# Contrastive Language–Structure Pre-training Driven by Materials Science Literature

**Yuta Suzuki**[1] **Tatsunori Taniai**[2] **Ryo Igarashi**[2] **Kotaro Saito**[3,4] **Naoya Chiba**[5]
**Yoshitaka Ushiku**[2] **Kanta Ono**[4]
[1]Toyota Motor Corporation [2]OMRON SINIC X Corporation [3]Randeft, Inc.
[4]Osaka University [5]Tohoku University

## Abstract

Understanding structure–property relationships is an essential yet challenging aspect of materials discovery and development. To facilitate this process, recent studies in materials informatics have sought latent embedding spaces of crystal structures to capture their similarities based on properties and functionalities. However, abstract feature-based embedding spaces are human-unfriendly and prevent intuitive and efficient exploration of the vast materials space. Here we introduce Contrastive Language–Structure Pre-training (CLaSP), a learning paradigm for constructing cross-modal embedding spaces between crystal structures and texts. CLaSP aims to achieve material embeddings that 1) capture property- and functionality-related similarities between crystal structures and 2) allow intuitive retrieval via user-provided description texts as queries. To compensate for the lack of sufficient datasets linking crystal structures with textual descriptions, CLaSP leverages a dataset of over 400,000 published crystal structures and corresponding publication records, including paper titles and abstracts, for training. We demonstrate the effectiveness of CLaSP through text-based crystal structure screening and embedding space visualization.

## 1 Introduction

The properties of materials, ranging from low-level properties such as bandgap and formation energy to high-level functionalities such as superconductivity, are determined by their crystal structures [5, 8]. Thus, unlocking the structure–property relationships of materials is key to accelerating materials discovery and development.

AI-driven materials science pursues this ambition through the use of machine learning (ML). One area of research has focused on predicting material properties using graph neural networks [32, 7, 6, 17] and transformers [34, 28], leveraging large-scale crystal structure datasets annotated with properties simulated by first-principles calculations. Although this approach has shown success, the models are specialized for specific simulatable properties, such as bandgap, and are unable to provide a comprehensive view of materials with diverse properties and functionalities.

Other studies have explored developing embedding spaces of crystal structures to capture their similarities based on properties and functionalities [33, 26, 16, 22]. However, these efforts are limited by the lack of dedicated training datasets with diverse property and functionality annotations, resulting in abstract, unannotated embedding spaces that are not easily navigable for materials discovery.

The annotation cost and model interpretability are common problems in ML, leading to the exploration of learning paradigms that use natural language text descriptions, instead of class labels, for supervision. The seminal work, CLIP (Contrastive Language–Image Pre-Training) [23], pioneered this approach by using contrastive learning between image and description text pairs. By learning to

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

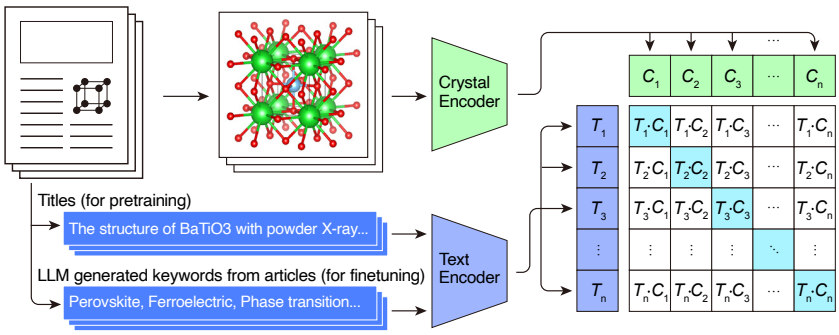


Figure 1: Contrastive learning paradigm of CLaSP in two stages. (1) Pre-training using pairs of crystal structures and publication titles. (2) Fine-tuning using pairs of crystal structures and keywords that are generated from the titles and abstracts using an LLM.

align two embedding spaces across the two modalities, CLIP enables cross-modal retrieval between images and texts, and zero-shot recognition of images using text-based prompts.

The success of CLIP has inspired language-supervised representation learning for molecular structures [35, 31, 18, 25, 27, 12] and crystal structures [19, 20]. However, existing methods for materials use textual descriptions about structural features rather than properties [19, 20], thus limiting the ability to capture high-level information such as material functionalities.

To overcome this limitation, we introduce Contrastive Language–Structure Pre-training (CLaSP) (Fig. 1). CLaSP leverages a large-scale dataset of published crystal structures and corresponding article information retrieved from the Crystallography Open Database (COD) [10]. Specifically, we utilize publication titles for pre-training and a combination of titles and abstracts for fine-tuning. We hypothesize that these textual sources offer a comprehensive representation of material characteristics. Extensive analyses demonstrate that CLaSP effectively learns structure embeddings that capture abstract and complex material concepts, such as 'superconductor' and 'metal-organic frameworks'.

## 2 Contrastive Language–Structure Pre-training

We propose using a dataset of crystal structures paired with their publication information, such as titles and abstracts, for language–structure contrastive learning. By assuming that these texts convey material characteristics, this approach aims to link crystal structure embeddings with material properties and functionalities through human-interpretable linguistic semantics.

We used a total of 406,048 crystal structures associated with paper titles retrieved from the COD, as detailed in Appendix A. CLaSP uses these captioned structures to jointly train a crystal encoder and a text encoder. For each training iteration, the crystal encoder transforms a batch of crystal structures into embeddings $\{\boldsymbol{c}_i\}$, whereas the text encoder transforms the paired caption texts into embeddings $\{\boldsymbol{t}_i\}$. CLaSP aligns the two encoders by minimizing the large margin cosine loss function [30]:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s(\cos(\boldsymbol{c}_i, \boldsymbol{t}_i) - m))}{\exp(s(\cos(\boldsymbol{c}_i, \boldsymbol{t}_i) - m)) + \sum_{j=1, j\neq i}^{N} \exp(s\ \cos(\boldsymbol{c}_i, \boldsymbol{t}_j))}, \tag{1}$$

where $N$ is the batch size, $s > 0$ is a scaling hyperparameter that amplifies cosine similarities to make the loss function more sensitive to similarity differences, enhancing training effectiveness, and $m \in [0, 1]$ is a margin hyperparameter that enforces a gap between positive-pair similarity $\cos(\boldsymbol{c}_i, \boldsymbol{t}_i)$ and negative-pair similarities $\cos(\boldsymbol{c}_i, \boldsymbol{t}_j)$. This loss function is the same as the cross entropy loss in CLIP [23] when $m = 0$. We found that incorporating the margin leads to better generalization in downstream tasks, as shown in hyperparameter studies in Appendix B.

We consider keyword-based crystal structure screening as a demonstrative downstream task and hence perform fine-tuning using keyword captions instead of titles. To this end, we identified abstracts for 80,813 entries of the training set and generated keywords for these entries from their title–abstract pairs using a large language model (LLM). We used Meta's Llama3 (70B Instruct) [1] to generate up to 10 keywords, such as 'visible light photocatalysis' and 'narrow bandgap', for each crystal structure. The overall dataset generation procedure is detailed in Appendix A.

# 3 Experiments

## 3.1 Encoder and training details

A CGCNN model [32] was trained from scratch to serve as the crystal encoder. Additionally, a frozen pre-trained SciBERT model [4] was used for the text encoder, followed by a three-layer multilayer perceptron (MLP) fed with the `CLS` token embedding. We used the embedding dimensionality of 768 for both modalities, and implemented the networks using PyTorch [2] and PyTorch Geometric [9].

We divided the dataset into training, validation, and test splits in an 8:1:1 ratio, and also divided its keyword-based subset accordingly. We optimized the loss function with scaling factor $s$ of 3 and margin $m$ of $0.5$, using stochastic gradient descent with global batch size $N$ equal to 16,384 (2,048 $\times$ 8 GPUs). Title-based pre-training was performed for a total of 2000 epochs, followed by keyword-based fine-tuning for additional 50 epochs. We used the AdamW optimizer [13] with a constant learning rate of $2 \times 10^{-5}$ for pre-training and $1 \times 10^{-6}$ for fine-tuning. Training was performed on a single server with eight NVIDIA A100 GPUs (80GB VRAM), taking approximately 16 hours overall.

## 3.2 Zero-shot crystal structure screening by text

To evaluate CLaSP's ability to link crystal structures with textual property descriptions, we performed crystal structure retrieval using keywords representing material functionalities (*e.g.*, 'thermoelectric' and 'superconductor'). Given the embedding of a query keyword, we retrieved structure embeddings from the test set that showed high cosine similarities with the keyword embedding. We regarded a structure to possess a queried property if the associated paper title contained the keyword or its variations (*e.g.*, for 'superconductor,' the terms 'superconductive' and 'superconductivity' were also considered correct). The trade-off between true positives and false positives was evaluated using the ROC (receiver operatorating characteristic) and ROC-AUC (area under the ROC curve).

Figure 2 shows ROC curves for six query keywords before and after fine-tuning. While the zero-shot prediction using the pre-trained model (Fig. 2, left) already demonstrates good performance, with an average ROC-AUC of 0.7185, fine-tuning (Fig. 2, right) further improved it to 0.7804. These results highlight the ability of the CLaSP models to capture complex structure–property relationships across diverse material functionalities by analyzing crystal structures alone.

To further validate the retrieval performance, we retrieved the top-100 materials using keywords related to bandgap—specifically, 'narrow-bandgap material' and 'insulator'—and analyzed their bandgap distributions. Since the COD does not provide property labels, we predicted bandgaps of materials by using a state-of-the-art property prediction model, Crystalformer [28], with pretrained model weights (specifically the seven-block model trained on the JARVIS-DFT 3D 2021 dataset) provided by the authors. Figure 3 shows violin plots of the bandgaps for the retrieved materials. These distributions successfully reflect the expected bandgap ranges for narrow-bandgap materials (*i.e.*, bandgaps smaller than $1.1$ eV) and insulators (*i.e.*, large bandgaps), compared to the random sampling distribution.

## 3.3 Embedding space visualization

To demonstrate how the proposed language–structure embedding can intuitively navigate the materials space, we created several visualizations of the structure embeddings from the test set using t-SNE.

First, we created a *world map* of COD materials to analyze the alignment of the learned materials space and semantics. We grouped the structure embeddings into 20 clusters using k-means++ and assigned an LLM-generated keyword label to each cluster that summarizes the associated paper titles, as detailed in Appendix C. The resulting map in Fig. 4a shows a meaningful distribution of materials, forming lands of clusters of similar materials, such as an organic materials land, a complexes land, and an inorganic materials land. The map suggests that the model recognizes material similarities that are intuitive to human. In contrast, embeddings learned without textual information often fail to capture such high-level semantic relationships, as shown in a comparison provided in Appendix D.

Furthermore, cosine similarity-based heat maps allow us to easily identify regions relevant to a given text query. Figure 4b shows that 'superconductor' is highly correlated with intermetallic compounds and oxides, while Fig. 4c shows 'metal-organic frameworks' is aligned with organic compounds.

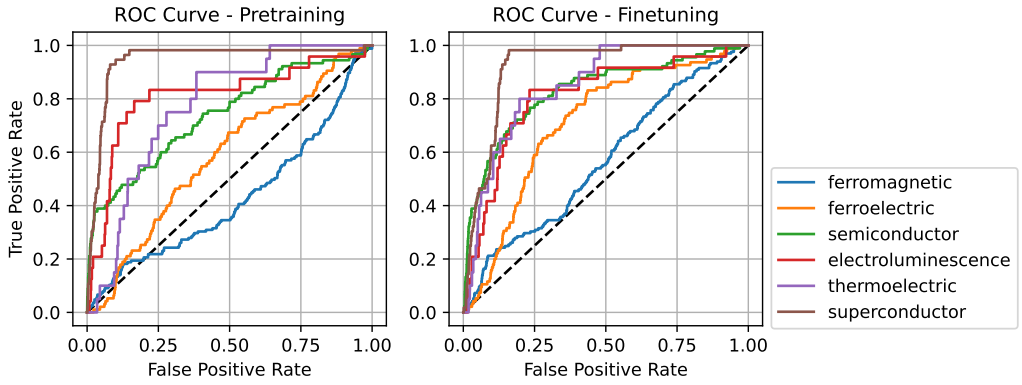


Figure 2: ROC curves of keyword-based crystal structure retrieval. The zero-shot results with only pre-training (left) show good performance, and fine-tuning leads to further improvements (right).

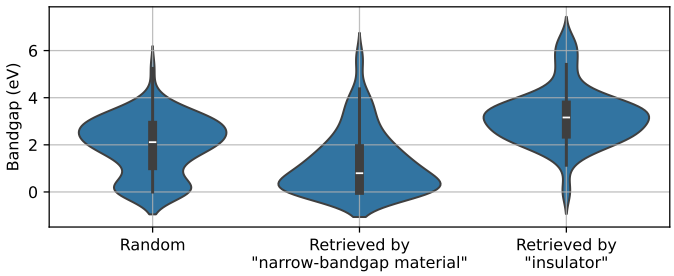


Figure 3: Violin plots of bandgaps for crystals retrieved via keyword searches. The distributions reflect the expected bandgap ranges for narrow bandgap materials and insulators, successfully demonstrating the retrieval of materials with targeted properties.

Finally, we verified the alignment between material properties and the map constitution by overlaying predicted bandgaps on the map. As done in Sec. 3.2, we used the pretrained Crystalformer [28] to predict the bandgaps of the COD materials. The resulting bandgap distribution in Fig. 4d shows consistencies with the map (Fig. 4a). For example, the right part in Fig. 4d with larger bandgaps corresponds to organic compounds in Fig. 4a, and the bottom part with near-zero bandgaps corresponds to intermetallic compounds. These results suggest that the embeddings not only capture intuitive semantics of materials but also reflect their similarities in terms of material properties.

## 4 Discussion and limitations

The results in Sec. 3 have confirmed that the publication information can provide a strong supervision in learning crystal structure embeddings and linking them to material properties. However, titles in the materials science literature tend to highlight a specific and potentially intriguing aspect of the reported materials, rather than provide a comprehensive description. For example, in Fig. 2a, the retrieval result with ‘superconductor’ outperforms the results with ‘semiconductor’ and the others, despite the more complex structure–property relationships involved. We hypothesize that this is due to the high co-occurrence of the presence of superconductivity in the reported materials and ‘superconductor’ in titles, whereas ‘semiconductor’ is less prioritized in titles. A similar issue may explain the relatively low retrieval accuracy for ‘ferromagnetic’ in Fig. 2a. Since ferromagnetism is a major characteristic of Fe, and iron-based materials are widely utilized, paper titles may often omit ‘ferromagnetic.’ This could introduce noise into the title-based supervision and hinder the learning of this material concept. Meanwhile in Fig. 2b, fine-tuning using keywords derived from titles and abstracts improved the overall retrieval performance, suggesting that abstracts convey more comprehensive information about the materials. This implies further possible improvements to CLaSP by incorporating richer training data sources beyond publication titles and abstracts, such as full texts, figures, and tables. For

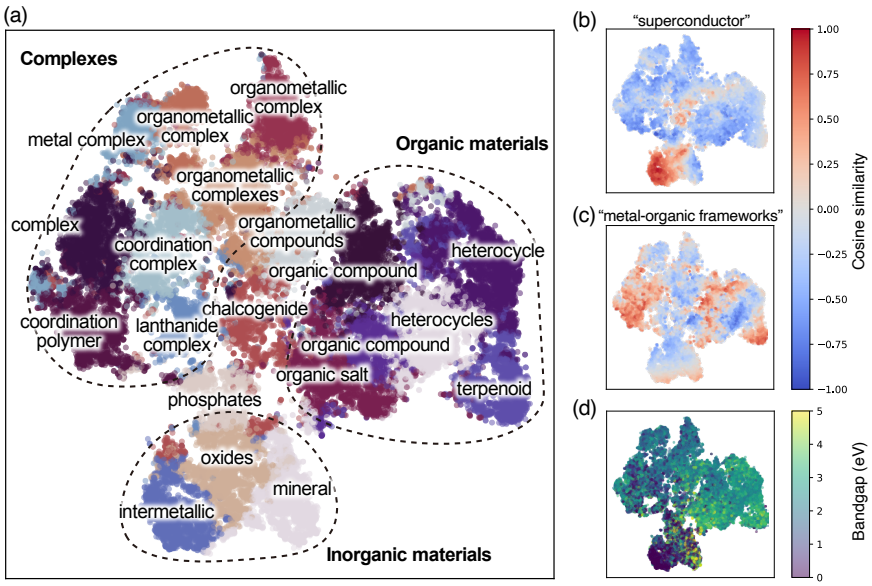


Figure 4: t-SNE visualization of crystal structure embeddings. (a) *World map* of COD materials. The embeddings are grouped into 20 clusters and assigned keywords that represent the paper titles associated with the clusters. (b, c) Heat maps showing cosine similarities between the structure embeddings and query-text embeddings ('superconductor' and 'metal-organic frameworks'). (d) Bandgap distribution of crystal structure embeddings. Predicted bandgaps trend to reflect known properties of material clusters in (a), such as larger bandgaps for organic compounds and near-zero bandgaps for intermetallic compounds.

example, when papers cite another publication that reports a specific crystal structure, their citation contexts may provide meaningful text descriptions for the structure.

We also analyzed the source journals used in the COD and found a potential bias in the dataset towards crystallography and chemistry publications (Appendix E). This bias suggests the need for more comprehensive and diverse data sources encompassing a broader range of materials and properties. Given the limited availability of large materials databases with publication records beyond the COD, we could augment it by utilizing external sources, such as citation contexts and Wikipedia entries related to materials science, in an approach similar to retrieval-augmented generation (RAG) [15] in LLM applications. We leave such extensions as future work.

## 5 Conclusion and broader impacts

In this study, we introduced CLaSP, a literature-driven learning paradigm for constructing cross-modal embedding spaces that connect crystal structures with their textual property descriptions. We demonstrated its effectiveness in learning structure embeddings that capture functionality-level material similarities and in enhancing the materials space with intuitive linguistic semantics.

These promising results indicate the potential to transform how we explore the vast materials space. Potential applications include crystal structure screening and tagging via text prompts. Furthermore, inspired by text-to-image generation [24], CLaSP's text embeddings could guide crystal structure generation models [11, 36], enabling innovative applications such as text-to-crystal generation. These advancements promise a more intuitive and efficient approach to exploring the materials space.

## Acknowledgments and Disclosure of Funding

R. I. is partly supported by JSPS KAKENHI Grant Number 24K23910. N. C. is partly supported by JSPS KAKENHI Grant Number 21K14130. Y. U. is partly supported by JST-Mirai Program Grant Number JPMJMI21G2 and JST Moonshot R&D Program Grant Number JPMJMS2236.

## References


[1] AI@Meta. Llama 3 model card. 2024.

[2] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, Asplos '24, pages 929–947, New York, NY, USA, 2024. Association for Computing Machinery.

[3] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620. Association for Computational Linguistics.

[5] William D Callister and David G Rethwisch. *Materials Science and Engeneering*. John Wiley and Sons, January 2010.

[6] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.*, 2(11):718–728, 2022.

[7] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.*, 31(9):3564–3572, May 2019.

[8] Marc De Graef and Michael E McHenry. *Structure of Materials*. An Introduction to Crystallography, Diffraction and Symmetry. Cambridge University Press, October 2012.

[9] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[10] Saulius Gražulis, Daniel Chateigner, Robert T. Downs, A. F. T. Yokochi, Miguel Quirós, Luca Lutterotti, Elena Manakova, Justas Butkus, Peter Moeck, and Armel Le Bail. Crystallography Open Database – an open-access collection of crystal structures. *J. Appl. Crystallogr.*, 42(4):726–729, 2009.

[11] Rui Jiao, Wenbing Huang, Peijia Lin, Jiaqi Han, Pin Chen, Yutong Lu, and Yang Liu. Crystal structure prediction by joint equivariant diffusion. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 17464–17497. Curran Associates, Inc., 2023.

[12] Benjamin Kaufman, Edward C. Williams, Carl Underkoffler, Ryan Pederson, Narbe Mardirossian, Ian Watson, and John Parkhill. COATI: Multimodal Contrastive Pretraining for Representing and Traversing Chemical Space. *J. Chem. Inf. Model.*, 64(4):1145–1157, 2024.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[14] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances*

*in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

[16] Qinyang Li, Rongzhi Dong, Nihang Fu, Sadman Sadeed Omee, Lai Wei, and Jianjun Hu. Global Mapping of Structures and Properties of Crystal Materials. *J. Chem. Inf. Model.*, 63(12):3814–3826, June 2023.

[17] Yuchao Lin, Keqiang Yan, Youzhi Luo, Yi Liu, Xiaoning Qian, and Shuiwang Ji. Efficient approximations of complete interatomic potentials for crystal property prediction. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[18] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nat Mach Intell*, 5(12):1447–1457, December 2023.

[19] Viggo Moro, Charlotte Loh, Rumen Dangovski, Ali Ghorashi, Andrew Ma, Zhuo Chen, Samuel Kim, Peter Y. Lu, Thomas Christensen, and Marin Soljačić. Multimodal Learning for Materials, 2024. *arXiv preprint arXiv:2312.00111*.

[20] Keisuke Ozawa, Teppei Suzuki, Shunsuke Tonogai, and Tomoya Itakura. Graph-text contrastive learning of inorganic crystal structure toward a foundation model of inorganic materials. *STAM Methods*, 0:2406219, 2024.

[21] Pavlin G. Poličar, Martin Stražar, and Blaž Zupan. openTSNE: A modular python library for t-SNE dimensionality reduction and embedding. *Journal of Statistical Software*, 109(3):1–30, 2024.

[22] Jiaxing Qu, Yuxuan Richard Xie, Kamil M. Ciesielski, Claire E. Porter, Eric S. Toberer, and Elif Ertekin. Leveraging language representation for materials exploration and discovery. *npj Comput. Mater.*, 10(1):1–14.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, April 2022. *arXiv preprint arXiv:2204.06125*.

[25] Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing Activity Prediction Models in Drug Discovery with the Ability to Understand Human Language, June 2023. *arXiv preprint arXiv:2303.03363*.

[26] Yuta Suzuki, Tatsunori Taniai, Kotaro Saito, Yoshitaka Ushiku, and Kanta Ono. Self-supervised learning of materials concepts from crystal structures via deep neural networks. *Mach. Learn.: Sci. Technol.*, 3(4):045034, February 2022.

[27] Seiji Takeda, Indra Priyadarsini, Akihiro Kishimoto, Hajime Shinohara, Lisa Hamada, Hirose Masataka, Junta Fuchiwaki, and Daiju Nakano. Multi-modal Foundation Model for Material Design. In *AI for Accelerated Materials Design - NeurIPS 2023 Workshop*, November 2023.

[28] Tatsunori Taniai, Ryo Igarashi, Yuta Suzuki, Naoya Chiba, Kotaro Saito, Yoshitaka Ushiku, and Kanta Ono. Crystalformer: Infinitely connected attention for periodic structure encoding. In *The Twelfth International Conference on Learning Representations*, 2024.

[29] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *J. Mach. Learn. Res.*, 9(86):2579–2605, January 2008.

[30] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[31] Jie Wang, Zihao Shen, Yichen Liao, Zhen Yuan, Shiliang Li, Gaoqi He, Man Lan, Xuhong Qian, Kai Zhang, and Honglin Li. Multi-modal chemical information reconstruction from images and texts for exploring the near-drug space. *Brief. Bioinform.*, 23(6):bbac461, 2022.

[32] Tian Xie and Jeffrey C Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.*, 120(14):145301, April 2018.

[33] Tian Xie and Jeffrey C Grossman. Hierarchical visualization of materials space with graph convolutional neural networks. *The Journal of Chemical Physics*, 149(17):174111, November 2018.

[34] Keqiang Yan, Yi Liu, Yuchao Lin, and Shuiwang Ji. Periodic graph transformers for crystal material property prediction. In *Advances in Neural Information Processing Systems*, volume 35, pages 15066–15080. Curran Associates, Inc., 2022.

[35] Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nat. Commun.*, 13(1):862, February 2022.

[36] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Ryota Tomioka, and Tian Xie. Mattergen: a generative model for inorganic materials design, 2024. *arXiv preprint arXiv:2312.03687*.

# Appendix

## A Details for dataset preparing

### Data retrieval

This study used the Crystallography Open Database (COD) [10] as the source of crystal structure data. Compared to other crystal structure databases, the COD is particularly well-suited for our purposes, as it provides publication information (including titles and DOIs) for each crystal structure entry and is available in the public domain. From the COD, we retrieved 512,312 pairs of crystal structures and their corresponding publication records as of March 2024. Using the DOIs from these records, we further extracted the abstracts of the papers via the Crossref API. This process collected abstracts for 141,311 entries, representing approximately 27.6% of the entire dataset.

### Data preprocessing and spliting

We filtered out the entries with structures containing more than 500 atomic sites, resulting in a dataset of 406,048 crystal structures with corresponding paper titles and DOIs. We randomly split the dataset into training, validation, and test sets in an 8:1:1 ratio, yielding 324,838 entries for training, 40,604 for validation, and 40,606 for testing. We used the train set for title-based pre-training, the validation set for selecting a model checkpoint during pre-training or fine-tuning, and the test set for evaluating the ROC (Fig. 2) and visualizing the embedding space (Fig. 4). We also used this dataset to generate the keyword-captioned dataset for fine-tuning, as explained next.

### Keyword dataset generation for fine-tuning

We derived the keyword-captioned dataset from the main dataset by removing entries without abstract from each split, ensuring no mixing across the splits. For each entry with a title and abstract pair, we generated up to 10 representative keywords using an LLM, specifically Meta's Llama3 (70B Instruct) [1]. The prompt template used for keyword generation is listed below. The keyword generation process took 36 hours using a server equipped with eight NVIDIA A100 GPUs (80GB VRAM) and an efficient LLM inference framework, vLLM [14]. Finally, we removed generated keywords if they were unrelated to material properties, such as 'crystal structure', 'X-ray diffraction', 'neutron diffraction', 'powder diffraction', and 'single-crystal X-ray diffraction'. Entries without any remaining keywords were also removed from the dataset. This process resulted in 80,813 entries for the training set, which was used to fine-tune the pre-trained model for the keyword-based retrieval task. The remaining two sets, containing 10,134 entries for validation and 10,197 for testing, were never used in this study. Note that the validation during fine-tuning was done based on the average ROC-AUC score, instead of the validation loss, for the validation set of the main dataset.

```
def prompt_format_func(material_id, title, abstract):
    """Formats the prompt for the Gemini model."""
```

```
prompt_template = """Below are title-abstract pairs for materials science papers dealing with crystal
    ↪ structures. For each paper, list up to 10 keywords in English that describe the features,
    ↪ functions, or applications of the material discussed. Focus on the material itself, and do not
    ↪  include general terms or measurement techniques (e.g., Crystal Structure, Crystal Lattice, X-
    ↪ ray diffraction, Neutron Diffraction, Powder Diffraction). Return the results in json format
    ↪ with the following schema.

**Example input 1:**

‘‘‘
ID: 0001
Title: Enhancement of Critical Temperature in Layered Copper Oxide Superconductors via Lattice
    ↪ Compression Techniques
Abstract: Superconductivity in copper oxides (cuprates) offers vast potential for technological
    ↪ applications due to their high critical temperatures (Tc). Our research presents a novel
    ↪ approach to enhance Tc in layered cuprate materials through the controlled application of
    ↪ lattice compression. Using advanced crystallographic methods, we systematically altered the
    ↪ interlayer spacing and analyzed the resultant changes in electronic properties. Our findings
    ↪ demonstrate a significant improvement in superconducting behavior at elevated temperatures,
    ↪ further supporting the unconventional mechanisms underpinning superconductivity in these
    ↪ materials.
‘‘‘

**Example output 1:**

‘‘‘json
[ {
    "ID": "0001",
    "Keywords": [
      "High-Tc",
      "Cuprate Superconductors",
      "Lattice Compression",
      "Electronic Properties",
      "Layered Structures",
      "Superconducting Phase",
      "Temperature Enhancement",
      "Unconventional Superconductivity"
    ]
  }]
‘‘‘

**Example input 2:**

‘‘‘
ID: 0002
Title: Advancements in Biodegradable Polymers for Sustained Drug Delivery Systems
Abstract: The development of biocompatible and biodegradable materials is critical in the field of
    ↪ medical implants and drug delivery systems. This paper examines the latest advancements in
    ↪ biodegradable polymers tailored for sustained release of therapeutic agents. We analyze
    ↪ various polymer compositions that provide controlled degradation rates and compatibility with
    ↪ a range of drugs. Our results show promising applications in long-term treatments, reducing
    ↪ the need for repeated administration and improving patient compliance.
‘‘‘

**Example output 2:**

‘‘‘json
[ {
    "ID": "0002",
    "Keywords": [
      "Biomaterials",
      "Biodegradable Polymers",
      "Sustained Release",
      "Drug Delivery Systems",
      "Biocompatibility",
      "Controlled Degradation",
      "Therapeutic Agents",
      "Medical Implants",
      "Long-Term Treatment"
    ]
  }]
‘‘‘
"""
prompt = prompt_template + f"""
**Input :**

‘‘‘
ID: {material_id}
Title: {title}
Abstract: {abstract}
‘‘‘
```

```
**Output :**

‘‘‘json
"""
return prompt
```

## B Hyperparameter studies

We investigated the dependency of the model's performance on the margin and scale hyperparameters in the loss function (Eq. 1). Note that our loss, adapted from CosFace [30], coincides with the loss in CLIP [23] when the margin is set to zero. We trained the model with various combinations of margin $m \in \{0, 0.3, 0.5\}$ and scale $s \in \{1.0, 1.5, 2.0, 2.5, 3.0, 3.5\}$, and evaluated the average ROC-AUC scores both before and after fine-tuning.

The results in Table 1 indicate that both parameters impact performance. Particularly, a higher margin tends to increase validation scores after fine-tuning, suggesting that the margin loss promotes better generalization in downstream tasks. We further analyzed the ROC curves of the best CLIP loss model ($m = 0$ and $s = 2.0$) in Fig. 5, and compared them with the best CosFace loss model ($m = 0.5$ and $s = 3.0$) in Fig. 2. The comparison shows that introducing a margin leads to well balanced performance across various keywords.

Table 1: ROC-AUC comparison of keyword-based crystal structure retrieval. The numbers in **bold** indicate the best results and the numbers with underline indicate the second best results.

| Loss | Margin | Scale | Pre-trained (val) | Fine-tuned (val) | Fine-tuned (test) |
|---|---|---|---|---|---|
| CLIP [23] | 0.0 | 1.0 | 0.6310 | 0.6943 | - |
| | 0.0 | 1.5 | 0.6526 | 0.6521 | - |
| | 0.0 | 2.0 | **0.7285** | **0.7837** | 0.7778 |
| | 0.0 | 2.5 | 0.6553 | 0.6791 | - |
| | 0.0 | 3.0 | 0.6946 | 0.7227 | - |
| | 0.0 | 3.5 | 0.6053 | 0.6856 | - |
| CosFace [30] | 0.3 | 1.0 | 0.5156 | 0.6495 | - |
| | 0.3 | 1.5 | 0.7170 | 0.7273 | - |
| | 0.3 | 2.0 | 0.6074 | 0.6701 | - |
| | 0.3 | 2.5 | 0.6925 | 0.7365 | - |
| | 0.3 | 3.0 | 0.6223 | 0.7395 | - |
| | 0.3 | 3.5 | 0.6498 | 0.7496 | - |
| | 0.5 | 1.0 | 0.6282 | 0.6994 | - |
| | 0.5 | 1.5 | 0.7164 | 0.7763 | - |
| | 0.5 | 2.0 | 0.5832 | 0.7006 | - |
| | 0.5 | 2.5 | 0.6347 | 0.7778 | - |
| | 0.5 | 3.0 | 0.7185 | 0.7828 | **0.7804** |
| | 0.5 | 3.5 | 0.6764 | 0.7031 | - |

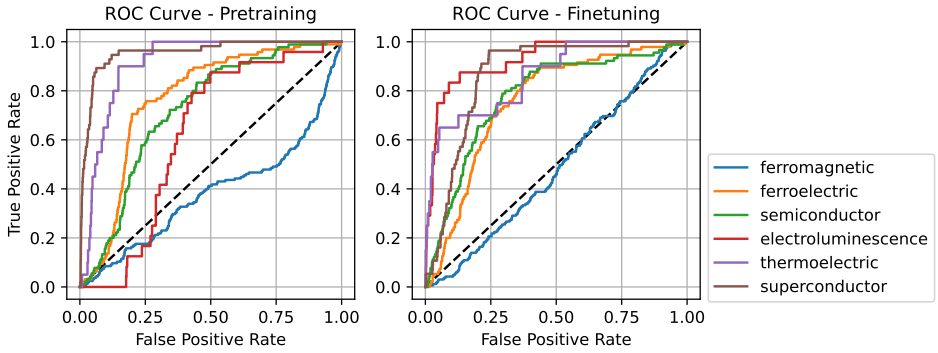


Figure 5: ROC curves of keyword-based crystal structure retrieval on the test set (CLIP loss).

## C Details for embedding visualization

The t-SNE algorithm [29] implemented in openTSNE [21] was used for embedding space visualization, and k-means++ [3] was used for embedding clustering. Google Gemini 1.5 Pro with the default temperature parameter of 1.0 was used to generate the cluster keywords in Fig. 4.

## D Comparison with an existing approach

We compared CLaSP with an existing crystal embedding learning approach called Contrastive Materials Metric Learning (CMML) [26]. CMML also employs contrastive learning, but it aligns the embeddings of two complementary structural representations: crystal structures and their corresponding X-ray diffraction (XRD) patterns. Since XRD patterns can be easily simulated from crystal structures, CMML is regarded as a self-supervised learning approach that only requires a collection of unannotated crystal structures for training. However, unlike CLaSP, which learns structure–language relationships, contrastive learning with purely structural data (structure and XRD pattern pairs) produces abstract embeddings that limit human-intuitive understanding. This distinction also prevents a quantitative comparison between CLaSP and CMML on the text-based retrieval task in Sec. 3.2.

To assess how well CLaSP and CMML embeddings capture high-level semantics of materials, we visually examined how these methods map semantically similar materials in their respective embedding spaces. Specifically, we generated embeddings of the crystal structures in the COD test set using both the CLaSP model and the pretrained CMML model (trained on the Materials Project dataset) publicly provided by the authors. We then created t-SNE visualizations of these embeddings, highlighting entry points whose corresponding publication titles included specific keywords—specifically, 'superconductor' and 'metal-organic framework.'

The results in Fig. 6 show that, while CMML randomly scatters these keyword-specified entries across the map (left), CLaSP highly concentrates these entries in specific areas in the map (right). These results highlight a key advantage of CLaSP. By incorporating textual information during training, CLaSP learns to recognize similarities between materials based not only on their structures but also on their properties and functionalities through text-based supervision. In contrast, CMML, which relies solely on structural data, struggles to capture these high-level relationships among materials, particularly when they exhibit diverse structures or compositions.

## E Potential biases in the dataset

This study utilized the Crystallography Open Database (COD) [10] as the source of the dataset. Although the COD is the world's largest database of experimentally determined crystal structures, it is not systematically constructed to encompass a wide variety of materials. Consequently, the dataset may contain biases. To investigate this, we analyzed the breakdown of journals that served as data sources for the COD dataset used in this study.

The resulting histogram in Fig. 7 reveals that over 80% of the 406,048 entries in the COD dataset originate from just 18 journals, the majority of which are related to crystallography and chemistry.

This dataset trend is reasonable considering the historical context of the COD, which has been maintained through the voluntary efforts of researchers interested in crystallography [10]. The presence of this bias suggests the importance of incorporating additional data sources to enhance the diversity of the COD, as well as the need to adapt ML models trained on the COD for target domains.

## F Potential negative societal impacts

A potential risk is the misuse of CLaSP for designing harmful materials. However, since does not directly synthesize substances, the risk is comparable to other computational methods in materials development, such as simulations.

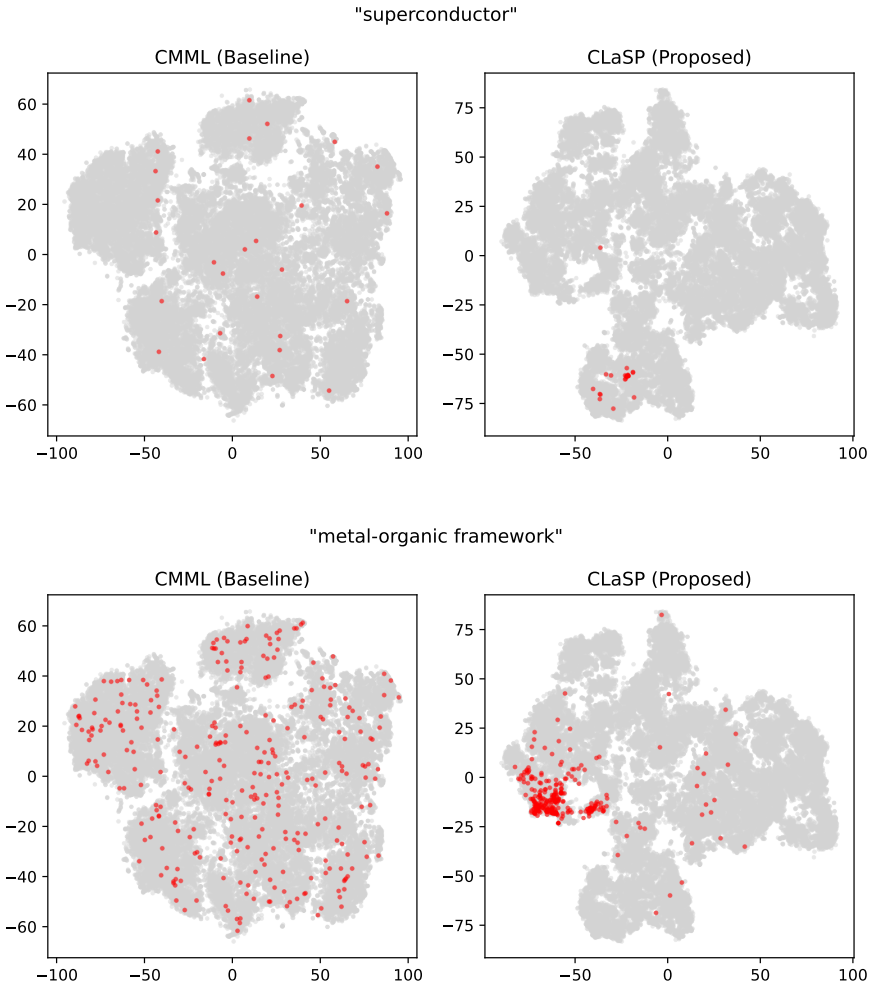


Figure 6: t-SNE visualizations of crystal structure embeddings generated by CLaSP and CMML [26]. Material entries with publication titles that include the keywords ‘superconductor’ (top row) or ‘metal-organic framework’ (bottom row) are highlighted in red.

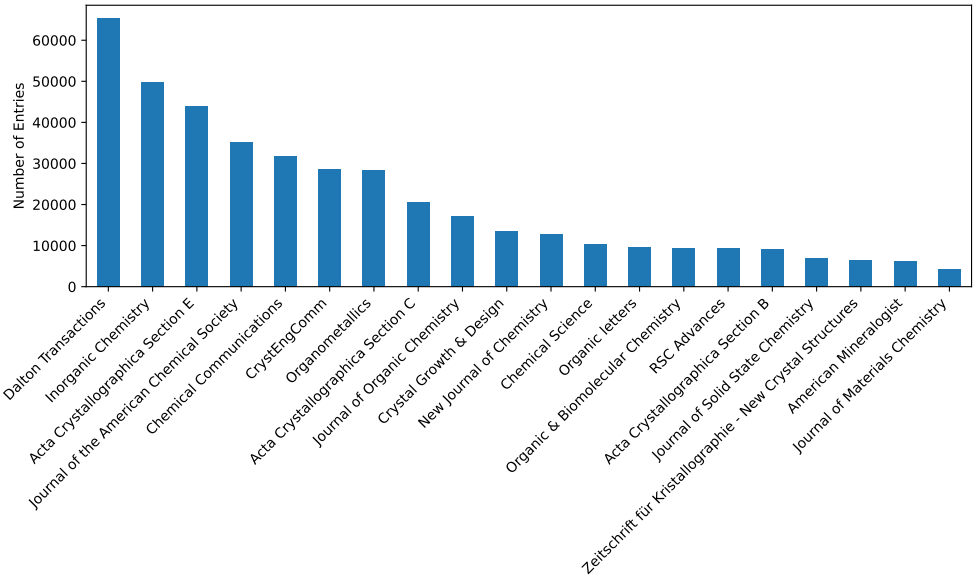


Figure 7: Top 20 journals contributing to the COD dataset.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly and accurately summarize the main contributions of the paper, including the key findings, methodology, and scope of the work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed possible limitations of this study in Section 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We did not conduct a theoretical study in this work.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide necessary information to reproduce the main claims in Sections 2 and 3, and Appendices A and C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data and code are not currently publicly available as we are preparing an extended version of this work for journal submission. However, we have provided the main code for reviewers. We plan to release both the data and code upon publication in a journal.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide necessary information in Section 3 and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: While we acknowledge that the paper does not report error bars or other measures of statistical significance due to computational cost and time constraints, we believe this does not undermine the core contribution of our work. The primary contribution lies in the problem formulation, rather than in demonstrating superior performance against existing methods through quantitative comparisons. Therefore, the absence of statistical significance analysis does not impact the main claims and conclusions. Future work will include a more thorough evaluation, incorporating statistical significance analysis.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computing resource information in Section 3 and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We reviewed the NeurIPS Code of Ethics and conducted this research following the code.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed positive impacts in Section 5 and potential social risks in Appendix F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: As discussed in Appendix F, the proposed method is a representation learning framework and does not directly synthesize substances that could potentially cause harm. The estimated risk associated with this work is considered comparable to other computational approaches for materials development, such as Ab initio simulations.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have provided appropriate citations for all the resources used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This work did not create new assets except the experimental results and code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing or human subjects were involved in this study.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects were involved in this study.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.