Benchmarking LLMs for atomic-level geometric manipulation in crystals

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent advancements with video generators, language aligned robotics models and tool-augmented design frameworks suggest that large language models (LLMs) may soon no longer struggle with 3D spatial reasoning. To bring these developments into the material sciences, we present AtomWorld, a data generator and benchmark that evaluates LLMs on atomic-level operations (e.g. insert, move, rotate atoms) in CIF files. This benchmark was tested across major chat models, finding these models to generally take an algorithmic approach - which yielded successful completion of simple tasks such as adding and moving atoms, but struggled with more complex tasks such as rotating around an atom. LLM inaptitude with spatial reasoning limits their usefulness in crystallography - addressing this problem is a necessary first step towards enabling higher level tasks such as seeing motifs, symmetries, repairing or validating complex structures, and proposing novel structures.

1 Introduction

A Crystallographic Information File (CIF) [1] is the standard format for storing crystallographic structural data. Suppose that there are three stages for an LLM to reason with CIF files: motor skills, perceptual skills and cognitive skills. Motor skills are about the mechanics of geometry — being able to add, move, rotate, or insert atoms consistently within a structure. Perceptual skills are about recognising patterns — seeing motifs like octahedra, channels, or layered frameworks, and detecting symmetry or connectivity. Cognitive skills are about reasoning and creativity — engaging in hypothesis-driven modifications and proposing novel structures.

LLMs for crystallography would primarily benefit researchers at the cognitive stage, however challenges such as hypothesis-driven modification require LLMs to also be strong at the motor and perceptual stages. In current literature, perceptual skills have been tested through question-answer (QA) style benchmarks e.g. LLM4Mat-Bench [2], but less attention has been given to testing motor skills. To address this gap, our research question asks: how can we measure and improve LLM "crystallographic motor skills", i.e. ability to manipulate atoms in crystal structures? We present the following contributions:

1. AtomWorld Playground: A scalable data generator and benchmark that evaluates LLMs on atomic-level operations (e.g. add, move, rotate, insert atoms) in CIF files.
2. Obtained benchmark results across several frontier chat models. We found these models to generally take an algorithmic approach - which yielded successful completion of simple tasks such as adding and moving atoms, but struggled with more complex tasks such as rotating around an atom.

To the best of our knowledge, we are the first benchmark to evaluate LLM motor skills in crystallography. While these tasks are trivially solved via software or packages such as Ovito[3] and

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

Atomic Simulation Environment(ASE)[4], installing this capability in LLMs can help unlock the more valuable downstream cognitive skills. Traditionally, LLMs have struggled with spacial reasoning tasks - but this may soon change with rapid advancements in tool-augmented design [5], diffusion LLMs [6, 7], and as language aligned video generation [8, 9] and robotics [10] models become increasingly capable. We hope that our AtomWorld playground can play a foundational role in testing the understanding of 3D CIF environments in tomorrow's LLMs.

2 Related Work

LLMs for crystallography. LLMs have been primarily explored for their capabilities in CIF generation and QA. LLMs have been demonstrated to hold an innate ability to generate crystal structures when pretrained on millions of CIF files [11]. This process may be further reinforced through evolutionary search frameworks [12]. However, as LLMs are pattern predictors, the search space is fundamentally limited by the scope of the pretraining data. LLMs can also be instruction fine-tuned to predict crystal properties or provide general QA responses from CIF, e.g. AlchemBERT[13], NatureLM[14], Darwin 1.5 [15], etc[16, 17]. Crystallography QA is well benchmarked, with the most comprehensive being LLM4Mat-Bench [2], consisting of approximately 2 million composition-structure-description pairs. Tool-augmented LLMs such as OSDA Agent [5] improve structure generation through coupling computational chemistry tools to LLMs. These tool-augmented design frameworks are able to address the lack of in-depth chemistry knowledge of LLMs without expensive (and not always effective) fine-tuning. LLMs may be able to reliably handle geometric CIF modification through tool-augmentation.

Multimodal reasoning. Approaches such as multimodal chain-of-thought (Multimodal-CoT) [18], visualization-of-thought (VoT) [19] add image modalities to the reasoning trace rather than pure textual chain-of-thought. In particular, Multimodal-CoT with under 1 billion parameters achieved state of the art in state-of-the-art performance on the ScienceQA benchmark, outperforming larger models like GPT-3.5. As CIF describes a 3D challenge, these results suggest that multimodal reasoning approaches can be highly applicable to improving LLM ability on CIF geometry tasks, as well as reasoning-intensive QA and structure generation/modification tasks. Approaches to multimodal representation may also be influenced from developments in video generation and robotics, where models such as Genie 3 [9] and V-JEPA 2 [10] are increasingly capable of understanding real-world physics and integrating this with natural language input/output. Finally, with the training objective of diffusion LLMs [6, 7] to be noise reversal, they have an advantage in understanding structural text compared to autoregressive LLMs - with LLaDA [6] surpassing GPT-4o in a reversal poem completion task. This also suggests diffusion LLMs may be inherently capable of differentiating between valid and invalid modifications to CIF - important for geometric modification tasks. Developments in multimodal reasoning and diffusion suggest that LLMs may be on the cusp of being able to grasp the 3D CIF environment, making it important to benchmark this progress.

3 Playground Design: AtomWorld

At its core, AtomWorld is designed as a scalable data generator which can be used for both benchmarking and training LLMs. The data generated follows a three-part structure: two CIF files of "before" and "after" states, and an action prompt describing the change - with the goal of the LLM to yield the "after" state, given the "before" state and action. A flowchart of the benchmarking workflow is presented in Figure 1. Detailed descriptions and examples of all supported actions prompts are found in Appendix A.1.

4 Experiments & Discussion

We benchmarked a selection of state-of-the-art LLMs, including variants from Gemini, GPT, Qwen, Deepseek, and LLaMA families. The results are summarized in Figure 2a and b. According to the evaluation metrics, it is evident that LLMs exhibit varying levels of performance across tasks. Simpler operations such as `add` are performed more consistently, whereas more spatially demanding manipulations, particularly `rotate around`, remain highly challenging. Overall, the relative task difficulty can be ordered as: `add` < `move` < `move_towards` < `insert_between` <

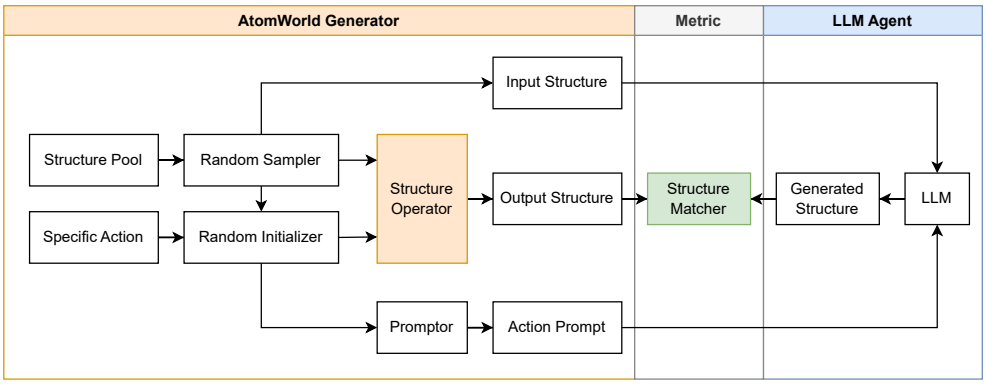

Figure 1: AtomWorld benchmark flowchart. The AtomWorld generator follows a structured data flow: the random sampler selects a structure from a predefined structure pool (in this work, a subset of CIF files from the Materials Project database[20]); the random initializer parametrizes the chosen action template by assigning atom indices and/or positions; the structure operator applies the instantiated action to the original structure to obtain the target structure; and the prompter generates a natural language description aligned with the action. The resulting (input structure, action prompt) pairs are then fed into the LLM agent system, whose generated structure is compared against the target structure using the `StructureMatcher` from *pymatgen*[21] to compute the evaluation metric (see Appendix A.2).

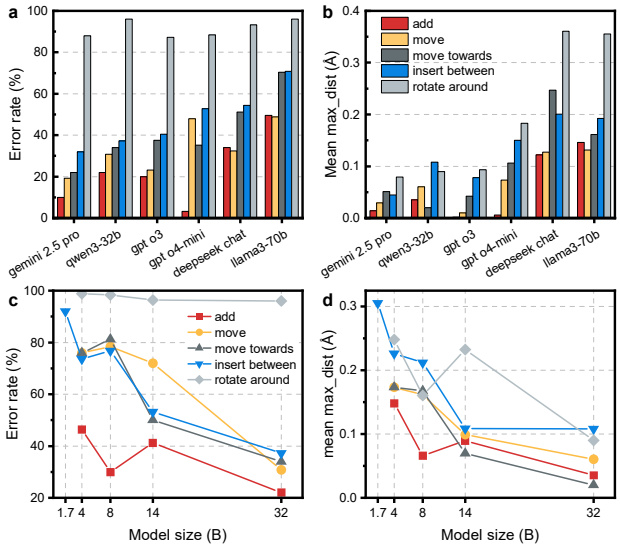

Figure 2: Evaluation results. **a** and **b** demonstrate the error rate and mean `max_dist` metrics for different actions. **c** and **d** demonstrate the change in performance with model sizes, tested using the Qwen3 series.

`rotate_around`. Gemini 2.5 Pro achieves the strongest performance across the evaluated tasks, showing particularly low error rates and displacement values in the `move`, `move_towards`, and `insert_between` tasks.

Geometric operation difficulty. To measure the inherent difficulty of each geometric operation, we tested Gemini 2.5 Pro and Deepseek V3-0324 on simplified point-based tasks, with results listed in Table 1. The models were given a set of points in three-dimensional space, expressed in raw coordinate format like "$[[x_1, y_1, z_1], [x_2, y_2, z_2]]$". The models were then asked to apply similar geometric operations directly on these points and return the transformed coordinates. This setting

removes the complexities of CIF files and serves as a controlled test of whether the LLM can handle spatial transformations at all. The results from this setting reflects the task difficulty found in AtomWorld benchmark results, observing that models perform well on simple actions like `move`, `move_towards`, and `insert_between`, but found the `rotate_around` action is significantly more difficult. The former could be solved with straightforward numerical calculations (e.g., addition or weighted averaging), which LLMs can handle reliably. In contrast, models often attempted to compute a rotation matrix for the `rotate_around` action and failed to apply it consistently, leading to high mean `max_dist`.

Table 1: Model performances on simplified point-based tasks. Error rate indicates the ratio of unreadable outputs from LLMs. Mean `max_dist` is calculated by the maximum distance between generated and target points after Hungarian sort.

	Gemini 2.5 Pro (50 frames)		Deepseek V3-0324 (250 frames)	
Action	Error rate (%)	mean `max_dist` (Å)	Error rate	mean `max_dist`
`move`	0.00	0.0000	0.00	0.0000
`move_towards`	2.00	0.0045	0.00	0.3172
`insert_between`	0.00	0.0051	21.2	0.0642
`rotate_around`	2.00	16.168	0.00	14.058

Paramater scaling. Qwen3-32B ranks second overall and is especially notable for its efficiency: despite having only 32B parameters, it outperforms or matches larger models (e.g., GPT o3, LLaMA3-70B) on several tasks. Figure 2c and d illustrate how parameter scaling of the Qwen3 series affects accuracy across tasks. In general, larger models tend to achieve lower error rates and smaller displacements, confirming that scaling improves spatial reasoning capabilities. This pattern is further supported by the Chemical Competence Score (CCS)[22], which increases with model size and highlights Qwen3-32B outperforming LLaMA3-70B. (See Appendix B.3) Nonetheless, the marginal benefits decrease with increasing model size, and for the `rotate_around` task, the improvements remain limited. These observations suggest that architectural design and training strategies play an equally important role as model scale in enabling atomic-level reasoning.

Solution approaches. Chat models generally approached these geometric challenges through generating the necessary linear algebra algorithms to solve. Failures across most CIF actions could be attributed to context-rot, as the chat models lost their train of thought across large reasoning traces. In Table 1, we found an interesting case where the Deepseek V3 model has an abnormally high error rate in the simplified `insert_between` tasks. A closer look at the wrong responses reveals that Deepseek often attempted to write a Python script to compute the coordinates, rather than directly performing the calculation.

5 Future Work & Conclusion

In this paper we presented AtomWorld as the first benchmark that evaluates LLM motor skills in crystallography. In general, we found that chat models took an algorithmic approach to solving the geometric tasks of our benchmark. With this approach, simpler operations such as add could be performed more consistently, whereas more spatially demanding manipulations, particularly rotations, remain highly challenging. These tasks are solved trivially via crystallography software, but for LLMs are an important first stage to enabling higher level tasks such as seeing motifs, symmetries, repairing or validating complex structures, and proposing novel structures.

In future work, we would like to increase the depth of our evaluation beyond frontier chat models. A stronger conclusion may be drawn about LLM capabilities through also evaluating specialised LLMs for material science, and tool-augmented LLMs. Future versions of the AtomWorld playground would likely see an expanded set of actions, prompt templates and evaluation metrics. A richer structure of modalities may also be included - e.g. graphs or visual depictions for input into multimodal LLMs.

LLMs have traditionally struggled with spacial reasoning tasks, however this may be soon to change with recent developments in tool-augmented design, diffusion, video generation and language aligned

robotics models [5, 7, 9, 10]. We hope that our AtomWorld playground can play a foundational role in helping researchers of tomorrow test LLM understanding of 3D CIF environments.

References

[1] S. R. Hall, F. H. Allen, and I. D. Brown. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A*, 47 (6):655–685, 1991. doi: https://doi.org/10.1107/S010876739101067X. URL https://onlinelibrary.wiley.com/doi/abs/10.1107/S010876739101067X. tex.eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1107/S010876739101067X.

[2] Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bousso Dieng. LLM4Mat-bench: Benchmarking large language models for materials property prediction. *Machine Learning: Science and Technology*, 6(2):020501, May 2025. ISSN 2632-2153. doi: 10.1088/2632-2153/add3bb. Publisher: IOP Publishing.

[3] Alexander Stukowski. Visualization and analysis of atomistic simulation data with OVITO-the open visualization tool. *MODELLING AND SIMULATION IN MATERIALS SCIENCE AND ENGINEERING*, 18(1), January 2010. ISSN 0965-0393. doi: 10.1088/0965-0393/18/1/015012. Number: 015012 tex.eissn: 1361-651X tex.orcid-numbers: Stukowski, Alexander/0000-0001-6750-3401 tex.researcherid-numbers: Stukowski, Alexander/G-9695-2017 tex.unique-id: ISI:000272791800012.

[4] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, June 2017. ISSN 0953-8984. doi: 10.1088/1361-648X/aa680e. URL https://dx.doi.org/10.1088/1361-648X/aa680e. Publisher: IOP Publishing.

[5] Zhaolin Hu, Yixiao Zhou, Zhongan Wang, Xin Li, Weimin Yang, Hehe Fan, and Yi Yang. OSDA agent: Leveraging large language models for de novo design of organic structure directing agents. In *The thirteenth international conference on learning representations*, 2025. URL https://openreview.net/forum?id=9YNyiCJE3k.

[6] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large Language Diffusion Models, February 2025. URL http://arxiv.org/abs/2502.09992. arXiv:2502.09992 [cs].

[7] Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, Yuwei Fu, Jing Su, Ge Zhang, Wenhao Huang, Mingxuan Wang, Lin Yan, Xiaoying Jia, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Yonghui Wu, and Hao Zhou. Seed Diffusion: A Large-Scale Diffusion Language Model with High-Speed Inference, August 2025. URL http://arxiv.org/abs/2508.02193. arXiv:2508.02193 [cs].

[8] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing Efficient Video Production for All, December 2024. URL http://arxiv.org/abs/2412.20404. arXiv:2412.20404 [cs].

[9] Google DeepMind. Genie 3: A new frontier for world models, May 2025. URL https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models.

[10] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba, Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning, June 2025. URL http://arxiv.org/abs/2506.09985. arXiv:2506.09985 [cs].

[11] Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1):10570, December 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54639-7.

[12] 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, February 2025. URL http://arxiv.org/abs/2502.20933. arXiv:2502.20933 [cond-mat].

[13] Xiaotong Liu, Yuhang Wang, Tao Yang, Xingchen Liu, and Xiaodong Wen. AlchemBERT: Exploring Lightweight Language Models for Materials Informatics, February 2025. URL https://chemrxiv.org/engage/chemrxiv/article-details/6781a6b481d2151a02a3212e.

[14] Yingce Xia, Peiran Jin, Shufang Xie, Liang He, Chuan Cao, Renqian Luo, Guoqing Liu, Yue Wang, Zequn Liu, Yuan-Jyue Chen, Zekun Guo, Yeqi Bai, Pan Deng, Yaosen Min, Ziheng Lu, Hongxia Hao, Han Yang, Jielan Li, Chang Liu, Jia Zhang, Jianwei Zhu, Ran Bi, Kehan Wu, Wei Zhang, Kaiyuan Gao, Qizhi Pei, Qian Wang, Xixian Liu, Yanting Li, Houtian Zhu, Yeqing Lu, Mingqian Ma, Zun Wang, Tian Xie, Krzysztof Maziarz, Marwin Segler, Zhao Yang, Zilong Chen, Yu Shi, Shuxin Zheng, Lijun Wu, Chen Hu, Peggy Dai, Tie-Yan Liu, Haiguang Liu, and Tao Qin. Nature language model: Deciphering the language of nature for scientific discovery, 2025. URL https://arxiv.org/abs/2502.07527. arXiv: 2502.07527 [cs.AI].

[15] Tong Xie, Yuwei Wan, Yixuan Liu, Yuchen Zeng, Shaozhou Wang, Wenjie Zhang, Clara Grazian, Chunyu Kit, Wanli Ouyang, Dongzhan Zhou, and Bram Hoex. DARWIN 1.5: Large language models as materials science adapted learners, 2025. URL https://arxiv.org/abs/2412.11970. arXiv: 2412.11970 [cs.CL].

[16] Joren Van Herck, María Victoria Gil, Kevin Maik Jablonka, Alex Abrudan, Andy S. Anker, Mehrdad Asgari, Ben Blaiszik, Antonio Buffo, Leander Choudhury, Clemence Corminboeuf, Hilal Daglar, Amir Mohammad Elahi, Ian T. Foster, Susana Garcia, Matthew Garvin, Guillaume Godin, Lydia L. Good, Jianan Gu, Noémie Xiao Hu, Xin Jin, Tanja Junkers, Seda Keskin, Tuomas P. J. Knowles, Ruben Laplaza, Michele Lessona, Sauradeep Majumdar, Hossein Mashhadimoslem, Ruaraidh D. McIntosh, Seyed Mohamad Moosavi, Beatriz Mouriño, Francesca Nerli, Covadonga Pevida, Neda Poudineh, Mahyar Rajabi-Kochi, Kadi L. Saar, Fahimeh Hooriabad Saboor, Morteza Sagharichiha, K. J. Schmidt, Jiale Shi, Elena Simone, Dennis Svatunek, Marco Taddei, Igor Tetko, Domonkos Tolnai, Sahar Vahdatifar, Jonathan Whitmer, D. C. Florian Wieland, Regine Willumeit-Römer, Andreas Züttel, and Berend Smit. Assessment of fine-tuned large language models for real-world chemistry and material science applications. *Chemical Science*, 16(2):670–684, 2025. doi: 10.1039/D4SC04401K. Publisher: The Royal Society of Chemistry.

[17] Andrea Madotto Nate Gruver, Anuroop Sriram and Zachary Ward Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *International conference on learning representations 2024*, 2024.

[18] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models, May 2024. URL http://arxiv.org/abs/2302.00923. arXiv:2302.00923 [cs].

[19] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's Eye of LLMs: Visualization-of-Thought Elicits Spatial Reasoning in Large Language Models, October 2024. URL http://arxiv.org/abs/2404.03622. arXiv:2404.03622 [cs].

[20] 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. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. ISSN 2166532X. doi: 10.1063/1.4812323.

[21] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, February 2013. ISSN 0927-0256. doi: 10.1016/j.commatsci.2012.10.028.

[22] Andres M. Bran, Tong Xie, Shai Pranesh, Jeremy Goumaz, Xuan Vu Nguyen, David Ming Segura, Ruizhi Xu, Jeffrey Meng, Dongzhan Zhou, Wenjie Zhang, and Philippe Schwaller. MiST: Understanding the Role of Mid-Stage Scientific Training in Developing Chemical Reasoning Models. In *FM4LS 2025: Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences at ICML 2025*, July 2025.

[23] Alex M. Ganose and Anubhav Jain. Robocrystallographer: automated crystal structure text descriptions and analysis. *MRS Communications*, 9(3):874–881, September 2019. ISSN 2159-6859, 2159-6867. doi: 10.1557/mrc.2019.94. URL http://link.springer.com/10.1557/mrc.2019.94.