MatExpert: Decomposing Materials Discovery By Mimicking Human Experts

Qianggang Ding[1,2], **Santiago Miret**[3], **Bang Liu**[1,2]*
[1] DIRO & Institut Courtois, Université de Montréal
[2] Mila - Quebec AI Institute
[3] Intel Labs
{qianggang.ding, bang.liu}@umontreal.ca
santiago.miret@intel.com

Abstract

Material discovery is a critical research area with profound implications for various industries. In this work, we introduce MatExpert, a novel framework that leverages Large Language Models (LLMs) and contrastive learning to accelerate the discovery and design of new materials. Inspired by the workflow of human material experts, our approach integrates three key stages: retrieval, transition, and generation. In the initial retrieval stage, MatExpert identifies an existing material that closely matches the desired criteria. Subsequently, in the transition stage, MatExpert outlines the necessary modifications to transform this material into one that meets specific requirements. Finally, in the generation state, MatExpert handles the detailed computations and structural generation. Our experimental results demonstrate that MatExpert outperforms state-of-the-art methods in material generation tasks, achieving superior performance across various metrics including validity, distribution, and stability. As such, MatExpert represents a meaningful advancement in computational material discovery using modern machine learning.

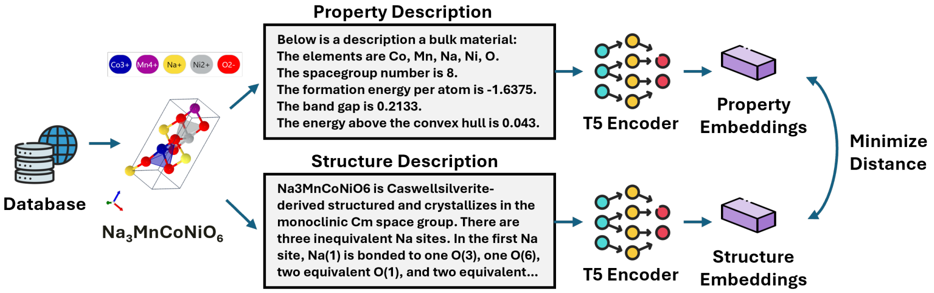
1 Introduction

The discovery of new materials is fundamental to advancements in various industries, including energy, automotive, aerospace, and others. Traditional material discovery methods often involve labor-intensive experimental and computational procedures, making the process time-consuming and costly [12]. The advent of artificial intelligence, particularly the development of Large Language Models (LLMs), presents new opportunities to streamline and accelerate this process [11, 9].

In recent years, large language models (LLMs) such as GPT-4 [13] have demonstrated exceptional capabilities in understanding and generating human language in various domains, including material science [4, 5, 15]. In this work, we introduce MatExpert, a novel framework designed to emulate the workflow of a material expert when discovering or designing new materials [20]. MatExpert integrates three critical stages: retrieval, transition, and generation. During the retrieval stage, we utilize contrastive learning to identify the most similar material from the material database based on user-defined criteria. This stage ensures that the selected material serves as a robust reference point for further modification. The transition stage leverages an LLM to generate detailed pathways for modifying the retrieved material. This involves outlining specific structural or compositional changes required to achieve the desired properties, effectively bridging the gap between the existing material and the target material. In the final generation stage, MatExpert produces lattice vectors and atomic coordinates, and a crystallographic information file (CIF) [7] for the designed material. This stage

[*] Corresponding Author. Canada CIFAR AI Chair.

AI4Mat Workshop on 38th Conference on Neural Information Processing Systems (NeurIPS 2024).



Below is a description a bulk material:
The elements are Co, Mn, Na, Ni, O.
The spacegroup number is 8.
The formation energy per atom is -1.6375.
The band gap is 0.2133.
The energy above the convex hull is 0.043.

Na3MnCoNiO6 is Caswellsilverite-derived structured and crystallizes in the monoclinic Cm space group. There are three inequivalent Na sites. In the first Na site, Na(1) is bonded to one O(3), one O(6), two equivalent O(1), and two equivalent...

Figure 1: We utilize a contrastive learning framework to train the material retrieval model. For a given sample material (e.g. Na$_3$MnCoNiO$_6$), we extract both its property description and structural description. The model employs two T5-based encoders, which are trained to minimize the distance between these two representations.

involves fine-tuning the LLM on multi-turn datasets, encompassing both the transition and generation stages, to generate novel and stable materials.

We evaluate the performance of MatExpert through extensive experiments on benchmark datasets from the Materials Project database [10]. Our results demonstrate that MatExpert significantly outperforms existing state-of-the-art methods in material generation tasks, achieving superior performance across various metrics, including validity, distribution, and stability.

2 Background

Large Language Models (LLMs) LLMs are deep learning models using Transformer architecture [19] to process and generate human language. Trained on vast text data, they produce contextually relevant responses from a prompt and instruction. Recent work has proposed diverse architectures specialized to materials science, including BERT-based models [16] and billion-parameters LLMs for materials science question-answering tasks [17, 22, 21], property prediction [2] and crystal structure generation [6]. In this work, we created a multi-turn dataset with structured prompts for material generation to train the specific MatExpert.

Contrastive Learning & LoRA Fine-Tuning Contrastive learning [3] is a self-supervised method for learning representations by bringing positive pairs closer and pushing negative pairs apart in the embedding space. We apply it to material retrieval by embedding material properties and structures in a shared space, enabling efficient retrieval of materials that best match specific queries. LoRA [8] reduces the computational overhead of fine-tuning LLMs by injecting trainable low-rank matrices into each Transformer layer, keeping the original weights fixed. This approach accelerates fine-tuning while adapting LLMs efficiently to material science tasks.

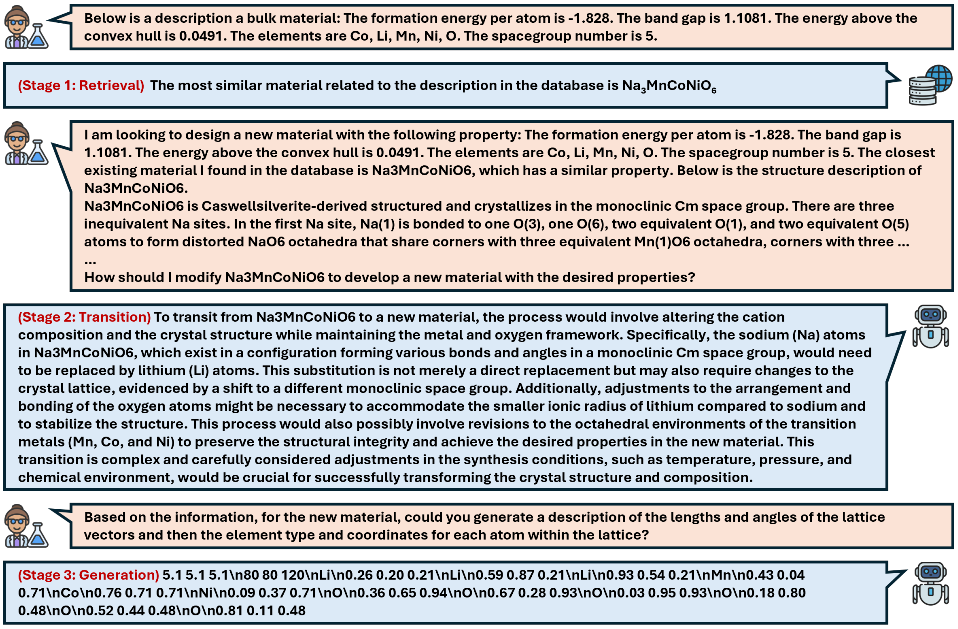
3 Methodology

Our framework addresses material discovery by imitating expert processes through three stages: retrieval, transition, and generation, using a T5-based [14] contrastive learning framework and an LLM for material retrieval and material design, respectively.

3.1 First Stage: Retrieval

The retrieval stage identifies the most similar material from the material database based on user criteria, using contrastive learning. Two T5-based encoders [14] embed the property descriptions and structure descriptions into a shared space. The objective is to minimize the distance between the property embedding ($\mathbf{q}_i$) and the structure embedding ($\mathbf{m}_i$) while maximizing the distance from other materials ($\mathbf{m}_k$) (See Figure 1). A contrastive loss function is applied:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(\mathbf{q}_i, \mathbf{m}_i)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\mathbf{q}_i, \mathbf{m}_k)/\tau)}. \tag{1}$$



Below is a description a bulk material: The formation energy per atom is -1.828. The band gap is 1.1081. The energy above the convex hull is 0.0491. The elements are Co, Li, Mn, Ni, O. The spacegroup number is 5.

(Stage 1: Retrieval) The most similar material related to the description in the database is $Na_3MnCoNiO_6$

I am looking to design a new material with the following property: The formation energy per atom is -1.828. The band gap is 1.1081. The energy above the convex hull is 0.0491. The elements are Co, Li, Mn, Ni, O. The spacegroup number is 5. The closest existing material I found in the database is Na3MnCoNiO6, which has a similar property. Below is the structure description of Na3MnCoNiO6.
Na3MnCoNiO6 is Caswellsilverite-derived structured and crystallizes in the monoclinic Cm space group. There are three inequivalent Na sites. In the first Na site, Na(1) is bonded to one O(3), one O(6), two equivalent O(1), and two equivalent O(5) atoms to form distorted NaO6 octahedra that share corners with three equivalent Mn(1)O6 octahedra, corners with three ...
...
How should I modify Na3MnCoNiO6 to develop a new material with the desired properties?

(Stage 2: Transition) To transit from Na3MnCoNiO6 to a new material, the process would involve altering the cation composition and the crystal structure while maintaining the metal and oxygen framework. Specifically, the sodium (Na) atoms in Na3MnCoNiO6, which exist in a configuration forming various bonds and angles in a monoclinic Cm space group, would need to be replaced by lithium (Li) atoms. This substitution is not merely a direct replacement but may also require changes to the crystal lattice, evidenced by a shift to a different monoclinic space group. Additionally, adjustments to the arrangement and bonding of the oxygen atoms might be necessary to accommodate the smaller ionic radius of lithium compared to sodium and to stabilize the structure. This process would also possibly involve revisions to the octahedral environments of the transition metals (Mn, Co, and Ni) to preserve the structural integrity and achieve the desired properties in the new material. This transition is complex and carefully considered adjustments in the synthesis conditions, such as temperature, pressure, and chemical environment, would be crucial for successfully transforming the crystal structure and composition.

Based on the information, for the new material, could you generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice?

(Stage 3: Generation) 5.1 5.1 5.1\n80 80 120\nLi\n0.26 0.20 0.21\nLi\n0.59 0.87 0.21\nLi\n0.93 0.54 0.21\nMn\n0.43 0.04 0.71\nCo\n0.76 0.71 0.71\nNi\n0.09 0.37 0.71\nO\n0.36 0.65 0.94\nO\n0.67 0.28 0.93\nO\n0.03 0.95 0.93\nO\n0.18 0.80 0.48\nO\n0.52 0.44 0.48\nO\n0.81 0.11 0.48

Figure 2: Pipeline of MatExpert: Given a description of the desired material, MatExpert first retrieves the most similar material from the database (e.g., $Na_3MnCoNiO_6$). Next, the LLM provides transition pathways to modify the retrieved material into the desired material (e.g., replacing Na with Li). Finally, the LLM generates the detailed structural information of the desired material ($Li_3MnCoNiO_6$).

This stage outputs the best matching material for further refinement.

3.2 Second Stage: Transition

The transition stage focuses on developing a detailed and viable method for modifying the retrieved material to meet the desired properties, which bridges the gap between the existing material and the target material by outlining specific structural or compositional changes required to achieve the desired specifications.

Prompt Construction: To facilitate the generation of transition pathways, we construct a dataset designed to simulate an interactive dialogue between a researcher and the model. This approach ensures that the LLM can interpret and respond to complex queries regarding material modification. See Figure 2 for example.

LLM-Guided Transition Pathways: To generate ground-truth outputs for fine-tuning the LLM, we utilize GPT-4 by providing it with carefully crafted prompts. These prompts elicit detailed explanations of the structural or compositional modifications necessary to achieve the desired material properties. An example template of prompt is:

```
I have two materials:  <formula_closest> and <formula_desired>.  Based
on the descriptions and properties of the two materials below, can
you summarize the main reasons for the differences in properties when
transitioning from <formula_closest> to <formula_desired>?  The description
of <formula_closest>:  <description_closest>.  The description of
<formula_desired>:  <description_desired>.
```

3.3 Third Stage: Generation

In the final stage, the framework generates a language representation of the material, including lattice vectors and atomic coordinates, which is then converted to CIF format [7]. The LLM is actually fine-tuned with LoRA [8] on the multi-turn datasets, including both the second stage and the third stage (See Figure 2).

Table 1: MatExpert shows general outperformance compared to baseline methods for unconditional generation on Materials Project based on various metrics like validity checks, coverage, property distribution, and stability. The best results are in **bold**, while the second-best results are underlined.

Method	Validity Check		Coverage		Distribution		Metastable	Stable
	Struc↑	**Comp↑**	**Rec↑**	**Prec↑**	**wdist(ρ)↓**	**wdist(N_{el})↓**	**M3GNet↑**	**DFT↑**
CDVAE	**1.00**	0.867	0.991	**0.995**	0.68	1.43	28.8%	5.4%
LM-CH	0.848	0.835	0.992	0.978	0.86	0.13	n/a	n/a
LM-AC	0.958	0.889	**0.996**	0.985	0.69	0.09	n/a	n/a
Crystal-LLM with LLaMA-2								
7B ($\tau = 1.0$)	0.918	0.879	0.969	0.960	3.85	0.96	35.1%	6.7%
7B ($\tau = 0.7$)	0.964	0.933	0.911	0.949	3.61	1.06	35.0%	6.2%
70B ($\tau = 1.0$)	0.965	0.863	0.968	0.983	1.72	0.55	35.4%	10.0%
70B ($\tau = 0.7$)	0.996	0.954	0.858	0.989	0.81	0.44	49.8%	10.6%
MatExpert with LLaMA-2								
7B ($\tau = 1.0$)	0.920	0.908	0.984	0.994	0.49	0.11	37.5%	7.8%
7B ($\tau = 0.7$)	0.946	0.943	0.952	0.986	0.51	0.13	36.2%	8.0%
70B ($\tau = 1.0$)	0.968	0.878	0.982	0.986	0.23	0.06	50.2%	11.3%
70B ($\tau = 0.7$)	0.998	0.959	0.969	0.993	0.23	0.07	50.9%	11.8%
MatExpert with LLaMA-3								
8B ($\tau = 1.0$)	0.933	0.910	0.992	**0.995**	0.36	0.07	39.1%	8.1%
8B ($\tau = 0.7$)	0.975	0.959	0.988	0.990	0.38	0.10	39.5%	8.4%
70B ($\tau = 1.0$)	0.970	0.880	0.980	0.988	0.21	0.05	50.5%	11.0%
70B ($\tau = 0.7$)	0.998	**0.961**	0.986	0.991	**0.18**	**0.04**	**51.0%**	**12.0%**

4 Experiments

We perform comparison experiments to evaluate the effectiveness of the MatExpert framework in the material generation task on benchmark datasets.

Datasets and Baselines We leverage datasets from the Materials Project database similar to prior studies [20, 5]. The MP-20 dataset [20] for unconditional generation contains 45,231 stable materials, filtered to exclude structures with over 30 atoms per unit cell for computational efficiency. Our baseline include: **Crystal-LLM [5]:** A state-of-the-art approach fine-tuning large language models (LLaMA-2 [18]) for generating stable inorganic materials. It excels in generating valid crystal structures and serves as a primary baseline due to its pioneering use of LLMs for material generation. **CDVAE [20]:** A Crystal Diffusion Variational Autoencoder tailored for periodic material generation. It combines generative models within a continuous VAE latent space, widely used for generating stable materials and considered a strong benchmark. **LM-CH and LM-AC [4]:** Two Transformer-based models for 3D crystal structure generation. LM-CH uses character-level tokenization, while LM-AC uses atom and coordinate-level tokens, both demonstrating strong performance in generating valid and diverse crystal structures.

4.1 Comparison Results

The results for all baselines and MatExpert are shown in Table 1. As evident from the table, MatExpert family with LLaMA-3 [1] outperforms all other methods across all metrics except metrics of coverage. Traditional models such as CDVAE, LM-CH, and LM-AC excel in coverage and structural validity. When compared to the state-of-the-art LLM-based model Crystal-LLM, our MatExpert method consistently outperforms Crystal-LLM across all settings, including the same model size and temperature as well as DFT-based stability. Notably, MatExpert achieves remarkable results in distribution metrics, benefiting from the transition from an existing material in the database.

5 Conclusion

In this work, we introduced MatExpert, a novel framework leveraging LLMs and contrastive learning for material discovery process. By emulating the workflow of a human materials science expert, MatExpert integrates retrieval, transition, and generation stages to design new materials. Our experiments show that MatExpert significantly outperforms state-of-the-art methods in material generation tasks, thereby exhibiting its potential to become scalable tool for accelerating material discovery with higher quality crystal generation.

References

[1] Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com. `https://ai.meta.com/blog/meta-llama-3/`. [Accessed 06-09-2024].

[2] Nawaf Alampara, Santiago Miret, and Kevin Maik Jablonka. Mattext: Do language models need more than text & scale for materials modeling? In *AI for Accelerated Materials Design-Vienna 2024*, 2024.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

[4] Daniel Flam-Shepherd and Alán Aspuru-Guzik. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as xyz, cif, and PDB files. *CoRR*, abs/2305.05708, 2023.

[5] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary W. Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *ICLR*. OpenReview.net, 2024.

[6] Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary Ward Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023.

[7] Sydney R Hall, Frank H Allen, and I David Brown. The crystallographic information file (cif): a new standard archive file for crystallography. *Foundations of Crystallography*, 47(6):655–685, 1991.

[8] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022.

[9] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.

[10] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

[11] Santiago Miret and NM Krishnan. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*, 2024.

[12] Santiago Miret, NM Anoop Krishnan, Benjamin Sanchez-Lengeling, Marta Skreta, Vineeth Venugopal, and Jennifer N Wei. Perspective on ai for accelerated materials design at the ai4mat-2023 workshop at neurips 2023. *Digital Discovery*, 2024.

[13] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[14] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[15] Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T Koch, José A Márquez, and Kevin Maik Jablonka. From text to insight: Large language models for materials science data extraction. *arXiv preprint arXiv:2407.16867*, 2024.

[16] Yu Song, Santiago Miret, and Bang Liu. Matsci-nlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. *arXiv preprint arXiv:2305.08264*, 2023.

[17] Yu Song, Santiago Miret, Huan Zhang, and Bang Liu. Honeybee: Progressive instruction finetuning of large language models for materials science. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, 2023.

[18] 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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

[19] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[20] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi S. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *ICLR*. OpenReview.net, 2022.

[21] Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, et al. Darwin series: Domain specific large language models for natural science. *arXiv preprint arXiv:2308.13565*, 2023.

[22] Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based agent system for materials science. *arXiv preprint arXiv:2409.00135*, 2024.