Cross Modal Predictive architecture for Material Property prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work, we propose CrossModal Predictive Architecture(**X-MoPA**), a multimodal learning model that combines crystal structure graphs, X ray diffraction (XRD) patterns, and text based structural descriptions to improve materials property prediction. Unlike prior multimodal approaches that rely on heavy attention mechanisms or simple concatenation, X-MoPA leverages lightweight predictors to learn a joint latent space through cross-modal prediction. For each training instance, we select two modalities and predict the third one in latent space. This formulation captures complementary information across modalities while avoiding reconstruction inefficiencies and contrastive memory bottlenecks. We train and evaluate the model on Matbench for several key properties, Band Gap, Shear Modulus, Bulk Modulus and formation energy for Perovskites. X-MoPA consistently outperforms state of the art(SOTA) models, with error reductions ranging from 16% to 60% across four key properties, while matching the best baseline on Shear Modulus.Beyond Matbench, X-MoPA achieves SOTA performance on AFLOW band gap prediction, showing that the learned cross-modal representations transfer well across datasets with different sampling strategies and property distributions.

1 Introduction

The dominant strategy in multimodal learning has been to use pretrained language and vision models and then align them during the training process. This *pretrain-then-transfer* relies on modality specific encoders trained on large datasets, producing strong but often narrow representations. Such representations may struggle to generalize in domains like materials science where modalities such as text, crystal graphs, and diffraction patterns encode complementary but structurally distinct information.

Self supervised learning(SSL) provides an alternative by constructing surrogate tasks that exploit data structure without requiring manual labels. Contrastive methods, such as InfoGraph, have demonstrated the ability to learn rich representations by maximizing agreement across augmented views[13]. However, these approaches often demand large memory banks and rely on data perturbations, which are not suited to scientific modalities like crystal graphs where small structural changes can drastically alter meaning[7]. Generative methods, in turn, require reconstruction in input space, which can lead to inefficiencies and overfitting to low-level details. To overcome these challenges, LeCun et al. introduced the Joint Embedding Predictive Architecture (JEPA), which emphasizes prediction in latent space rather than raw reconstruction[5]. This perspective allows models to focus on schematic information while discarding extraneous detail. We draw inspiration from this principle to design a multimodal model that predicts one modality from the others in latent space, aiming to capture complementary structure across text, graphs, and diffraction data.

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

Recent work in multimodal learning combines text and graph embeddings via concatenation to predict material properties[2, 8]. However, this simple fusion does not effectively utilize complimentary information from different modalities and lacks interpretability. This challenge is addressed in models like UniMat[10] and CAST[6] which use cross attention mechanisms. However, these models are computationally expensive making them difficult to scale or deploy.

In this work, we propose a Cross Modal Predictive Architecture (X-MoPA), a multimodal predictive architecture that integrates three complimentary modalities, crystal graph embeddings (CGCNN)[15], contextual text embeddings (MatSciBERT)[4], and spectral features from XRD patterns using lightweight MLPs instead of expensive transformer based predictors. We use two modalities to predict the third one in latent space. Material modalities have high redundancy because they describe the same underlying system. This is exploited in this framework to ensure that the trained model is able to learn this system which can be used for downstream tasks. We demonstrate state of the art performance on multiple properties in the Material Project dataset. Moreover, we provide information theoretic bounds in the Appendix.

2 Proposed Model Architecture

Crystallographic structures from CIF files are represented as graphs using the CGCNN framework. Atoms are nodes with features including group, periodic table position, electronegativity, first ionization energy, covalent radius, valence electron count, electron affinity, and atomic number. Bonds form edges, and atom features are updated through localized message passing. Textual descriptions are generated with RoboCrystallographer and encoded with MatSciBERT, a 12-layer transformer with 12 attention heads per layer and hidden size 768, pretrained on 3.17B words from materials science literature. Text input is processed by WordPiece tokenization, positional and segment encoding, multi-head attention, residual connections, layer normalization, and feedforward layers with ReLU. The [CLS] token embedding is projected to 150 dimensions. XRD spectra are encoded with a 1D CNN consisting of a convolutional layer, max-pooling, and two fully connected layers to obtain spectral embeddings. Each modality (text, graph, XRD) is encoded separately and projected into a shared joint embedding space. Cross-modal predictors map pairs of modalities to reconstruct the third. Prediction error is measured with L2 loss, summed over all three modalities. Lightweight MLPs are used for prediction. A variance-invariance-covariance regularization term is added to stabilize the latent space and prevent overfitting. The total loss is the sum of cross-modal prediction loss and regularization. The figure1 shows a schematic of the model. The model architecture has been described in detail in the Appendix section.

3 Proposed Methodology

We propose a self-supervised learning framework inspired by the Joint-Embedding Predictive Architecture (JEPA) paradigm [1], designed to learn semantically meaningful and modality-aligned representations from multimodal crystallographic data. The framework incorporates the graph encoder, text encoder and the XRD encoder, the three complementary encoders discussed above. The objective of the proposed framework is to learn **Cross Modality Prediction**. For each training instance, we randomly select two modalities and predict the third one in latent space. Let m_1, m_2 be the input modalities and m_3 the target modality. Then the predicted latent representation is obtained using a MLP based lightweight joint projection network. The advantages of prediction in the latent space are also clearly explained in the Appendix.

$$\hat{z}_{m_3} = h([z_{m_1}, z_{m_2}]) \tag{1}$$

Mean squared error (MSE) between the predicted and actual latent representation is used as the prediction loss

$$L_{pred} = ||\hat{z_{m_3}} - z_{m_3}||_2^2 \tag{2}$$

Further, to prevent **representational collapse** which result in trivial or degenerate outputs, we employ a variance-invariance-covariance (VIC) regularizer[1]. This regularizer operates on the latent features.

To encourage semantically aligned and disentangled representations across modalities, we use a contrastive loss in the latent space[11].

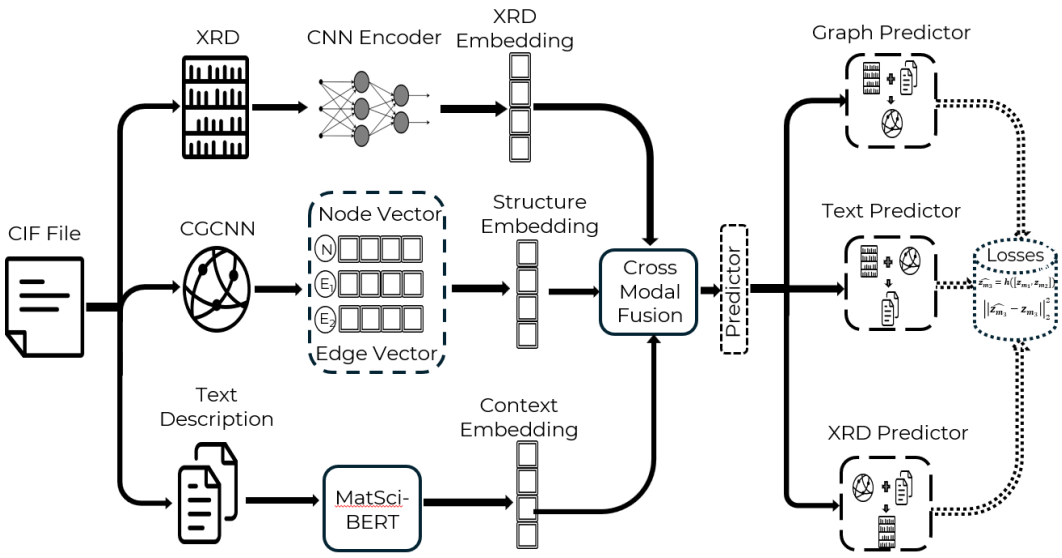

Figure 1: An overview of the proposed X-MoPA. Each modality is encoded separately and projected into a shared joint embedding space. Cross-modal predictors map pairs of modalities to reconstruct the third. Prediction error is measured with L2 loss, summed over all three modalities.

The full loss function combines all the above losses. The model is trained using AdamW optimizer with weight decay[9]. A cosine annealing scheduler is used for stable convergence.

$$L_{total} = L_{pred} + \alpha L_{VIC} + \beta L_{contrast} \tag{3}$$

4 Experimentation

We use a Nvidia RTX 4090 graphics processing unit (GPU) to run our experiments. The framework is implemented using the Pytorch library version[12]. We have used the Matbench dataset consists of 13 benchmark tasks for evaluating predictive models in materials science, covering regression and classification with predefined 5 fold train test splits. In this work we have focused on 5 crystal property prediction tasks with CIF structures as input. We generate text descriptions and simulate XRD patterns using the CIF. The process is explained in detail in the Appendix.

5 Results

In this section, we evaluate how the knowledge from the pretraining using cross modal prediction in latent space compares to other SOTA material property predictors. We are using fine tuning of only the final prediction heads while keeping the weights of the model frozen. We choose three state of the art models with different architectures, CoGCNN, CGCNN and ALIGNN. The results of these models are taken from the official Matbench leaderboard. In table1 report mean absolute error(MAE) of the predicted and actual value of a particular property to compare the performance of different algorithms. For each property, we have used the same train and test splits as given in MatBench. We observe that the proposed algorithm outperforms the other algorithms across all the properties. In addition to Matbench, we evaluate X-MoPA on the prediction of Band Gap in the AFLOW database[3], which provides a large-scale set of DFT-computed material properties. This allows us to test whether the learned cross-modal representations generalize across datasets with different sampling and property distributions. The results are presented in the Appendix.

Unlike cross-attention multimodal models such as LXMERT[14] with around 305 million trainable parameters or CAST[6] with 200 to 300 million, X-MoPA has a total of 111 million parameters. Of these, only 0.6% are updated during finetuning on downstream tasks. The MatSciBERT encoder, which makes up most of the parameters, is kept frozen during training. This design keeps the

Table 1: Benchmarking model performance. The lower the error the better the model performance.

	Mean Absolute Error(MAE)			
	CGCNN	CoGN	ALIGNN	X-MoPA
Dielectric	0.5988	0.3088	0.3449	**0.1238**
Shear Modulus	0.0895	**0.0689**	0.0715	0.069
Bulk Modulus	0.0712	0.0535	0.0568	**0.0385**
Bandgap(MBJ)	0.2972	0.1559	0.1861	**0.0661**
Perovskites	0.0452	0.0269	0.0288	**0.0225**

model lightweight and efficient, reducing the computational cost of both training and transfer to new properties.

6 Discussion

Crystal structures, despite being described in high-dimensional spaces of atomic coordinates or diffraction patterns, inherently lie on structured low-dimensional manifolds due to periodicity, space-group symmetries, and local coordination environments. Encoders map these structures into latent manifolds where nearby points correspond to structurally and chemically similar crystals.

In this setting, property prediction becomes a mapping over a smooth manifold rather than raw high-dimensional noise. The key mathematical requirement is Lipschitz continuity: small perturbations in latent space such as slight bond length variations or symmetry preserving lattice distortion must lead to proportionally small changes in predicted properties. This ensures stable optimization, prevents variance amplification, and improves generalization across material classes. By grounding property prediction in latent manifolds shaped by crystallography and enforcing Lipschitz continuity X-MoPA achieves stable and physically consistent learning.

7 Conclusion

In this work, we introduced X-MoPA, a multimodal framework that predicts material properties by operating in a shared latent space. The model learns from three complementary inputs: crystal graphs for local bonding, XRD spectra for global structure, and text for literature-driven context. Instead of reconstructing raw inputs or relying on contrastive memory banks, X-MoPA uses lightweight MLP predictors to take any two modalities and predict the third in latent space. This design makes the model more efficient while still forcing it to capture the connections between different scientific representations.

Our experiments on the Matbench benchmarks show that this approach not only improves accuracy over existing methods like CGCNN and ALIGNN, but also does so with lower computational cost. For example, we see significant reductions in MAE for properties such as band gap and bulk modulus. Because the model is trained to align modalities in latent space, the learned representations are more stable and interpretable. Overall, X-MoPA shows that lightweight cross-modal prediction in latent space is a scalable way to predict properties in materials science.

Limitations and Future scope of work: Missing or corrupted data in one modality might not align with structural information in another modality. Moreover, prediction in the latent space might bias the model to focus on short range correlation than learning long-range dependencies. Incorporating additional modalities, such as material characterization data, could provide complementary information.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[2] Kishalay Das, Pawan Goyal, Seung-Cheol Lee, Satadeep Bhattacharjee, and Niloy Ganguly. Crysmmnet: multimodal representation for crystal property prediction. In *Uncertainty in Artificial Intelligence*, pages 507–517. PMLR, 2023.

[3] Eric Gossett, Cormac Toher, Corey Oses, Olexandr Isayev, Fleur Legrain, Frisco Rose, Eva Zurek, Jesús Carrete, Natalio Mingo, Alexander Tropsha, et al. Aflow-ml: A restful api for machine-learning predictions of materials properties. *Computational Materials Science*, 152:134–145, 2018.

[4] Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, 2022.

[5] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022.

[6] Jaewan Lee, Changyoung Park, Hongjun Yang, Sungbin Lim, and Sehui Han. Cast: Cross attention based multimodal fusion of structure and text for materials property prediction. *arXiv preprint arXiv:2502.06836*, 2025.

[7] Namkyeong Lee, Junseok Lee, and Chanyoung Park. Augmentation-free self-supervised learning on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7372–7380, 2022.

[8] Youjia Li, Vishu Gupta, Muhammed Nur Talha Kilic, Kamal Choudhary, Daniel Wines, Weikeng Liao, Alok Choudhary, and Ankit Agrawal. Hybrid-llm-gnn: integrating large language models and graph neural networks for enhanced materials property prediction. *Digital Discovery*, 2025.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[10] Janghoon Ock, Joseph Montoya, Daniel Schweigert, Linda Hung, Santosh K Suram, and Weike Ye. Unimat: Unifying materials embeddings through multi-modal learning. *arXiv preprint arXiv:2411.08664*, 2024.

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

[12] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *Neural Information Processing Systems*, 2017.

[13] Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. *arXiv preprint arXiv:1908.01000*, 2019.

[14] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[15] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

Appendix: Cross Modal Predictive architecture for Material Property prediction

Anonymous Author(s)
Affiliation
Address
email

A Proposed Model Architecture

Let us describe the architecture of our multimodal framework. Given a dataset of inorganic crystals denoted by $D = [(S, T, X), P]$ where S, T, X and P denote the structure information in CIF format, the text description, the XRD spectra and the material property respectively. The model trains the parameters of the XRD Encoder (X_θ), Graph encoder (G_θ) and the BERT encoder (B_θ) to learn the function $f_\theta \rightarrow P$.

In our multimodal framework, the crystallographic structure, provided in Crystallographic Information File (CIF) format, is represented as a graph $G(V, E)$ using the Crystal Graph Convolutional Neural Network (CGCNN) architecture, where atoms correspond to nodes V and interatomic bonds are represented as edges E. Node attributes encode physicochemical properties of each atom, including group, periodic table position, electronegativity, first ionization energy, covalent radius, valence electron count, electron affinity, and atomic number. The CGCNN convolution operation updates each atom's feature vector based on its neighbors $j \in N(i)$, enabling localized message passing over the crystal graph. Complementing this structural representation, natural language descriptions of crystal structures are generated using a template, which extracts and summarizes symmetry information and structural motifs from CIF files. These descriptions are processed using MatSciBERT [3], a BERT-base–style transformer with 12 encoder layers, 12 attention heads per layer, and a hidden size of 768, pretrained on 3.17 billion words from materials science literature. Input text is tokenized using the WordPiece algorithm, embedded with positional and segment encodings, and transformed through stacked self-attention layers, where multi-head attention is computed followed by residual connections, layer normalization, and position-wise feedforward networks with ReLU activation. The [CLS] token embedding from the final layer is linearly projected to a fixed 150-dimensional representation for compatibility with the multimodal fusion stage. Additionally, X-ray diffraction (XRD) spectra are encoded via a 1D convolutional neural network comprising a convolutional layer, max-pooling layer, and two fully connected layers, which progressively extract local diffraction patterns, reduce dimensionality, and produce a compact spectral embedding. These modality-specific embeddings are subsequently integrated in the fusion module for downstream predictive tasks. Each encoder is described in further detail in the supplementary information.

A.1 Cross Modal Joint Embedding Predictive Architecture

The three modalities are first processed by their own encoders as described in the previous section. For instance, the Text: $x_t \rightarrow z_t = f_t(x_t)$, Graph: $x_g \rightarrow z_g = f_g(x_g)$ and XRD: $x_r \rightarrow z_r = f_r(x_r)$.

These embeddings are then projected into a shared latent space which is known as *Joint Embedding Space*.

$$\tilde{z}_t = P_t(z_t), \quad \tilde{z}_g = P_g(z_g), \quad \tilde{z}_r = P_r(z_r) \tag{1}$$

Each modality is predicted in this shared latent space by using the other two modalities. This makes the model focus on the underlying physical system and ignore the modality specific features. For

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

each modality $m \in \{t, g, r\}$, let $\{m_1, m_2\} = \mathcal{M} \setminus \{m\}$ denote the other two modalities. We define a predictor h_m:

$$\hat{z}_m = h_m([\tilde{z}_{m_1}, \tilde{z}_{m_2}]) \tag{2}$$

The L2 norm is used to calculate the cross modality prediction loss L is:

$$\mathcal{L}^{(m)}_{\text{X-MoPA}} = \|\hat{z}_m - \tilde{z}_m\|_2^2 \tag{3}$$

We use three lightweight MLP predictors, one for predicting each modality. The total Cross Modal JEPA loss is calculated as the sum of all three predictors. We

$$\mathcal{L}_{\text{X-MoPA}} = \sum_{m \in \{t,g,r\}} \mathcal{L}^{(m)}_{\text{X-MoPA}} \tag{4}$$

Further, to prevent **representational collapse** which result in trivial or degenerate outputs, we employ a variance-invariance-covariance (VIC) regularizer[1]. This regularizer operates on the latent features. The *Variance loss* maintains a minimum variance along each latent dimension, the *invariance loss* encourages invariance between positive pairs. Finally, the *Covariance loss* minimizes redundancy between dimensions.

$$\mathcal{L}_{\text{VIC}} = \lambda_v \sum_{d=1}^{D} \max\left(0, \gamma - \text{Var}(Z_{:,d})\right) + \lambda_c \sum_{i \neq j} \left(\text{Cov}(Z_{:,i}, Z_{:,j})\right)^2 \tag{5}$$

To encourage semantically aligned and disentangled representations across modalities, we use a contrastive loss ion the latent space. Positive pairs are constructed from different modalities of the same instance, and negative pairs from different instances. A temperature scaled InfoNCE loss is used for contrastive loss[8].

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp\left(\frac{\text{sim}(z_i, z_j)}{\tau}\right)}{\sum_{k=1}^{2B} 1_{[k \neq i]} \exp\left(\frac{\text{sim}(z_i, z_k)}{\tau}\right)} \tag{6}$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{X-MoPA}} + \lambda \cdot \mathcal{L}_{\text{VIC}}, \quad \lambda \ll 1 \tag{7}$$

B Information Theoretic Bounds

Let X, Y and Z represent the text ,graph and XRD modalities respectively. Let the target property by P and the the latent representations be represented by h_X, h_Y and h_Z.

Theorem B.1 *The use of two input modalities to predict the third modality is possible if and only if I(Z;X,Y) >= H(Z)-ϵ*

Theorem B.1 *It directly follows from the mathematical formulation that for perfect prediction $H(Z|x, y) = 0$. Assuming ϵ is an acceptable level of error then $H(z|X, Y) <= \epsilon$ By information theory, we know that $I(Z; X, Y) = H(Z) - H(Z|X, Y)$ Therefore, it must be true that I(Z;X,Y) >= H(Z)-ϵ*

Theorem B.2 *All modalities describe the same inorganic crystal thus they have an underlying structure S therefore, $I(X; Y; Z) >= I(X; S) + I(Y; S) + I(Z; S) - 2 * H(S)$*

Theorem B.3 *The pre train using a cross modal predictive architecture and then transfer to a downstream tasks has the following generalization bound for downstream tasks.*
$E[L_{downstream}] =< E[L_{JEPA}] + \lambda D_{KL}(P_{pretrain}||P_{downstream}) + O(\sqrt{d/n})$
where, L_{JEPA} denotes the cross modal prediction loss D_{KL} is the KL divergence between the pretraining and downstream tasks. d denotes the dimension of the representation. n is the number of training examples. and λ denotes the transfer coefficient

Corollary B.3.1 *From the above theoretical bound, we have the corollary that if the cross modal L_{JEPA} loss $\leq \delta$ then the downstream loss satisfies.*
$L_{downstream} \leq \delta + C * \sqrt{log(1/\delta)/n}$
where C depends on the Lipschitz constant of the downstream task.

B.1 Latent Space prediction

The latent space is a lower dimensional representation learned from high dimensional data using encoders. If the encoder is trained well, the latent space captures abstract, disentangled and structured features. Thus, the prediction occurs over structured manifolds instead of high-dimensional noise. In high dimensional spaces, the learned mapping between the input and target might be ill-conditioned resulting in high variance. For well trained encoders, the mapping between the latent space and target is Lipschitz-continuous which leads to better learning and generalization.

$$||g(z_1) - g(z_2)|| \leq L||z_1 - z_2|| \tag{8}$$

The manifold hypothesis states that real world data lies in a low-dimensional smooth manifold embedded in high dimensional space. The encoders learn to unwrap these manifolds and thus, the function defined over latent spaces follows geodesics. Thus, they tend to be Lipschitz-continuous.

Let:

- $\mathbf{x} \in \mathbb{R}^n$: input in raw space
- $\mathbf{z} = f_{\text{enc}}(\mathbf{x}) \in \mathbb{R}^d$: latent representation, where $d \ll n$
- $\hat{\mathbf{y}} = g(\mathbf{z})$: prediction made in latent space
- $\hat{\mathbf{y}} = g(f_{\text{enc}}(\mathbf{x}))$: full composition in raw space

Lipschitz Continuity

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be **Lipschitz continuous** if there exists a constant $L > 0$ such that:

$$\|f(\mathbf{x}_1) - f(\mathbf{x}_2)\| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n \tag{9}$$

Jacobian-Based View

For a differentiable function f, the Lipschitz constant is bounded by the operator norm (spectral norm) of the Jacobian:

$$L_f \leq \sup_{\mathbf{x}} \|\nabla f(\mathbf{x})\|_2 \tag{10}$$

Let us now consider the composition:

$$f(\mathbf{x}) = g(f_{\text{enc}}(\mathbf{x})) = g(\mathbf{z}) \tag{11}$$

By the chain rule:

$$\nabla f(\mathbf{x}) = \nabla g(\mathbf{z}) \cdot \nabla f_{\text{enc}}(\mathbf{x}) \tag{12}$$

Hence, the Lipschitz constant of the full model is bounded by:

$$L_f \leq \sup_{\mathbf{x}} \|\nabla g(\mathbf{z})\|_2 \cdot \|\nabla f_{\text{enc}}(\mathbf{x})\|_2 \tag{13}$$

This gives insight into how latent spaces help:

- $\|\nabla f_{\text{enc}}(\mathbf{x})\|_2$: well-trained encoders map high-dimensional, noisy inputs into a smooth, structured space, often with regularized Jacobians.
- $\|\nabla g(\mathbf{z})\|_2$: predictors in latent space are often more stable and operate on disentangled features, reducing the gradient norm.

C Dataset

The Matbench dataset, introduced as part of the Matbench benchmark suite in the Materials Machine Learning (matminer) ecosystem, comprises 13 distinct tasks for evaluating predictive models in materials science. These tasks span regression and classification problems, each with predefined

train-test splits (5-fold) to ensure reproducibility.[4]. For our study, we focus on 13 properties with unique crystal structures in Crystallographic Information File (CIF)

To construct the input modalities required by our X-MoPA model, we process each CIF file through three parallel pipelines. For the graph-based modality, each crystal structure is transformed into an undirected graph using the CGCNN architecture, where atoms serve as nodes and interatomic interactions define the edges. Node features encode essential atomic attributes such as atomic number, electronegativity, ionization energy, and group number[10]. These features are propagated through a message-passing neural network to produce a structure-aware embedding that captures local atomic environments.

For the text-based modality, the CIF files are processed using a template for text generation, which generates structured textual descriptions that summarize structural motifs such as coordination geometries, symmetry operations, and lattice parameters. These descriptions are encoded using the pretrained MatSciBERT language model, which captures domain-specific contextual knowledge from scientific literature and converts the input text into a dense, fixed-length vector representation.

For the spectral modality, we simulate X-ray diffraction (XRD) patterns from the CIF structures using the Pymatgen diffraction module[7]. Each diffraction pattern is computed over a 2θ range of 5° to 90° and discretized into a fixed-length 1D intensity array. These spectra encode long-range order, phase symmetry, and crystallographic fingerprints, which are essential for distinguishing between polymorphs and identifying structural characteristics beyond the atomic neighborhood.

D AFLOW Benchmarking

For further validation, we also evaluate X-MoPA on the AFLOW database, which contains over 60,000 materials. We first split the dataset into 80% training, 10% validation, and 10% test, and pre-trained X-MoPA on the training set using the proposed cross-modal prediction loss in a fully self-supervised manner, without using property labels. To assess downstream performance on the prediction of Band Gap, we then selected a subset of 5,000 samples, again splitting them into train/validation/test following the same 80/10/10 protocol. The model was fine-tuned on the labeled training data and evaluated on the corresponding test set to measure property prediction accuracy. The reference values for the SOTA models have been taken from literature[2].

Table 1: Benchmarking model performance. The lower the error the better the model performance.

	Band Gap(E_g)(eV)			
	SchNet[9]	ElemNet[5]	MPNN[6]	X-MoPA
MAE	0.235	0.515	0.180	**0.161**
RMSE	0.489	0.816	0.399	**0.275**

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[2] Tim Bechtel, Daniel T Speckhard, Jonathan Godwin, and Claudia Draxl. Band-gap regression with architecture-optimized message-passing neural networks. *Chemistry of Materials*, 37(4):1358–1369, 2025.

[3] Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102, 2022.

[4] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. The materials project: A materials genome approach to accelerating materials innovation, apl mater. *Applied Physics Letter*, 2013.

[5] Dipendra Jha, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. Elemnet: Deep learning the chemistry of materials from only elemental composition. *Scientific reports*, 8(1):17593, 2018.

[6] Peter Bjørn Jørgensen, Karsten Wedel Jacobsen, and Mikkel N Schmidt. Neural message passing with edge updates for predicting properties of molecules and materials. *arXiv preprint arXiv:1806.03146*, 2018.

[7] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

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

[9] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018.

[10] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.