An Autoregressive Approach For Material Generation

Niklas Dobberstein [1 2] **Jan Hamaekers** [1]

[1]*Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Schloss Birlinghoven, 53757 Sankt Augustin, Germany* [2]*University Münster, Institute for Organic Chemistry, Corrensstraße 40, 48149 Münster, Germany*. Correspondence to: Niklas Dobberstein niklas.dobberstein@uni-muenster.de.

1. Introduction

Discovering and designing materials with specific properties is of high relevance for industries ranging from energy to electronics. However, exploring the vast combinatorial chemical space remains a significant challenge. While *in silico* methods such as Density Functional Theory (DFT) have enabled systematic screening, their computational cost remains a bottleneck.

In this work, we introduce Materium, an autoregressive transformer for crystal structure generation. Unlike diffusion approaches, Materium places atoms at precise fractional coordinates in a single pass. Our method discretizes continuous 3D material representations into token sequences, enabling the model to be trained in a few hours on a single GPU, orders of magnitude faster than current state-of-the-art diffusion models.

2. Materium Framework and Evaluation

Materium is a decoder-only transformer based on the Llama 2 architecture [1], comprising approximately 43 million parameters. The core innovation lies in our tokenization strategy, which converts an inherently unordered set of atoms into an ordered, one-dimensional sequence.

2.1 Tokenization and Quantization

We represent a crystal structure as a sequence of discrete tokens including atomic elements, oxidation states, fractional coordinates, and lattice parameters. To handle continuous values, we employ a quantization technique where fractional coordinates $r_{\text{frac}} \in [0, 1)^3$ are discretized into 1024 bins. Lattice parameters are similarly normalized and quantized.

A fundamental challenge is defining the order of atoms. We investigated several deterministic strategies and found that sorting atoms by the heaviest element first (M_{high}) yielded the most stable results. The final sequence structure is represented as:

$$[\text{SOS}] \ldots [\text{Elem}|\text{Ox}][x][y][z] \ldots [\text{LATTICE}] \ldots [\text{EOS}] \quad (1)$$

This design allows the model to learn precise 3D structures without the complex post-processing required by invariant representations like SLICES.

2.2 Conditional Generation and Training

To enable controlled generation, we condition the model on various properties, including band gap, magnetic density, space group, and density. Property values are projected into embeddings and prepended to the material token sequence. We utilize a conditional dropout strategy during training to ensure robustness against missing data.

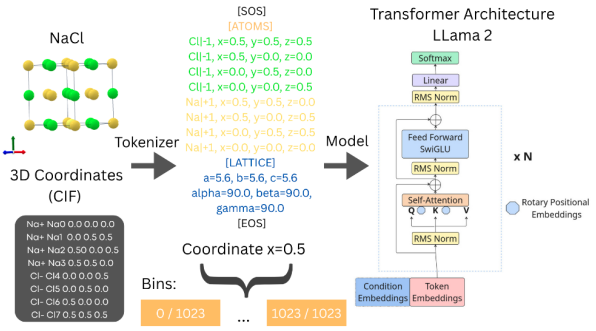

Fig. 1: Visualization of the Materium tokenizer. Continuous coordinates are discretized into 1024 bins, creating a unified sequence of atomic and lattice tokens.

We trained Materium on the Alex-MP-20 dataset [2] (approx. 600k materials) using a single A100 40GB GPU. Due to the token processing and fast algorithms like FlashAttention [3], the model can be trained in a couple of hours.

2.3 Results

We evaluated the model on 1024 generated structures. First, we compared the match rate using pymatgen's StructureMatcher [4] to measure how many generated structures successfully matched their relaxed counterparts. As shown in Table 1, Materium (M_{high}) achieves a high match rate of 85.3%, indicating that it can successfully generate new structures near their local energy minima.

Table 1: Comparison of structural match rates (relaxed vs. generated). Note that baselines were trained on MP20, while Materium utilized Alex-MP-20.

Model	Training Data	Match Rate (%)
Materium (Ours)	Alex-MP-20	85.30
FlowMM	MP20	61.39
DiffCSP	MP20	51.49
CDVAE	MP20	45.31

We further assessed the quality of generated structures using the Stability, Uniqueness, and Novelty

(S.U.N.) metrics. As detailed in Table 2, Materium (M_{high}) demonstrated excellent uniqueness (99.8%) and good stability (73.9%). However, the novelty rate (46.5%) and the combined S.U.N. score (22.7%) are lower than the state-of-the-art diffusion model MatterGen. While Materium generates fewer novel stable crystals, it offers a significant advantage in training and generation speed.

Table 2: S.U.N. metrics comparison. Materium achieves high stability and uniqueness, though with lower novelty than diffusion baselines.

Model	Stab. (%)	Uniq. (%)	Nov. (%)	S.U.N. (%)
Materium (Ours)	73.9	99.8	46.5	22.7
MatterGen [2]	73.2	99.8	61.6	38.9

For conditional generation, Materium performs consistently well on fundamental properties like density and space group. We observed that the model generalizes reasonably well even in data-sparse regions, such as high magnetic density or large band gaps, though variance increases.

2.4 Related Work

Recently, generative models have emerged as powerful tools for inverse design. Diffusion-based approaches, such as MatterGen [2], have shown great promise but rely on iterative denoising steps that are computationally expensive to train and sample. Alternatively, language model-based methods like CrystaLLM [5] generate structures using text-based formats (e.g., CIF), but often suffer from inefficient tokenization where coordinates are split into multiple character tokens.

In terms of structural match rates, we compared our approach to recent models including FlowMM [6], DiffCSP [7], and CDVAE [8]. While these baselines were trained on the MP20 dataset (unlike Alex-MP-20 used here), the comparison highlights the effectiveness of our specialized tokenization strategy in generating physically valid structures.

Acknowledgments

This work was supported in parts by the Fraunhofer Internal Programs under Grant No. PREPARE 40-08394 (ESPINN).

References

[1] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[2] 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.

[3] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022.

[4] 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.

[5] Luis M. Antunes, Keith T. Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Nature Communications*, 15(1), December 2024.

[6] Benjamin Kurt Miller, Ricky T. Q. Chen, Anuroop Sriram, and Brandon M Wood. Flowmm: Generating materials with riemannian flow matching, 2024.

[7] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion, 2024.

[8] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation, 2022.