale Schuurmans, Igor Mordatch, and Ekin D. Cubuk. Generative hierarchical materials search, 2024. URL https://arxiv.org/abs/2409.06762.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. *CoRR*, abs/2210.03629, 2022. doi: 10.48550/ARXIV.2210.03629. URL https://doi.org/10.48550/arXiv.2210.03629.

Samuel Yuan and Sasa V Dordevic. Diffusion models for conditional generation of hypothetical new families of superconductors. *Scientific Reports*, 14(1):10275, 2024.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, January 2025. doi: 10.1038/s41586-025-08628-5. URL http://dx.doi.org/10.1038/s41586-025-08628-5.

Duo Zhang, Xinzijian Liu, Xiangyu Zhang, Chengqian Zhang, Chun Cai, Hangrui Bi, Yiming Du, Xuejian Qin, Anyang Peng, Jiameng Huang, et al. DPA-2: a large atomic model as a multi-task learner. *npj computational materials*, 10(1):293, 2024.

A Run Families and Evaluation Details

This appendix records the exact run families used in the paper: the open-ended matched base run `r64e2_A`; the single-ReAct open-ended run `r64e2_A_plannerapi_single_r200`; the stronger but less matched open-ended baseline `r64e2`; and the superconductivity runs `r64e2`, `r64e2_batchtc`, and `r64e2_superconduct_ds_qwen3_batchtc`. These superconductivity runs share the same `r64e2` generator family and differ only in the superconductivity-side control or tuning configuration.

Table 6: Run families used in the paper.

Run family	Setting	Key components
`r64e2_A`	open-ended matched base	plain run from the shared `r64e2_A` family used in the main matched open-ended comparison
`r64e2_A_plannerapi_single_r200`	open-ended harness	fixed LLM-diffusion generator wrapped by the single-ReAct harness
`r64e2`	appendix open-ended baseline	alternative base configuration used in the selected-run comparison
`r64e2`	superconductivity base	same base generator family as the open-ended `r64e2`, evaluated under the superconductivity task setting
`r64e2_batchtc`	superconductivity harness	single-ReAct harness, frozen TcProxy guidance, retrieved knowledge pack, final reporting with the structural verifier without explicit composition features
`r64e2_superconduct_ds_qwen3_batchtc`	superconductivity harness + tuned generator	same harness, frozen TcProxy, same knowledge pack, superconductivity-tuned generator, final reporting with the structural verifier without explicit composition features

Round sizes and superconductivity-search settings. Open-ended generation updates the planner every 200 accepted samples; superconductivity search uses 100. In the main `batchtc` launch scripts, the harnessed runs score up to eight decoded LLM candidates per prompt with the local TcProxy and update the planner once per 100 accepted samples. The main open-ended and

controlled superconductivity experiments use `deepseek-v4-pro` as the planner, with a 180s per-call timeout, two-stage critique→rewrite JSON, knowledge-pack top-$k = 4$, and keep-fraction $1/4$ (top 25% retained per round).

A.1 Discovery and Compositional-Distance Metrics

Our CHGNet discovery metrics are computed with a MatterGen-compatible evaluation wrapper that directly reuses MatterGen's reference-dataset and structure-metric utilities (Zeni et al., 2025). We evaluate novelty and uniqueness against the official `MP2020correction` reference dataset through MatterGen's `ReferenceDataset`, `FracNovelStructures`, `FracUniqueStructures`, and `DefaultDisorderedStructureMatcher` components. For ordered structures, MatterGen groups samples and reference entries by reduced formula; a generated structure is `Unique` only if it is the first representative of its matcher-equivalence class, and `Novel` if it has no matcher hit in `MP2020correction`. The default matcher uses lattice tolerance `ltol`$= 0.2$, site tolerance `stol`$= 0.3$, angle tolerance $5°$, primitive-cell reduction enabled, disallowing supercell/subset matching on the ordered path. Each CIF is relaxed with CHGNet before hull evaluation; samples whose terminal systems are missing from the reference set are excluded. We report SUN with $E_{\text{hull}} \leq 0.0$ and MSUN with $E_{\text{hull}} \leq 0.1$ eV/atom.

Compositional distance. Let $E(x)$ be the set of elements present in candidate x and $\mathcal{D}_{\text{train}}$ the 27,136 structures in the MP-20 training split. We compute

$$d_{\text{Jac}}(x) = 1 - \max_{y \in \mathcal{D}_{\text{train}}} \frac{|E(x) \cap E(y)|}{|E(x) \cup E(y)|}.$$

The distance lies in $[0, 1]$ and depends on element presence. Duplicate element sets in the reference data leave the maximum unchanged. Table 5 counts candidates meeting both a distance threshold and the TCPROXY ≥ 20 threshold, averaged over six seeds.

Joint screening metrics. We intersect the CHGNet-based SUN or MSUN indicator with the structural verifier's predicted-T_c threshold for each CIF. Counts refer to the 3000 accepted candidates in a run, and joint fractions divide by 3000. The four-seed joint results in Table 4 use seeds 22, 33, 55, and 66, matching its discovery columns.

A.2 Selected Open-Ended Runs

Table 7 reports the best-seed runs from nearby configurations. The selected harness and base have SUN $= 0.4296$ and 0.2369, respectively. These selected-run values use a different aggregation from the three matched-seed means in Table 1.

Table 7: Selected best-seed open-ended runs from nearby generator configurations; planner: `deepseek-v4-pro`.

Method	comp.	struct.	valid	W_{dens}	W_{elem}	recall	prec.	SUN	MSUN
`r64e2`	89.20	99.97	89.17	0.1765	0.0903	93.57	99.83	0.2369	0.3437
`r64e2_A_plannerapi_single_r200`	83.70	99.87	83.57	1.382	0.258	91.57	99.67	0.4296	0.5260

B Frozen TCProxy Predictor Details

The search-time TCPROXY is a frozen predictor trained beforehand and reused without online updates. Training data is constructed by cross-referencing the SuperCon database (Stanev et al., 2018) with the Materials Project (Jain et al., 2013), using T_c values from SuperCon as labels and retrieving relaxed structures from MP by formula. The dataset is split 80:10:10 via a scikit-learn `GroupShuffleSplit` grouped by Materials Project `material_id` (seed 42), so no individual structure appears in more than one split, and records appearing in multiple source tables are deduplicated by `material_id`. Polymorphs with the same reduced formula and different `material_id`s can occur across splits. For each material we extract a 100-dimensional composition vector plus zero-shot structural embeddings from UMA (Wood et al., 2025), ORB (Neumann et al., 2024), and

DPA (Zhang et al., 2024). An MLP with residual blocks fuses the four features to predict T_c under MSE loss. The stored checkpoint reaches test $R^2 = 0.921$, MAE $= 5.406$ K, RMSE $= 9.863$ K on the SuperCon–MP aligned data.

C Structural Verifier Without Explicit Composition Features and Multi-Proxy Agreement

The structural verifier uses the same SuperCon–MP data, material-identity split, and UMA/ORB/DPA features as TcProxy, with the explicit composition vector removed. It achieves held-out $R^2 = 0.655$, MAE 7.76 K, and RMSE 13.11 K. Table 8 reports predictions and agreement with TcProxy. Using the structural verifier's $T_c \geq 20$ indicator as a label, TcProxy achieves AUROC values of 0.85, 0.69, and 0.79 for Base, Harness, and Harness + tuned, respectively. The corresponding average precision values are 0.28, 0.49, and 0.82. These comparisons measure agreement between predictors that share data and representations.

Table 8: Final-reporting summary and multi-proxy-agreement statistics under the structural verifier without explicit composition features (UMA/ORB/DPA features, no composition inputs; *single matched run*).

Method	Mean T_c (K)	Median T_c (K)	$T_c \geq 20$	$T_c \geq 30$	Pearson r	Spearman ρ
Base (`r64e2`)	6.69	3.13	226	131	0.34	0.40
Harness (`r64e2_batchtc`)	17.05	11.78	859	520	0.40	0.31
Harness + tuned generator	27.37	23.77	1596	1340	0.56	0.55

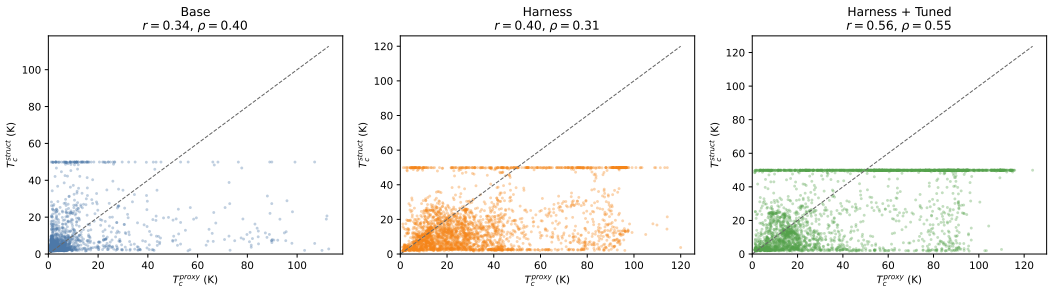

Figure 4: Agreement between TcProxy and the structural verifier on three runs of 3000 candidates. Both predictors use UMA/ORB/DPA features; the structural verifier omits the explicit composition vector.

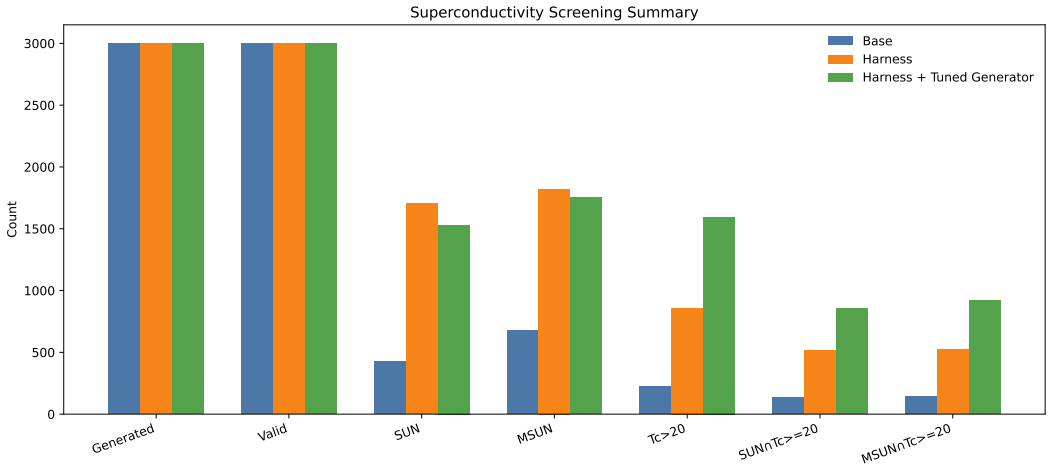

Figure 5: Numbers of candidates meeting the validity, discovery, predicted-T_c, and joint screening criteria in three matched runs. SUN and MSUN use CHGNet and structure matching; T_c uses the structural verifier. The joint columns count candidates satisfying both criteria.

D Component Removal in a Four-Prompt Controller

These experiments use an earlier controller that maintains four prompt branches with a GPT-5.4 planner. They are reported separately from the single-prompt experiments in the main text. Table 9 summarizes all available repeated runs. Removing knowledge or TCPROXY does not consistently reduce the final screening yield in this configuration; the full controller also has substantial run-to-run variation.

Table 9: Component removal in the four-prompt controller. Means $\pm$ sample standard deviations over available repeated runs; the memory-removal configuration has two runs, and the others have three. Each run returns 3000 candidates. SUN uses CHGNet and structure matching, and T_c uses the structural verifier.

Configuration	Runs	SUN	# $T_c \geq 20$	# SUN $\cap$ ($T_c \geq 20$)
Full controller	3	0.193 ± 0.069	625.7 ± 563.2	182.3 ± 166.3
No knowledge pack	3	0.161 ± 0.026	855.0 ± 120.4	258.3 ± 85.9
No TCPROXY	3	0.315 ± 0.078	776.7 ± 451.7	338.7 ± 190.7
No planner memory	2	0.150 ± 0.004	625.5 ± 51.6	190.5 ± 24.7

E Sensitivity Sweep

All tested settings exceed 1300 candidates with TCPROXY $\geq$ 20 K, compared with the Best-of-k six-seed mean of 1232. Removing memory lowers mean predicted T_c by 5.78 K. Increasing the retained fraction from 25% to 50% raises it by 5.39 K. Each setting is varied separately using seed 22.

Table 10: Sensitivity to planner settings (seed 22; TCPROXY evaluated after diffusion refinement). The default uses round size 100, four retrieved cards, memory, and a 25% retained fraction. Each row changes one setting; Δ is relative to the default.

Setting	Value	Mean T_c (K)	# $T_c \geq 20$	# $T_c \geq 50$	Δ mean T_c (K)
Default	As above	27.95	1728	334	0
Round size	50	27.17	1591	399	-0.78
Round size	200	28.85	1614	516	$+0.91$
Round size	400	23.70	1385	293	-4.25
Retrieved cards	2	23.23	1361	178	-4.72
Retrieved cards	6	26.36	1604	276	-1.59
Memory	Disabled	22.17	1305	175	-5.78
Retained fraction	10%	25.56	1424	328	-2.39
Retained fraction	20%	30.51	1939	398	$+2.56$
Retained fraction	50%	33.34	2019	626	$+5.39$

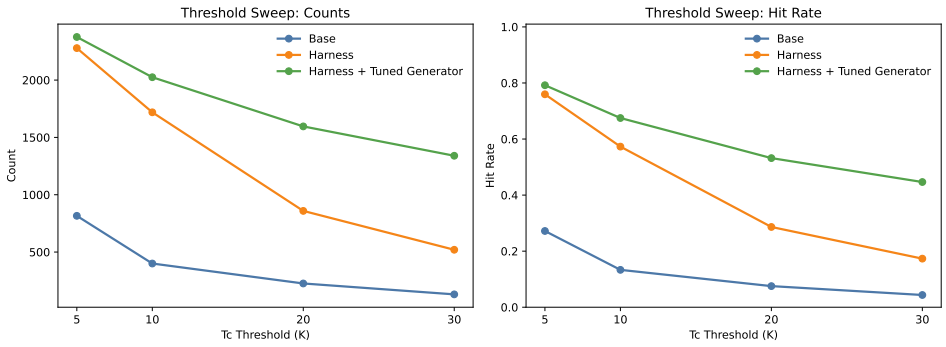

Figure 6: Candidate counts above predicted-T_c thresholds of 5, 10, 20, and 30 K for the three matched configurations, evaluated with the structural verifier. At 30 K, the harness yields 520 candidates and the harness with a tuned generator yields 1340.

F COMPUTE BUDGET

Table 11: Candidate budgets and compute costs (six-seed means ± standard deviations). Each search policy has a budget of 24,000 LLM proposals and retains 3000 candidates for diffusion refinement. React and Control-only use the planner; Base and Fine-tuning-only generate 3000 candidates without a verifier loop. The final column gives wall-clock seconds per accepted candidate.

Policy	Proposal budget	# retained	Retained fraction	Planning rounds	Planner tokens	Time (s)	s/candidate
Base	3000	3000	1.00	0	0	—	—
Finetune-only	3000	3000	1.00	0	0	—	—
Random	24000	3000	0.125	0	0	5909 ± 248	1.97 ± 0.08
Verifier-only	24000	3000	0.125	0	0	7301 ± 142	2.43 ± 0.05
Best-of-k	24000	3000	0.125	0	0	7329 ± 131	2.44 ± 0.04
Knowledge-only	24000	3000	0.125	0	0	10114 ± 222	3.37 ± 0.07
Control-only	24000	3000	0.125	30	$553{,}402 \pm 13{,}366$	13145 ± 412	4.38 ± 0.14
React	24000	3000	0.125	30	$552{,}652 \pm 11{,}364$	13400 ± 512	4.47 ± 0.17

Each search policy budgets 24,000 LLM proposals and retains 3000 for diffusion refinement. This budget counts eight proposal slots for each of 3000 selections; unparseable proposals are excluded from scoring. Mean verifier-call counts are 21,939 for React, 21,217 for Control-only, 23,556 for Best-of-k, 23,594 for Verifier-only, 23,672 for Random, and 22,416 for Knowledge-only. React and Control-only each complete 30 planning rounds per run. Each round consists of a critique request followed by a rewrite request, giving 60 requests when both stages succeed. The recorded token total, approximately 553K per run, includes both stages. React takes 4.47 ± 0.17 seconds per accepted candidate, compared with 2.44 ± 0.04 for Best-of-k and 3.37 ± 0.07 for Knowledge-only. These times cover generation, planning, in-loop scoring, and diffusion refinement. Subsequent CHGNet screening and structural-verifier evaluation are outside this timing.

G PER-SEED DISCOVERY RESULTS

Table 12 lists the four seed-level SUN values summarized in Table 4. React's lowest value is 0.3940; the highest value among the other policies is 0.2675.

Table 12: Seed-level SUN for React and three alternative control policies.

Seed	React	Best-of-k	Verifier-only	Random
22	0.3940	0.2675	0.2458	0.1512
33	0.4743	0.2583	0.2502	0.1360
55	0.5023	0.2481	0.2411	0.1455
66	0.4498	0.2557	0.2256	0.1649

H GENERATOR FINE-TUNING AND CONTROL

Table 13 compares three configurations using six-seed TCPROXY rescoring. Fine-tuning alone changes mean predicted T_c from 8.50 to 8.09 K and the count above 20 K from 306.7 to 289.2. Adding React to the frozen generator raises the count to 1701.5. The controller-plus-tuning experiment in Table 2 uses the structural verifier.

Table 13: Generator fine-tuning and control comparisons, evaluated by TCPROXY rescoring over six seeds.

Configuration	Generator	Harness	Mean T_c (K)	# $T_c \geq 20$
Base (control-off)	frozen	off	8.50 ± 0.33	306.7 ± 17.2
Fine-tuning-only	tuned	off	8.09 ± 0.12	289.2 ± 12.2
React	frozen	on	$\mathbf{27.83 \pm 3.36}$	$\mathbf{1701.5 \pm 218.7}$

I REMOVE-ONE-FAMILY-OUT ANALYSIS

We remove each family card from the Knowledge-only policy and evaluate seeds 11 and 44 (Table 14). With the complete pack, this policy produces a mean of 1981.5 candidates above 20 K on the same seeds. Removing the cuprate or nickelate card lowers this count to 736 or 1228, respectively. Removing the kagome, MgB_2-like, or superhydride card gives counts between 1974.5 and 1996. These comparisons measure the effect of knowledge-pack content under fixed prompting.

Table 14: Family-card removal under Knowledge-only prompting, evaluated with TCPROXY. All rows are means over seeds 11 and 44 with 3000 accepted candidates per run. The full-pack row uses the same seeds and policy.

Family removed	Mean T_c (K)	# $T_c \geq 20$	# $T_c \geq 50$
None (full pack)	34.83	1981.5	695.5
Cuprate layered	14.93	736	95.5
Iron-based pnictide/chalcogenide	24.03	1469.5	333.5
Nickelate (infinite-layer/square-planar)	22.61	1228	321
Layered TMD/intercalate	27.17	1549	478.5
Kagome AV_3Sb_5	34.16	1974.5	657.5
MgB_2-like boride	34.86	1985	680.5
Superhydride/clathrate	34.79	1996	684

J KNOWLEDGE-PACK RETRIEVAL AND INJECTION

Each round retrieves a subset of the superconductivity knowledge pack: round-level and global summaries extract focus elements; family cards are ranked by element-role overlap; motif cards are ranked by association with the selected families; a small prefix of heuristics is retained; and the subset is injected into planner prompts and memory as a soft prior. The retrieved material provides context for planning and does not impose hard acceptance constraints. The pack (`superconductivity_knowledge_pack_v1.json`) contains 7 family cards, 7 motif cards, 9 search heuristics, 5 element-role prior groups, and 3 usage notes, covering cuprate layered, iron-based pnictide/chalcogenide, infinite-layer and square-planar nickelates, layered transition-metal dichalcogenides, kagome AV_3Sb_5, MgB_2-like borides, and hydrogen-rich superhydrides.

K PLANNER OUTPUTS AND UPDATE RULES

At each update, the planner first assesses the recent proposals and chooses how to revise the search. A second request turns this decision into prompt conditions for the next round.

Critique output. For open-ended discovery the critique stage returns keys `{explore_mode, candidate_policy, active_profiles, stability_mode, critique, action, mutation_direction}`. For superconductivity it uses `{explore_mode, active_profiles, stability_mode, critique, action, mutation_direction, memory_update}`, where `memory_update` may contain `{mode, family_hypotheses, element_role_priors, successful_patterns, failed_patterns}`. The permitted decisions are: `explore_mode`$\in$ {conservative, balanced, exploratory}; `candidate_policy`$\in$ {balanced, coverage_first, novelty_first, diversity_first}; `stability_mode`$\in$ {guardrail_only, cautious}; `action`$\in$ {keep, mutate_small, mutate_large}; `mutation_direction`$\in$ {stabilize, increase_coverage, increase_novelty, increase_task_focus, rebalance}.

Prompt revision. For open-ended: `{prompt_plan: {profile, num_elems, element_hints, guidance}}` with `num_elems`$\in \{1, 2, 3, 4+\}$ and at most two

`element_hints`. For superconductivity: `{prompt_plan: {family_focus, preferred_element_roles, motif_preferences, avoid_patterns, preferred_num_elements, search_mode, guidance}}`. The rendered block is inserted above the base user prompt via a literal `[PLANNER]...[/PLANNER]` wrapper.

Round update rules. Let e_0 be the base exploration ratio before mode-specific clamping; the normalized ratio is $\mathrm{clip}(\min(e_0, 0.34), 0.18, 0.40)$ for conservative, $\mathrm{clip}(e_0, 0.22, 0.48)$ for balanced, and $\mathrm{clip}(e_0{+}0.05, 0.28, 0.55)$ for exploratory. Then `candidate_policy` modifies rerank weights: `coverage_first` enforces $\lambda_{\mathrm{cov}} \geq 3.3$, $\lambda_{\mathrm{profile}} \geq 1.4$; `novelty_first` enforces $\lambda_{\mathrm{nov}} \geq 2.2$, $\lambda_{\mathrm{rep}} \geq 0.95$; `diversity_first` enforces $\lambda_{\mathrm{div}} \geq 2.2$, $\lambda_{\mathrm{rep}} \geq 0.95$. The superconductivity controller instead updates a compact memory (family hypotheses with evidence count, top-10 T_c mean, hit rate above 10 K, recent trend, focus elements; successful/failed patterns; a memory mode: `anti_collapse` on repeated-system failure, `exploit` if top10_tc_mean exceeds both 10 K and the global mean, `local_expand` if the hit rate above 10 K is nonzero, `exploratory` otherwise). The mutation direction updates `search_mode`, `preferred_num_elements`, family focus, motif preferences, and avoid-patterns.

Candidate selection. Online novelty rewards formulas and element sets absent from the selected history; diversity rewards a previously unseen structure signature. Geometric plausibility rewards density between 1 and 12 g/cm^3, volume per atom between 6 and 45 Å^3, regular cell angles, and moderate lattice aspect ratios. These feedback signals are computed on provisional structures. In open-ended search the candidate is sampled from a top-k pool (sizes 3, 4, and 6 for conservative, balanced, and exploratory modes, respectively) using

$$w(x) = 1 + \lambda_{\mathrm{cov}} b_{\mathrm{cov}}(x) + \lambda_{\mathrm{profile}} b_{\mathrm{profile}}(x) + \lambda_{\mathrm{nov}} b_{\mathrm{nov}}(x) + \lambda_{\mathrm{div}} b_{\mathrm{div}}(x) + \lambda_{\mathrm{ws}} b_{\mathrm{ws}}(x) - \lambda_{\mathrm{rep}} p_{\mathrm{rep}}(x) - 0.1\,\mathrm{rank}(x).$$

In the superconductivity setting, accepted candidates are chosen greedily by the lexicographic tuple $(\text{valid}, T_c, \text{tc_hit_bonus}, b_{\mathrm{profile}}, b_{\mathrm{ws}}, b_{\mathrm{nov}}{+}b_{\mathrm{div}}, -p_{\mathrm{rep}}, \text{blocked}, \text{comp_valid}, \text{struct_valid})$.

L Backbone Diffusion Model

Our starting point is a frozen two-stage crystal generator, CrysLLMGen. Let $\mathcal{M} = \{\mathbf{M}_i\}_{i=1}^m$ be a dataset of stable crystals with each material $\mathbf{M} = (\mathbf{A}, \mathbf{X}, \mathbf{L})$, where $\mathbf{A} \in \{0,1\}^{N\times k}$ is a one-hot matrix of atom types, $\mathbf{X} \in \mathbb{R}^{N\times 3}$ fractional coordinates, and $\mathbf{L} \in \mathbb{R}^{3\times 3}$ the lattice matrix. Following the hybrid decomposition of (Khastagir et al., 2025), the first stage proposes atom types, coordinates, and lattice parameters with an LLM; we retain its predicted atom types $\mathbf{A}$; the second stage refines $\mathbf{X}$ and $\mathbf{L}$ with an equivariant diffusion model conditioned on $\mathbf{A}$. Because coordinates lie in the periodic quotient space $\mathbb{R}^{N\times 3}/\mathbb{Z}^{N\times 3}$, we use a Wrapped Normal forward process with truncation $f_w(\mathbf{X}) = \mathbf{X} - \lfloor \mathbf{X} \rfloor$, while the lattice uses a standard DDPM schedule $\bar{\alpha}_t$:

$$\mathbf{X}_t = f_w(\mathbf{X}_0 + \sigma_t \boldsymbol{\epsilon}^X) \quad \text{and} \quad \mathbf{L}_t = \sqrt{\bar{\alpha}_t}\mathbf{L}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}^L \tag{2}$$

where $\boldsymbol{\epsilon}^X, \boldsymbol{\epsilon}^L \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The coordinate target is the normalized negative wrapped-normal score,

$$\mathbf{u}_t = -c_t^{-1/2} \nabla_{\mathbf{X}_t} \log q(\mathbf{X}_t \mid \mathbf{X}_0),$$

where c_t is the scheduler's estimate of the per-coordinate second moment of the unnormalized score. The equivariant network predicts $\hat{\mathbf{u}}_t$ and lattice noise $\hat{\boldsymbol{\epsilon}}^L$, with loss

$$\mathcal{L}_{\mathrm{Diff}} = \mathbb{E}_{t\sim\mathcal{U}(1,T)}\left[\mathrm{MSE}(\mathbf{u}_t, \hat{\mathbf{u}}_t) + \mathrm{MSE}(\boldsymbol{\epsilon}^L, \hat{\boldsymbol{\epsilon}}^L)\right], \tag{3}$$

where MSE averages over the entries of each target. The controller leaves these backbone parameters fixed. Section 4.3 separately evaluates a tuned generator.

M Broader Impact

Feedback-guided crystal generation can help prioritize candidates for further study. Surrogate errors can also direct resources toward unsuitable structures. Generated candidates should undergo expert inspection, higher-fidelity simulation, and experimental validation before being treated as discoveries.