Boosting Superconductor Discovery: CrystalGAT Achieves High-Accuracy Critical Temperature Prediction Through Local-Global Crystal Graph Learning

Yixiang Zeng [1, a], Ruijie Liu [2], Yongquan Jiang [3], Zigang Deng [4]

[1] The Party School of the CPC Zigong Municipal Committee, ZiGong, 643010, China

[2] SWJTU-LEEDS Joint School, Southwest Jiaotong University, Chengdu, 611756, China

[3] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, 611756, China

[4] State Key Laboratory of Rail Transit Vehicle System, Southwest Jiaotong University, Cheng du, 610031, China

[a] yz259@foxmail.com

Abstract. This study aims to predict the critical temperature (Tc) of superconducting materials and determine their superconducting state using our novel graph neural network model, CrystalGAT. This model skillfully integrates graph attention networks and convolutional neural layers to effectively capture local interactions between atoms and global material properties. Our research is based on the 3DSC_MP dataset, which contains three-dimensional crystal structures of superconductors and their corresponding Tc data. The CrystalGAT model processes atomic features, bond features, angular features, and chemical composition features (MAGPIE features) to learn complex interactions between atoms and bonds, as well as global properties of the entire chemical composition. Experimental results demonstrate that CrystalGAT achieves significant improvements in predictive performance compared to traditional models, exhibiting lower error metrics and higher R^2 values. This strongly highlights the immense potential of integrating detailed atomic-level and structural characteristics into superconductor prediction models. Our research provides new insights and tools for the discovery and design of high-temperature superconducting materials.

Keywords: Graph Neural Network; Critical Temperature Prediction; High-temperature Superconducting Materials.

1. Introduction

In recent years, machine learning (ML) techniques have emerged as powerful tools for materials discovery and property prediction [1]. Various ML approaches, including linear regression, random forests, and neural networks, have been applied to the problem of Tc prediction [2,3]. However, these methods often face limitations in capturing the complex relationships between atomic structure, electronic properties, and superconductivity. Many existing models rely on pre-engineered features or simplified representations of materials, potentially missing crucial structural information that determines superconducting behavior [4]. Furthermore, the inherent complexity of superconducting mechanisms, especially in HTS, poses significant challenges for conventional ML approaches [5]. The interplay between crystal structure, electronic correlations, and quantum effects in superconductors necessitates more sophisticated modeling techniques that can handle multi-scale interactions and learn from raw structural data [6]. To address these limitations, our research aims to develop a novel machine learning framework that can accurately predict the Tc of superconducting materials while incorporating detailed structural and compositional information. We propose CrystalGAT, a graph neural network model that combines the power of graph attention mechanisms [7] with convolutional neural networks to process and learn from three-dimensional crystal structures directly. Our approach leverages the 3DSC_MP dataset [8], which provides a comprehensive collection of superconductor crystal structures and their corresponding Tc values. By integrating atomic features, bond characteristics, angular

information, and global compositional attributes, CrystalGAT aims to capture the multi-scale interactions that govern superconductivity.

2. Data and Methods

2.1 Dataset

2.1.1 Introduction to the 3DSC_MP Dataset

This study utilizes the 3DSC_MP [8] dataset as the primary data source. This dataset represents the first comprehensive compilation that integrates both the critical temperature (Tc) and the three-dimensional crystal structures of superconductors, offering an invaluable resource for advanced research into superconducting materials. The 3DSC_MP dataset encompasses the chemical formulas of various superconductors, their critical temperatures measured in Kelvin, and the corresponding .cif file paths, which provide access to their 3D crystal structures. Additionally, it incorporates supplementary data derived from the original Materials Project database, including crucial electronic properties such as bandgap and Fermi energy.

The 3DSC dataset is publicly accessible via Figshare and can be extracted using specialized commands, streamlining data processing and analysis for this research. Furthermore, the 3DSC repository on GitHub provides comprehensive code and documentation, facilitating operations on the dataset, including statistical chart generation and the training of machine learning models. The scope and detail of this dataset are unparalleled in superconductor research, establishing a robust foundation for this study.

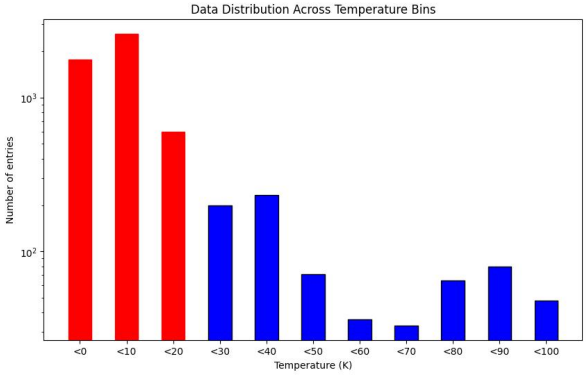
2.1.2 Data Preprocessing and Oversampling Methods

During the exploration of the 3DSC_MP dataset, it was observed that the distribution of superconducting critical temperature (Tc) data was uneven. To address this and ensure effective model learning across the entire Tc range, an oversampling technique was implemented to balance the dataset.

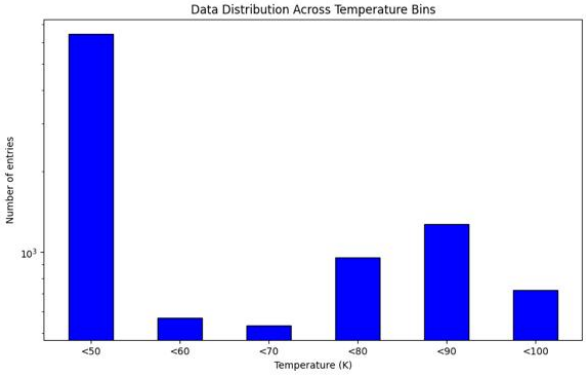
A function named oversample_data was defined to achieve this. This function takes a dataset and a threshold as input parameters. First, it counts the number of high Tc samples exceeding the threshold (set at 40K in this study) and the number of low Tc samples below the threshold. Based on the number of low Tc samples, the function increases the high Tc samples through random selection until their count matches that of the low Tc samples. This ensures the final dataset is balanced in terms of high and low Tc samples.

Implementing this oversampling strategy enhances the model's performance and generalization ability across different Tc ranges. By comparing the lengths of the original and processed datasets, we confirmed that oversampling was effective. For example, passing the training dataset through the oversample_data function (e.g., train_dataset = oversample_data (train_dataset, 40)) ensures that the data used for model training is balanced. This balance is crucial for improving the model's prediction accuracy and robustness.

This method quantitatively assesses samples with Tc above the threshold (40K) against those below it. After this assessment, a random replication strategy increases the number of high Tc samples until their count matches that of the low Tc samples. This results in a balanced dataset representing both high and low Tc instances. The unsampled and sampled data are shown on Fig. 1.



(a) Unsampled data



(b) Sampled data

Fig.1 Unsampled data and sampled data

Overall, this approach significantly enhances the model's performance and ability to generalize across a wide range of Tc values. The effectiveness of the oversampling method is further verified by comparing the dataset sizes before and after processing.

2.2 Methods

2.2.1 Feature extraction

In the process of extracting features from the 3DSC_MP dataset to predict the superconducting critical temperatures of materials, a comprehensive feature engineering approach is adopted. This method integrates atomic and structural characteristics of the crystal structures of superconducting materials and is structured as follows:

(a) **Atomic Initialization:** Atomic types are transformed into feature vectors using a predefined embedding file, such as onehot-embedding.json. This process involves loading a JSON file containing embeddings for various chemical elements and mapping each atomic type within the crystal structure to its corresponding feature vector, known as atomic features.

(b) **Radial Basis Function (RBF) Expansion:** The RBFExpansion class is employed to project interatomic distances into a higher-dimensional feature space, which captures more complex interatomic relationships. This transformation applies a Gaussian function to each distance, effectively converting scalar distances into a set of feature vectors based on predefined radial basis functions. These transformed distances are referred to as bond features.

(c) **Angle Calculation**: The cosine of the angle between atomic vectors is computed to analyze the spatial arrangement of atoms within the crystal structure. This information is essential for understanding bonding geometries and electronic interactions that influence superconducting properties. The resulting data are termed angle features.

(d) **Graph Construction**: A graph-based representation of each crystal structure is generated using libraries such as PyTorch Geometric. In this graph, each node corresponds to an atom,

represented by its feature vector, while the edges represent atomic bonds characterized by distance-based and angle-based features. This graph-based representation is critical for capturing the intricate interactions within crystal structures that influence superconducting behavior.

(e) **MAGPIE Feature Extraction**: In addition to atomic- and bond-level features, the MAGPIE library is utilized to compute global descriptors of the crystal structure. These descriptors include properties such as valence configurations, elemental characteristics, stoichiometry, and ionicity. These features provide a comprehensive overview of the material's chemical composition and its potential impact on superconductivity.

(f) **Data Preprocessing for Efficient Training**: To streamline the machine learning model training process, all extracted features (including atomic features, bond features, angle features, and MAGPIE features) are preprocessed and stored in an optimized format. This preprocessing involves iterating over each dataset entry, constructing the graph representation, computing MAGPIE features, and serializing this information for efficient access during model training.

In summary, the features of the crystal structures are derived from four main perspectives: atomic features, bond features, angle features, and MAGPIE features. This multi-dimensional feature engineering approach facilitates a comprehensive characterization of superconducting materials, thereby enhancing the prediction accuracy of their critical temperatures.

2.2.2 CrystalGAT network

We proposed a model named CrystalGAT, designed to predict the critical temperature of superconducting materials by integrating the functionalities of a Graph Attention Network (GAT) and a one-dimensional Convolutional Neural Network (1DCNN). This model incorporates atomic features, bond features, angular features, and chemical composition features (MAGPIE features) to understand the complex relationship between crystal structure and superconductivity. The overall network result is shown in Fig. 2.

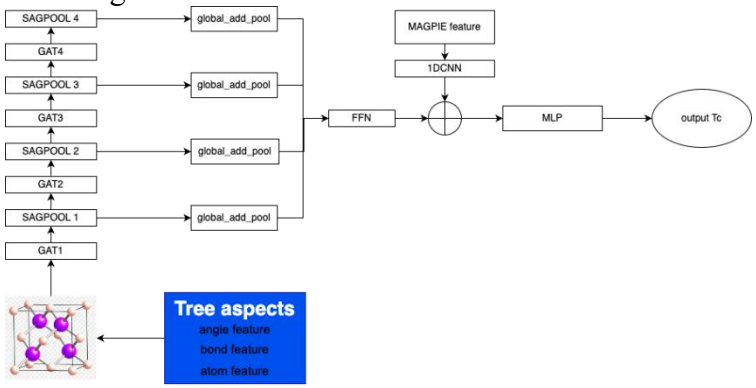


Fig. 2 CrystalGAT structure

- GAT Layer

The GAT layer in CrystalGAT includes multiple sub-layers that handle atomic, positional embedding (PE), edge, and angular embedding features. The output dimensions of each sub-layer increase progressively from the initial 112, 41, and 12 dimensions to 1024 dimensions in the hidden layer. This layer consists of four GAT convolutional layers and four self-attention graph pooling (SAGPooling) layers. Each layer uses eight attention heads to extract and integrate information from neighboring nodes while reducing the number of nodes to enhance the model's focus and computational efficiency. The detailed parameters are shown in Table 1.

Table 1 GAT Layers

Sub-layer	Input Dimension	Output Dimension	Activation	Additional Parameters
Atomic Embedding	112	1024	ReLU	-
- Linear Layer 1	112	256	ReLU	-

- Linear Layer 2	256	512	ReLU	-
- Linear Layer 3	512	1024	-	-
Positional Embedding (PE)	3	1024	ReLU	-
- Linear Layer 1	3	256	ReLU	-
- Linear Layer 2	256	512	ReLU	-
- Linear Layer 3	512	1024	-	-
Edge Embedding	41	1024	ReLU	-
- Linear Layer 1	41	1024	ReLU	-
- Linear Layer 2	1024	1024	-	-
Angular Embedding	12	1024	ReLU	-
- Linear Layer 1	12	1024	ReLU	-
- Linear Layer 2	1024	1024	-	-
GAT Convolution Layer 1	1024	1024	-	8 attention heads
GAT Convolution Layer 2	1024	1024	-	8 attention heads
GAT Convolution Layer 3	1024	1024	-	8 attention heads
GAT Convolution Layer 4	1024	1024	-	8 attention heads
SAGPooling Layer 1	GraphConv	1024	-	Ratio 0.2, multiplier 1.0
SAGPooling Layer 2	GraphConv	1024	-	Ratio 0.2, multiplier 1.0
SAGPooling Layer 3	GraphConv	1024	-	Ratio 0.2, multiplier 1.0
SAGPooling Layer 4	GraphConv	1024	-	Ratio 0.2, multiplier 1.0
Feedforward Neural Network	1024	512	ReLU	Dropout layer, p=0.2
- Linear Layer 1	1024	512	ReLU	Dropout layer, p=0.2
- Linear Layer 2	512	512	ReLU	Dropout layer, p=0.2
- Linear Layer 3	512	512	ReLU	Dropout layer, p=0.2

- Convolutional Layer

The second branch of the model, the ConvLayer, uses three one-dimensional convolutional layers to process MAGPIE features, capturing the global information of the material. Specifically, the first layer is a one-dimensional convolutional layer with an input channel of 1, an output channel of 64, a kernel size of 3, a stride of 1, and padding of 1. The second layer has 64 input channels, 128 output channels, and the same kernel size, stride, and padding. The third layer has 128 input channels, 256 output channels, and the same parameters as the previous layers. This branch then uses two fully connected layers to align the output of the convolutional layer with the output of the GAT layer. The detailed parameters are shown in Table 2.

Table 2 Conv Layers

Sub-layer	Input Dimension	Output Dimension	Activation	Additional Parameters
Conv Layer 1	1	64	-	Kernel size=3, stride=1, padding=1
Conv Layer 2	64	128	-	Kernel size=3, stride=1, padding=1
Conv Layer 3	128	256	-	Kernel size=3, stride=1, padding=1
Fully Connected 1	37120	1024	-	-
Fully Connected 2	1024	512	-	-

- Output Layer

The final output stage of the model integrates the outputs of the GAT layer and the Conv layer and processes them through a Multilayer Perceptron (MLP) to produce the predicted critical temperature value. This MLP consists of three fully connected layers with ReLU activation functions and Dropout layers to prevent overfitting. Specifically, the first fully connected layer reduces the input features from 1024 to 512 dimensions, followed by ReLU activation and a Dropout layer. The second fully connected layer further reduces the features from 512 to 256 dimensions, repeating the activation and Dropout process. The final fully connected layer reduces the features from 256 to 128 dimensions, producing the final prediction. The detailed parameters are shown in Table 3.

Table 3 Output Layers

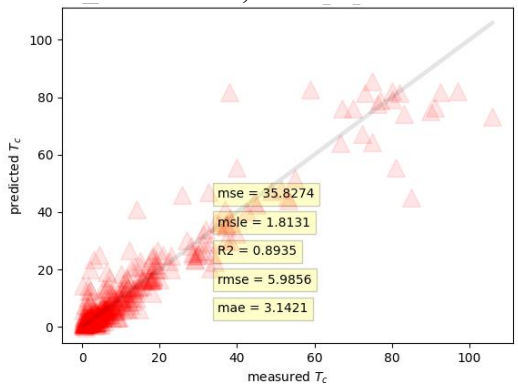
Sub-layer	Input Dimension	Output Dimension	Activation	Additional Parameters
Fully Connected	512	256	ReLU	Dropout layer, p=0.2
Linear Layer 1	512	256	ReLU	Dropout layer, p=0.2
Linear Layer 2	256	128	ReLU	Dropout layer, p=0.2
Linear Layer 3	128	1	-	-

3. Results and discussion

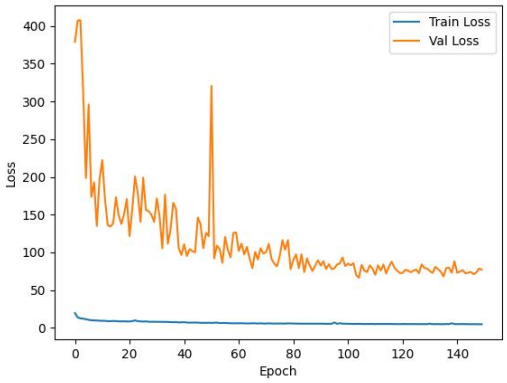
3.1 Results

The loss function dynamics observed during the training and validation phases of the CrystalGAT model, as illustrated in Fig. 3, provide significant insights into the model's performance. During the training phase, the Train Loss shows a consistent downward trend, indicating the model's increasing proficiency in capturing and learning the underlying features and patterns present in the training dataset. This trend suggests that the model is effectively minimizing the error between its predictions and the actual values, thereby improving its accuracy on the training data.

In the performance evaluation on the 3DSC_MP test set, the CrystalGAT model demonstrated various error metrics: Mean Absolute Error (MAE) of 3.1421, Mean Square Error (MSE) of 35.8274, Mean Square Log Error (MSLE) of 1.8131, Root Mean Square Error (RMSE) of 5.9856, and an R-squared (coefficient of determination) of 0.8935. These metrics collectively indicate the model's robust predictive accuracy and its ability to generalize well to unseen data. Compare with other network which also use 3DSC_MP dataset, as shown in Table 4.



(a) Predicted vs measured plot



(b) Loss curve

Fig. 3 Train and validation loss curve

Table 4 Experimental results of different graph network models

MODEL	MAE	RMSE	R^2
CGCNN [9]	3.459	6.534	0.9076
ALIGNN [10]	3.632	6.927	0.8958
BSGNN [11]	3.466	6.673	**0.912**
CrystalGAT (Ours)	**3.16**	**6.050**	0.898

3.2 Discussion

The performance metrics of the CrystalGAT model, as observed during both training and validation, demonstrate its strong predictive capabilities. The consistent decrease in training loss throughout the training phase indicates that the model effectively learns and captures the underlying features of the dataset. This steady improvement in accuracy highlights the model's proficiency in minimizing prediction errors over time.

When evaluated on the 3DSC_MP test set, the CrystalGAT model achieves a Mean Absolute Error (MAE) of 3.1421, a Mean Square Error (MSE) of 35.8274, a Mean Square Log Error (MSLE) of 1.8131, a Root Mean Square Error (RMSE) of 5.9856, and an R-squared value of 0.8935. These metrics collectively underscore the model's robust performance and its capacity to generalize to previously unseen data, a critical requirement for practical applications in material property prediction.

In comparison with other models applied to the 3DSC_MP dataset, as illustrated in Table 4.1, the CrystalGAT model exhibits competitive performance. With an MAE of 3.16, RMSE of 6.050, and R-squared value of 0.898, CrystalGAT slightly outperforms both the CGCNN and ALIGNN models in terms of MAE, while maintaining a lower RMSE. Although the BSGNN model achieves a marginally higher R-squared value, CrystalGAT's favorable balance between MAE and RMSE suggests it offers an advantageous trade-off between predictive accuracy and error distribution.

These results affirm the CrystalGAT model's effectiveness in capturing complex patterns within the data, underscoring its potential as a reliable predictive tool for crystal property estimation. Future research could explore optimizing the model architecture further or applying it to additional datasets to enhance its robustness and broaden its applicability in the field of materials science.

4. Conclusion

In this study, we introduce a novel graph neural network model, CrystalGAT, which integrates the functionalities of graph attention networks (GAT) and one-dimensional convolutional neural networks (1DCNN) to effectively leverage both microstructural and atomic-level information in materials. Experimental results on the 3DSC_MP dataset demonstrate that CrystalGAT outperforms

traditional models, achieving a significant reduction in mean absolute error (MAE) to 3.16. This improvement underscores the model's capability in capturing complex relationships between crystalline structure, electronic properties, and superconducting performance.

In the design of CrystalGAT, a multilayer perceptron (MLP) is incorporated for final feature integration. This design balances the local atomic interaction information extracted by the GAT branch with the global chemical composition information captured by the 1DCNN branch, leading to more precise predictions of critical temperature. The MLP's superior non-linear expression capabilities and flexible network structure make it an ideal tool for integrating features across different granularities, enhancing the model's overall predictive accuracy.

Reference

[1] Butler K T, Davies D W, Cartwright H, et al. Machine learning for molecular and materials science. Nature, 2018, 559(7715): 547-555.

[2] Stanev V, Oses C, Kusne A G, et al. Machine learning modeling of superconducting critical temperature. npj Computational Materials, 2018, 4: 29.

[3] Hamidieh K. A data-driven statistical model for predicting the critical temperature of a superconductor. Computational Materials Science, 2018, 154: 346-354.

[4] Schmidt J, Marques M R G, Botti S, et al. Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials, 2019, 5: 83.

[5] Dagotto E. Complexity in strongly correlated electronic systems. Reviews of Modern Physics, 1994, 66(3): 763-840.

[6] Lee P A, Nagaosa N, Wen X G. Doping a Mott insulator: Physics of high-temperature superconductivity. Reviews of Modern Physics, 2006, 78(1): 17-85.

[7] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. arXiv preprint, 2017: arXiv:1710.10903.

[8] Sommer T, Willa R, Schmalian J, et al. 3DSC - A new dataset of superconductosincluding crystal structures. figshare, 2022. [Online]. Available: https://arxiv.org/abs/2212.06071

[9] Xie T, Grossman J C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical Review Letters, 2018, 120(14): 145301.

[10] Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. arXiv preprint, 2021: arXiv:2106.01829. [Online]. Available: https://doi.org/10.48550/arXiv.2106.01829

[11] Gu L, Zhang Y, Wang J, et al. Searching high temperature superconductors with the assistance of graph neural networks. arXiv preprint, 2023: arXiv:2308.11160. [Online]. Available: https://doi.org/10.48550/arXiv.2308.11160