

# Text to Band Gap: Pre-trained Language Models as Encoders for Semiconductor Band Gap Prediction

Ying-Ting Yeh,<sup>†</sup> Janghoon Ock,<sup>†</sup> Shagun Maheshwari,<sup>‡</sup> and Amir Barati  
Farimani\*,<sup>¶</sup>

<sup>†</sup>*Department of Chemical Engineering, Carnegie Mellon University, 5000 Forbes Avenue,  
Pittsburgh, PA 15213, USA*

<sup>‡</sup>*Department of Material Science Engineering, Carnegie Mellon University, 5000 Forbes  
Avenue, Pittsburgh, PA 15213, USA*

<sup>¶</sup>*Department of Mechanical Engineering, Carnegie Mellon University, 5000 Forbes Avenue,  
Pittsburgh, PA 15213, USA*

E-mail: barati@cmu.edu

## Abstract

We investigate the use of transformer-based language models, RoBERTa, T5, and LLaMA, for predicting the band gaps of semiconductor materials directly from textual representations that encode key material features such as chemical composition, crystal system, space group, number of atoms per unit cell, valence electron count, and other relevant electronic and structural properties. Quantum chemistry simulations such as Density Functional Theory (DFT) provide accurate predictions but are computationally intensive, limiting their feasibility for large-scale materials screening. Shallow machine learning (ML) models offer faster alternatives but typically require extensive

data preprocessing to convert non-numerical material features into structured numerical inputs, often at the cost of losing critical descriptive information. In contrast, our approach leverages pretrained language models to process textual data directly, eliminating the need for manual feature engineering. We construct material descriptions in two formats: structured strings that combine key features in a consistent template, and natural language narratives generated using the ChatGPT API. For each model, we append a custom regression head and perform task-specific finetuning on a curated dataset of inorganic compounds. Our results show that finetuned language models, particularly the decoder-only LLaMA-3 architecture, can outperform conventional approaches in prediction accuracy and flexibility, achieving a mean absolute error (MAE) of 0.248 eV, root mean squared error (RMSE) of 0.345 eV, and  $R^2$  of 0.891, compared to the best shallow ML baseline (XGBoost), which achieved an MAE of 0.318 eV, RMSE of 0.537 eV, and  $R^2$  of 0.838. Notably, LLaMA-3 achieves competitive accuracy with minimal fine-tuning, suggesting its architecture enables more transferable representations for scientific tasks. This work demonstrates the effectiveness of finetuned language models for scientific property prediction and provides a scalable, language-native framework for materials informatics.

## Introduction

The band gap of semiconductor materials is a fundamental property that directly impacts their electronic and optical behaviors. This parameter dictates crucial attributes such as conductivity, light absorption, and emission, making it essential for the performance of various electronic, optoelectronic, and photovoltaic devices.<sup>1</sup> Therefore, the precise prediction and control of the band gap are vital for optimizing semiconductor applications in these fields.<sup>2</sup>

Band gaps are determined primarily through experimental methods, with UV-visible absorption spectroscopy and photoluminescence spectroscopy being the most commonly used

techniques.<sup>3</sup> However, these experimental methods can only measure the band gaps of synthesized materials and are not applicable to new materials designed theoretically. On the computational side, Density Functional Theory (DFT) has been the primary tool for studying the electronic structure of materials.<sup>4,5</sup> While DFT can accurately simulate electronic properties such as band structures and band gaps, its high computational cost and resource-intensive nature make it less practical for high-throughput material screening, especially for complex systems.<sup>6</sup>

Machine learning (ML) methods have become powerful tools for addressing the computational challenges of DFT. Shallow ML models, such as Random Forest and Support Vector Regression, are commonly used to predict materials properties like band gaps based on material descriptors.<sup>7-9</sup> These models offer a cost-effective alternative to solving the full quantum mechanical equations, significantly reducing computational overhead. However, these models often struggle with non-numerical features, requiring extensive preprocessing to convert material properties into numerical formats. This reliance on extensive preprocessing and feature engineering not only adds complexity but also risks discarding nuanced or qualitative information, such as symmetry, bonding environments, or textual metadata, that could be valuable for accurate property prediction.<sup>10,11</sup> Recent advances in deep learning, particularly Graph Neural Networks (GNNs),<sup>12</sup> have improved the ability to capture the interconnectivity of atomic structures by representing them as graphs. However, this approach requires the conversion of atomic arrangements into graph representations, which adds another layer of preprocessing. Furthermore, GNNs still face limitations in integrating non-numerical properties, such as compound names, into the training process without additional preprocessing steps. These challenges underscore the need for approaches that can seamlessly handle both numerical and non-numerical features in material property predictions while minimizing complex preprocessing requirements.

Language models offer unique advantages by directly utilizing human-readable text data, eliminating the need for elaborate pre-processing while preserving critical information embed-

ded in material descriptions.<sup>13–16</sup> This streamlines the prediction process. Recent advances in natural language processing, particularly with large language models (LLMs), have introduced transformative possibilities for materials science. This simplifies the prediction process compared to conventional ML approaches, which typically require precise atomic coordinates or extensive preprocessing to generate numerical features. In contrast, LLMs can directly process text-based descriptions. Leveraging this capability, we predict band gap values directly from text-formatted input, bypassing the need for detailed structural data and eliminating complex feature engineering.

Recent advances in natural language processing, particularly with LLMs, have introduced transformative opportunities for materials science. For instance, AlloyBERT demonstrates the potential of transformer-based models to predict material properties from descriptive text.<sup>17</sup> Similarly, AMGPT showcases the benefits of using composition-based input strings and finetuned LLMs, enabling accurate and efficient predictions for materials science tasks.<sup>18,19</sup> Additionally, CatBERTa, a RoBERTa-based predictive model, has been developed to predict adsorption energy in catalyst systems.<sup>13,14</sup>

In this study, we explore the use of transformer-based language models, RoBERTa, T5, and LLaMA-3, to predict the band gaps of semiconductor materials directly from textual descriptions. These models enable the direct encoding of structured or natural language representations of materials, such as chemical composition, crystal symmetry, and electronic features, without requiring conventional feature engineering. While pretrained language models possess strong linguistic priors, we emphasize that fine-tuning on domain-specific objectives is essential for adapting them to materials property prediction. We implement task-specific regression heads on top of each model and finetune them on a curated dataset of inorganic compounds. This approach allows models to learn mappings from text-based input to scalar band gap values. These models provide a flexible and generalizable framework for property prediction from textual materials data, extending the application of language models beyond conventional natural language tasks into scientific domains such as materials

informatics.

## Methods

### RoBERTa

RoBERTa (Robustly Optimized BERT Approach) is a transformer-based encoder-only language model that builds upon the BERT architecture by refining pretraining strategies and removing restrictive design choices.<sup>20</sup> It is widely recognized for its ability to capture rich contextual relationships between tokens through deep bidirectional self-attention, making it highly effective for a range of natural language processing tasks.

RoBERTa is pretrained using the masked language modeling (MLM) objective, where a subset of tokens is masked and the model learns to predict them based on surrounding context. Unlike BERT, RoBERTa omits the next-sentence prediction task and instead trains on longer, contiguous sequences of text.<sup>20</sup> It also introduces dynamic masking, larger batch sizes, and improved sentence packing strategies, which collectively enhance the generality and robustness of the learned representations. Tokenization is performed using a byte-level Byte-Pair Encoding (BPE) tokenizer, which allows the model to effectively handle multilingual text, rare subwords, and non-standard Unicode characters.

In this study, we leverage the RoBERTa-base variant, comprising 12 transformer layers, 12 attention heads per layer, and a hidden dimension of 768. While RoBERTa provides strong pretrained representations, our framework emphasizes the necessity of objective-specific fine-tuning to adapt the model to domain-specific tasks such as band gap prediction. The model’s encoder-only structure offers advantages in processing structured text inputs efficiently, particularly in cases where outputs are scalar values rather than generated sequences.

## T5

T5 (Text-to-Text Transfer Transformer) is a unified encoder-decoder architecture that re-frames all NLP tasks—including classification, translation, and regression—as text-to-text problems.<sup>21</sup> In this study, we utilize the T5-small variant, which consists of 6 encoder and 6 decoder layers, each with a hidden size of 512 and 8 attention heads. Unlike encoder-only models such as RoBERTa, T5 processes input and output sequences through separate but coordinated transformer stacks, allowing for flexible sequence-to-sequence modeling and richer bidirectional context capture within the encoder.

T5 is pretrained using a span corruption objective, in which contiguous spans of tokens in the input are replaced by unique sentinel tokens, and the model learns to reconstruct the missing content. This span-level objective fosters stronger global reasoning compared to token-level masking and supports robust performance on tasks requiring understanding of longer context windows. Architectural enhancements include the use of relative positional embeddings rather than absolute positions, parameter sharing between input and output embeddings, and simplified layer normalization without additive bias. These design choices collectively reduce model complexity and improve generalization.

Tokenization in T5 is performed using SentencePiece, a subword tokenizer that supports language-agnostic processing and eliminates the need for whitespace-based token segmentation.<sup>22,23</sup> It is particularly useful for scientific or symbolic inputs, as it can encode rare tokens and domain-specific terms with high fidelity. During training, subword regularization introduces stochasticity in tokenization, enhancing robustness by exposing the model to a diverse range of token combinations.

While T5’s encoder-decoder structure is typically leveraged for generative tasks, it is also adaptable for scalar prediction tasks like band gap regression. In our implementation, we repurpose the encoder to process material descriptions and append a custom regression head to the encoder output.

## LLAMA-3

LLaMA-3.2-1B is a lightweight, decoder-only transformer model derived from Meta’s LLaMA-3 family of large language models.<sup>24,25</sup> Designed to provide a strong performance-efficiency trade-off, this version incorporates approximately 1 billion parameters and benefits from structured pruning and knowledge distillation from larger models. The architecture emphasizes both representational depth and inference efficiency, making it suitable for scalable applications beyond conventional text generation tasks.

Unlike encoder-only models like RoBERTa or encoder-decoder frameworks such as T5, LLaMA-3 operates in an autoregressive, decoder-only configuration. Despite its unidirectional nature, LLaMA-3 can learn rich representations through deeply stacked attention layers and efficient architectural enhancements. The model employs several innovations that distinguish it from earlier transformers: it uses RMSNorm for pre-normalization to improve training stability, SwiGLU as a more expressive activation function, and Rotary Positional Embeddings (RoPE) to better generalize to longer sequences.<sup>26</sup> In some variants, Grouped Query Attention (GQA) is used to reduce computational overhead while maintaining strong modeling capacity.

LLaMA-3 tokenization is performed using BPE via SentencePiece,<sup>22</sup> which enables robust handling of diverse character sets and subword units. Numerals are decomposed into individual digit tokens, and rare or unknown characters are represented using byte-level encodings. These choices ensure flexibility in modeling scientific texts that may include specialized symbols, units, or chemical formulas.

Although LLaMA-3 was originally developed for autoregressive text generation, its modular architecture and robust token-level representation capabilities make it a versatile candidate for regression tasks. The model is particularly well-suited for structured and semi-structured textual input, where consistent formatting allows it to encode domain-relevant relationships. In our context, a custom regression head is appended to the decoder’s final hidden states to adapt the model for scalar band gap prediction. As with other language

models in this study, fine-tuning on task-specific objectives is necessary to align the model’s pretrained linguistic knowledge with the specialized needs of materials property prediction.

## Dataset

In this study, we utilized the AFLOW database, a comprehensive open repository for computational materials science that contains extensive information on inorganic crystalline materials and their properties.<sup>27,28</sup> Band gap calculations in AFLOW combine first-principles methods with empirical corrections through an automated workflow. The framework uses VASP to perform DFT calculations with the GGA-PBE functional for standard compounds while applying the GGA+U method for strongly correlated systems containing d- and f-block elements. To address GGA’s tendency to underestimate band gaps, AFLOW employs an empirical correction scheme based on a linear regression model derived from 102 benchmark compounds with known experimental values.<sup>27,29,30</sup> This systematic approach, along with the database’s vast size and rich feature space, makes AFLOW particularly well-suited for ML tasks aimed at material property prediction.

For our specific analysis, we selected a subset of 27,600 materials with band gap values ranging between 0 and 5 eV (inclusive). This range was chosen because it encompasses the most relevant band gap values for semiconductors, which are of particular interest in materials science and electronic applications. The lower bound of 0 eV represents materials with metallic behavior, where there is no electronic band gap. Materials that have band gap higher than 5 eV are insulating materials, which have too large band gaps and do not conduct electricity under normal conditions.<sup>3,31</sup> By focusing on materials within this range of 0-5 eV band gaps, we ensure that our model targets materials with practical applications in electronics and optoelectronics.

The dataset was divided into training, validation, and test sets to ensure reliable evaluation and optimization of the model. Specifically, 10% of the data was reserved for the test set to evaluate the final performance of the model. The remaining 90% was further split into 80%



for training and 20% for validation, ensuring sufficient data for model training while retaining a representative validation set. This splitting strategy ensured the distribution of band gap values across all subsets, minimizing sampling bias and enhancing representativeness.

## Text Data Format

To investigate the impact of input data representation on model performance, we employed two formats for encoding material property information as text. The first format consists of *structured strings*, where material attributes, such as chemical composition, crystal structure, and electronic features, were compiled into a consistent, template-based layout. This format emphasizes uniformity and feature alignment across samples, providing a well-controlled structure for the language models to process.

The second format consists of *natural language descriptions* generated using OpenAI’s GPT-3.5 Turbo API. The same core features were provided to the API to produce narrative-style descriptions, introducing greater linguistic variability and a more conversational tone. Prompts were configured to ensure descriptions remained within a 512-token limit to maintain compatibility with the tokenizer constraints of RoBERTa, and to accommodate input length limits for T5 and LLaMA-3 as well.

Both formats were applied uniformly across all three models: RoBERTa, T5, and LLaMA-3. The models processed these inputs through their native tokenization pipelines, without additional handcrafted feature engineering. For each model, we appended a custom regression head to enable scalar prediction of band gap values and performed fine-tuning on the downstream task using the corresponding textual inputs.

## Input Features

We carefully selected features that capture both the chemical composition and structural properties of the materials, ensuring a comprehensive understanding of their electronic characteristics, especially the band gap. The selected features include chemical formula, atomic

species, valence electron count, crystal symmetry, and magnetic properties, all of which are known to play critical roles in determining the electronic structure and band gap of materials. A complete list of the 23 selected features, categorized by their respective domains, is provided in Table 1.

The chemical formula represents the basic building blocks of the material, providing critical information about its stoichiometry and composition. The nature and type of atoms constituting the material greatly influence its electronic properties.<sup>9,32,33</sup> For example, the number of valence electrons of each species is crucial for band gap predictions.<sup>34</sup> The total number of valence electrons per unit cell helps predict how these electrons will contribute to the band structure, which is directly related to the band gap.

Additionally, we paid particular attention to structural features, including crystal class, family, and system, as well as lattice parameters, which include the dimensions and angles of the unit cell. These factors not only shape the arrangement of atoms and their interactions but also define the symmetry and geometric properties of the crystal, directly influencing the distribution of electronic states within the energy bands.<sup>30,35,36</sup> The space group and point group information were included to account for the effects of symmetry on electronic states and band splitting. Magnetic properties, such as atom magnetic moments and cell magnetization, were also considered due to their relationship with spin distribution and electronegativity.<sup>3</sup> We specifically chose properties derived from relaxed structures in which the structural configurations have been optimized to minimize energy and stress, ensuring that the atoms are in their equilibrium positions.

## Results and Discussion

### Framework

We developed a language model-based framework to predict the band gaps of semiconductor materials using transformer-based language models: RoBERTa, T5, and LLaMA-3. As

Table 1: Selected feature list. Each feature is accompanied by a specific description explaining its physical significance and contribution to material characterization.

Feature	Description
Compound	Chemical formula of the material, representing its chemical composition
Species	List of atomic species constituting the material
Composition	Proportion of each element in the material
Valence cell (iupac)	Total number of valence electrons in the unit cell, calculated according to IUPAC standards
Species pseudopotential	Type of atomic pseudopotentials used for calculations
Crystal class	Describing the symmetry properties of the crystal
Crystal family	Indicating the basic geometric features of the crystal
Crystal system	Describing the shape and symmetry of the unit cell
Fractional coordinates	Representing the relative positions of atoms in the unit cell
Lattice parameters	The edge lengths and angles of the unit cell
Lattice system	Describing the basic geometric features of the unit cell
Lattice variation	Providing a more detailed description of the lattice
Space group of the structure	Describing the symmetry of the crystal
Space group change loose	Space group determined under looser conditions for crystal structure relaxation, potentially leading to larger symmetry changes
Space group change tight	Space group determined under stricter conditions for structure relaxation, resulting in fewer symmetry changes
Point group orbifold	Describing the topological properties of the point group
Point group order	Indicating the number of symmetry operations in the point group
Point group structure	Describing the geometric features of the point group
Point group type	Classifying the symmetry properties of the point group
Magnitude of magnetic moment for each atom	Describing the local magnetism of the material
Magnetization intensity of each atom	Representing magnetism at the atomic scale
Total magnetization intensity of the entire unit cell	Describing the overall magnetism of the material
Density	Density of the material

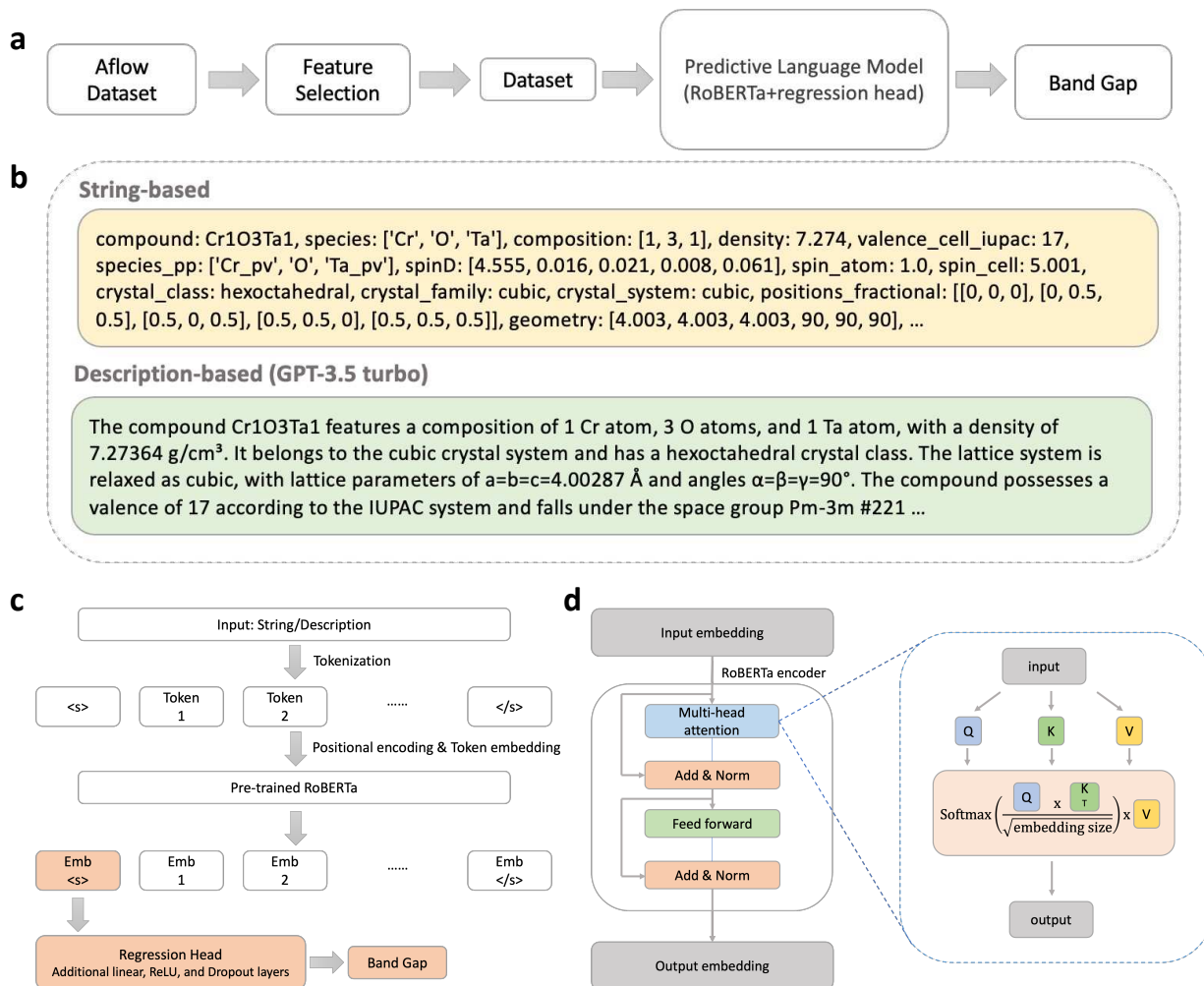


Figure 1: Overview of the proposed band gap prediction framework. **a** The pipeline starts from the AFLOW dataset, followed by feature selection, dataset preparation, and LLM model training for final band gap prediction. **b** Two input formats are illustrated. string-based representation using direct feature values and description-based format generated by GPT-3.5 turbo. **c** Visualization of the finetuning process. The input text undergoes tokenization and embedding through a pretrained LLM, followed by a custom regression head for prediction. **d** Demonstrates the Transformer encoder and the multi-head attention mechanism with Query (Q), Key (K), and Value (V) operations.

shown in Figure 1a, we constructed the dataset from the AFLOW database by extracting relevant material features and transforming them into textual formats suitable for language model inputs. Two types of textual representations were used (Figure 1b): a structured string format that followed a fixed template, and a more flexible natural language description generated using the GPT-3.5 Turbo API. These formats enabled us to assess how each model handles both highly regular and semantically rich input styles. Additional examples of both the structured strings and natural language descriptions are provided in the Supporting Information.s

At the core of our approach are transformer-based language models, each differing in architecture and tokenization strategy. While RoBERTa uses an encoder-only structure, T5 adopts an encoder-decoder design, and LLaMA-3 operates with a decoder-only architecture. Each model tokenizes the input text using its native tokenizer, such as byte-level BPE or SentencePiece, and processes the sequence to generate contextual embeddings (Figure 1c). These embeddings are passed through a custom regression head to produce a scalar band gap prediction, as shown in Figure 1d. The design of this framework allows for direct comparison across model types and input formats while minimizing the need for handcrafted features or domain-specific encodings.

## Model Performance

We evaluated the performance of three transformer-based language models, RoBERTa, T5, and LLaMA-3, on the task of predicting semiconductor band gaps from text-based material descriptions. Each model was finetuned with a custom regression head and assessed using two types of input formats: a structured string representation and a natural language description generated via GPT-3.5 Turbo. These models were compared against shallow ML baselines, which also use the same input data format), including support vector regression (SVR), random forest, and XGBoost. The accuracy of the model was quantified using three metrics: The mean absolute error (MAE), the root mean square error (RMSE), and the coefficient of

determination ( $R^2$ ), as shown in Table 2, with the parity plots visualized in Figure 2.

Table 2: Comparison of model performance across different ML approaches. Each transformer-based model was evaluated using both structured string and description input formats. Best performance per metric is shown in bold.

Model	Model Type	MAE (eV)	RMSE (eV)	$R^2$
SVR	Shallow ML	$0.601 \pm 0.010$	$0.844 \pm 0.008$	$0.600 \pm 0.008$
Random Forest	Shallow ML	$0.385 \pm 0.006$	$0.609 \pm 0.006$	$0.792 \pm 0.005$
XGBoost	Shallow ML	$0.318 \pm 0.005$	$0.537 \pm 0.005$	$0.838 \pm 0.004$
RoBERTa <sub>(string)</sub>	LLM (Encoder)	$0.325 \pm 0.006$	$0.447 \pm 0.005$	$0.855 \pm 0.004$
RoBERTa <sub>(description)</sub>	LLM (Encoder)	$0.421 \pm 0.007$	$0.590 \pm 0.006$	$0.797 \pm 0.006$
T5 <sub>(string)</sub>	LLM (Encoder-Decoder)	$0.301 \pm 0.007$	$0.448 \pm 0.006$	$0.861 \pm 0.005$
T5 <sub>(description)</sub>	LLM (Encoder-Decoder)	$0.446 \pm 0.011$	$0.615 \pm 0.008$	$0.759 \pm 0.008$
LLaMA-3 <sub>(string)</sub>	LLM (Decoder)	<b><math>0.248 \pm 0.006</math></b>	<b><math>0.345 \pm 0.005</math></b>	<b><math>0.891 \pm 0.004</math></b>
LLaMA-3 <sub>(description)</sub>	LLM (Decoder)	$0.335 \pm 0.008$	$0.473 \pm 0.006$	$0.843 \pm 0.054$

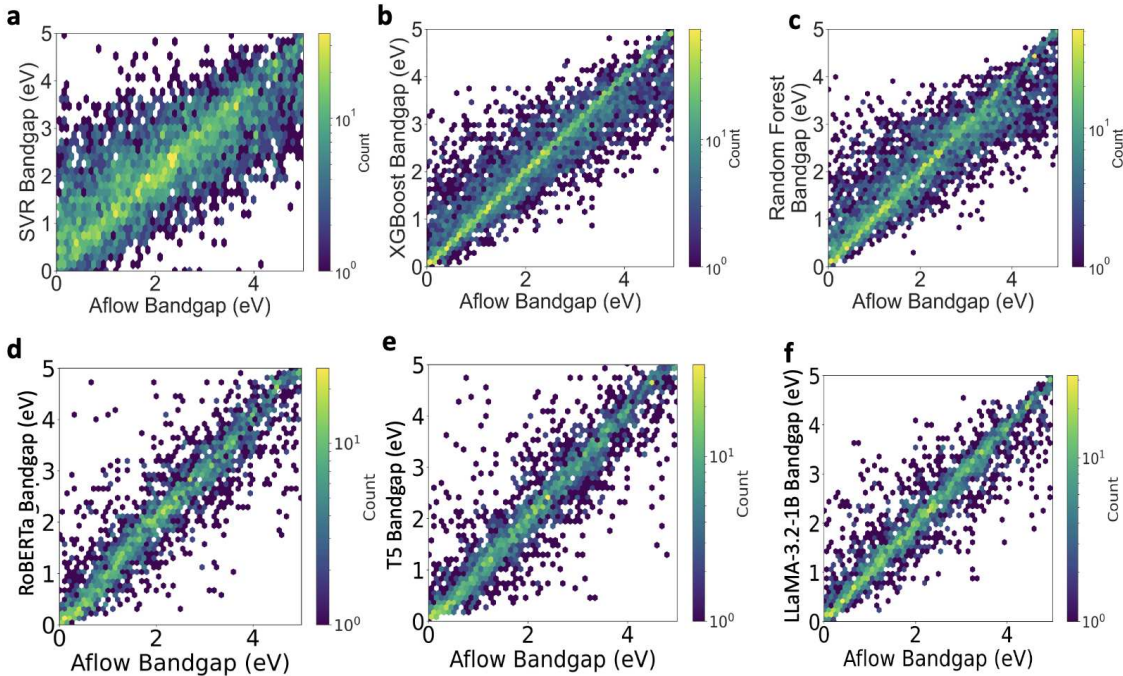


Figure 2: Parity plots for band gap predictions across models: **a** SVR, **b** XGBoost, **c** Random Forest, **d** RoBERTa, **e** T5, **f** LLaMA-3

Across all models, inputs in structured string format consistently outperformed descriptive natural language inputs. Among the transformer-based models, LLaMA-3 with structured input achieved the best overall performance, with a mean absolute error of 0.248 eV and

$R^2$  of 0.891. While RoBERTa and T5 also outperformed the shallow ML models, their performance was lower when using GPT-generated descriptions compared to structured string inputs. However, the ability of all three language models to still learn meaningful correlations from unstructured natural language input underscores a key strength of this approach. Rather than viewing this variability as a limitation, we interpret it as evidence of the models’ capacity to generalize from flexible, human-like descriptions of materials. Notably, even in its smallest size, LLaMA-3 demonstrated strong capacity for property prediction when paired with fine-tuning, indicating that decoder-only architectures can serve as lightweight yet powerful alternatives for scientific regression tasks.

## Layer Freezing Analysis

To assess the extent to which each model’s pretrained representations contribute to band gap prediction, we conducted layer freezing experiments for RoBERTa, T5, and LLaMA-3 using structured string inputs. In each experiment, we progressively froze layers and finetuned only a portion of the model. The goal was to determine how much domain-specific tuning is necessary to achieve high predictive accuracy. Table 3 report the MAE, RMSE, and  $R^2$  values for each model under various freezing configurations. In all cases, a custom regression head was trained on top of the partially or fully frozen transformer backbone.

Throughout this study, we refer to the original language models, trained on general text corpora prior to any materials-specific adaptation, as “pretrained” models. After supervised training on the band gap prediction task, we refer to the models as “finetuned.”

In all three models, predictive performance improved progressively as more transformer layers were unfrozen, highlighting the importance of task-specific fine-tuning in scientific regression applications. This trend highlights a key insight: while LLMs capture broadly useful representations during pretraining, these representations alone are insufficient to achieve optimal accuracy in domain-specific prediction tasks without additional adaptation. In the case of RoBERTa and T5, the most notable performance gains occurred when the final three

Table 3: Comparison of layer freezing strategies across RoBERTa, T5, and LLaMA-3 using structured string inputs. The fully finetuned (non-frozen) results are included for reference. All values are reported as mean  $\pm$  standard deviation.

Model	Freezing Strategy	MAE (eV)	RMSE (eV)	$R^2$
RoBERTa	Fully finetuned (no freezing)	$0.325 \pm 0.006$	$0.447 \pm 0.005$	$0.855 \pm 0.004$
	Freeze first layer	$0.328 \pm 0.008$	$0.448 \pm 0.006$	$0.848 \pm 0.005$
	Freeze all but final 3 layers	$0.388 \pm 0.009$	$0.510 \pm 0.007$	$0.817 \pm 0.006$
	Freeze all but final layer	$0.509 \pm 0.012$	$0.648 \pm 0.009$	$0.721 \pm 0.009$
	Freeze all layers	$0.663 \pm 0.016$	$0.826 \pm 0.011$	$0.563 \pm 0.013$
T5	Fully finetuned (no freezing)	$0.301 \pm 0.007$	$0.448 \pm 0.006$	$0.861 \pm 0.005$
	Freeze first layer	$0.350 \pm 0.008$	$0.504 \pm 0.007$	$0.849 \pm 0.006$
	Freeze all but final 3 layers	$0.367 \pm 0.009$	$0.516 \pm 0.007$	$0.832 \pm 0.006$
	Freeze all but final layer	$0.598 \pm 0.014$	$0.784 \pm 0.011$	$0.619 \pm 0.011$
	Freeze all layers	$0.792 \pm 0.019$	$0.981 \pm 0.013$	$0.420 \pm 0.014$
LLaMA-3	Fully finetuned (no freezing)	<b><math>0.248 \pm 0.006</math></b>	<b><math>0.345 \pm 0.005</math></b>	<b><math>0.891 \pm 0.004</math></b>
	Freeze first layer	$0.279 \pm 0.007$	$0.426 \pm 0.006$	$0.878 \pm 0.004$
	Freeze all but final 3 layers	$0.318 \pm 0.008$	$0.474 \pm 0.006$	$0.851 \pm 0.005$
	Freeze all but final layer	$0.424 \pm 0.010$	$0.576 \pm 0.008$	$0.793 \pm 0.007$
	Freeze all layers	$0.716 \pm 0.017$	$0.893 \pm 0.012$	$0.518 \pm 0.013$

layers of the model were unfrozen. This suggests that the highest-level contextual embeddings learned during pretraining are not directly aligned with the band gap target property but can be effectively adapted through partial retraining.

LLaMA-3, on the other hand, exhibited relatively strong performance even when the majority of its layers were frozen. For example, when only the final three layers were unfrozen during fine-tuning, LLaMA-3 achieved  $R^2$  values comparable to those of fully fine-tuned RoBERTa and T5 models. This observation points to its efficient architecture and inductive biases, which may make its learned representations more transferable to new scientific tasks. In particular, the decoder-only architecture of LLAMA-3, combined with its use of Rotary Positional Embeddings and SwiGLU activation functions, appears to confer greater flexibility in adapting to structured input formats such as the material strings used in this study. Furthermore, the ability of LLAMA-3 to achieve high  $R^2$  values with only the final layer trainable supports its utility in resource-constrained scenarios, such as edge inference or few-shot scientific learning contexts.



Collectively, these results reinforce the conclusion that while pretrained LLMs provide a valuable foundation, their effective use in materials science and other physical domains requires carefully calibrated fine-tuning strategies. Strategic layer freezing not only reduces computational burden during training but also offers insight into which parts of the model architecture contribute most directly to scientific prediction tasks. This understanding is critical for efficiently deploying transformer-based models in practical, data-limited, or performance-critical settings.

## Embedding Analysis

We conducted embedding space analysis using t-SNE visualizations to investigate how each model organizes material representations before and after fine-tuning. As shown in Figure 3, we extracted the first-token embeddings from the pretrained and finetuned versions of RoBERTa, T5, and LLaMA-3 using structured string inputs, and colored the points by crystal system. In the “pretrained” condition, which corresponds to the original language model weights before finetuning on the band gap prediction task, the embeddings reflect structural signals learned from general language corpora.

In the pretrained state, the embeddings of all three models exhibit meaningful clustering based on the crystal system. This structure-aware behavior is likely driven by explicit features in the input, such as crystal class, space group, and lattice parameters, information that correlates with symmetry. However, while these features can separate materials by structural family, they do not offer sufficient signal to accurately predict electronic properties like band gap. Without finetuning, the models lack the task-specific supervision needed to connect these latent structural cues to quantitative property outcomes. Among the models, LLaMA-3 showed the most distinct clustering by crystal system even before finetuning, particularly in distinguishing high-symmetry groups like cubic and trigonal. RoBERTa and T5 also demonstrate some clustering by structural categories, albeit with greater overlap and more diffuse boundaries. These patterns indicate that pretrained language models are capa-

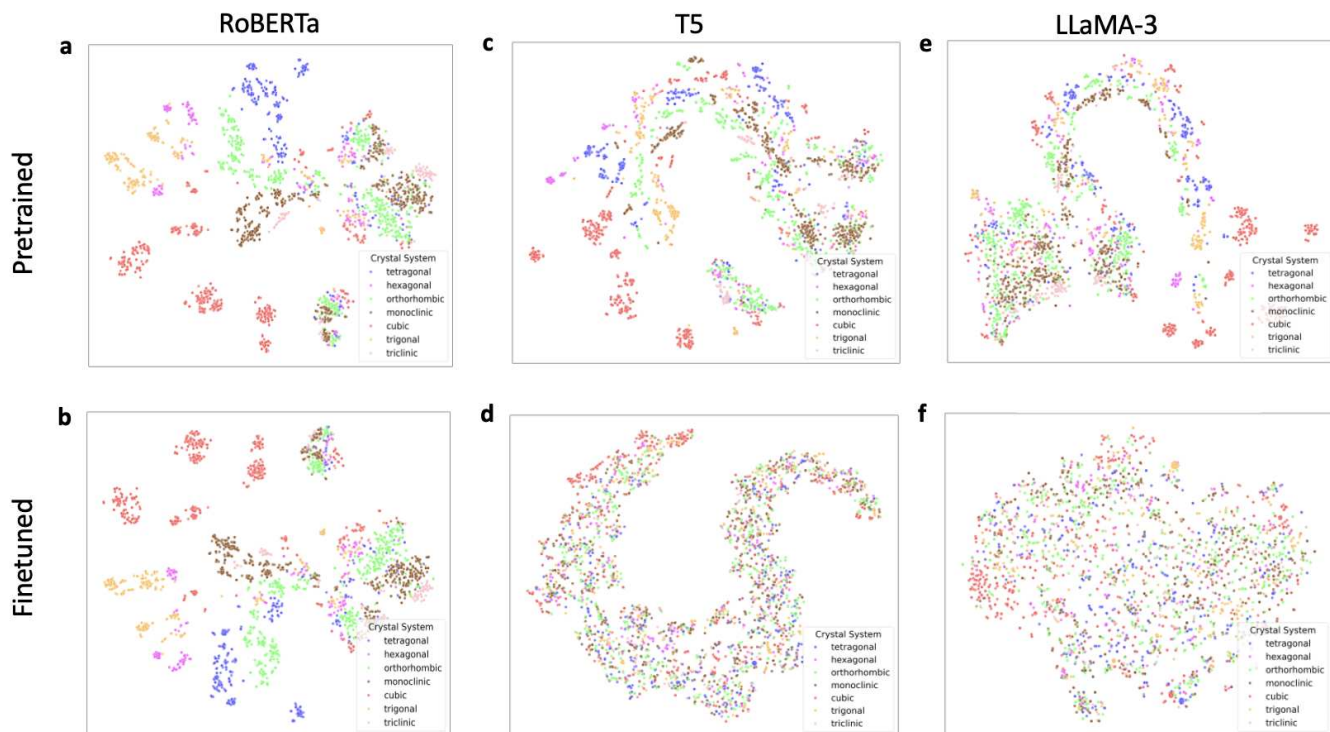


Figure 3: t-sne visualizations of embedding analysis with crystal systems. **a** pretrained RoBERTa, **b** finetuned RoBERTa, **c** pretrained T5, **d** finetuned T5, **e** pretrained LLaMA-3, **f** finetuned LLaMa-3

ble of learning latent structural relationships from text, despite not being trained specifically for materials science.

Once finetuned on the band gap prediction task, however, the clustering behavior undergoes a clear transition: embeddings now group primarily by band gap values rather than by crystal system, as visualized in Figure 4. This shift reveals that the fine-tuning process successfully reorients the model’s latent space to emphasize property-relevant dimensions over purely structural ones. The result is an embedding space where materials with similar band gaps are positioned closer together, regardless of their crystallographic classification. This transformation illustrates the effectiveness of supervised training in realigning general-purpose language representations toward targeted scientific objectives.

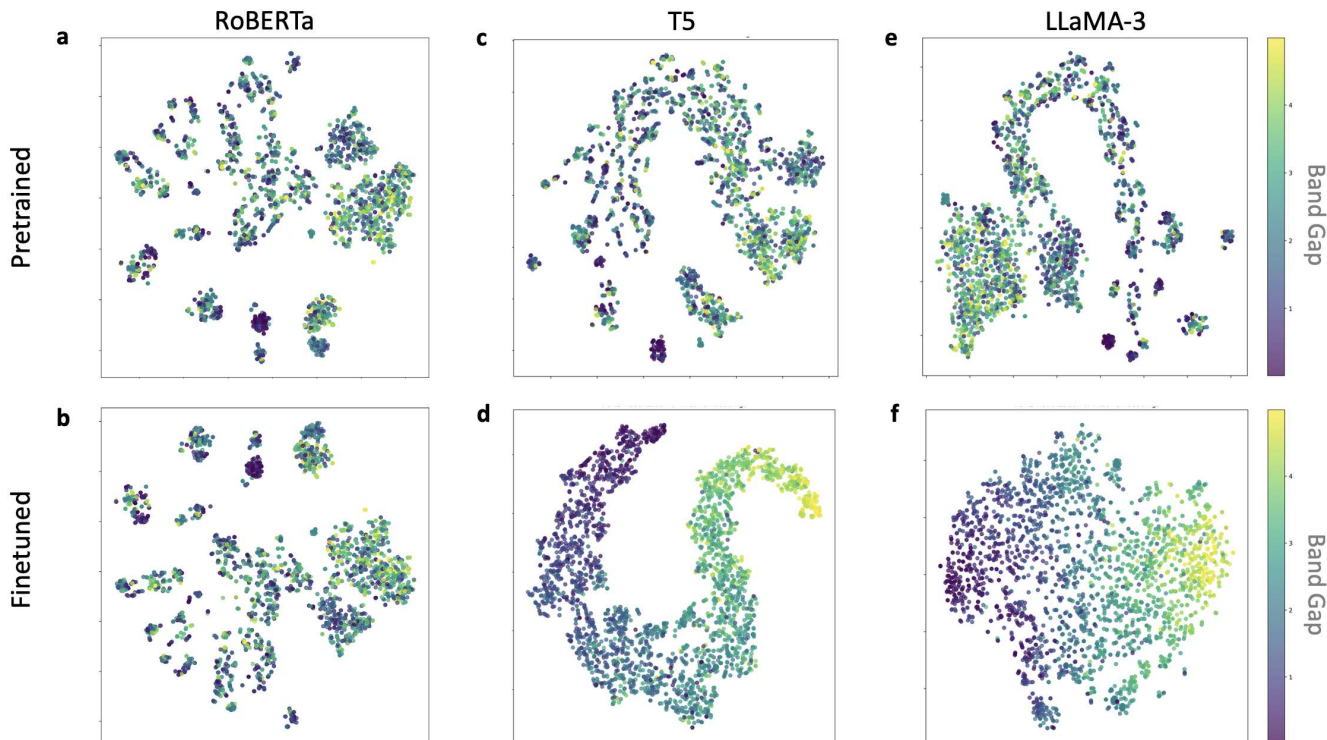


Figure 4: t-SNE visualizations of embedding analysis for bandgap. **a** pretrained RoBERTa, **b** finetuned RoBERTa, **c** pretrained T5, **d** finetuned T5, **e** pretrained LLaMa-3, **f** finetuned LLaMa-3

## Conclusion

In this study, we investigated the use of transformer-based language models, RoBERTa, T5, and LLaMA-3, for predicting the band gaps of semiconductor materials directly from textual inputs. We demonstrated that these models can learn meaningful structure–property relationships without relying on handcrafted features or complex numerical descriptors.

Finetuned language models consistently outperformed shallow ML baselines, with LLaMA-3 achieving the highest accuracy using structured inputs (MAE 0.248 eV,  $R^2$  0.891). Even with natural language descriptions, the models captured relevant patterns, highlighting their flexibility for scenarios lacking structured data. Additionally, layer freezing analysis showed that strong predictive performance can be achieved with minimal finetuning. These findings underscore the adaptability of language models for scientific regression tasks.

Overall, our results establish transformer-based language models, particularly compact, decoder-only architectures like LLaMA-3, as promising tools for text-driven materials property prediction, offering scalability, interpretability, and efficiency for future materials informatics pipelines.

## Code Availability Statement

The Python code in this study is available on GitHub at the following link: [https://github.com/yingtiny/bandgap\\_prediction\\_RoBERTa](https://github.com/yingtiny/bandgap_prediction_RoBERTa).

## References

- (1) Yu, P. Y.; Cardona, M. *Fundamentals of Semiconductors: Physics and Materials Properties*, 4th ed.; Springer, 2010.
- (2) Kim, S.; Lee, M.; Hong, C.; Yoon, Y.; An, H.; Lee, D.; Jeong, W.; Yoo, D.; Kang, Y.; Youn, Y.; others A band-gap database for semiconducting inorganic materials calculated with hybrid functional. *Scientific Data* **2020**, *7*, 387.
- (3) Masood, H.; Sirojan, T.; Toe, C. Y.; Kumar, P. V.; Haghshenas, Y.; Sit, P. H.; Amal, R.; Sethu, V.; Teoh, W. Y. Enhancing prediction accuracy of physical band gaps in semiconductor materials. *Cell Reports Physical Science* **2023**, *4*.
- (4) Koch, W.; Holthausen, M. C. *A chemist's guide to density functional theory*; John Wiley & Sons, 2015.
- (5) Kohn, W.; Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical review* **1965**, *140*, A1133.
- (6) Schleder, G. R.; Padilha, A. C.; Acosta, C. M.; Costa, M.; Fazzio, A. From DFT to

- machine learning: recent approaches to materials science—a review. *Journal of Physics: Materials* **2019**, *2*, 032001.
- (7) Wang, T.; Tan, X.; Wei, Y.; Jin, H. Accurate bandgap predictions of solids assisted by machine learning. *Materials Today Communications* **2021**, *29*, 102932.
- (8) Rajan, A. C.; Mishra, A.; Satsangi, S.; Vaish, R.; Mizuseki, H.; Lee, K.-R.; Singh, A. K. Machine-learning-assisted accurate band gap predictions of functionalized MXene. *Chemistry of Materials* **2018**, *30*, 4031–4038.
- (9) Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *The journal of physical chemistry letters* **2018**, *9*, 1668–1673.
- (10) Faber, F. A.; Hutchison, G. R.; Huang, B.; von Lilienfeld, O. A. Prediction errors of molecular machine learning models lower than hybrid DFT error. *Journal of Chemical Theory and Computation* **2017**, *13*, 5255–5264.
- (11) Choudhary, K.; DeCost, B. The Atomistic Line Graph Neural Network for improved materials property predictions. *npj Computational Materials* **2021**, *7*, 185.
- (12) Taniguchi, T.; Hosokawa, M.; Asahi, T. Graph comparison of molecular crystals in band gap prediction using neural networks. *ACS omega* **2023**, *8*, 39481–39489.
- (13) Ock, J.; Guntuboina, C.; Barati Farimani, A. Catalyst Energy Prediction with CatBERTa: Unveiling Feature Exploration Strategies through Large Language Models. *ACS Catalysis* **2023**, *13*, 16032–16044.
- (14) Ock, J.; Badrinarayanan, S.; Magar, R.; Antony, A.; Barati Farimani, A. Multimodal language and graph learning of adsorption configuration in catalysis. *Nature Machine Intelligence* **2024**, 1–11.
- (15) Guntuboina, C.; Das, A.; Mollaei, P.; Kim, S.; Barati Farimani, A. PeptideBERT: A

- Language Model Based on Transformers for Peptide Property Prediction. *The Journal of Physical Chemistry Letters* **2023**, *14*, 10427–10434, PMID: 37956397.
- (16) Pak, P.; Farimani, A. B. AdditiveLLM: Large Language Models Predict Defects in Additive Manufacturing. 2025; <https://arxiv.org/abs/2501.17784>.
- (17) Chaudhari, A.; Guntuboina, C.; Huang, H.; Farimani, A. B. AlloyBERT: Alloy property prediction with large language models. *Computational Materials Science* **2024**, *244*, 113256.
- (18) Jacobs, R.; Polak, M. P.; Schultz, L. E.; Mahdavi, H.; Honavar, V.; Morgan, D. Regression with Large Language Models for Materials and Molecular Property Prediction. 2024; <https://arxiv.org/abs/2409.06080>.
- (19) Chandrasekhar, A.; Chan, J.; Ogoke, F.; Ajenifujah, O.; Barati Farimani, A. AMGPT: A large language model for contextual querying in additive manufacturing. *Additive Manufacturing Letters* **2024**, *11*, 100232.
- (20) Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* **2019**, *364*.
- (21) Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **2020**, *21*, 1–67.
- (22) Google SentencePiece: A language-independent subword tokenizer. 2020; <https://github.com/google/sentencepiece>.
- (23) HuggingFace T5 Model Documentation. 2024; [https://huggingface.co/docs/transformers/en/model\\_doc/t5](https://huggingface.co/docs/transformers/en/model_doc/t5).
- (24) Meta AI LLaMA 3.2 Model Overview. 2024; <https://arxiv.org/pdf/2407.21783>.

- (25) Meta AI LLaMA 3.1 Technical Report. 2023; <https://arxiv.org/pdf/2302.13971>.
- (26) Meta AI LLaMA 3.2: Connect 2024 Vision for Edge and Mobile Devices. 2024; <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>.
- (27) Gossett, E.; Toher, C.; Oses, C.; Isayev, O.; Legrain, F.; Rose, F.; Zurek, E.; Carrete, J.; Mingo, N.; Tropsha, A.; others AFLOW-ML: A RESTful API for machine-learning predictions of materials properties. *Computational Materials Science* **2018**, *152*, 134–145.
- (28) Taylor, R. H.; Rose, F.; Toher, C.; Levy, O.; Yang, K.; Nardelli, M. B.; Curtarolo, S. A RESTful API for exchanging materials data in the AFLOWLIB. org consortium. *Computational materials science* **2014**, *93*, 178–192.
- (29) Setyawan, W.; Gaume, R. M.; Lam, S.; Feigelson, R. S.; Curtarolo, S. High-throughput combinatorial database of electronic band structures for inorganic scintillator materials. *ACS combinatorial science* **2011**, *13*, 382–390.
- (30) Wang, T.; Zhang, K.; Thé, J.; Yu, H. Accurate prediction of band gap of materials using stacking machine learning model. *Computational Materials Science* **2022**, *201*, 110899.
- (31) Tripathy, S. K.; Pattanaik, A. Optical and electronic properties of some semiconductors from energy gaps. *Optical Materials* **2016**, *53*, 123–133.
- (32) He, Y.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. Metallic metal–organic frameworks predicted by the combination of machine learning methods and ab initio calculations. *The journal of physical chemistry letters* **2018**, *9*, 4562–4569.
- (33) Khan, A.; Tayara, H.; Chong, K. T. Prediction of organic material band gaps using graph attention network. *Computational Materials Science* **2023**, *220*, 112063.

- (34) Huang, Y.; Yu, C.; Chen, W.; Liu, Y.; Li, C.; Niu, C.; Wang, F.; Jia, Y. Band gap and band alignment prediction of nitride-based semiconductors using machine learning. *Journal of Materials Chemistry C* **2019**, *7*, 3238–3245.
- (35) Zheng, X.; Cohen, A. J.; Mori-Sánchez, P.; Hu, X.; Yang, W. Improving band gap prediction in density functional theory from molecules to solids. *Physical review letters* **2011**, *107*, 026403.
- (36) Na, G. S.; Jang, S.; Lee, Y.-L.; Chang, H. Tuplewise material representation based machine learning for accurate band gap prediction. *The Journal of Physical Chemistry A* **2020**, *124*, 10616–10623.



# Supporting Information:

## Text to Band Gap: Pre-trained Language Models as Encoders for Semiconductor Band Gap Prediction

Ying-Ting Yeh,<sup>†</sup> Janghoon Ock,<sup>†</sup> Shagun Maheshwari,<sup>‡</sup> and Amir Barati  
Farimani\*,<sup>¶</sup>

<sup>†</sup>*Department of Chemical Engineering, Carnegie Mellon University, 5000 Forbes Avenue,  
Pittsburgh, PA 15213, USA*

<sup>‡</sup>*Department of Material Science Engineering, Carnegie Mellon University, 5000 Forbes  
Avenue, Pittsburgh, PA 15213, USA*

<sup>¶</sup>*Department of Mechanical Engineering, Carnegie Mellon University, 5000 Forbes Avenue,  
Pittsburgh, PA 15213, USA*

E-mail: barati@cmu.edu

## Contents

S1 String and Description Examples

S2

## S1 String and Description Examples

CrO<sub>3</sub>Ta

### String:

compound: Cr1O3Ta1, species: ['Cr', 'O', 'Ta'], composition: [1, 3, 1], density: 7.274, valence\_cell\_iupac: 17, species\_pp: ['Cr\_pv', 'O', 'Ta\_pv'], spinD: [4.555, 0.016, 0.021, 0.008, 0.061], spin\_atom: 1.0, spin\_cell: 5.001, crystal\_class: hexoctahedral, crystal\_family: cubic, crystal\_system: cubic, positions\_fractional: [[0, 0, 0], [0, 0.5, 0.5], [0.5, 0, 0.5], [0.5, 0.5, 0], [0.5, 0.5, 0.5]], geometry: [4.003, 4.003, 4.003, 90, 90, 90], lattice\_system\_relax: cubic, lattice\_variation\_relax: CUB, spacegroup\_relax: 221, sg: ['Pm-3m #221', 'Pm-3m #221', 'Pm-3m #221'], sg2: ['Pm-3m #221', 'Pm-3m #221', 'Pm-3m #221'], point\_group\_orbifold: \*432, point\_group\_order: 48, point\_group\_structure: 2\_x\_symmetric, point\_group\_type: centrosymmetric

### Description:

The compound CrO<sub>3</sub>Ta features a composition of 1 Cr atom, 3 O atoms, and 1 Ta atom, with a density of 7.27364 g/cm<sup>3</sup>. It belongs to the cubic crystal system and has a hexoctahedral crystal class. The lattice system is relaxed cubic with lattice parameters  $a = b = c = 4.00287$  Å and angles  $\alpha = \beta = \gamma = 90^\circ$ .

The compound has a valence of 17 according to the IUPAC system and crystallizes in the space group Pm-3m #221. It exhibits spin values of 1.00013 at the atomic level and 5.00063 at the cell level. The structure is highly symmetric, belonging to the \*432 point group with an order of 48 and a centrosymmetric configuration.

The atoms are located at fractional coordinates: (0, 0, 0), (0, 0.5, 0.5), (0.5, 0, 0.5), (0.5, 0.5, 0), and (0.5, 0.5, 0.5). The atomic species present are Cr, O, and Ta, with the pseudopotential designations Cr\_pv, O, and Ta\_pv, respectively.

**String:**

compound: Bi1Dy1Ni1, species: ['Bi', 'Dy', 'Ni'], composition: [1, 1, 1], density: 10.599, valence\_cell\_iupac: 13, species\_pp: ['Bi\_d', 'Dy\_3', 'Ni\_pv'], spinD: [0, 0, 0], spin\_atom: 0.0, spin\_cell: 0.0, crystal\_class: tetrahedral, crystal\_family: cubic, crystal\_system: cubic, positions\_fractional: [[0, 0, 0], [0.5, 0.5, 0.5], [0.25, 0.25, 0.25]], geometry: [4.568, 4.568, 4.568, 60, 60, 60], lattice\_system\_relax: cubic, lattice\_variation\_relax: FCC, spacegroup\_relax: 216, sg: ['F-43m #216', 'F-43m #216', 'F-43m #216'], sg2: ['F-43m #216', 'F-43m #216', 'F-43m #216'], point\_group\_orbifold: \*332, point\_group\_order: 24, point\_group\_structure: symmetric, point\_group\_type: none

**Description:**

This material is a cubic compound with the chemical formula BiDyNi. It has a density of 10.5987 g/cm<sup>3</sup> and a valence of 13 according to the IUPAC system. The crystal structure is tetrahedral within the cubic crystal family and system. The lattice system is relaxed cubic with a face-centered cubic (FCC) lattice variation. The space group is F-43m #216, and the point group is \*332 with an order of 24, showing a symmetric structure. The atomic positions in the unit cell are at (0,0,0), (0.5,0.5,0.5), and (0.25,0.25,0.25). The atomic species present are Bi, Dy, and Ni, with spins of 0 for each atom. The geometry of the unit cell is characterized by lattice parameters of  $a = b = c = 4.567913 \text{ \AA}$  and  $\alpha = \beta = \gamma = 60^\circ$ . The species have the configurations Bi\_d, Dy\_3, and Ni\_pv respectively.

**String:**

compound: Au<sub>2</sub>Bi<sub>2</sub>Li<sub>4</sub>, species: ['Au', 'Bi', 'Li'], composition: [2, 2, 4], density: 8.068, valence\_cell\_iupac: 24, species\_pp: ['Au', 'Bi\_d', 'Li\_sv'], spinD: [0, 0, 0, 0, 0, 0, 0, 0], spin\_atom: 0.0, spin\_cell: 0.0, crystal\_class: orthorhombic-bipyramidal, crystal\_family: orthorhombic, crystal\_system: orthorhombic, positions\_fractional: [[0, 0, 0], [0, 0, 0.5], [0.662, 0.662, 0.25], [0.338, 0.338, 0.75], [0.474, 0.12, 0.25], [0.526, 0.88, 0.75], [0.12, 0.474, 0.25], [0.88, 0.526, 0.75]], geometry: [5.563, 5.563, 5.638, 90, 90, 97.937], lattice\_system\_relax: orthorhombic, lattice\_variation\_relax: ORCC, spacegroup\_relax: 63, sg: ['Cmcm #63', 'Cmcm #63', 'Cmcm #63'], sg2: ['Cmcm #63', 'Cmcm #63', 'Cmcm #63'], point\_group\_orbifold: \*222, point\_group\_order: 8, point\_group\_structure: 2\_x\_dihedral, point\_group\_type: centrosymmetric

**Description:**

The compound Au<sub>2</sub>Bi<sub>2</sub>Li<sub>4</sub> features a unique orthorhombic crystal structure with a crystal class of orthorhombic-bipyramidal, belonging to the orthorhombic crystal family and system. The material has a density of 8.06805 g/cm<sup>3</sup> and a valence cell of 24. The chemical composition consists of 2 atoms of Au, 2 atoms of Bi, and 4 atoms of Li. The lattice system is orthorhombic, with lattice parameters  $a = b = 5.562936$  Å,  $c = 5.638398$  Å, and angles  $\alpha = \beta = 90^\circ$ ,  $\gamma = 97.937^\circ$ . The space group is Cmcm #63, with a relaxed lattice system of orthorhombic and lattice variation of ORCC.

The atoms are positioned in the crystal structure at fractional coordinates: (0, 0, 0), (0, 0, 0.5), (0.662, 0.662, 0.25), (0.338, 0.338, 0.75), (0.474, 0.12, 0.25), (0.526, 0.88, 0.75), (0.12, 0.474, 0.25), and (0.88, 0.526, 0.75).

The point group characteristics include a point group orbifold of \*222, an order of 8, a structure of 2\_x\_dihedral, and a centrosymmetric type. Spin properties indicate zero spin for both individual atoms and the overall unit cell.

## Ag<sub>2</sub>CrPt

### String:

compound: Ag2Cr1Pt1, species: ['Ag', 'Cr', 'Pt'], composition: [2, 1, 1], density: 12.174, valence\_cell\_iupac: 20, species\_pp: ['Ag', 'Cr\_pv', 'Pt'], spinD: [0, 0, 0, 0], spin\_atom: 0.0, spin\_cell: 0.0, crystal\_class: tetrahedral, crystal\_family: cubic, crystal\_system: cubic, positions\_fractional: [[0, 0, 0], [0.25, 0.25, 0.25], [0.5, 0.5, 0.5], [0.75, 0.75, 0.75]], geometry: [4.469, 4.469, 4.469, 60, 60, 60], lattice\_system\_relax: cubic, lattice\_variation\_relax: FCC, spacegroup\_relax: 216, sg: ['F-43m #216', 'F-43m #216', 'F-43m #216'], sg2: ['F-43m #216', 'F-43m #216', 'F-43m #216'], point\_group\_orbifold: \*332, point\_group\_order: 24, point\_group\_structure: symmetric, point\_group\_type: none

### Description:

The material is a compound with the chemical formula Ag<sub>2</sub>CrPt, consisting of silver (Ag), chromium (Cr), and platinum (Pt) in a ratio of 2:1:1. It has a density of 12.1739 g/cm<sup>3</sup> and a valence of 20 according to the IUPAC standard. The crystal structure is cubic with a tetrahedral arrangement. The lattice system is cubic with a face-centered cubic (FCC) lattice variation, and the space group is F-43m #216. The point group is \*332, indicating a symmetric structure with 24-fold rotational symmetry.

The system exhibits no net magnetic moment, with spin values of 0.0 at both the atom and cell levels. Atomic positions are defined by fractional coordinates: (0, 0, 0), (0.25, 0.25, 0.25), (0.5, 0.5, 0.5), and (0.75, 0.75, 0.75) within the unit cell. The crystal geometry is characterized by lattice parameters  $a = b = c = 4.4694 \text{ \AA}$  and angles  $\alpha = \beta = \gamma = 60^\circ$ .