LapidaryEngine: Feedback-Guided Iterative Text-to-Crystal Structure Generation

Yusei Ito [1,2] **Yuta Suzuki** [1] **Tomoya Murata** [1] **Masaki Adachi** [1]

[1]*Lattice Lab, Toyota Motor Corporation* [2]*The University of Osaka*. Correspondence to: Tomoya Murata tomoya_murata_aa@mail.toyota.co.jp.

1. Introduction

AI for materials research has experienced rapid growth driven by recent progress in artificial intelligence, with generative models increasingly applied to materials discovery and design [1, 2, 3, 4, 5]. These approaches promise to accelerate exploration of the vast chemical and structural design space.

In parallel, large language models (LLM) have demonstrated strong capabilities in reasoning and problem formulation, enabling scientific tasks to be expressed directly in natural language [6, 7, 8, 9, 10, 11, 12]. This development opens a new paradigm for AI in materials science, in which users specify desired properties or functions in text and obtain candidate crystal structures as outputs. Such text-to-crystal structure generation enables the expression of ambiguous or qualitative desiderata that are difficult to capture using explicit physical property values alone.

Early work has demonstrated the feasibility of this idea. GenMS and Chemeleon show that crystal structures can be generated from textual descriptions [13, 14]. However, these approaches rely on carefully crafted prompts and produce results in a single step, which limits their usefulness in realistic design settings. In practice, materials discovery is iterative and uncertain: users rarely know their exact targets at the outset, and design goals evolve as new constraints emerge. One-shot generation therefore conflicts with the inherently iterative nature of materials discovery.

Enabling a conversational, multi-round interface requires models that can interpret and refine previously generated structures. However, existing methods lack this bidirectional capability because paired data between crystal structures and text are scarce. We address this gap with *LapidaryEngine*, the first framework for fully conversational crystal structure refinement.

Our key idea is a *pivot representation* that bridges text and crystal structures, enabling bidirectional mapping between crystal structures and natural language without relying on directly paired training data. We evaluated LapidaryEngine on a verifiable task and confirmed that our method can incorporate user desiderata through dialogue.

2. LapidaryEngine

Just as two languages without direct parallel data (*e.g.*, Kiswahili and Japanese) can communicate through a shared pivot language like English, we introduce a pivot representation to connect text and crystal structures. Although no generative model currently supports *crystal* → *text*, there exists a *rule-based* text generator for crystal structures: Robocrystallographer [15]. This expression is not a property of the crystal structure; rather, it is a text representation that focuses solely on the structure itself and can be generated in a rule-based manner.

As illustrated in Fig. 1(a), our workflow proceeds as follows. We first map the user's imprecise natural-language prompt to a precise pivot description using an LLM. We then employ a graph neural network (GNN)-based diffusion model [14], trained on paired (pivot, crystal) data, to generate candidate structures. After generating a candidate structure, we convert the crystal back into its pivot description and update it based on user feedback using an LLM. The refined pivot is then decoded again into a new crystal structure. This closed-loop pipeline resolves the second limitation—one-directional text-only generation—and makes iterative, conversation-style crystal design possible.

Detailed descriptions of the methods are provided in Appendix A.1.

3. Experiments

We demonstrated LapidaryEngine in Fig. 1(b) to show whether crystal structures can be designed from human feedback. Using the insulator (*i.e.*, large bandgap materials) discovery task as an example, we illustrated how the crystal structure evolves over three rounds of feedback. Each generated structure is evaluated by predicting its bandgap using the property predictor CrystalFramer [16]. In the first round, we simply provided predicted property values obtained from the property predictor. The model then modified the structure while preserving the composition, successfully widening the bandgap. In the second round, we instructed the model to replace strontium with a more readily available element. Although the bandgap slightly decreased, strontium was replaced with calcium, which is abundant on earth. In the third round, we asked the model to increase the distance between structural units to disrupt conduction pathways. The model responded by editing the structure while maintaining the composition, resulting in enlarged inter unit distances. Overall, these results demonstrate that our framework supports an iterative and conversational design process, in which initially vague user desiderata are gradually refined into more concrete design constraints, including restrictions on elemental composition, rather than being treated as a fixed property optimization problem.

4. Discussions

In this study, we demonstrated that introducing a linguistic description of structure as a pivot rep-

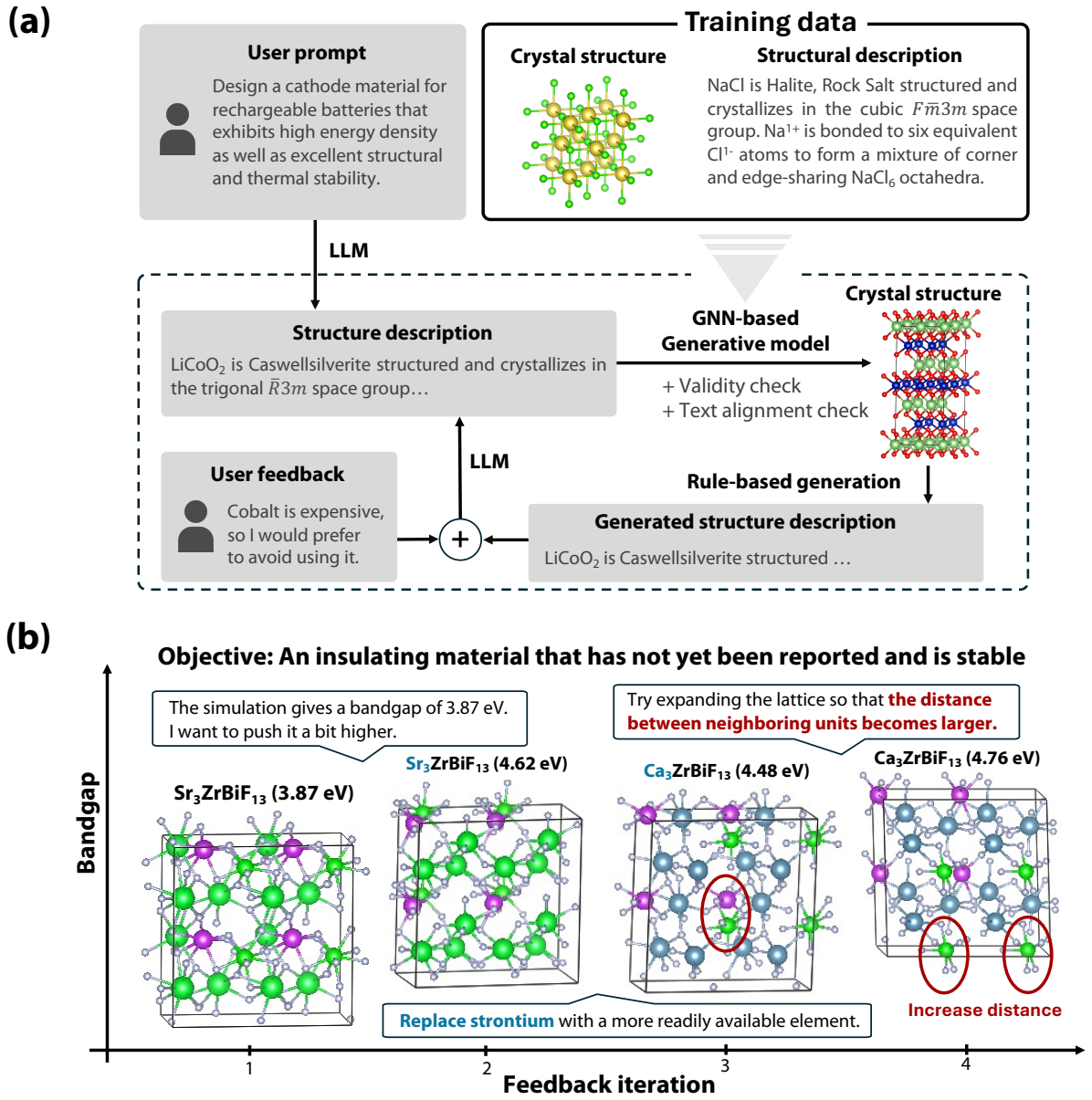

Fig. 1: (a) Overview of LapidaryEngine, feedback-enabled framework for text-guided crystal structure generation. (b) Examples of crystal structures generated by our framework.

resentation makes it possible to connect text and crystal structures in a feedback-capable manner. We adopted the notation of Robocrystallographer since it provides the most intuitive linguistic expression [15]. However other approaches such as SLICES [17] have also been developed to describe crystal structures linguistically [18]. A comparative investigation of these representations will be left for future work.

Additionally, we focused on quantifiable properties such as formation energy and bandgap, which can be evaluated through DFT calculations or their surrogate machine learning models. Meanwhile, the performance required in practical materials development is often not something that can be directly and quantitatively assessed; rather, it tends to involve a combination of diverse and sometimes qualitative requirements. Demonstrating our method in an actual materials development context would demand advanced expertise and practical knowledge specific to the field, and thus was not performed in this work. The potential of this approach to extend to complex materials development remains an interesting open question.

5. Conclusions

This study presents LapidaryEngine, which bridges the gap between traditional trial-and-error–based materials development and current ML methods that typically follow a “generate-once” paradigm. By leveraging the pivot structural descriptions provided by Robocrystallographer, we develop a feedback-driven framework that enables crystal structure generation guided by expert input. The proposed framework paves the way for introducing a new paradigm of human–AI collaboration in materials design.

References

[1] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P. Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620:47–60, 2023.

[2] Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2020.

[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

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

[5] 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, 2025.

[6] OpenAI. GPT-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).

[7] Gemini Team. Gemini: A family of highly capable multimodal models. Preprint at https://doi.org/10.48550/arXiv.2312.11805 (2023).

[8] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. Preprint at https://doi.org/10.48550/arXiv.2505.09388 (2025).

[9] Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022.

[10] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. Preprint at https://doi.org/10.48550/arXiv.2408.06292 (2024).

[11] Abbi Abdel-Rehim, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J. Collins, Elizabeth Bourne, Gareth W. Fearnley, Emma Tate, Holly X. Smith, Larisa N. Soldatova, and Ross King. Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. *Journal of The Royal Society Interface*, 22(227):20240674, 2025.

[12] Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, 646(8085):716–723, 2025.

[13] Sherry Yang, Simon Batzner, Ruiqi Gao, Muratahan Aykol, Alexander L Gaunt, Brendan McMorrow, Danilo Jimenez Rezende, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Generative hierarchical materials search. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

[14] Hyunsoo Park, Anthony Onwuli, and Aron Walsh. Exploration of crystal chemical space using text-

guided generative artificial intelligence. *Nat. Commun.*, 16(1), 2025.

[15] Alex M. Ganose and Anubhav Jain. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9(3):874–881, 2019.

[16] Yusei Ito, Tatsunori Taniai, Ryo Igarashi, Yoshitaka Ushiku, and Kanta Ono. Rethinking the role of frames for se(3)-invariant crystal structure modeling. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025.

[17] Hang Xiao, Rong Li, Xiaoyang Shi, Yan Chen, Liangliang Zhu, Xi Chen, and Lei Wang. An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning. *Nature Communications*, 14, 2023.

[18] Shuyi Jia, Aamod Varma, Pranav Manivannan, Dhruva Chayapathy, and Victor Fung. Benchmarking text representations for crystal structure generation with large language models. In *AI for Accelerated Materials Design - ICLR 2025*, 2025.

[19] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. Preprint at `https://doi.org/10.48550/arXiv.2110.06197` (2021).

[20] Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024.

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

[22] Daniel W. Davies, Keith T. Butler, Adam J. Jackson, Jonathan M. Skelton, Kazuki Morita, and Aron Walsh. Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software*, 4(38):1361, 2019.

[23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, pages 7514–7528, 2021.

[24] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. Preprint at `https://doi.org/10.48550/arXiv.2210.11427` (2022).

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 6840–6851, 2020.

[26] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems (NeurIPS 2021)*, 34:17981–17993, 2021.

[27] Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1), 2022.

[28] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. Preprint at `https://doi.org/10.48550/arXiv.2207.12598` (2022).

[29] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[30] 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), 2019.

[31] Matthew K. Horton, Patrick Huck, Ruo Xi Yang, Jason M. Munro, Shyam Dwaraknath, Alex M. Ganose, Ryan S. Kingsbury, Mingjian Wen, Jimmy X. Shen, Tyler S. Mathis, Aaron D. Kaplan, Karlo Berket, Janosh Riebesell, Janine George, Andrew S. Rosen, Evan W. C. Spotte-Smith, Matthew J. McDermott, Orion A. Cohen, Alex Dunn, Matthew C. Kuner, Gian-Marco Rignanese, Guido Petretto, David Waroquiers, Sinead M. Griffin, Jeffrey B. Neaton, Daryl C. Chrzan, Mark Asta, Geoffroy Hautier, Shreyas Cholia, Gerbrand Ceder, Shyue Ping Ong, Anubhav Jain, and Kristin A. Persson. Accelerated data-driven materials science with the materials project. *Nature Materials*, 2025.

[32] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *The Third International Conference on Learning Representations (ICLR 2015)*, 2015.

Appendix A. Method details

A.1 Details of LapidaryEngine

Algorithm A.1 summarizes the proposed framework described in Sec. 2. The initially provided target property description p_1 is converted into a Robocrystallographer-format structural description d_1 by using an LLM, and a pure noise state is prepared for crystal generation via GNN-based diffusion model. Then the loop begins, consisting of two main stages. The first stage (lines 4–7) generates a crystal structure from the structural description, while the second stage (lines 8–13) receives feedback on the generated crystal structure and, based on the result, produces the next structural description and the noise state to be denoised during generation. We adopted Alibaba's Qwen3-Next-80B-A3B-Thinking model [8] as the LLMs. Each stage is described in detail in the following Sec. A.1.1 and Sec. A.1.2.

A.1.1 Crystal structure generation from structural descriptions

We employed Chemeleon [14] to generate crystal structures from Robocrystallographer format descriptions [15]. Details of the model are provided in Sec. A.2. To encourage the model to generate crystals that are physically plausible and more faithful to the given text, we generated $N = 10$ candidate samples (*i.e.*, each starting from a different pure noise state). Among them, only the samples satisfying both structural and compositional validity were retained, following the validity criteria commonly used in previous studies [19, 20, 21]. Specifically, structural validity was determined by ensuring that no pair of atoms was closer than 0.5 Å, while compositional validity was checked by confirming overall charge neutrality using the SMACT library [22]. Finally, the structure with the highest alignment score—analogous to the CLIP score [23], which indicates the degree of semantic consistency between the generated structure and the text description—was selected as the final output. This alignment score was computed by encoding the crystal structures and the text with encoders trained during the contrastive learning stage of Chemeleon (see Sec. A.2), and then measuring the cosine similarity between their embeddings. Note that if none of the structures passes the validity check, the iteration is reset and repeated.

A.1.2 Crystal structure refinement

To design the next crystal structure, feedback $p_k^{\mathrm{fb}} = \texttt{Feedback}(c_k)$ is provided based on the generated crystal structure c_k. Although only c_k appears as an argument, it should be noted that the feedback is not derived solely from c_k itself, but from various results (*e.g.*, simulation and experimental results). We then obtain the Robocrystallographer representation of the previously generated crystal structure and supply it to the LLM along with feedback. Rather than initiating the next crystal structure generation from pure noise state, the diffusion process starts from a partially noised version of the original structure, ensuring that the refinement stays guided by the initial configuration. We control the level of noise through the denoising strength parameter $\alpha \in (0, 1]$, a technique commonly used in image-to-image tasks [24].

In this approach, crystal structure generation from feedback begins not from pure noise state but from a partially noisy state of the previously generated crystal structure corresponding to the diffusion time step $T \times \alpha$, where T denotes the fully noisy state and α controls the diffusion strength. This partial noising process enables the model to refine or adjust the output while preserving its overall character.

A.2 Crystal structure generation model

We strictly followed Chemeleon for the model that generates crystal structures from linguistic structural descriptions. After briefly describing the model architecture, we provide a detailed explanation of the dataset and training details. For a more comprehensive description of the architecture, please refer to the original paper [14].

A.2.1 Model architecture

Chemeleon is a text-guided generative model that learns to generate crystal structures through a denoising diffusion process conditioned on text embeddings obtained from a pretrained text encoder. During training, Gaussian noise is added to crystal structures, and the model learns by performing a denoising task to predict the added noise. During inference, the model starts from pure noise state and progressively denoises it to generate complete crystal structures. For the lattice constants and atomic positions, it follows the framework of Denoising Diffusion Probabilistic Models (DDPM) [25], while for the atomic species, it adopts the Discrete Denoising Diffusion Probabilistic Models (D3PM) framework [26].

Chemeleon comprises two key elements. The first component is Crystal CLIP, a cross modal contrastive learning module for pretraining the text encoder MatTPUSciBERT [27] by aligning its embeddings with the corresponding crystal structure embeddings produced by the GNN. By bringing positive text–crystal pairs closer together and pushing negative pairs farther apart, Crystal CLIP learns a shared latent space where textual representations reflect structural geometric information.

The second element is a classifier-free guided denoising diffusion model [28] that predicts the noise added to each variable of the crystal structure (*i.e.*, lattice matrices, atomic coordinates, and atom types), conditioned on the text embeddings produced by the text encoder of Crystal CLIP. The denoising network builds upon the DiffCSP framework [29], which was originally developed for crystal structure prediction tasks.

By aligning linguistic embeddings with the geomet-

Algorithm A.1 LapidaryEngine

Require:
Input:
Target property text p_1 (*e.g.*, "high electrical conductivity")
Structural description-conditioned crystal diffusion model $\mathcal{G}$
Number of feedback iterations K
Number of generations per iteration N (default: 10)
Denoising strength $\alpha \in (0, 1]$ (default: 0.1)
Ensure: Optimized crystal structure c^*

$d_1 \leftarrow \texttt{LLM\_interpret}(p_1)$
Initialize $\{\mathcal{Z}_i\}_{i=1}^{N}$ with pure noise state for the diffusion model
for $k = 1$ to K **do**
  Generate structure c_k from structural description: ▷ Details in Sec. A.1.1
    Sample N candidate structures $\{c_k^{(i)}\}_{i=1}^{N}$ using the diffusion model $\mathcal{G}$

$$c_k^{(i)} = \mathcal{G}(\mathcal{Z}_i, \text{condition} = d_k)$$

    Evaluate structural and compositional validity, and filter out invalid samples:

$$\{c_k^{(i)}\} \leftarrow \{c_k^{(i)} \mid \texttt{isValid}(c_k^{(i)}) = \texttt{True}\}$$

    Compute text–structure alignment scores and pick best:

$$c_k \leftarrow \arg\max_i \texttt{AlignmentScore}(c_k^{(i)}, d_k)$$

  Feedback and refinement: ▷ Details in Sec. A.1.2
    Provide feedback based on the generated structure $p_k^{\text{fb}} \leftarrow \texttt{Feedback}(c_k)$
    Convert c_k to structural description $d_k^{\text{gen}} \leftarrow \texttt{Robocrystallographer}(c_k)$
    Update structural description with feedback p_k^{fb}:

$$d_{k+1} \leftarrow \texttt{LLM\_refine}(p_k^{\text{fb}}, d_k^{\text{gen}})$$

    Initialize next step from partially noised state:

$$\{\mathcal{Z}_i\}_{i=1}^{N} \leftarrow \texttt{Add\_noise}(c_k, \text{strength} = \alpha)$$

    Denoise from $\{\mathcal{Z}_i\}_{i=1}^{N}$ in the next iteration
end for
return $c^* \leftarrow c_K$

ric information of crystal structures through CLIP and training the denoising model conditioned on these embeddings, the model can generate crystal structures that follow textual instructions.

A.2.2 Dataset and training details

We used the MEGNet dataset [30], which is a snapshot of the Materials Project database [31]. Following the official split, the dataset was divided into 60,000, 5,000, and 4,239 samples for training, validation, and testing, respectively. After generating textual descriptions using Robocrystallographer, we trained the model on this dataset.

The training of Chemeleon consists of two stages: (1) contrastive pretraining of the Crystal CLIP module, and (2) text-conditioned diffusion model training for crystal generation.

In the contrastive learning stage, the text encoder and the GNN based crystal encoder are trained together so that their embeddings align within a shared latent space, which helps the text embedding capture geometric information. The text embeddings are obtained from the `[CLS]` token of the text encoder output, and the crystal embeddings are produced by averaging the node features from the GNN-based crystal structure encoder. The training objective combines text-to-graph and graph-to-text cross-entropy losses with a symmetric contrastive formulation. A batch size of 128 is used with the Adam optimizer [32], where the learning rates for the text and graph encoders are set to 1×10^{-5} and 1×10^{-4}, respectively. Training proceeds for up to 1,000 epochs, employing early stopping if the validation loss does not improve for 300 epochs. A learning rate scheduler with `ReduceLROnPlateau` (patience = 200 epochs) is applied for stability.

During diffusion model training, the text encoder is kept frozen, and the pretrained Crystal CLIP embeddings are used as conditional inputs. The denoising network is optimized using the Adam optimizer with a learning rate of 1×10^{-3}, maintaining the same batch size and scheduling settings as in the contrastive learning stage. The loss function consists of three components: atom species, lattice, and coordinate denoising losses. Both trainings were performed on four NVIDIA H200 (141 GB) GPUs and took 30 hours for contrastive learning and 20 hours for diffusion model training.