Predicting Many Crystal Properties via an Adaptive Transformer-Based Framework

Haosheng Xu[1,2†], Dongheng Qian[1,2†], and Jing Wang[1,2,3,4*]

[1]*State Key Laboratory of Surface Physics and Department of Physics, Fudan University, Shanghai 200433, China*
[2]*Shanghai Research Center for Quantum Sciences, Shanghai 201315, China*
[3]*Institute for Nanoelectronic Devices and Quantum Computing, Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai 200433, China*
[4]*Hefei National Laboratory, Hefei 230088, China*

(Received 1 April 2025; accepted manuscript online 21 May 2025)

Machine learning has revolutionized many fields, including materials science. However, predicting the properties of crystalline materials using machine learning faces challenges in input encoding, output versatility, and interpretability. We introduce CrystalBERT, an adaptable transformer-based framework integrating space group, elemental, and unit cell information. This novel structure can seamlessly combine diverse features and accurately predict various physical properties, including topological properties, superconducting transition temperatures, dielectric constants, and more. CrystalBERT provides insightful interpretations of features influencing target properties. Our results indicate that space group and elemental information are crucial for predicting topological and superconducting properties, underscoring their intricate nature. By incorporating these features, we achieve 91% accuracy in topological classification, surpassing prior studies and identifying previously misclassified materials. This research demonstrates that integrating diverse material information enhances the prediction of complex material properties, paving the way for more accurate and interpretable machine learning models in materials science.

DOI: 10.1088/0256-307X/42/7/070706 CSTR: 32039.14.0256-307X.42.7.070706

1. Introduction. Machine learning (ML) has emerged as a prominent research field in computer science, extending its applications to various domains within physics.[1] A significant application is in uncovering patterns and predicting properties from large datasets, which might be previously unknown or difficult for humans to calculate.[2–4] In materials science and condensed matter physics, ML models have been developed to predict a wide range of material properties without the need for specific, arduous calculations.[5–23]

When designing ML models, several challenges typically arise. The first challenge involves input encoding, which concerns how to represent a specific material as vectors. Previous strategies can be broadly categorized into three approaches: manual feature selection,[24,25] text description, which describes a material using a human-readable sentence for input into a large language model,[26–30] and physically motivated methods such as crystal graphs.[31–33] A pertinent question is whether a general method exists to integrate all these types of features into a single model. Recent work has begun to explore this issue by applying multimodal learning.[34–36] The second challenge is output versatility. Most models excel at certain tasks but perform poorly on others. Given the goal of predicting various properties of a material, it is worth exploring whether a general model can perform well across a wide range of tasks. Data utilization presents the third significant challenge. For specific tasks, the number of labeled materials might be relatively small compared to the total number of known materials. A model that effectively utilizes vast amounts of unlabeled data would be desirable and potentially more powerful, making pretrained models a promising solution. Interpretability is the final aspect. Beyond knowing what the model predicts, it is essential to understand why it makes certain predictions. While deep learning models are often viewed as black boxes, making it challenging to derive a concrete formula relating input to output, there is still a need to identify the most relevant features that influence a specific property.

In this Letter, we introduce CrystalBERT (XBERT), a novel ML model designed to potentially address all the aforementioned challenges. The model structure is illustrated in Fig. 1(a). We utilize the bidirectional encoder representation from transformers (BERT) architecture[37] for pretraining our model, treating the prediction of various physically important properties as different downstream tasks. During pretraining, we leverage large amounts of crystalline materials to develop a sophisticated representation of materials. This pretraining process

†These authors contributed equally to this work.
*Corresponding author. Email: wjingphys@fudan.edu.cn
© 2025 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

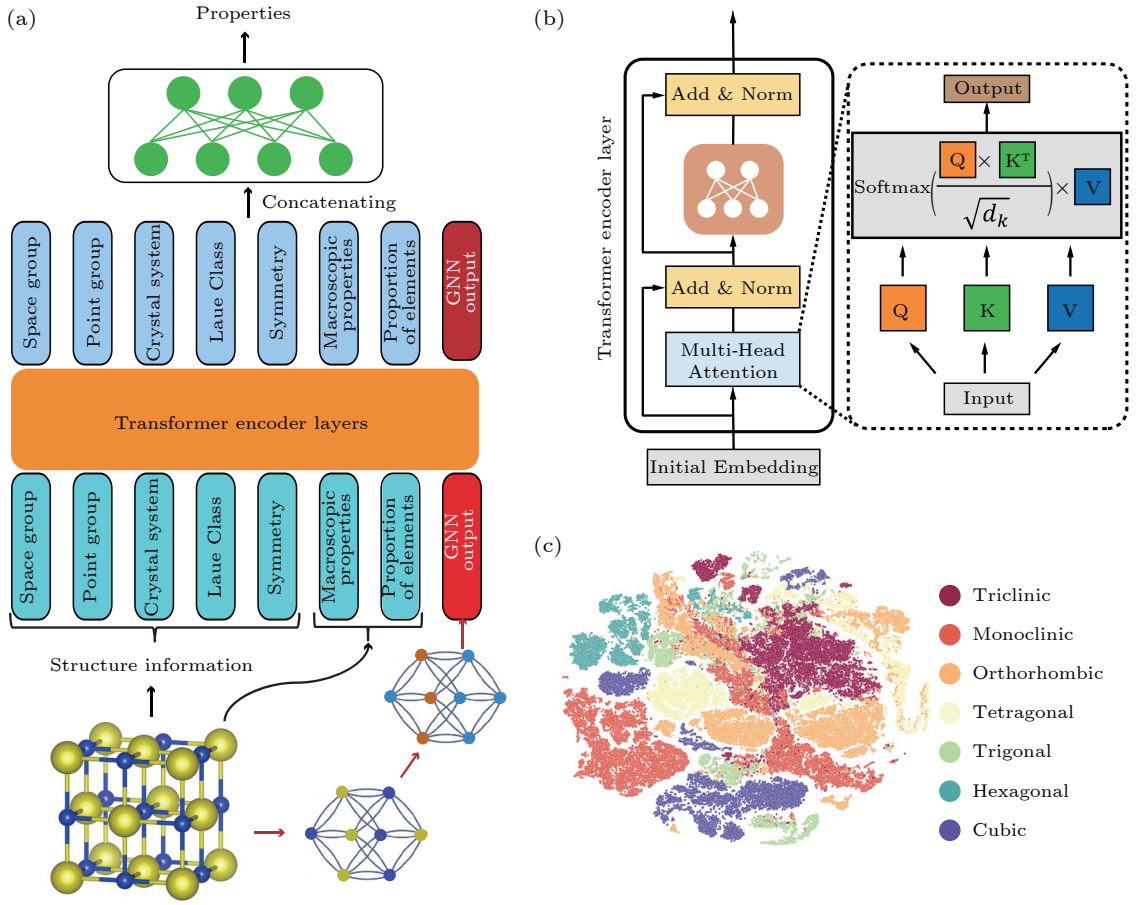

Fig. 1. XBERT's structure and t-SNE result. (a) XBERT's structure. A crystalline material's structural and elemental information are encoded into eight tokens. The last token is derived from the output of a GNN, which captures unit cell information. These tokens are then fed into the transformer encoder, and a single fully connected layer is applied to the output tokens, with outputs depending on the specific task. The symbols positioned above the encoder layers indicate that each output token primarily retains information derived from its corresponding input token. (b) The detailed structure of a transformer encoder layer. We choose "Scaled Dot-Product Attention" with queries, where keys and values are from different embeddings of the input vector. (c) 2D representation of the crystal feature vector by t-SNE.

introduces a beneficial bias, which facilitates faster convergence when subsequently trained on smaller, labeled datasets for specific tasks. It has also been shown that pretraining is effective in enhancing model performance, particularly for predicting complex properties.[38,39] To our knowledge, XBERT presents a novel model structure by being the first to integrate graph neural network (GNN) outputs with global properties using a transformer architecture. We treat space group information and other global properties as tokens, and distinctively incorporate the output of the crystal graph convolutional neural network (CGCNN)[31] or the atomistic line graph neural network (ALIGNN)[32] as an additional input token. This method allows for a more comprehensive and detailed encoding of materials, addressing the key challenges of input encoding, output versatility, data utilization, and interpretability in ML models for materials science.

We deliberately select a diverse set of properties to predict, including topological properties, superconducting transition temperatures, dielectric constants, and more. Our model's performance is better than or comparable to the underlying GNN results across all these tasks. Remarkably, we achieve around 91% accuracy rate for topological property prediction, outperforming all previous results,[24,40–46] and even identifying some previously misclassified topological materials. We further provide insightful physical interpretations regarding the features that most significantly influence the target properties. These findings pave the way for developing more powerful ML models to predict material properties and are of significant physical importance.

2. XBERT Model. To construct the input to our XBERT model, each crystalline material is described by eight tokens. The first five tokens, inspired by Ref. [26], encode space group information, including space group symbol, point group, crystal system, Laue class, and symmetry. For example, if the space group of a material is 225, then the five tokens are $[F\frac{4}{m}\bar{3}\frac{2}{m}]$, $[m\bar{3}m]$, [cubic], $[m\bar{3}m]$, and [centrosymmetric]. As is commonly done in natural language processing, we counted all the "words" that appeared in the tokens derived from the 230 space groups and removed the duplicates to get a dictionary of length 280.

Then each "word" was assigned a number (from 0 to 279), which can be easily embedded into a 256-dimensional vector. The sixth token represents global properties; in this work, we include only the total electron number in a unit cell (N_{e}), though it is straightforward to add additional properties such as the volume of the unit cell and the number of atoms in a unit cell. N_{e} is represented by the digits in its binary representation. The largest N_{e} in our dataset is 11448, which can be represented by a 14-dimensional vector $[1,0,1,1,0,0,1,0,1,1,1,0,0,0]$. Thus, for each material, N_{e} is described by a 14-dimensional vector. Leading zeros are added to the vector if its binary representation is less than 14 bits. The seventh token encodes the element information directly in the form of $\sum_\mu f_\mu \boldsymbol{v}_\mu$, where f_μ and $\boldsymbol{v}_\mu$ represent the proportion of element μ in a material and the element vector of element μ taken from Ref. [31], respectively. The eighth "token" represents the output of a GNN model. We primarily consider using the CGCNN model.[31] Initially, each atom in the unit cell is assigned a vector $\boldsymbol{v}_i^0$, which is solely determined by the element type. The edge between every two atoms is described by a vector $\boldsymbol{u}_{ij}$, encoding the distance between the two atoms. Subsequently, through the message passing mechanism, information from the surrounding atoms will be transmitted to the target atom, resulting in the updating of the atom's vector,

$$\boldsymbol{v}_i^{t+1} = \boldsymbol{v}_i^t + \sum_{j,k} \sigma\left[\boldsymbol{z}_{(i,j)_k}^t \boldsymbol{W}_f^t + \boldsymbol{b}_f^t\right] \odot g\left[\boldsymbol{z}_{(i,j)_k}^t \boldsymbol{W}_s^t + \boldsymbol{b}_s^t\right], \tag{1}$$

$$\boldsymbol{z}_{(i,j)k}^t = \boldsymbol{v}_i^t \oplus \boldsymbol{v}_j^t \oplus \boldsymbol{u}_{(i,j)_k}, \tag{2}$$

where $\odot$ and $\oplus$ denote element-wise multiplication and concatenation of feature vectors, respectively. The functions σ and g correspond to the sigmoid and softplus activation functions, respectively. After L rounds of message passing, the vector on each atom is updated to be $\boldsymbol{v}_i^L$. We then take the average of the vectors of all atoms within the unit cell to obtain a representation of the entire unit cell,

$$\boldsymbol{v} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{v}_i^L. \tag{3}$$

The parameters in CGCNN are also updated during training. Every token is embedded into a vector of length 256 before being fed into the transformer encoder. This method effectively combines the three aforementioned encoding approaches in a concrete and flexible manner. To demonstrate the flexibility of XBERT and to achieve superior performance, we train a model in which CGCNN is replaced by a more advanced graph neural network, ALIGNN.[32] We anticipate that employing a more powerful GNN model will further enhance the performance of our XBERT model, as discussed later. It is also instructive to compare XBERT with a pure large language model, which encodes all material information through text rather than using tokens generated by a GNN model. As highlighted by Ref. [47], large language models encounter difficulties in accurately extracting unit cell information, particularly the spatial arrangement of atoms, potentially due to the limitations of text in precisely describing atomic positions. Although one could directly input a crystallographic information file (CIF) into such a model, its capacity to interpret the file remains insufficient.[48] This further supports the argument that combining a GNN model with the BERT architecture is advantageous, yielding results that surpass the performance of either model in isolation.

The architecture of our XBERT model is based on the encoder side of the transformer,[49] with the core being the self-attention layers as shown in Figs. 1(a) and 1(b). Three different embeddings [queries (Q), keys (K) and values (V)] are obtained from the input tokens. The attention score is then calculated using the formula:

$$\text{Attention}\,(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{4}$$

where d_k is the dimension of keys K. XBERT utilizes 5 attention layers with 4 heads. The output vectors $\{\boldsymbol{x}_i\}$ from the last layer are concatenated into a single long vector, referred to as the crystal feature vector. These crystal feature vectors serve as the input to the subsequent fully connected layer, whose outputs are then used to predict various properties depending on the specific downstream task.

During pretraining, we use approximately 150000 crystalline materials from the Materials Project.[50] The pretraining task involves predicting the lattice parameters of the unit cell, which aids our model in capturing the complex interplay between structural and elemental information. In Fig. 1(c), we apply t-distributed stochastic neighbor embedding (t-SNE) dimension reduction to the crystal feature vectors output by the pretrained model and plot them in a 2D space.[51] Different colors represent different crystal systems, and materials belonging to the same crystal system cluster together, demonstrating that the pretraining task has been effectively accomplished. We anticipate that utilizing a larger open-source dataset for pretraining, such as MatSci ML,[52] would further enhance XBERT's robustness, enabling it to better manage noise and improving its adaptability across various downstream tasks. For fine-tuning, we utilize the weights from the pretrained model as the initial weights. Detailed training methods can be found in the Supplementary Materials (SM).

3. Performance and Analysis. To evaluate XBERT's performance and demonstrate its generality, we selected ten different material properties that encompass various aspects of crystalline materials as prediction targets, as detailed in the SM. We assessed the model's performance by removing different input features to understand their influence on specific properties. Additionally, we explored various model structures, such as replacing the transformer with a simple fully connected layer and examining whether normalizing each output token vector using a softmax layer affects the results. The final outcomes are presented in Table 1.

Table 1. Summary of the prediction performance for ten different models on different properties.[53] The values represent accuracy for topological classification and mean absolute error (MAE) for other tasks. The MAE values are obtained from 5-fold cross-validation and rounded to three significant figures. The digits in parentheses represent the standard error. Different models are defined as follows: XBERT_{FC} refers to the full model (with CGCNN); I, II, III, and IV exclude the transformer, macroscopic properties, element information, and CGCNN, respectively; V is XBERT_{FC} without the softmax layer; and VI is II without the softmax layer. XBERT_{FA} refers to the full model (with ALIGNN). TopoA, TopoB, 3DSC, Perovskites, Dielectric, GVRH, JDFT2D, KVRH, Band Gap, Phonons, and E-Form are datasets representing topological materials in Refs. [54–56], in Refs. [57–59], superconducting transition temperature[60], formation energy of perovskites,[61] refractive index,[62] shear modulus,[21,63] exfoliation energy,[64] bulk modulus,[63] band gap,[50] last phonon density of states peak,[65] and formation energy,[50] respectively. The best results achieved in XBERT_{FC}-based models are shown in bold.

Model	TopoA ↑	TopoB ↑	3DSC ↓	Perovskites ↓	Dielectric ↓	GVRH ↓
CGCNN	85.49%	86.28%	3.17(46)	0.048(3)	0.541(66)	0.086(2)
XBERT_{FC}	**89.76%**	90.91%	**2.69(25)**	0.157(32)	**0.378(66)**	0.094(4)
I	86.76%	89.10%	3.06(33)	0.291(11)	0.400(41)	0.090(1)
II	88.14%	87.85%	2.74(36)	0.143(14)	0.388(24)	0.091(2)
III	89.74%	90.86%	2.76(35)	0.169(36)	0.397(51)	0.093(3)
IV	88.83%	89.84%	2.90(34)	0.457(12)	0.435(103)	0.121(4)
V	89.56%	**91.19%**	2.85(49)	0.079(3)	0.488(57)	0.088(2)
VI	87.52%	88.17%	2.74(31)	**0.076(3)**	0.455(54)	**0.085(3)**
ALIGNN	87.91%	88.36%	2.67(42)	0.029(1)	0.385(32)	0.072(1)
XBERT_{FA}	89.90%	91.33%	2.59(44)	0.033(1)	0.335(49)	0.071(1)

Model	JDFT2D ↓	KVRH ↓	Band gap ↓	Phonons ↓	E-Form ↓
CGCNN	46.7(6.1)	0.070(4)	0.263(8)	56.1(4.8)	0.042(3)
XBERT_{FC}	51.7(5.1)	0.078(7)	0.244(8)	93.0(8.6)	0.047(2)
I	64.0(3.7)	0.073(3)	0.284(5)	157.0(7.8)	**0.044(3)**
II	52.3(7.1)	0.077(1)	0.273(10)	99.2(15.0)	0.049(4)
III	53.7(3.6)	0.079(4)	**0.235(6)**	109.8(20.0)	0.048(3)
IV	54.8(4.2)	0.096(6)	0.321(6)	98.6(10.1)	0.105(4)
V	49.1(3.3)	0.071(3)	0.262(21)	78.8(10.0)	0.045(4)
VI	**48.2(8.4)**	**0.069(3)**	0.258(10)	**61.7(3.9)**	0.046(4)
ALIGNN	42.9(9.9)	0.058(1)	0.191(4)	40.5(7.4)	0.021(1)
XBERT_{FA}	40.5(10.9)	0.059(3)	0.211(11)	51.1(10.1)	0.034(2)

The XBERT-based model outperforms CGCNN in seven tasks, and the model with all input features outperforms CGCNN in five tasks. As shown in Table S4 of SM, the token corresponding to CGCNN has the highest sum of absolute values before normalization. Thus, removing the normalization layer amplifies the impact of CGCNN and improves model performance for tasks that heavily depend on unit cell information. The inferior performance of the full model in some tasks can be attributed to less important features detracting from the overall performance of our model, despite its ability to prioritize relevant features. Therefore, better performance is expected with fewer input features. For predicting bulk modulus, removing the electron number feature yields better results, which is intuitive since the electron number is irrelevant to modulus. Similarly, removing the elemental feature achieves the best performance for the electron gap task. For tasks related to phonons, the properties are primarily dependent on unit cell information, resulting in our model's performance being slightly inferior to CGCNN.

Notably, in predicting topological properties and superconducting properties, XBERT shows clear advantages over CGCNN. The substantial decrease in performance upon removing either space group, elemental, or unit cell information indicates that these properties result from the intricate interplay between structural and elemental information. To further enhance performance, we also trained an XBERT model by replacing CGCNN with ALIGNN, expecting improved outcomes. As anticipated, the performance significantly exceeded that achieved with CGCNN across all tasks, and our model outperformed ALIGNN on six tasks. These findings suggest that the XBERT-based model would likely benefit further from more advanced GNN architectures, and it is expected to maintain its advantages in predicting complex properties over GNN-based models.

To better interpret the model, we first simplify the encoder layer to a single layer with one attention head. This simplification is intended to prevent excessive mixing of information between tokens. Since attention scores undergo softmax normalization and residual connections are applied, the output tokens are predominantly influenced by their corresponding input tokens. Consequently, each output token can be regarded as primarily encapsulating the feature of its respective input token. Then, we can investigate its workings through two steps. First, we aim to understand what information the output token encodes. We simplify the fine-tuning layer by mapping each token vector $\boldsymbol{x}_i$ to a single value s_i and consider the final output as $y = \sum_i s_i$. Taking the topological property task

as an example, we focus on the value $s_{\rm sg}$ corresponding to the space group token and aim to understand its physical meaning. By averaging this value over materials in the same space group, we obtain the mean feature value $\overline{s}_{\rm sg}$ for each space group. As illustrated in Fig. 2(a), this value actually reflects the proportion of topological materials within each space group. In Ref. [24], it was found that this proportion is a better encoding of space group information and improves performance. Importantly, this feature is automatically learned by our model rather than being manually instructed by human intelligence. Furthermore, we found that higher group numbers usually have a higher value of topological contribution, which may be due to the fact that more symmetry provides a greater ability to protect certain kinds of topology. To visualize this relationship, we present in Fig. 3 a color-coded representation of the $\overline{s}_{\rm sg}$ values, where darker colors represent higher $\overline{s}_{\rm sg}$ values. Space groups that appear fewer than 20 times in the materials were excluded from the analysis. From the figure, we observe a general trend where, within the same crystal system, the color tends to darken as the space group number increases.

Next, we aim to determine the influence of each feature on specific physical properties. Our model has only one linear layer connecting the output vector of the transformer encoder to the final output, represented by $y = \sum_{ij} w_{ij}x_{ij} + b_{ij}$. Here, i represents different tokens, j labels different components in the token vector, and x_{ij} is the crystal feature vector. The value $\sum_j |w_{ij}|$ reflects the influence that feature i has on the property, as shown in Figs. 2(b)–2(d). For the formation energy of perovskites, where unit cell information is commonly believed to be more important, the highest weight is on the last token. Remarkably, for topological properties and superconducting transition temperature, the weights are more concentrated on the space group and element tokens.

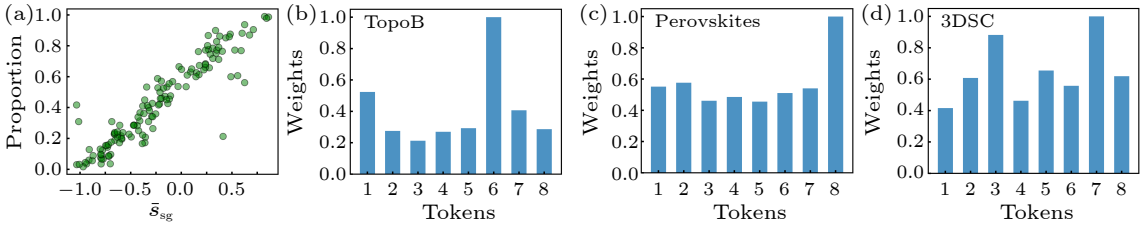

Fig. 2. Interpretation of the model. (a) The relationship between the mean space group feature value $\overline{s}_{\rm sg}$ and the proportion of topological materials within each space group. Each point represents a specific space group, exhibiting a clear linear correlation. This suggests that the output token reflects the frequency of topological materials within that space group to a significant extent.(b)–(d) Bar charts of weights $\sum_j |w_{ij}|$ assigned to different features for predicting topological property, formation energy, and superconducting transition temperature. It is important to note that tokens 1–5 capture space group information, tokens 6 and 7 capture elemental information, and token 8 captures unit cell information.

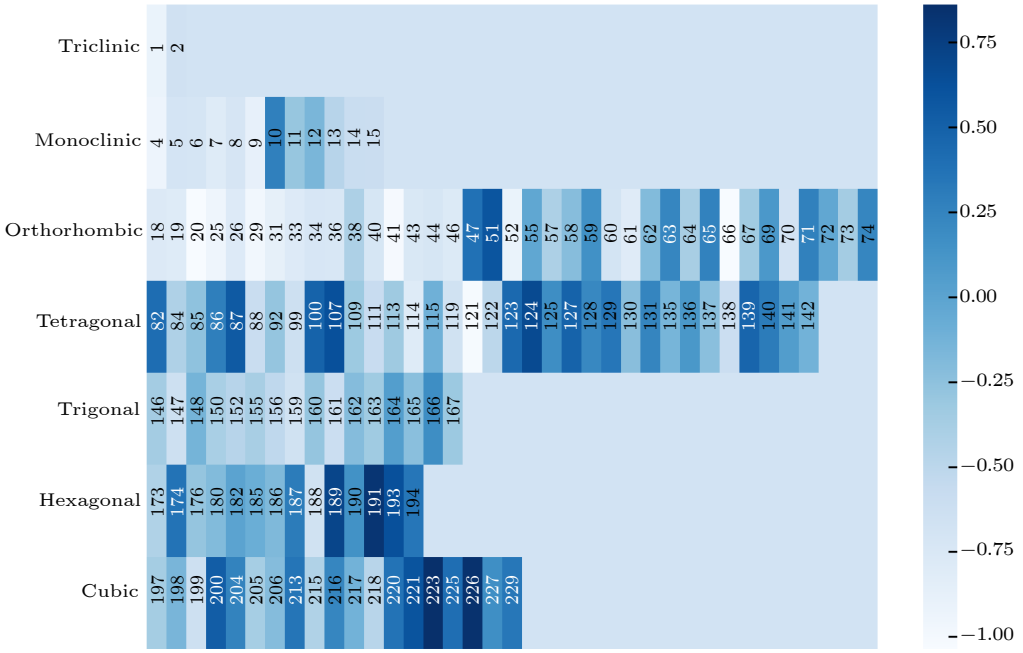

Fig. 3. The visualization diagram of $\overline{s}_{\rm sg}$. The numbers displayed represent space group numbers. It should be noted that certain space groups are not present due to the absence of corresponding materials in the TopoA database. Within a given crystal system, space groups with higher group numbers generally exhibit a greater number of symmetry operations and consequently possess higher overall symmetry.

This indicates that these features are more significant and emphasizes the intricate nature of these properties, which cannot be fully captured by CGCNN. Interestingly,the most important features identified through ablation studies align well with those highlighted by $\sum_j |w_{ij}|$. For instance, in the TopoB task, removing macroscopic properties leads to the greatest performance drop in the full model, which is consistent with Fig. 2(b), where macroscopic properties have the highest weight. Similarly, in the Perovskites task, removing CGCNN features degrades performance the most, aligning with the fact that CGCNN has the highest weight in Fig. 2(c).

4. Advantages. Since we have already demonstrated the generality and interpretability of XBERT, we now address two additional advantages of our model: efficiency and accuracy. First, the efficiency of XBERT is a notable advantage, which is commonly observed in pretrained models.[37] We compare the train loss reduction for different downstream tasks between XBERT and CGCNN. Across all tasks, XBERT demonstrates a faster reduction in loss, as shown in Fig. 4, highlighting its superior efficiency.

Second, XBERT achieves exceptional accuracy, particularly in predicting topological properties, with 89% accuracy for TopoA[54–56] and 91% for TopoB.[57–59] Additionally, Table 2 shows the precision, recall, and F1 score for the TopoA and TopoB tasks, with their definitions explained in the SM. We can easily find that XBERT performs better than CGCNN in all metrics. For the TopoB task, we used Model V, which performed best in the binary classification, to conduct a three-class classification of topological insulators, topological semimetals, and trivial materials. We compared this with the results of directly using CGCNN for the three-class classification. As shown in Table 3, XBERT achieved higher accuracy and F1 scores for each category compared to CGCNN. However, as indicated by the F1 scores, our model's ability to distinguish topological insulators was lower than its performance for topological semimetals and trivial materials, which is consistent with the findings in Ref. [24]. It is worth emphasizing that Ref. [46] tested various other graph neural networks, including GATGNN, DeeperGATGNN, and MEGNet, as well as their proposed models, for this three-class classification task. None of these models achieved an accuracy exceeding 80%. This benchmark indicates that combining additional information with the graph neural network through the transformer encoder layers can significantly improve the accuracy of the three-class classification.

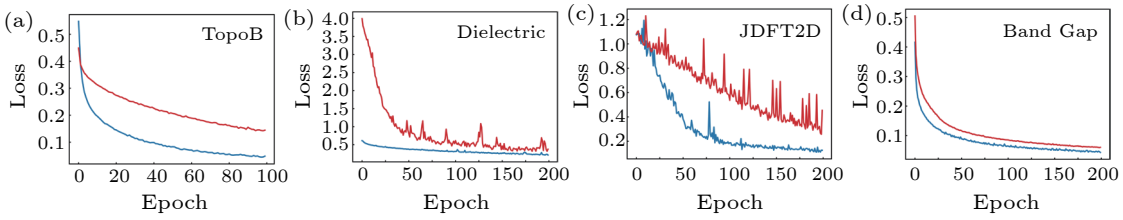

Fig. 4. Efficiency of XBERT. We selected four tasks: (a) topological property, (b) dielectric constant, (c) exfoliation energy, and (d) electronic band gap, and compared the reduction in training loss over time. The red and blue lines correspond to CGCNN and XBERT, respectively. XBERT's training loss converges in fewer epochs compared to CGCNN.

Table 2. Accuracy, precision, recall, and F1 score for the TopoA and TopoB tasks. The following results are the average values from cross-validation. The best results are shown in bold.

	TopoA			TopoB		
Model	Precision	Recall	F1 score	Precision	Recall	F1 score
CGCNN	85.03%	85.20%	84.87%	86.91%	87.08%	86.80%
XBERT_{FC}	**90.42%**	88.11%	89.10%	91.03%	**91.72%**	91.25%
I	86.69%	86.09%	86.18%	88.93%	90.54%	89.59%
II	88.15%	87.38%	87.60%	88.20%	88.84%	88.35%
III	90.21%	**88.48%**	**89.18%**	91.10%	91.52%	91.20%
IV	89.34%	87.42%	88.18%	89.93%	90.89%	90.27%
V	90.20%	87.97%	88.89%	**91.83%**	91.35%	**91.46%**
VI	87.27%	87.11%	87.01%	88.61%	88.27%	88.27%

Table 3. Performance of CGCNN and model V as measured by the accuracy and F1 score. The F1 score for three-class classification is calculated using the harmonic mean of precision and recall for each class.

Model	Accuracy	F1 Triv.	F1 TI	F1 TSM
CGCNN	74.03%	85.08%	42.60%	71.34%
V	85.71%	90.60%	63.51%	88.06%

This high level of accuracy allows us to identify previously undiscovered or misclassified topological materials. Since the symmetry indicator already serves as a quick and reliable check for topological materials,[66,67] our goal is to identify non-symmetry-diagnosable and previously undiscovered topological materials. Here we focus on non-magnetic topological insulators. To achieve this, we conducted screening on two different datasets. Firstly,

we employ XBERT to scan through 1433 materials that are undiagnosable by symmetry indicators, as outlined in Ref. [41], and identify 206 materials labeled as topological. We subsequently exclude those that are already known to be topological and those that are metals. As a result, we find that Ag_2HgSe_4Sn is a promising topological insulator candidate, which has a large band gap of about 16.2 meV. Figure 5(a) illustrates the crystal structure. Figures 5(b) and 5(c) reveal that spin–orbit coupling opens a gap at the Γ point, leading to nontrivial topology, which is confirmed by the Dirac surface state shown in Fig. 5(d). Interestingly, this material has already been experimentally fabricated [68] and was misclassified by the ML method in Refs. [41,45] since its topologicality is around −7, which should be regarded as quite trivial. Additionally, we observed that in the TopoB dataset, topological materials lacking inversion symmetry may not be identifiable by symmetry indicators, potentially leading to their misclassification as trivial. To address this, we conducted a screening of materials classified as trivial but lacking inversion symmetry, in search of potential topological insulators. This approach resulted in the identification of 10 additional topological insulators, which are listed in Table 4. Their band structures and Wilson loops, along with further training details, are provided in the SM.

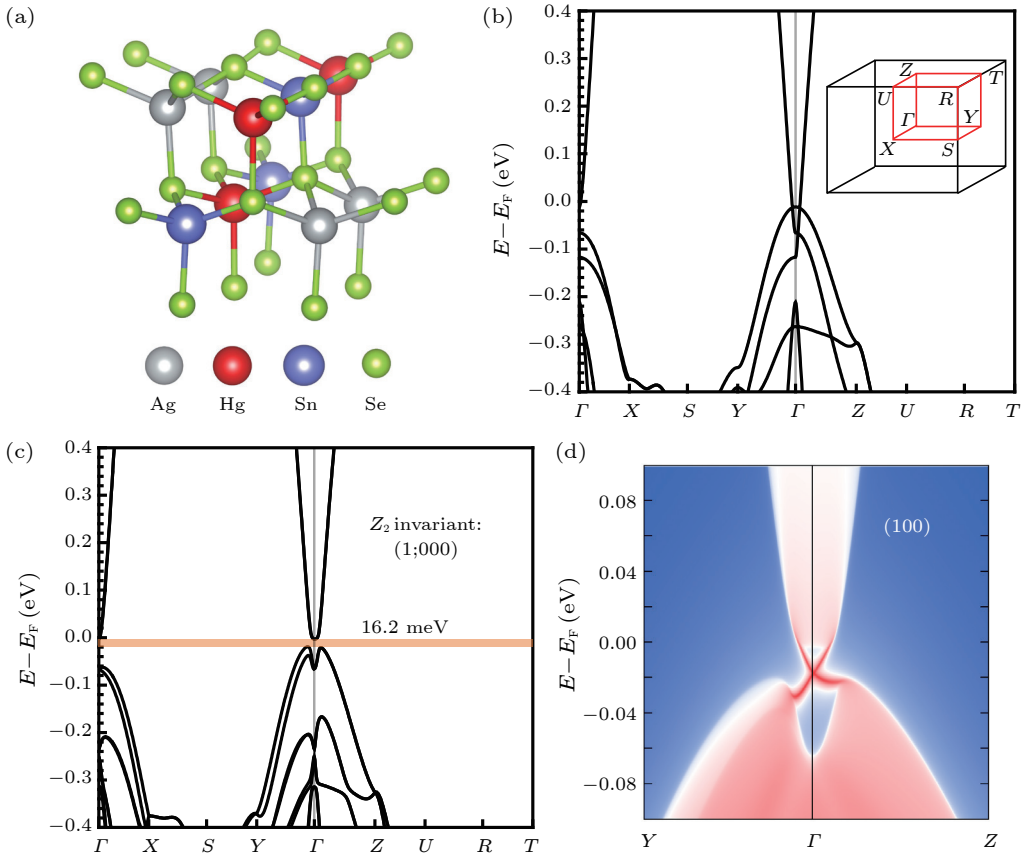

Fig. 5. Electronic structure and topological property of Ag_2HgSe_4Sn. (a) Crystal structure with space group $Pmn2_1$ (No. 31). (b) and (c) Band structure without and with SOC; the Z_2 invariant is (1; 000). [69] (d) A single Dirac cone surface state.

Table 4. The ten newly discovered strong topological insulators. These materials were not identified as topological in the database established in Refs. [57–59], but were discovered by our XBERT.

ICSD	Formula	Space group	Gap (meV)
619784	$Cd_1Cu_2Sn_1Se_4$	121	1.7
629097	$Zn_1Cu_2Sn_1Se_4$	121	7.4
93409	$Zn_1Cu_2Ge_1Se_4$	121	10.6
629293	$Zn_1Cu_2Sn_1Te_4$	121	2.8
656153	$Zn_1Cu_2Ge_1Te_4$	121	no full gap
656158	$Hg_1Cu_2Sn_1Te_4$	121	no full gap
250247	$In_2Cu_2Se_4$	1	5.8
674921	Ru_4H_{24}	6	23.6
160882	Cu_2SnTe_3	44	no full gap
161837	Cd_2O_2	186	5.4

5. Discussion. A key material feature absent from our model is the atomic forces and stresses, which have been shown to significantly enhance the prediction of formation energy.[70] This omission likely accounts for XBERT's suboptimal performance in this task. Future research could focus on incorporating this information into the model and evaluating its impact on other complex physical properties. Another crucial yet missing aspect is the magnetic properties of materials. While magnetic properties intuitively have a considerable influence on many physical properties, the challenge lies in the complexity of magnetic structures and the lack of relevant data.[71,72] Additionally, if we identify the same material with different magnetic orderings as different samples, we should reconsider pretraining tasks to better direct the model toward learning magnetic properties. An alternative potential approach involves treating the prediction of magnetic order as an auxiliary task, rather than incorporating it during the pretraining stage, a strategy that has been shown to effectively enhance model performance.[73]

Predicting vector properties, such as the density of states and the band structure, rather than just a single scalar property, presents a more interesting and challenging task and has been the focus of recent studies.[74–78] In the SM, we discuss combining XBERT with the recently developed Xtal2DOS model[74] to make predictions on the density of states. We believe that the band structure can be predicted in a similar way, albeit with greater difficulty, since the band structure is much more elaborate. Such a model would be highly valuable for the materials science and condensed matter physics community, given the computational intensity of calculating band structures.[79]

We have focused on the BERT structure since it is much more suitable for the pretraining–finetuning paradigm and has far fewer parameters than other common transformer-based architectures. We anticipate that incorporating our model with a further decoder layer would result in a competitive generative model for generating novel materials with specific properties. We leave this to future work.

Acknowledgements. We thank Xiao-Liang Qi for enlightening discussions. This work was supported by the Natural Science Foundation of China (Grant Nos. 12350404 and 12174066), the Innovation Program for Quantum Science and Technology (Grant No. 2021ZD0302600), the Science and Technology Commission of Shanghai Municipality (Grant Nos. 23JC1400600, 24LZ1400100, and 2019SHZDZX01), and was sponsored by "Shuguang Program" supported by Shanghai Education Development Foundation and Shanghai Municipal Education Commission.

Data availability. The datasets are available at https://doi.org/10.5281/zenodo.14917736.

References

[1] Carleo G, Cirac I, Cranmer K *et al.* 2019 *Rev. Mod. Phys.* **91** 045002
[2] Snyder J C, Rupp M, Hansen K, Müller K R, and Burke K 2012 *Phys. Rev. Lett.* **108** 253002
[3] Venderley J, Khemani V, and Kim E A 2018 *Phys. Rev. Lett.* **120** 257204
[4] Zhang Y and Kim E A 2017 *Phys. Rev. Lett.* **118** 216401
[5] Agrawal A and Choudhary A 2016 *APL Mater.* **4** 053208
[6] Choudhary K, DeCost B, Chen C *et al.* 2022 *npj Comput. Mater.* **8** 59
[7] Vasudevan R K, Choudhary K, Mehta A *et al.* 2019 *MRS Communications* **9** 821
[8] Schmidt J, Marques M R G, Botti S, and Marques M A L 2019 *npj Comput. Mater.* **5** 83
[9] Schütt K T, Sauceda H E, Kindermans P J, Tkatchenko A, and Müller K R 2018 *J. Chem. Phys.* **148** 241722
[10] Deringer V L, Pickard C J, and Csányi G 2018 *Phys. Rev. Lett.* **120** 156001
[11] Grisafi A, Wilkins D M, Csányi G, and Ceriotti M 2018 *Phys. Rev. Lett.* **120** 036002
[12] Magar R, Wang Y, Lorsung C *et al.* 2022 *Mach. Learn.: Sci. Technol.* **3** 045015
[13] Magar R, Wang Y, and Barati Farimani A 2022 *npj Comput. Mater.* **8** 231
[14] Cao Z, Magar R, Wang Y, and Barati Farimani A 2023 *J. Am. Chem. Soc.* **145** 2958
[15] Jinnouchi R, Lahnsteiner J, Karsai F, Kresse G, and Bokdam M 2019 *Phys. Rev. Lett.* **122** 225701
[16] Chen C, Ye W, Zuo Y, Zheng C, and Ong S P 2019 *Chem. Mater.* **31** 3564
[17] Ock J, Tian T, Kitchin J, and Ulissi Z 2023 *J. Chem. Phys.* **158** 214702
[18] Kazeev N, Al-Maeeni A R, Romanov I *et al.* 2023 *npj Comput. Mater.* **9** 113
[19] Yan K, Liu Y, Lin Y, and Ji S 2022 *Advances in Neural Information Processing Systems* **35** 15066
[20] Choudhary K, DeCost B, and Tavazza F 2018 *Phys. Rev. Mater.* **2** 083801
[21] Ward L, Dunn A, Faghaninia A *et al.* 2018 *Comput. Mater. Sci.* **152** 60
[22] Kostiuchenko T S, Shapeev A V, and Novikov I S 2024 *Chin. Phys. Lett.* **41** 066101
[23] Zhu C, Shen X, Zhu G, and Li B 2023 *Chin. Phys. Lett.* **40** 124402
[24] Claussen N, Bernevig B A, and Regnault N 2020 *Phys. Rev. B* **101** 245117
[25] Tao Q, Xu P, Li M, and Lu W 2021 *npj Comput. Mater.* **7** 23
[26] Huang H, Magar R, Xu C, and Farimani A B 2023 arXiv:2308.16259
[27] Choudhary K 2024 *J. Chem. Phys.* **15** 6909
[28] Gupta T, Zaki M, Krishnan N M A, and Mausam 2022 *npj Comput. Mater.* **8** 102
[29] Rubungo A N, Arnold C, Rand B P, and Dieng A B 2023 arXiv:2310.14029
[30] Olivetti E A, Cole J M, Kim E *et al.* 2020 *Appl. Phys. Rev.* **7** 041317
[31] Xie T and Grossman J C 2018 *Phys. Rev. Lett.* **120** 145301
[32] Choudhary K and DeCost B 2021 *npj Comput. Mater.* **7** 185
[33] Karamad M, Magar R, Shi Y, Siahrostami S, Gates I D, and Farimani A B 2020 *Phys. Rev. Mater.* **4** 093801
[34] Gong S, Yan K, Xie T *et al.* 2023 *Sci. Adv.* **9** eadi3245
[35] Gong S, Wang S, Zhu T, Shao-Horn Y, and Grossman J C 2023 arXiv:2309.04478)
[36] Huang H and Barati Farimani A 2024 *J. Appl. Phys.* **135** 165104
[37] Devlin J, Chang M W, Lee K, and Toutanova K 2019 *Proceedings of the 209 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* **1** 4171
[38] Zaidi S, Schaarschmidt M, Martens J *et al.* 2022 arXiv:2206.00133

[39] Lee K L K, Gonzales C, Spellings M *et al.* 2023 *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis* p. 51
[40] Schleder G R, Focassio B, and Fazzio A 2021 *Appl. Phys. Rev.* **8** 031409
[41] Ma A, Zhang Y, Christensen T *et al.* 2023 *Nano Lett.* **23** 772
[42] Andrejevic N, Andrejevic J, Bernevig B A *et al.* 2022 *Adv. Mater.* **34** 2204113
[43] Liu J, Cao G, Zhou Z, and Liu H 2021 *J. Phys. Condens. Matter.* **33** 325501
[44] Cao G, Ouyang R, Ghiringhelli L M *et al.* 2020 *Phys. Rev. Mater.* **4** 034204
[45] Xu H, Jiang Y, Wang H, and Wang J 2024 *Phys. Rev. B* **109** 035122
[46] Rasul A, Hossain M S, Dastider A G *et al.* 2024 *Sci. Rep.* **14** 31564
[47] Alampara N, Miret S, and Jablonka K M 2024 arXiv:2406.17295)
[48] Miret S and Krishnan N M A 2024 arXiv:2402.05200
[49] Vaswani A, Shazeer N, Parmar N *et al.* 2017 *Proceedings of the 31st International Conference on Neural Information Processing Systems* p. 6000
[50] Jain A, Ong S P, Hautier G *et al* 2013 *APL Mater.* **1** 011002
[51] van der Maaten L and Hinton G 2008 *J. Mach. Learn. Res.* **9** 2579
[52] Lee K L K, Gonzales C, Nassar M *et al.* 2023 arXiv:2309.05934
[53] Dunn A, Wang Q, Ganose A, Dopp D, and Jain A 2020 *npj Comput. Mater.* **6** 138
[54] Po H C, Vishwanath A, and Watanabe H 2017 *Nature Commun.* **8** 50
[55] Tang F, Po H C, Vishwanath A, and Wan X 2019 *Nature Phys.* **15** 470
[56] Tang F, Po H C, Vishwanath A, and Wan X 2019 *Nature* **566** 486
[57] Bradlyn B, Elcoro L, Cano J *et al.* 2017 *Nature* **547** 298
[58] Vergniory M G, Elcoro L, Felser C *et al.* 2019 *Nature* **566** 480
[59] Vergniory M G, Wieder B J, Elcoro L *et al.* 2022 *Science* **376** eabg9094
[60] Sommer T, Willa R, Schmalian J, and Friederich P 2023 *Sci. Data* **10** 816
[61] Castelli I E, Olsen T, Datta S *et al.* 2012 *Energy Environ. Sci.* **5** 5814
[62] Petousis I, Mrdjenovich D, Ballouz E *et al.* 2017 *Sci. Data* **4** 160134
[63] de Jong M, Chen W, Angsten T *et al.* 2015 *Sci. Data* **2** 150009
[64] Choudhary K, Kalish I, Beams R, and Tavazza F 2017 *Sci. Rep.* **7** 5179
[65] Petretto G, Dwaraknath S, P C Miranda H *et al.* 2018 *Sci. Data* **5** 180065
[66] Krutho J, de Boer J, van Wezel J, Kane C L, and Slager R J 2017 *Phys. Rev. X* **7** 041069
[67] Zhang T, Jiang Y, Song Z *et al.* 2019 *Nature* **566** 475
[68] Parasyuk O, Gulay L, Piskach L, and Kumanska Y O 2002 *J. Alloys Compd.* **339** 140
[69] Fu L, Kane C L, and Mele E J 2007 *Phys. Rev. Lett.* **98** 106803
[70] Riebesell J, E A Goodall R, Benner P *et al.* 2023 arXiv:2308.14920
[71] Elcoro L, Wieder B J, Song Z *et al.* 2021 *Nature Commun.* **12** 5965
[72] Choudhary K, Garrity K F, Jiang J, Pachter R, and Tavazza F 2020 *npj Comput. Mater.* **6** 49
[73] Liao Y L, Smidt T, Shuaibi M, and Das A 2024 arXiv:2403.09549
[74] Bai J, Du Y, Wang Y *et al.* 2023 arXiv:2302.01486
[75] Kong S, Ricci F, Guevarra D *et al.* 2022 *Nature Commun.* **13** 949
[76] Li H, Wang Z, Zou N *et al.* 2022 *Nature Comput. Sci.* **2** 367
[77] Li H, Tang Z, Fu J *et al.* 2024 *Phys. Rev. Lett.* **132** 096401
[78] Chen Z, Andrejevic N, Smidt T *et al.* 2021 *Adv. Sci.* **8** 2004214
[79] Giustino F 2017 *Rev. Mod. Phys.* **89** 015003