MATBIND: PROBING THE MULTIMODALITY OF MATERIALS SCIENCE WITH CONTRASTIVE LEARNING

Adrian Mirza
Helmholtz Institute for Polymers
in Energy Applications Jena
07743 Jena, Germany

Le Yang
Institute for Advanced Simulations (IAS-9)
Forschungszentrum Jülich GmbH
52425 Jülich, Germany

Anoop K. Chandran
Jülich Supercomputing Centre
Forschungszentrum Jülich GmbH
52425 Jülich, Germany

Jona Östreicher
Institute of Nanotechnology
Karlsruhe Institute of Technology
76131 Karlsruhe, Germany

Sebastien Bompas
Institute for Advanced Simulations (IAS-9)
Forschungszentrum Jülich GmbH
52425 Jülich, Germany

Bashir Kazimi
Institute for Advanced Simulations (IAS-9)
Forschungszentrum Jülich GmbH
52425 Jülich, Germany

Stefan Kesselheim
Jülich Supercomputing Centre
Forschungszentrum Jülich GmbH
52425 Jülich, Germany
Institute of Physics I, University of Cologne
50937 Köln
Helmholtz AI

Pascal Friederich
Institute of Nanotechnology,
Karlsruhe Institute of Technology
76131 Karlsruhe, Germany

Stefan Sandfeld
Institute for Advanced Simulations (IAS-9)
Forschungszentrum Jülich GmbH
52425 Jülich, Germany
RWTH Aachen University
52056 Aachen, Germany

Kevin Maik Jablonka
Center for Energy and Environmental Chemistry Jena
Jena Center for Soft Matter
Friedrich Schiller University Jena
Helmholtz Institute for Polymers in Energy Applications Jena
07743 Jena, Germany

ABSTRACT

Materials discovery depends critically on integrating information from multiple experimental and computational techniques, yet most tools today analyze these different data types in isolation. Here, we present MatBind, a model based on the ImageBind architecture that creates a unified embedding space across four key materials science modalities: density of states (DOS), crystal structures, text descriptions, and powder X-ray diffraction (pXRD) patterns. Using a hub-and-spoke architecture with crystal structure as the central modality, MatBind achieves cross-modal recall@1 performance of up to 97% between directly aligned modalities and up to 73% for pairs of modalities not explicitly trained together. Our model demonstrates the ability to make semantically meaningful connections across modalities, enabling researchers to query one type of materials data using another. Our analysis shows that combining multiple modalities can improve the model's ability to recognize important structural features like perovskite crystal systems. This approach lays the foundation for more integrated materials research platforms that can accelerate discovery by leveraging the collective knowledge encoded in materials databases.

1 INTRODUCTION

Modern materials discovery relies on synthesizing heterogeneous data modalities, from experimental fingerprints like powder X-ray diffraction (pXRD) and nuclear magnetic resonance spectroscopy (NMR) to simulated properties such as density of states (DOS) and band structures. Each modality acts as a distinct "lens," revealing partial facets of a material's behavior. Yet today's tools commonly force scientists to peer through these lenses one at a time. For instance, research data management systems (e.g., electronic lab notebooks (Jablonka et al., 2022; Scheidgen et al., 2023)) typically lack cross-modal semantic search capabilities, restricting queries to isolated metadata fields for one modality. Thus, it is commonly not possible to semantically search across modalities to, for instance, find a crystal structure that best matches a pXRD pattern or DOS. Such a feature, however, would lead to large increases in research productivity. Also, machine learning models typically focus only on one modality. For instance, by predicting properties based on pXRD patterns as the input (Khan & Moosavi, 2024; Jablonka et al., 2020; Lee et al., 2022). However, gains are to be expected by leveraging the latent synergies between different modalities.

Here, we report MatBind, a model based on the ImageBind architecture (Girdhar et al., 2023) that trains a shared embedding space across four modalities (DOS, crystal structures, text, and pXRD) using a hub-and-spoke architecture. The architecture scales modularly, allowing seamless integration of new modalities such as emerging spectroscopic techniques. MatBind achieves cross-modal recall@1 performance of up to 97 % and up to 73 % for pairs of modalities it has not seen in training. These results have very practical implications: they enable researchers to build systems that can semantically query for data across different modalities and, in this way, open the possibility of discovering previously unknown links and accelerating materials design and discovery.

2 RELATED WORK

Recent advances in multimodal learning have demonstrated the value of unified representations. Girdhar et al. (2023) established that contrastive alignment with a central modality can bind additional data types into a shared embedding space, inspiring domain-specific adaptations and extensions, for instance, by coupling the embeddings to large language models (LLM) (Su et al., 2023; Han et al., 2023). In materials science, Moro et al. (2023) reported contrastively trained multimodal models but focused on property prediction rather than cross-modal retrieval. Das et al. (2023) fused crystal graphs with textual descriptors to improve property prediction through global structural awareness, though their reliance on curated text limits scalability compared to unsupervised alignment methods. Based on a modified CLIP architecture (Radford et al., 2021), Seidl et al. (2023) demonstrated how textual task descriptions enhance few-shot drug discovery.

For molecular systems, Mirza et al. (2024) aligned chemical representations via contrastive learning, while Mirza & Jablonka (2024) combined multimodal embeddings with optimization for spectroscopy-structure mapping. Protein science has seen analogous progress through frameworks like (Flöge et al., 2024), which unified structural and sequence data, or Xiao et al. (2024) who unified protein data with molecules and language.

Our work adapts the contrastive paradigm by Girdhar et al. (2023) to materials-specific modalities while in contrast to Moro et al. (2023) focusing on retrieval.

3 METHODS

3.1 DATASET

We obtained our dataset from the Materials Project (Jain et al., 2013), which contains approximately 169,000 materials. We retrieve the crystal structure and DOS from the Materials Project. The DOS data is simplified to the total density of states. To generate textual descriptions of crystal structures, we use Robocrystallographer (Ganose & Jain, 2019). We generated pXRD patterns using the pipeline described in Schopmans et al. (2023).

3.2 MODEL ARCHITECTURE

MatBind is based on the ImageBind architecture. Figure 1 illustrates that we use encoders for different modalities and align them contrastively by training on pairs of modalities (black lines). Other links between modalities are emergent (red lines).

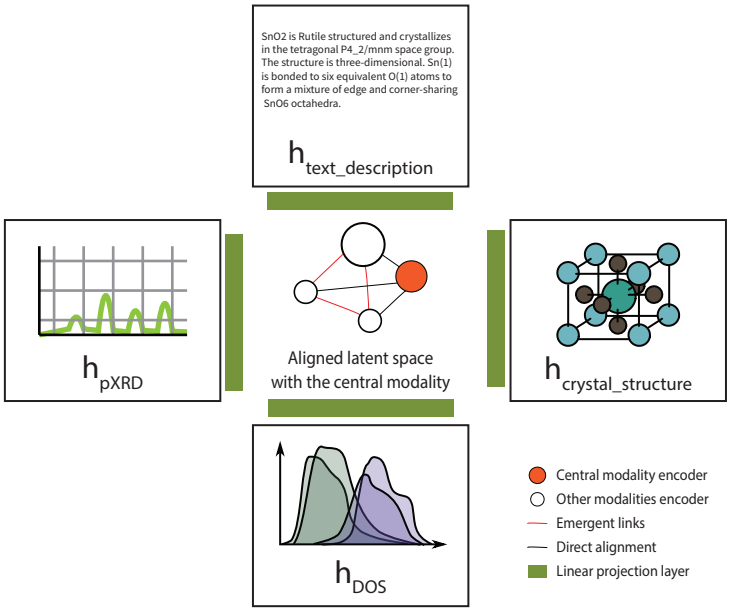

Figure 1: **Overview of the MatBind architecture.** $h_{<\text{modality}>}$ represents the embeddings generated by the encoder. All modalities are aligned with the central modality. We contrastively train on pairs of modalities (black lines), where one modality is always the central modality, $\mathcal{C}$. The size of the circles represents the parameter count of the different encoders but is not to scale (see Table 1 for detailed parameter counts). Embeddings obtained from each encoder are passed through a projection layer (green bars) to obtain vectors of fixed size.

3.2.1 ENCODER MODELS

Crystal Structure Encoder We encode crystal structures using a graph convolutional neural network (Kipf & Welling, 2016) with six sequential graph convolution layers. A site-specific atomic species encoding scheme (Taniai et al., 2024) is used for graph node representation where binary vectors for single-species sites and weighted vectors for mixed-occupancy sites, while edges connect nearest neighbors within an 8 Å cutoff radius. Edge attributes encode interatomic distances using a Gaussian radial basis expansion with 41 components. After message passing, node embeddings are aggregated via mean pooling across the crystal graph, producing a global material representation.

Powder X-Ray Diffractogram Encoder The pXRD encoder is a convolutional, ResNet-based architecture (He et al., 2015) that was successfully used before for the prediction of space groups associated with the diffractograms (Schopmans et al., 2023). It can be pre-trained using the space group prediction task using synthetic (i.e., simulated, plus synthetic noise) pXRD diffractogram derived from crystal structures in materials databases such as the Materials Project or the ICSD database (Hellenbrandt, 2004), but as shown in Schopmans et al. (2023), it can also be trained on fully synthetic crystal structures and simulated pXRD patterns.

Density of States (DOS) Encoder To efficiently encode the density of states into a latent space, we employ a Transformer model (Vaswani, 2017). The DOS data can be viewed as two-dimensional data, where one dimension represents energy and the other the density of states. However, the measured energy range varies across different materials. To address this, we incorporate not only

positional embeddings but also explicit energy values as additional input features. Our implementation follows the approach described by Moro et al. (2023). Additionally, we offset the energy values by subtracting the Fermi energy of the material, enabling the model to capture energy-related information better.

Text encoder Textual descriptors are tokenized using BERT Tokenizer and then encoded using MatBERT (Walker et al., 2021; Trewartha et al., 2022), a BERT-based model (Devlin et al., 2019) pretrained on scientific papers. We extract embeddings from the final hidden layer of the model, computing the material's representation as the attention-weighted average of all non-padding token vectors.

3.2.2 JOINT TRAINING

MatBind is based on the ImageBind framework (Girdhar et al., 2023) and aligns pairs of different modalities, with crystal structure ($\mathcal{C}$) as the anchor modality paired with either text, DOS, or pXRD ($\mathcal{E}$). For each pair, the respective encoders $\phi_{\mathcal{C}}$ and $\phi_{\mathcal{E}}$ transform the raw inputs into representations $\mathbf{a}_i = \phi_{\mathcal{C}}(a_i)$ and $\mathbf{b}_i = \phi_{\mathcal{E}}(b_i)$.

For a batch of pairs, we optimize two objectives: bringing matching representations closer together while pushing non-matching ones apart. The InfoNCE loss function (van den Oord et al., 2018) captures this dual aim:

$$L = \frac{1}{n} \sum_i \log \frac{\exp(\mathbf{a}_i^\top \mathbf{b}_i / \tau)}{\exp(\mathbf{a}_i^\top \mathbf{b}_i / \tau) + \sum_{j \neq i} \exp(\mathbf{a}_i^\top \mathbf{b}_j / \tau)}. \tag{1}$$

The temperature parameter τ controls the scaling of similarity scores, while $\mathbf{b}_i$ terms act as contrasting examples within each batch. During training, we maintain crystal structures ($\mathcal{C}$) as a constant presence across batches of paired modalities, ensuring consistent alignment across all modalities.

4 RESULTS AND DISCUSSION

4.1 RETRIEVAL METRICS

In Figure 2, we show the recall@1 for all combinations of models, where the central modality is the crystal structure. We find consistently very high retrieval performance for all pairs ($\mathcal{E}$, $\mathcal{C}$) except for pairs of crystal structures and pXRD patterns. In addition, we observe both successful and unsuccessful emergent links. For instance, we find very strong retrieval between DOS and text descriptions.

The addition of additional encoders sometimes improves retrieval performance for other modalities. This is particularly the case with the addition of the text encoder. For example, The DOS-crystal structure, DOS-pXRD, and crystal structure-pXRD retrieval are improved by adding a text encoder. But this is not always the case: for example, the text-DOS retrieval (0.643) in the four-modality model (crystal structure - DOS - text - pXRD) underperforms the three-modality model (0.731, crystal structure - DOS - text).

4.2 LATENT SPACE ANALYSIS

It is interesting to analyze how the addition of various encoders changes the structure of latent spaces. An embedding of a material is particularly useful if it can be used to distinguish materials of different classes and properties (Zhang et al., 2024; Isayev et al., 2015). Historically, scientists have hand-crafted heuristics such as tolerance factors to perform such analyses. One important tolerance factor is the Goldschmidt tolerance factor (Bartel et al., 2019; Kronmüller & Parkin, 2007) that can be used to compute the compatibility of ABX_3 with different structure types:

$$t = \frac{r_A + r_X}{\sqrt{2}\,(r_B + r_X)}, \tag{2}$$

where $r_{A,B,X}$ are the radii of the different ions.

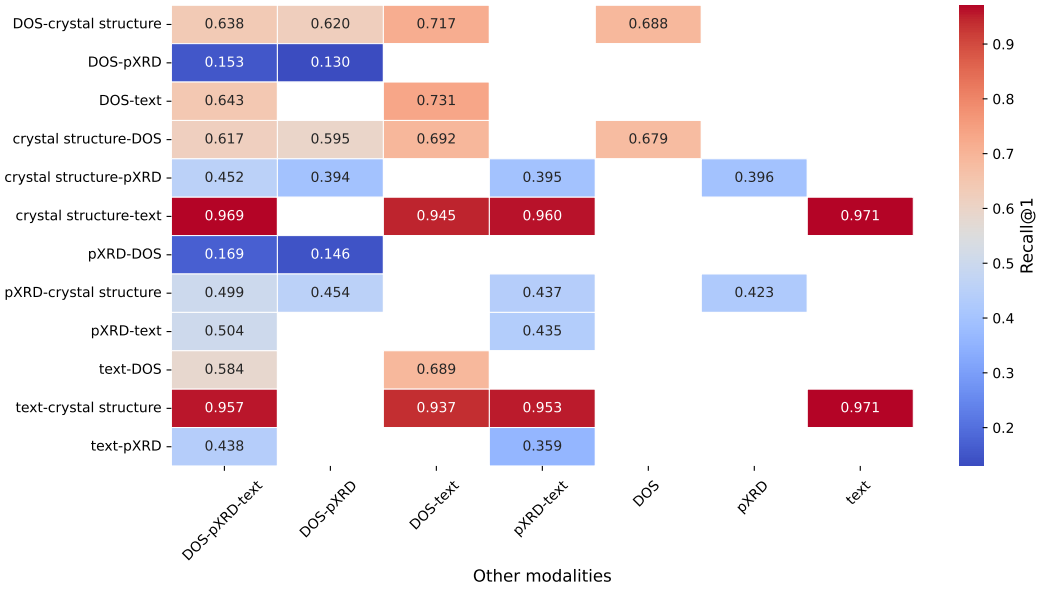

Figure 2: **Recall@1 performance heatmap for all possible modality combinations, where the central modality is the crystal structure.** The y-axis represents the query modality - key modality pairs (e.g., DOS-pXRD represents a pair of modalities where DOS embeddings are used as queries and pXRD embeddings are used as keys. The x-axis labels the modalities used in training besides the central modality. The white cells are impossible pairs for the respective models (because they are not included in the model).

In this case, the following condition should be respected for a material to be a perovskite:

$$\begin{cases} 0.9 \leq t \leq 1.09, & \text{if the formula contains S (sulfides)} \\ 1.01 \leq t \leq 1.05, & \text{if the formula contains Se (selenides).} \end{cases} \tag{3}$$

Based on the implementation of Schilling-Wilhelmi et al. (2025), we compute the tolerance factors and embed the crystal structures using the crystal encoder aligned with different combinations of other modalities. We find that all encoders show some ability to discriminate between perovskite and non-perovskite structures (a Kolgomorov-Smirnov test showed a significant difference for all pairs of distributions displayed in Figure 3), which indicates that our embeddings capture relevant information about the structure of materials. However, the discrimination is typically best when pXRD is included in the alignment training (see Figure 7). This observation highlights that different modalities can act as different "lenses" on materials and showcases the importance of combining different modalities.

5 Conclusion

Modern materials discovery relies on synthesizing information from multiple experimental and computational techniques. Each technique—from X-ray diffraction to density of states calculations—acts as a distinct "lens", revealing different aspects of a material's behavior. For simple systems, scientists can often mentally integrate these different views to form a complete understanding. However, as materials become more complex and the volume of data grows, this manual synthesis becomes increasingly challenging. Yet, the materials science community has generated vast amounts of multimodal data, implicitly encoding relationships between different measurement techniques and material properties.

Here, we have shown that this implicit knowledge can be leveraged through a contrastive learning framework that unifies different modalities in a shared embedding space. Our MatBind architecture achieves cross-modal recall performance of up to 97% ($k = 1$) and, remarkably, discovers emergent

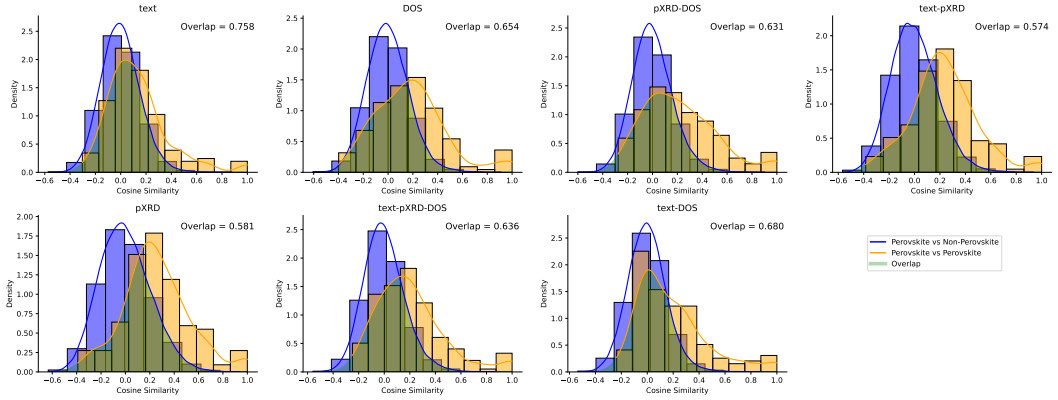

Figure 3: **Overlap between the distributions of cosine similarities within the set of perovskites vs. non-perovskite materials.** Embeddings of the validation set are used for these plots. The titles of the subplots are consistent with the x-axes in Figure 2. We show the value of the area of the overlap region between the two kernel density estimation (KDE) curves.

relationships between modalities that were never explicitly paired during training. The success of this approach, even with encoders trained from scratch, demonstrates the power of data-driven techniques to bridge different experimental and computational methods in materials science.

Our work highlights how modern machine learning techniques can solve fundamental challenges in materials informatics for which no reliable alternatives exist. By enabling semantic queries across different modalities, MatBind lays the foundation for more integrated materials research platforms that can accelerate discovery by leveraging the collective knowledge encoded in materials databases.

ACKNOWLEDGMENTS

This project is funded by the Helmholtz Foundation Model Initiative supported by the Helmholtz Association, and the HAICORE@JSC grant. Additional computational resources were provided by the AI service center WestAI.

REFERENCES

Christopher J Bartel, Christopher Sutton, Bryan R Goldsmith, Runhai Ouyang, Charles B Musgrave, Luca M Ghiringhelli, and Matthias Scheffler. New tolerance factor to predict the stability of perovskite oxides and halides. *Science advances*, 5(2):eaav0693, 2019.

Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. CrysMMNet: Multimodal representation for crystal property prediction. In Robin J. Evans and Ilya Shpitser (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 507–517. PMLR, 31 Jul-04 Aug 2023. URL `https://proceedings.mlr.press/v216/das23a.html`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*, 2019. doi: 10.18653/v1/N19-1423.

Klemens Flöge, Srisruthi Udayakumar, Johanna Sommer, Marie Piraud, Stefan Kesselheim, Vincent Fortuin, Stephan Günneman, Karel J van der Weg, Holger Gohlke, Alina Bazarova, and Erinc Merdivan. Oneprot: Towards multi-modal protein foundation models. *arXiv preprint arXiv: 2411.04863*, 2024.

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

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. *arXiv preprint arXiv: 2305.05665*, 2023.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv: 2309.03905*, 2023.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Computer Vision and Pattern Recognition*, 2015. doi: 10.1109/cvpr.2016.90.

Mariette Hellenbrandt. The inorganic crystal structure database (icsd)—present and future. *Crystallography Reviews*, 10(1):17–22, January 2004. ISSN 1476-3508. doi: 10.1080/08893110410001664882. URL http://dx.doi.org/10.1080/08893110410001664882.

Olexandr Isayev, Denis Fourches, Eugene N. Muratov, Corey Oses, Kevin Rasch, Alexander Tropsha, and Stefano Curtarolo. Materials cartography: Representing and mining materials space using structural and electronic fingerprints. *Chemistry of Materials*, 27(3):735–743, January 2015. ISSN 1520-5002. doi: 10.1021/cm503507h. URL http://dx.doi.org/10.1021/cm503507h.

Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. Big-data science in porous materials: Materials genomics and machine learning. *Chemical Reviews*, 120(16):8066–8129, June 2020. ISSN 1520-6890. doi: 10.1021/acs.chemrev.0c00004. URL http://dx.doi.org/10.1021/acs.chemrev.0c00004.

Kevin Maik Jablonka, Luc Patiny, and Berend Smit. Making the collective knowledge of chemistry open and machine actionable. *Nature Chemistry*, 14(4):365–376, April 2022. ISSN 1755-4349. doi: 10.1038/s41557-022-00910-7. URL http://dx.doi.org/10.1038/s41557-022-00910-7.

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.

Sartaaj Takrim Khan and Seyed Mohamad Moosavi. Connecting metal-organic framework synthesis to applications with a self-supervised multimodal model. October 2024. doi: 10.26434/chemrxiv-2024-mq9b4. URL http://dx.doi.org/10.26434/chemrxiv-2024-mq9b4.

Thomas Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations*, 2016.

Helmut Kronmüller and Stuart S. P. Parkin. *Handbook of magnetism and advanced magnetic materials*. J. Wiley & Sons, Chichester, 2007. ISBN 9780470022177.

Byung Do Lee, Jin-Woong Lee, Woon Bae Park, Joonseo Park, Min-Young Cho, Satendra Pal Singh, Myoungho Pyo, and Kee-Sun Sohn. Powder x-ray diffraction pattern is all you need for machine-learning-based symmetry identification and property prediction. *Advanced Intelligent Systems*, 4 (7), May 2022. ISSN 2640-4567. doi: 10.1002/aisy.202200042. URL http://dx.doi.org/10.1002/aisy.202200042.

Adrian Mirza and Kevin Maik Jablonka. Elucidating structures from spectra using multimodal embeddings and discrete optimization. *ChemRxiv*, 2024. doi: 10.26434/chemrxiv-2024-f3b18-v2. This content is a preprint and has not been peer-reviewed.

Adrian Mirza, Sebastian Starke, Erinc Merdivan, and Kevin Maik Jablonka. Bridging chemical modalities by aligning embeddings. In *AI for Accelerated Materials Design-Vienna 2024*, 2024.

Viggo Moro, Charlotte Loh, Rumen Dangovski, Ali Ghorashi, Andrew Ma, Zhuo Chen, Peter Y Lu, Thomas Christensen, and Marin Soljačić. Multimodal learning for crystalline materials. *arXiv preprint arXiv:2312.00111*, 2023.

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. *arXiv preprint arXiv: 2103.00020*, 2021.

Markus Scheidgen, Lauri Himanen, Alvin Noe Ladines, David Sikter, Mohammad Nakhaee, Ádám Fekete, Theodore Chang, Amir Golparvar, José A. Márquez, Sandor Brockhauser, Sebastian Brückner, Luca M. Ghiringhelli, Felix Dietrich, Daniel Lehmberg, Thea Denell, Andrea Albino, Hampus Näsström, Sherjeel Shabih, Florian Dobener, Markus Kühbach, Rubel Mozumder, Joseph F. Rudzinski, Nathan Daelman, José M. Pizarro, Martin Kuban, Cuauhtemoc Salazar, Pavel Ondračka, Hans-Joachim Bungartz, and Claudia Draxl. Nomad: A distributed web-based platform for managing materials science research data. *Journal of Open Source Software*, 8 (90):5388, 2023. doi: 10.21105/joss.05388. URL https://doi.org/10.21105/joss.05388.

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 chemical data extraction. *Chemical Society Reviews*, 2025. ISSN 1460-4744. doi: 10.1039/d4cs00913d. URL http://dx.doi.org/10.1039/D4CS00913D.

Henrik Schopmans, Patrick Reiser, and Pascal Friederich. Neural networks trained on synthetically generated crystals can extract structural information from icsd powder x-ray diffractograms. *Digital Discovery*, 2(5):1414–1424, 2023. ISSN 2635-098X. doi: 10.1039/d3dd00071k. URL http://dx.doi.org/10.1039/D3DD00071K.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30458–30490. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/seidl23a.html.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv: 2305.16355*, 2023.

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. URL https://openreview.net/forum?id=fxQiecl9HB.

Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 3(4):100488, April 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2022.100488. URL http://dx.doi.org/10.1016/j.patter.2022.100488.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

Nicholas Walker, Amalie Trewartha, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin Persson, Gerbrand Ceder, and Anubhav Jain. The impact of domain-specific pre-training on named entity recognition tasks in materials science. *Available at SSRN 3950755*, 2021.

Teng Xiao, Chao Cui, Huaisheng Zhu, and Vasant G Honavar. Molbind: Multimodal alignment of language, molecules, and proteins. *arXiv preprint arXiv:2403.08167*, 2024.

Xiaoqi Zhang, Kevin Maik Jablonka, and Berend Smit. Deep learning-based recommendation system for metal–organic frameworks (mofs). *Digital Discovery*, 3(7):1410–1420, 2024. ISSN 2635-098X. doi: 10.1039/d4dd00116h. URL http://dx.doi.org/10.1039/D4DD00116H.

A Appendix

A.1 Other retrieval metrics

In the figure below, we also provide the Recall@5 metric for all models shown in Figure 2.

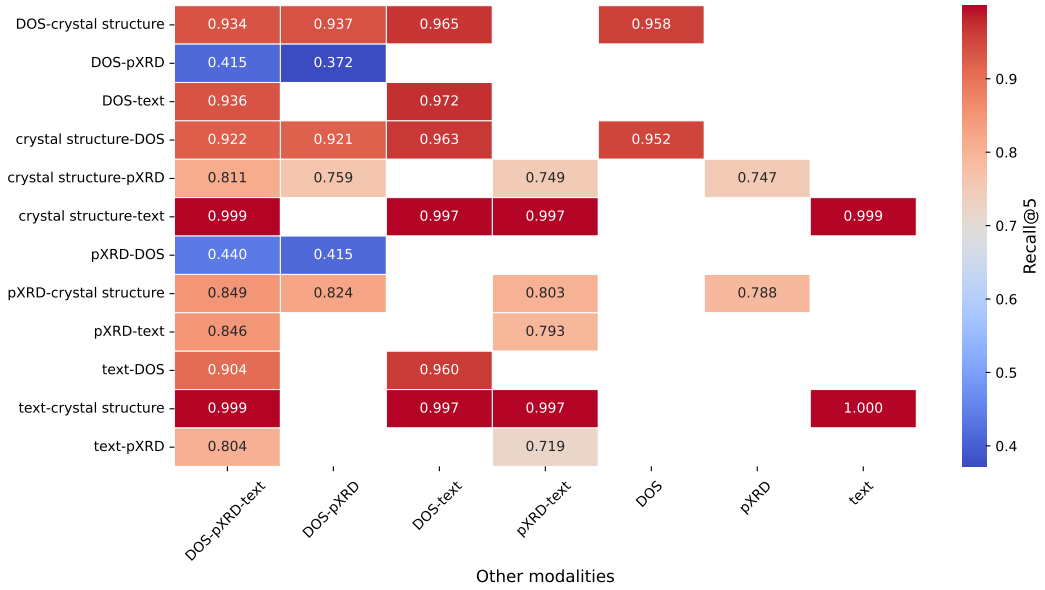

Figure 4: **Recall@5 for all combinations of modalities, where crystal structure acts as the central modality.** This figure is analogous to Figure 2.

A.2 Encoder details

In Table 1 we show the number of parameters for all the encoders used in this work.

Table 1: **Parameter count of the encoders used for the models described in Figure 1.** For the experiments shown in this work, most encoders, except the text encoder, were trained from scratch during the contrastive training.

Encoder	Parameter count
Density of states	6.4 M
Text description	110 M
Powder X-Ray Diffractogram	5 M
Crystal structure	1 M

A.3 Additional latent space visualizations

Here, we provide additional latent space visualization in lower dimensions. Figure 5 shows a two-dimensional TSNE plot. The default settings from `sklearn` are used. Figure 6 shows the distribution of theoretical perovskites in a two-dimensional principal component space.

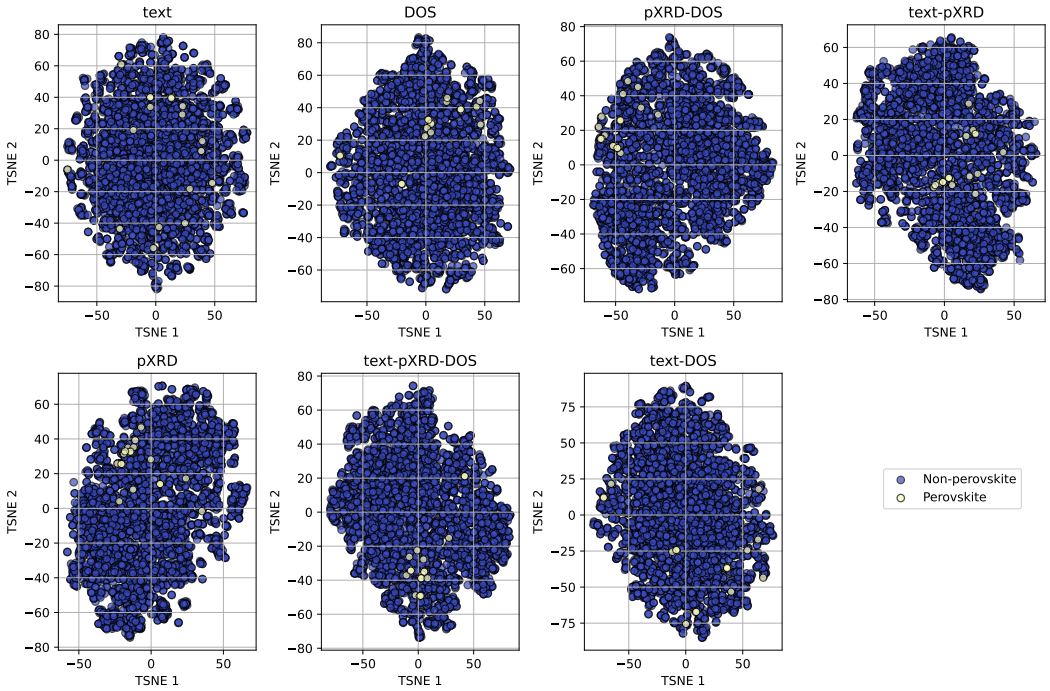

Figure 5: **TSNE component mapping for the crystal structure embeddings of every possible encoder combination.**

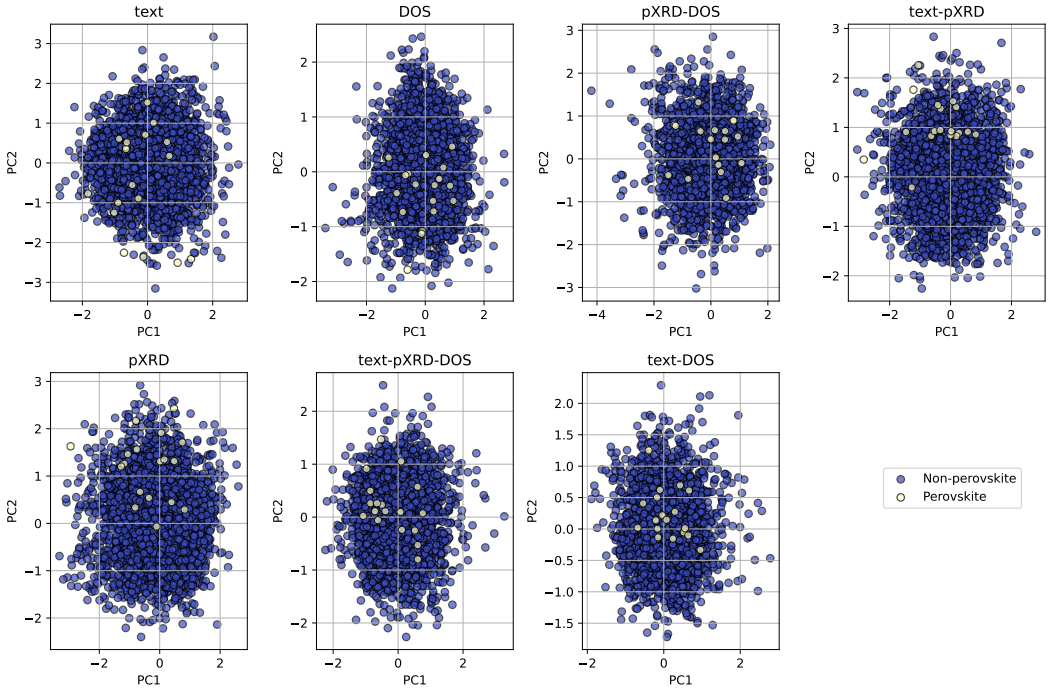

Figure 6: **PCA component mapping for the crystal structure embeddings of every possible encoder combination.**

A.4 pXRD VS. NON-pXRD MODELS

In the figure below, we provide the distribution of all embeddings across models that contain the pXRD encoder vs models that do not contain the pXRD encoder. The overlap area is significantly higher (circa 14%) for models not containing the pXRD encoder.

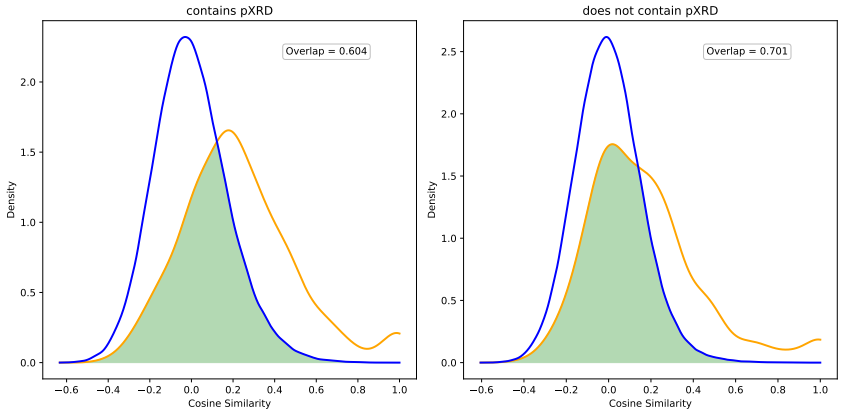

Figure 7: Overlap between the KDEs for perovskite vs. non-perovskite materials for models containing pXRD vs. models not containing pXRD.