scientific reports

Check for updates

OPEN

A machine learning method for predicting molecular antimicrobial activity

Bangjiang Lin[1,2]✉, Shujie Yan[1,2] & Bowen Zhen[1,2]

In response to the increasing concern over antibiotic resistance and the limitations of traditional methods in antibiotic discovery, we introduce a machine learning-based method named MFAGCN. This method predicts the antimicrobial efficacy of molecules by integrating three types of molecular fingerprints—MACCS, PubChem, and ECFP—along with molecular graph representations as input features, with a specific focus on molecular functional groups. MFAGCN incorporates an attention mechanism to assign different weights to the importance of information from different neighboring nodes. Comparative experiments with baseline models on two public datasets demonstrate MFAGCN's superior performance. Additionally, we conducted an analysis of the functional group distribution in both the training and test sets to validate the model's predictions. Furthermore, structural similarity analyses with known antibiotics are performed to prevent the rediscovery of established antibiotics. This approach enables researchers to rapidly screen molecules with potent antimicrobial properties and facilitates the identification of functional groups that influence antimicrobial performance, providing valuable insights for further antibiotic development.

Antibiotics, a pivotal class of drugs, are primarily used to treat infectious diseases by killing bacteria or inhibiting their growth[1]. The introduction of antibiotics marks a significant milestone in medical history, having saved countless lives. Despite their success, the rise of antibiotic resistance has become a major global health threat, as noted by the World Health Organization among the top ten global health threats[2]. In 2019, infections involving resistant bacteria resulted in approximately 1.27 million deaths worldwide, contributing to nearly 5 million additional deaths associated with antibiotic resistance. It is projected that by 2050, antibiotic resistance will cause an estimated 10 million direct deaths annually, paralleling the number of deaths from cancer in 2020[3]. The urgent need for new antibiotics to address this crisis is clear, yet developing them is inherently challenging. Traditionally, the path from discovery to clinical application spans over a decade and incurs costs in the hundreds of millions of dollars. This method relies on screening active compounds from natural sources such as soil, plants, or marine microorganisms. This process involves sampling and culturing microorganisms, screening for antimicrobial activity, isolation and identification of compounds, and activity verification and optimization. Although this method has historically yielded significant antibiotics like penicillin and streptomycin, its viability has diminished due to its time-consuming, costly nature, and frequent rediscovery of known compounds[4]. Consequently, identifying structurally unique candidate molecules early in the research phase is crucial to expedite the development process and reduce costs. As illustrated in Fig. 1, the development of a new drug typically involves the following four main stages:

- Early stage of new drug development: Target identification and validation, lead compound discovery and lead optimization.
- Preclinical research: Pharmaceutical research of raw materials and formulations, pharmacological efficacy and safety assessment in animals.
- Clinical trials Phase I, II, III.
- Regulatory review and post-marketing research and monitoring.

The application of machine learning in antibiotic discovery has garnered significant research interest, offering remarkable advantages in identifying molecules with antibiotic potential[4–10]. Machine learning can accelerate the screening process and significantly reduce the time required for initial preparation[11–15]. With the continuous growth and refinement of biological data, machine learning enables efficient and extensive screening of potential

[1]Quanzhou Institute of Equipment Manufacturing, Haixi Institutes, Chinese Academy of Sciences, Quanzhou 362216, China. [2]College of Electrical Engineering and Automation, Fuzhou University, Fuzhou 350108, China. ✉email: linbangjiang@fjirsm.ac.cn

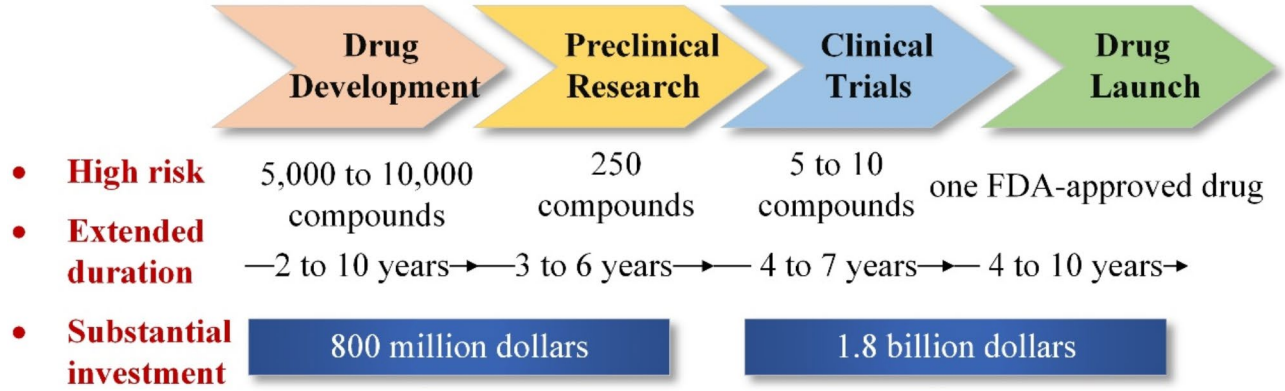

Fig. 1. The pharmaceutical drug development pipeline.

antimicrobial molecules. Furthermore, leveraging machine learning to identify antimicrobial compounds with structures distinct from known antibiotics helps avoid redundancy, saving both time and financial resources[16–20].

Early researchers utilized traditional approaches, extracting molecular fingerprints or hand-engineered features combined with basic machine learning models for prediction. For instance, as early as 2004, Murcia-Soler et al. employed topological indices derived from SMILES strings to represent information about atom types, chemical bonds, vertex degrees, and distances between atomic pairs[14]. These topological indices were then input into an ANN to distinguish compounds with antibacterial activity. Later, in 2014, Wang et al. applied Naive Bayes (NB), Support Vector Machine (SVM), Recursive Partitioning (RP) and k-Nearest Neighbors (kNN) four machine learning methods along with physicochemical descriptors and fingerprint features to construct predictive models[15]. Their work successfully identified potential compounds with activity against Staphylococcus aureus and explored the impact of different descriptors and models on predictive accuracy. With the advent of GNNs, researchers recognized that molecular graphs are naturally represented as graph-structured data, which traditional machine learning models had not fully leveraged. By using GNNs for predicting the antibacterial performance of molecules, both predictive accuracy and interpretability have significantly improved[16–25]. A notable study from the MIT, led by Stokes et al., employed an MPNN to predict antibacterial activity[4]. Among the top 99 compounds predicted by the model, 51 exhibited strong antibacterial properties during biological testing. Their predictions, applied across multiple chemical libraries, led to the discovery of a compound named Halicin. Structurally distinct from traditional antibiotics, Halicin demonstrated potent bactericidal activity against various bacteria, including drug-resistant pathogens. This groundbreaking work received widespread acclaim as it overcame the limitations of traditional antibiotic discovery methods and experimentally validated the feasibility of using machine learning to discover antibiotics. The size and quality of the dataset are also key factors affecting the performance of the model. In 2023, based on the research of Stokes et al., Gary Liu and colleagues from McMaster University screened 7,500 compounds and trained a GNN model using bacterial growth inhibition data[22]. By incorporating ensemble learning techniques to enhance model performance, they successfully identified a novel antibacterial compound named Abaucin. Remarkably, Abaucin demonstrated excellent anti-infective efficacy in a mouse wound model, highlighting its potential for clinical applications. These studies underscore the effectiveness of graph-based deep learning architectures, which have emerged as a popular and successful approach for molecular antibacterial activity prediction. Current research emphasizes the integration of multidisciplinary knowledge and experimental validation. In 2025, Si Zheng et al. developed a virtual screening workflow combining machine learning and deep learning models[40]. They screened 11,576 compounds from the DrugBank database and successfully identified two novel anti-tuberculosis drugs, aldoxorubicin and quarfloxin. Experimental validation demonstrated that both drugs exhibited potent inhibitory effects against drug-resistant strains. Furthermore, molecular docking, molecular dynamics simulations, and surface plasmon resonance experiments confirmed the direct binding of these compounds to Mycobacterium tuberculosis DNA gyrase. This study provides a new approach for the repurposing of antimicrobial drugs. Figure 2 provides a comprehensive overview of data and methods used in antibiotic discovery driven by machine learning. The drug development process is illustrated from left to right, emphasizing key elements. The initial stage of any machine learning-based project involves the collection of experimental data, which forms the foundation for model development. This data is subsequently transformed into a format compatible with machine learning algorithms. Various algorithms, ranging from traditional decision trees to advanced graph neural networks, are then employed to train the model. Once trained, these models can predict a range of properties, including the efficacy of antibiotics, the structures of novel compounds, and the identification of compounds with desired characteristics.

To overcome the limitations of existing machine learning-based methods for predicting molecular antimicrobial properties, which often rely on single-modal molecular representations, this study introduces a multimodal prediction model named MFAGCN. MFAGCN employs a hybrid molecular representation by integrating molecular graphs with three types of molecular fingerprints—MACCS, PubChem, and ECFP—as input features. The model utilizes a Graph Convolutional Network (GCN) to efficiently process molecular graph data and incorporates an attention mechanism to assign varying weights to information from different

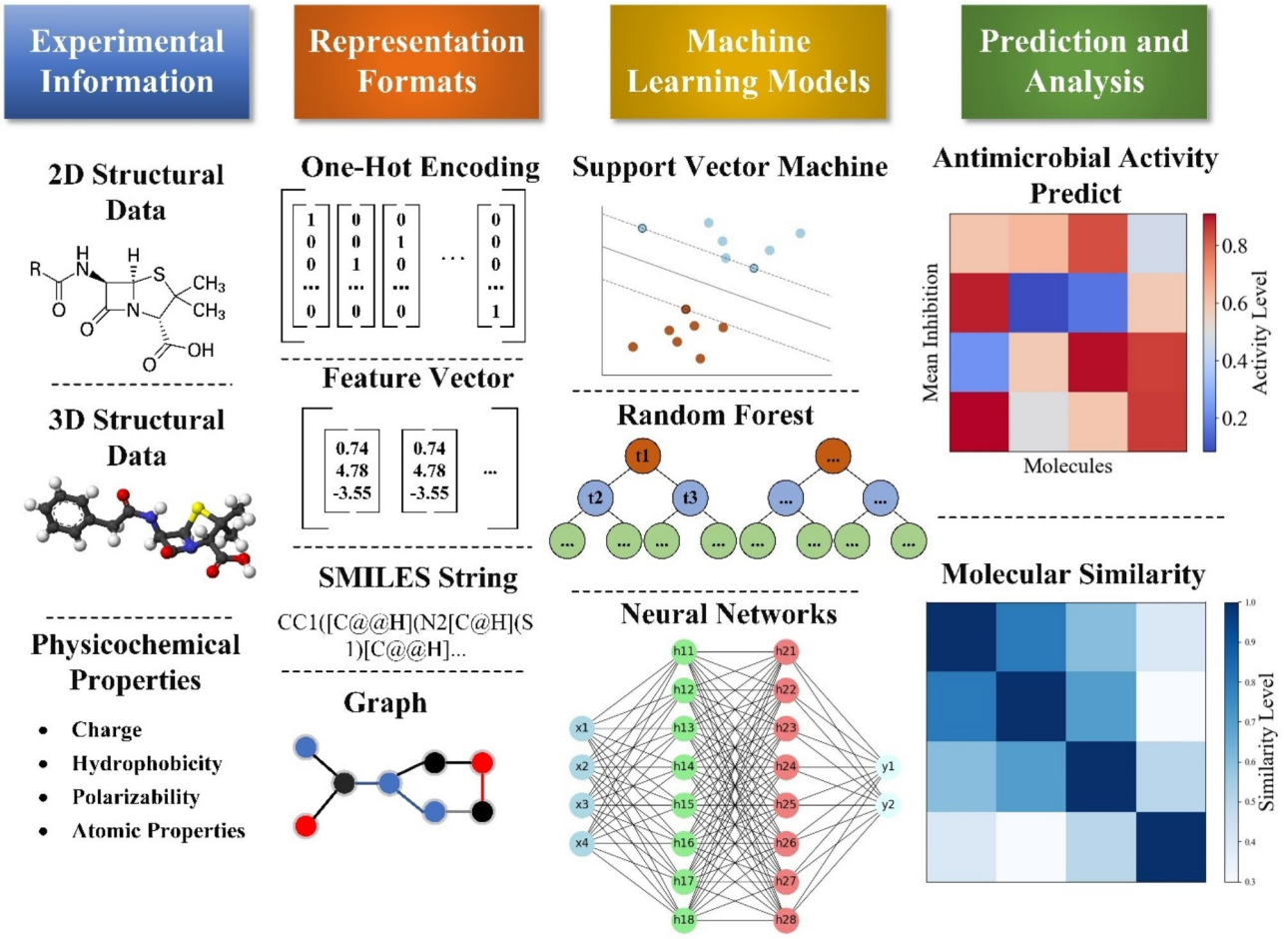

Fig. 2. The antibiotic discovery process and the baseline models utilized in this work.

neighboring nodes. Additionally, MFAGCN explicitly considers functional groups within molecular structures, and the analysis of functional group distributions further validates the model's reliability. Experimental results on two public datasets demonstrate that the proposed MFAGCN model outperforms existing methods in predicting the antimicrobial performance of molecules.

Method

Dataset preparation

We selected Escherichia coli and Acinetobacter baumannii as our primary datasets. Escherichia coli is widely used in antibacterial drug screening and research due to its low-cost and high-efficiency cultivation conditions, providing significant experimental value and promising application prospects. On the other hand, the incidence of Acinetobacter baumannii infections has been steadily increasing in recent years, particularly among patients with weakened immune systems, leading to severe complications and high mortality rates. Therefore, studying the pathogenic mechanisms, antibiotic resistance characteristics, and developing new antibacterial treatment strategies for Acinetobacter baumannii holds substantial clinical and societal significance.

The first dataset[4] provides growth inhibition data for *E. coli* BW25113, encompassing 1760 molecules with diverse structures and functions. Additionally, 800 natural products isolated from plants, animals, and microorganisms were included to further enhance chemical diversity. After removing duplicates, a total of 2,335 unique compounds were obtained. This dataset includes the SMILES (Simplified Molecular Input Line Entry System) representations of each compound, along with the growth inhibition rates of E. coli BW25113 in media containing these compounds. Binary classification was performed based on whether the inhibition rate was below 0.2, resulting in positive samples accounting for 5.14% of the dataset, indicating their effectiveness in inhibiting the growth of E. coli BW25113. To address the issue of class imbalance, we employed techniques such as class weight adjustment and balanced sampling during model training. These methods allow the model to better learn and recognize the characteristics of positive samples, thereby improving its predictive accuracy and generalization ability.

The second dataset[22] focuses on the growth inhibition of Acinetobacter baumannii, a common Gram-negative bacterium associated with hospital infections, particularly in intensive care units, and known for its antibiotic resistance due to misuse. A. baumannii has become a significant source of hospital-acquired infections. This

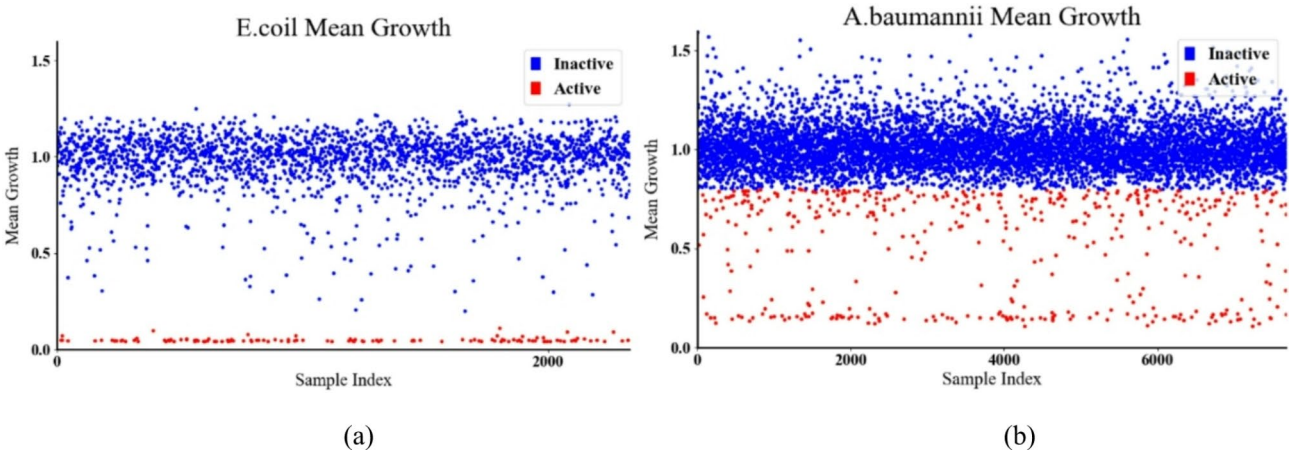

Fig. 3. Growth inhibition dataset distribution (**a**: *E. coli*, **b**: *A. baumannii*).

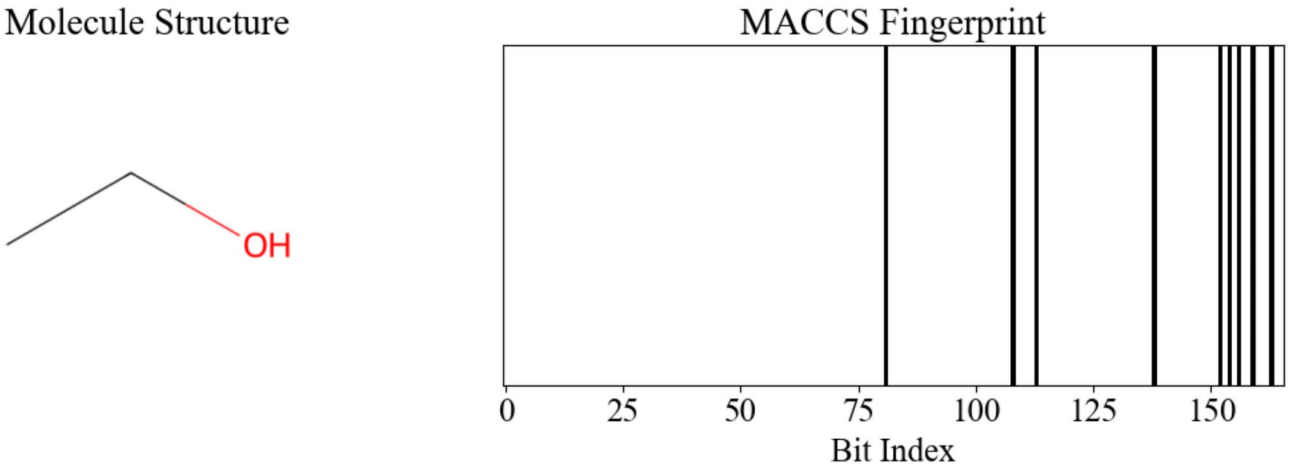

Fig. 4. The 2D structure of ethanol "CCO" and the binary heatmap of its MACCS fingerprint.

dataset provides SMILES representations for 7684 compounds along with their corresponding growth inhibition rates. Binary labels were assigned using a cutoff based on one standard deviation below the average growth rate for the entire dataset, with positive samples accounting for 6.24% of the dataset. Similar to the first dataset, we addressed class imbalance by applying class weight adjustments and balanced sampling techniques during model training. These binary labels indicate whether a compound inhibits the growth of *A. baumannii* in the medium.

The input data were partitioned using the Scaffold method, allocated in an 8:2 ratio into training and test sets, respectively. Scaffold splitting is a more sophisticated method typically used for handling data with complex chemical molecular structures. It ensures that samples with similar chemical structures, like those in this work, are grouped into the same dataset. This approach guarantees differentiation between the training and test sets, thereby enhancing the generalizability of the model. Several other partitioning ratios were considered for comparison, such as 7:3. After evaluating the performance of the model with these different ratios, the 8:2 split was selected as it provided the best balance between training set size and model generalization. The distribution information of both datasets is illustrated in the Fig. 3.

Molecular feature selection

The molecular features in this study comprise two key components: molecular fingerprints and molecular graph representations. Molecular fingerprints typically encapsulate a range of molecular properties, including general physical attributes, electrochemical characteristics, and electron distribution. In this work, we employ three types of fingerprints: MACCS, PubChem, and ECFP.

MACCS fingerprints consist of 166 predefined binary bits, each representing a specific structural fragment or chemical property within the molecule[38]. Each bit is assigned a value of 0 or 1, indicating the presence or absence of the corresponding atomic species, bonding information, or local atomic environment. Figure 4 shows the 2D structure of the ethanol molecule and the binary heatmap of its MACCS molecular fingerprint. PubChem fingerprints are designed to represent molecular substructures and properties with a predefined bit string that encapsulates the presence or absence of certain substructural features. These fingerprints are frequently used in

virtual screening and chemical informatics due to their comprehensiveness and reliability. ECFP (Extended-Connectivity Fingerprints)[39] are generated using the Morgan algorithm with a specific radius, typically set to 2. These fingerprints are based on circular substructures surrounding each atom and are represented as a binary vector of 2048 bits. ECFP fingerprints are particularly effective in capturing local atomic environments and chemical connectivity.

The other component involves molecular graph representations, where each molecule is depicted as a graph. In this graph, nodes correspond to atoms, and edges represent the chemical bonds between them. Each node can store detailed information about the atom, such as its type, charge, multiplicity, and mass. The edges contain information regarding the bonds, including bond type (single, double, triple) and bond order. Additionally, both nodes and edges can encode information about aromaticity and stereochemistry. The final input to the model integrates MACCS, PubChem, and ECFP fingerprints with molecular graph representations, enabling the model to comprehensively leverage molecular structural information. As illustrated in Fig. 5, we compared the model's performance with different types of molecular fingerprints as inputs. The results indicate that the model's performance varies depending on the type of fingerprint used, with the combination of MACCS, PubChem, and ECFP fingerprints yielding superior model performance.

Graph convolutional neural networks

Molecules inherently possess graph-structured data, and to predict molecular antibacterial performance, this work employs Graph Convolutional Neural Networks (GCNs)[24–26]. GCNs are a special type of graph neural network that extends the principles of convolutional neural networks (CNNs) to graph data, enabling direct processing of graph-structured data. GCNs can learn both node features and the information pertaining to edges between nodes, facilitating comprehensive utilization of molecular graph data. Spatial features in molecular graph data possess the following characteristics: each node has its own features, and there exist connections between nodes representing bond features. Thus, when handling molecular graph data, both node and structural information must be considered.

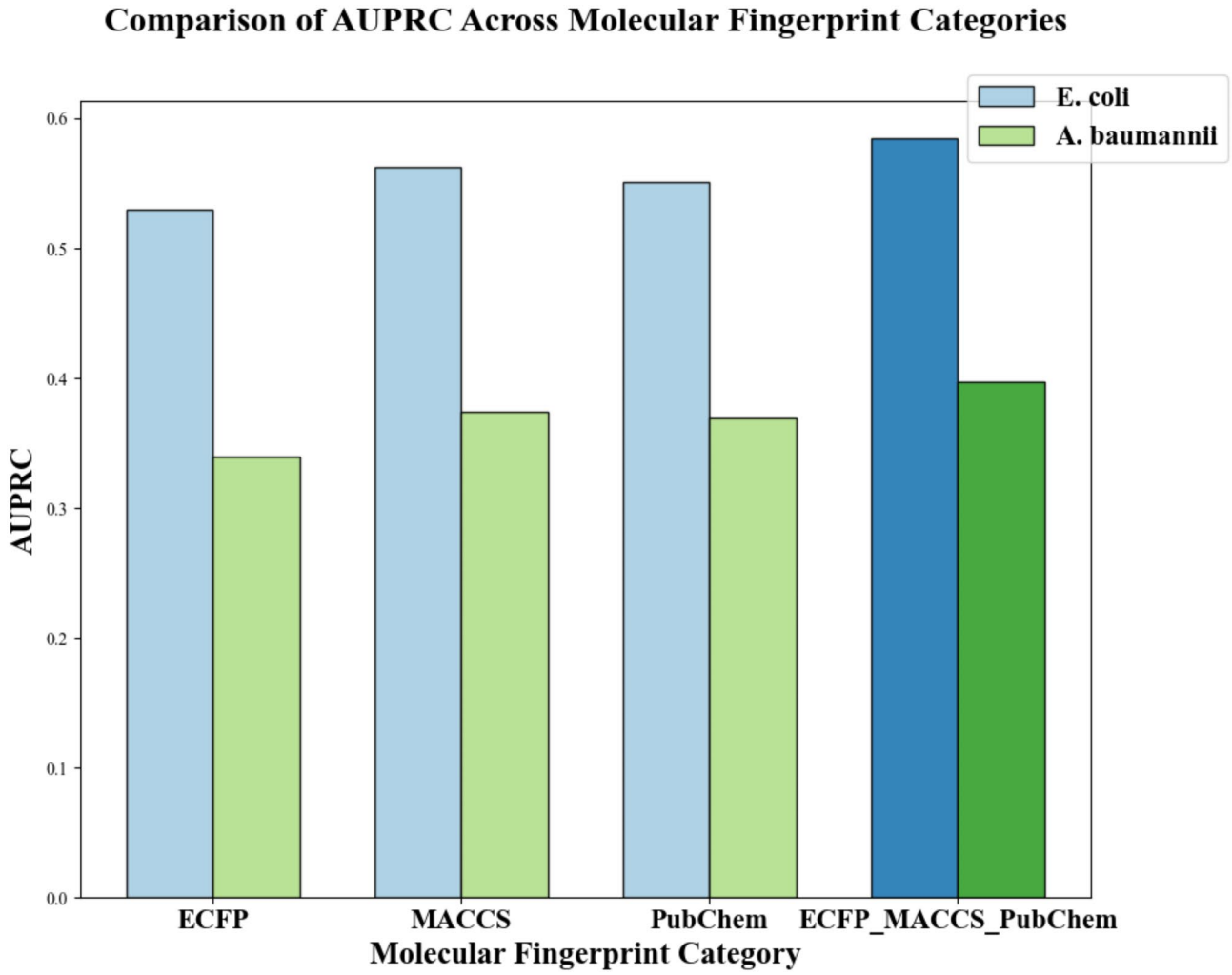

Fig. 5. Impact of different molecular fingerprint categories on model performance.

In this work, GCNs are utilized to extract spatial features of molecular graph structures. Convolutions are applied at each node, and based on the information from neighboring nodes, the feature representation of each node is updated. Specifically, for a molecular graph data with n nodes, each node has its feature vector, forming an $n \times d$ dimensional feature matrix $\mathbf{X}$. Additionally, the relationships between nodes form an $n \times n$ adjacency matrix $\mathbf{A}$. $\mathbf{X}$ and $\mathbf{A}$ serve as the inputs to the GCN model. Let $\mathbf{H}^{(l)}$ represent the feature matrix of all nodes in the l-th layer, $\mathbf{H}^0 = \mathbf{X}$,and $\mathbf{H}^{(l+1)}$ denote the feature matrix after one convolution operation. The formula for one convolution operation is as follows:

$$\mathbf{H}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right)$$

where $\mathbf{A}$ is the adjacency matrix, and $\mathbf{A}_{ij} \neq 0$ indicates that nodes *i* and *j* are neighbors, i.e., there exists an edge between *i* and *j*. $\tilde{\mathbf{A}}$represents the adjacency matrix $\mathbf{A}$ plus the identity matrix $\mathbf{I}$, i.e., adding self-connections to each node in the original graph. $\tilde{\mathbf{D}}$ represents the diagonal matrix obtained by adding the identity matrix $\mathbf{I}$ to the degree matrix $\mathbf{D}$. Each element $\tilde{\mathbf{D}}_{ii}$of the degree matrix is the degree of node *i*, i.e., the number of edges connected to node *i*. $\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{D}^{-\frac{1}{2}}$ denotes the normalization of the adjacency matrix A. $\mathbf{H}$ is the feature matrix of all nodes in each layer. σ represents a non-linear activation function, such as ReLU. $\mathbf{W}^{(l)}$ represents the trainable parameter matrix for the convolution transformation in the current layer. Through the aforementioned convolution operation, weighted summation of neighboring nodes is performed for each node in the graph, resulting in the generation of new feature representations for the nodes. By iteratively propagating through layers and performing convolution operations, the feature representation of each node is updated.

Attention mechanism

Incorporating an attention mechanism module into GCN enhances the model's ability to represent interactions between nodes. The attention mechanism dynamically allocates different weights based on the similarity between nodes, enabling the model to focus more on nodes that have a significant impact on the current task[24,25]. Therefore, in this work, GCN is coupled with an attention mechanism to better handle complex molecular graph data. For a given molecular graph data, attention coefficients can quantitatively characterize whether certain chemical fragments in the molecule contribute more to the prediction of molecular properties[27,28,33]. The attention coefficients can be calculated using the following formula:

$$e_j = \mathrm{LeakyReLU}(\mathbf{a} \cdot [\mathbf{W}_{\mathbf{h_i}} || \mathbf{W}_{\mathbf{h_j}}])$$

here, $\mathbf{h}_i$represents the feature vector of node *i*, which belongs to R^l space, indicating that the dimensionality of the feature vector is *l*. $\mathbf{a}$ is a parameter vector used to compute attention coefficients. During computation, the feature vectors of nodes *i* and its neighboring node *j* are first mapped through a transformation matrix $\mathbf{W}$, concatenated, and then passed through the parameter vector $\mathbf{a}$ to calculate the original attention coefficient e_j for each neighbor of node *i*. After obtaining the original attention coefficients for all neighbors, normalization is performed using the SoftMax function to derive the final attention weights a_{ij}. This weight signifies the relative importance of each neighboring node *j* in updating the feature of node *i*. The normalization formula is as follows:

$$a_{ij} = \frac{\exp(e_j)}{\sum_{k \in N(i)} \exp(e_k)}$$

thus, the new feature vector $\mathbf{h}_i'$ for each node *i* is obtained by weighted aggregation of the feature vectors of all its neighboring nodes, with the attention weight α_{ij}. Through this process, the model can dynamically allocate different weights based on the importance of neighboring nodes when updating node features. This approach enables the GCN model to better capture and utilize information about interactions between nodes in molecular graph data, thereby enhancing the prediction of molecular antibacterial performance. Integrating an attention mechanism into the GCN significantly enhances the model's ability to capture and represent the complex interactions between nodes in molecular graph data.

Figure 6 illustrates this enhancement by highlighting the relative importance of each atom within the molecule. These attention coefficients are calculated by the model and visually demonstrate the contribution of different structural components in the molecular graph to the prediction of molecular antibacterial performance. Atoms with higher attention weights are depicted in deeper red, indicating their substantial impact on the model's predictions. Conversely, atoms with lower attention weights are shown in lighter shades.

Molecular functional groups

Most existing graph neural network models designed to predict molecular antibacterial performance primarily focus on node-level or graph-level tasks, often neglecting the crucial information provided by molecular functional groups. Functional groups are fundamental components of molecular graphs, typically consisting of specific arrangements of atoms and bonds. Common functional groups include hydroxyl, carboxyl, ether bonds, aldehyde, carbonyl, and others. The understanding and prediction of molecular properties heavily rely on the recognition of these functional groups.

In this study, to fully exploit the influence of functional groups on molecular properties, functional group features were incorporated into the input data. In this context, if a specific functional group is present in the molecular graph, its feature value is set to 1; otherwise, it is set to 0. This binary representation allows the model to accurately capture the presence or absence of functional groups, thereby enhancing the prediction of molecular antibacterial performance.

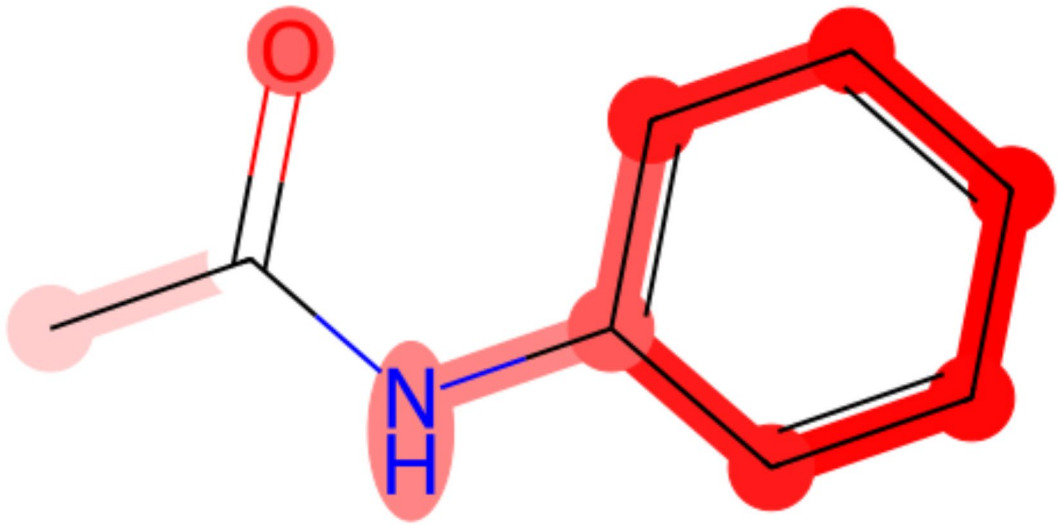

Fig. 6. Application of multi-head attention mechanism in molecular graphs.

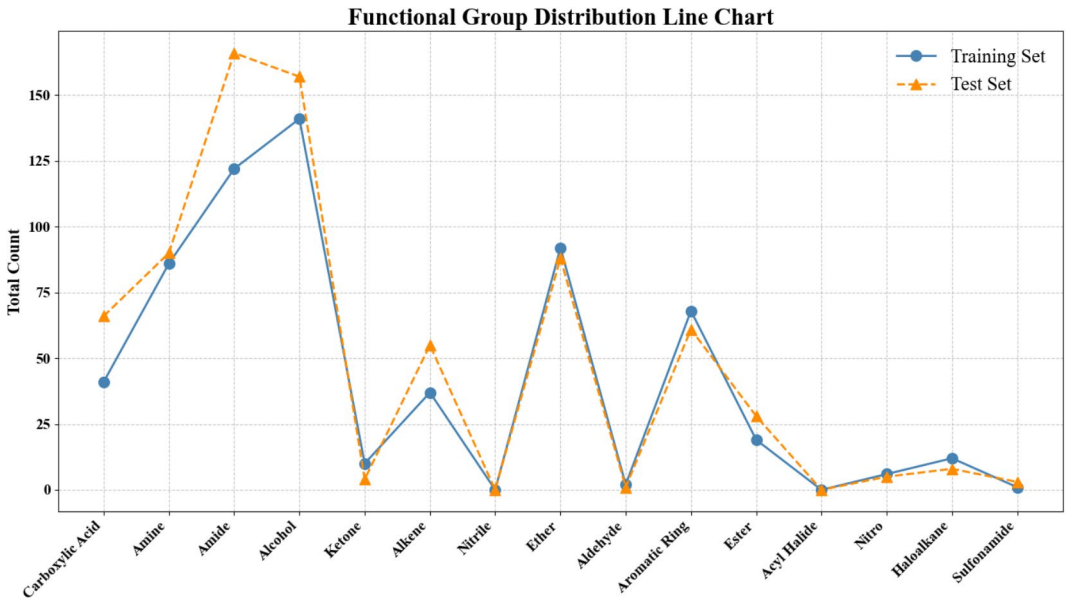

Fig. 7. Functional group distribution.

Figure 7 illustrates the distribution of functional groups in the top 100 molecules with the best antibacterial performance in the training set and the top 100 molecules with the highest predicted antibacterial efficacy in the test set. The high level of consistency in the functional group distributions between these two sets not only confirms the reliability of the model's predictions but also highlights its strong interpretability.

Visualizing the distribution of functional groups serves to validate the model's predictions while offering researchers valuable insights into the molecular structures that significantly impact antibacterial properties. For example, functional groups such as amines and alcohols exhibit high frequencies in both the training and test sets, suggesting that these groups may play a critical role in the antibacterial efficacy of the molecules.

Understanding the distribution and significance of these functional groups enables researchers to better identify key structural components that contribute to antibacterial activity. This knowledge is vital for the further development and optimization of antibiotics, helping to achieve more targeted and efficient drug design strategies.

Model architecture and interpretability

MFAGCN integrates multiple molecular fingerprints (MACCS, ECFP, PubChem), molecular graph and molecular functional group features to capture diverse aspects of molecular characteristics. During the feature extraction process, StandardScaler is employed to normalize different types of features, ensuring comparability

on a uniform scale. Concurrently, VarianceThreshold is utilized to eliminate constant features with zero variance, thereby reducing redundancy. Feature selection is performed using the SelectKBest method based on univariate analysis of variance to identify the most relevant features associated with the target variable. This approach not only enhances model performance but also improves feature interpretability, ensuring that the selected features possess clearer biological significance.

As shown in Fig. 8, the GCN layers are employed to capture the topological structure and node (atom) features of molecular graphs. Through multiple GCN layers and Batch Normalization, the model learns both local and global structural information of the molecules. Following the GCN layers, a Multihead Attention mechanism is introduced to enhance the model's ability to focus on different substructures. The attention weights derived from this mechanism can be used to interpret which key parts the model emphasizes during the decision-making process, thereby providing a degree of interpretability. Various fingerprint and functional group features are processed through independent embedding layers, comprising linear transformations, Batch Normalization, and ReLU activations, to convert them into uniformly dimensioned embedding vectors. These embedding vectors are then fused with the GCN outputs through weighted aggregation, ensuring that information from different sources is balanced and interpretable in the decision-making process. Global Mean Pooling is applied to the outputs of the graph convolutions, which are then added to the embedded features to form the final feature representation.

The model's decision-making process begins with the conversion of molecular SMILES strings into graph structures and multiple fingerprint and functional group features. The graph structures are processed through GCN layers to extract molecular structural features, while the fingerprint and functional group features are transformed into uniformly dimensioned vectors via independent embedding layers. The Multihead Attention mechanism enhances the model's focus on critical substructures within the molecule. The fingerprint and functional group features, once processed through their respective embedding layers, are fused with the GCN outputs through an additive operation to form a comprehensive feature representation. This fused feature vector is then passed through fully connected layers, Batch Normalization, activation functions, and Dropout layers, ultimately producing a binary classification output indicating whether the molecule is active or inactive. The Sigmoid activation function is used to convert the outputs into probability values, facilitating the classification decision. To address class imbalance, Focal Loss is employed, which adjusts the weights of easily classified samples, thereby increasing the model's focus on difficult-to-classify instances.

Evaluation metrics

In the context of using machine learning techniques to discover new antibiotics, datasets often exhibit imbalance, containing a large number of non-antibiotic molecules (negative samples) and relatively fewer known antibiotic molecules (positive samples). Therefore, using AUC as an evaluation metric may not effectively reflect the model's performance. The focus of this work is to identify molecules with antibacterial potential. In this scenario, the Area Under the Precision-Recall Curve (AUPRC) might be a more appropriate choice. Precision-recall curve evaluates the trade-off between precision and recall. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations, while recall is the ratio of correctly predicted positive observations to all actual positive observations. AUPRC calculates the area under this precision-recall curve, providing a measure of the model's ability to identify positive class samples. In the context of this work, a higher

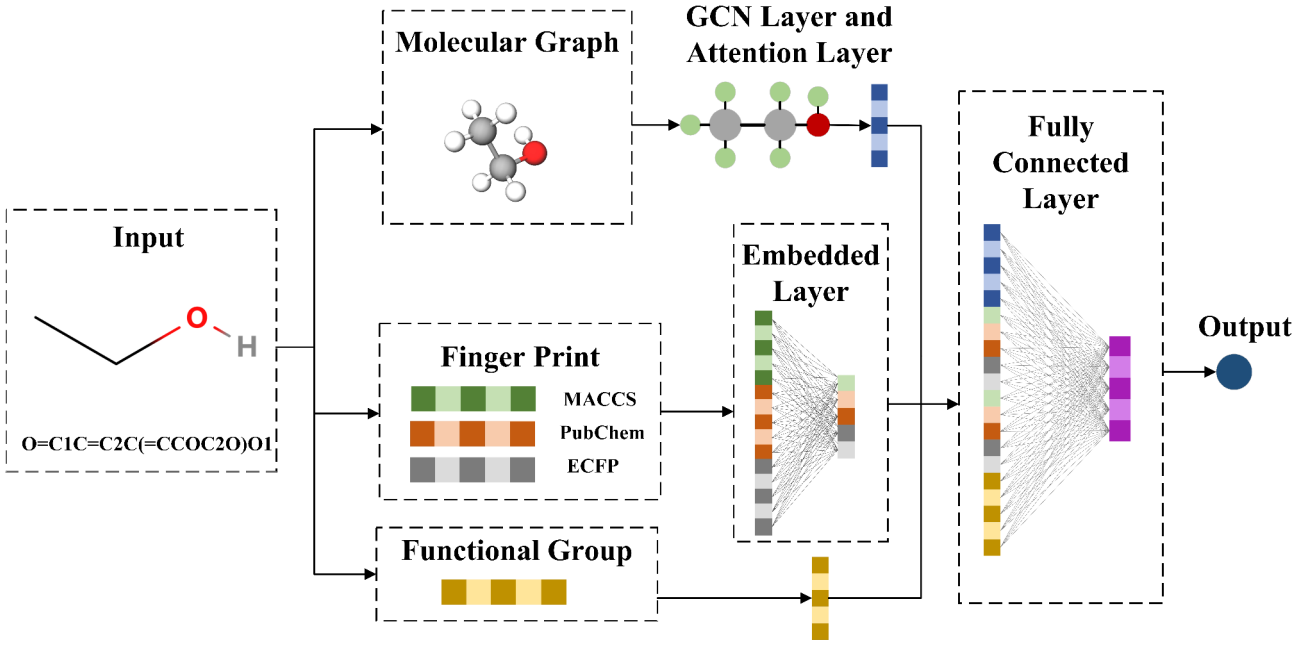

Fig. 8. MFAGCN model architecture diagram.

AUPRC indicates a stronger ability of the model to recognize molecules with antibacterial properties. Utilizing AUPRC as an evaluation metric can significantly reduce the experimental validation costs by selecting molecules identified by the model as having antibacterial potential.

Results and discussion

The experiments in this work were conducted using the PyTorch framework on a server equipped with a GeForce RTX 2080 Ti GPU. The neural networks were built using the open-source deep learning frameworks TensorFlow 2.16.1 and Keras. The specific configuration parameters of the MFAGCN model were determined through grid search and cross-validation to find the optimal values. These parameters are as follows: the dropout rate was set to 0.1, the batch size was 32, the learning rate(lr) was 0.001, weight decay was set to 0.01, the number of epochs was 500, and the AdamW optimizer was used.

Baseline models

To evaluate the performance of MFAGCN, the following models were employed as baselines:

- **RF**: A traditional machine learning method that constructs multiple decision trees and averages their results to improve prediction accuracy and model robustness. The following hyperparameters were set for the RF: "max_depth": 5627, "min_samples_leaf": 1, "min_samples_split": 2, "n_estimators": 3629.
- **GAT (Graph Attention Network)**: A popular model for predicting molecular properties, leveraging attention mechanisms to assign different weights to neighboring nodes and aggregate information effectively. The hyperparameters for GAT were optimized using grid search, with the number of attention heads set to 8 and the dropout rate set to 0.5.
- **CHEMPROP**: A widely-used machine learning software package for predicting chemical properties, utilizing message-passing neural networks for molecular property prediction. The model uses the default parameter configuration.
- **GRU (Gated Recurrent Unit)**: A type of recurrent neural network that is effective in processing sequential data, often used in molecular property prediction to model complex temporal dependencies and enhance prediction accuracy. The hyperparameters are set as follows: hidden_size = 128, num_layers = 2, and dropout = 0.5.
- **GIN (Graph Isomorphism Network)**: A specialized graph neural network designed to capture isomorphic graph features, showing strong performance in molecular property prediction tasks. The hyperparameters are set as follows: hidden_size = 128, num_layers = 3, and dropout = 0.4.
- **MFGCN**: A variant of the MFAGCN model without the attention mechanism, included to assess the impact of the attention mechanism on model performance. The hyperparameter settings are the same as those of MFAGCN.

As a binary classification task, the outputs of all algorithms in this work are continuous values ranging from 0 to 1, representing the probability that a sample belongs to the positive class. Using a threshold of 0.5, the continuous output values are binarized. Samples with predicted values exceeding the threshold are classified as positive samples, while those below the threshold are classified as negative samples.

Performance comparison of MFAGCN and other models

This study evaluates the antibacterial properties of E. coli and A. baumannii using various machine learning models, including MFAGCN, RF, GNN, CHEMPROP, GRU, GIN, and MFGCN, across two distinct datasets. The results, presented in Table 1 and Fig. 9a, reflect the highest AUPRC values observed within a credible range for the MFAGCN model and six baseline models. The MFAGCN model demonstrated superior performance, achieving AUPRC values of 0.5842 for E. coli and 0.3968 for A. baumannii. These results indicate that MFAGCN effectively captures the complex relationships between molecular structures and antibacterial activity, attributable to its multimodal feature integration and attention mechanisms, which allow the model to focus on the most relevant molecular features. In contrast, the RF model achieved an AUPRC of 0.4543 for E. coli in the smaller dataset, but its performance declined significantly in the larger dataset, yielding an AUPRC of only 0.2246 for A. baumannii. This decline suggests that RF may struggle with the increased complexity and variability present in larger datasets, possibly due to its inability to effectively model intricate molecular interactions.The GNN model recorded AUPRC scores of 0.4739 for E. coli and 0.3213 for A. baumannii, while CHEMPROP

	E. coli	A. baumannii
MFAGCN	0.5842	0.3968
RF	0.4543	0.2246
GNN	0.4739	0.3213
CHEMPROP	0.5263	0.3379
GRU	0.4552	0.2611
GIN	0.5217	0.3363
MFGCN	0.5333	0.3574

Table 1. Comparative AUPRC scores of MFAGCN and baseline models for inhibitory activity prediction against E. coli and A. baumannii.

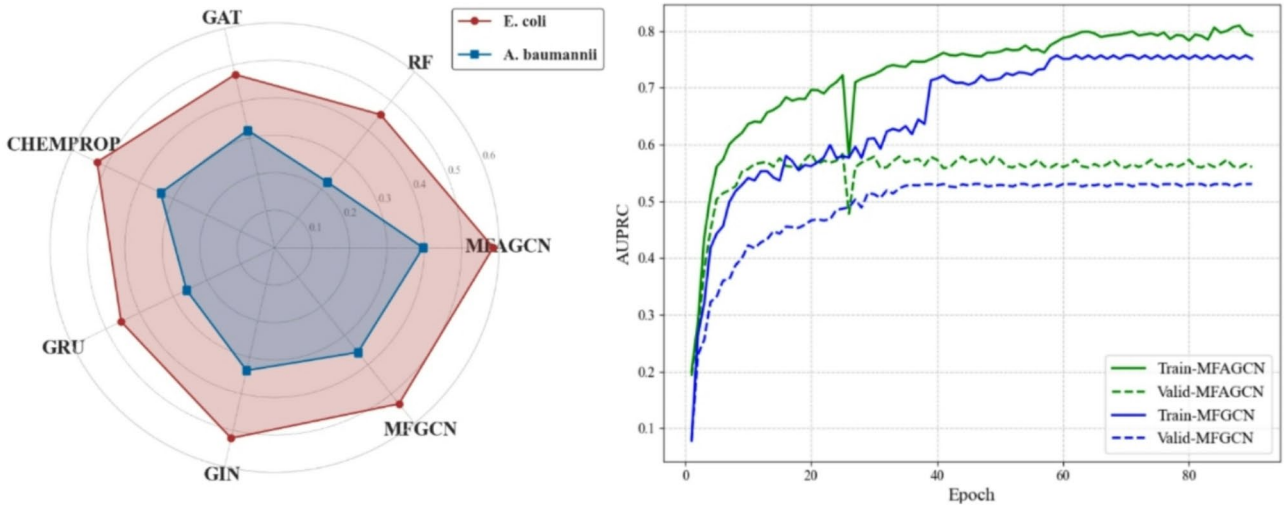

Fig. 9. Comparative analysis of model efficacy in predicting antibacterial properties (**a**: baseline model, **b**: ablation experiment of the attention mechanism).

obtained 0.5263 and 0.3379 respectively. These graph neural network models are effective to a certain extent, but they do not leverage the attention mechanisms employed by MFAGCN, which may explain their relatively lower performance.The GRU model showed AUPRC values of 0.4552 for E. coli and 0.2611 for A. baumannii, indicating moderate performance but still lagging behind MFAGCN. The GIN model achieved scores of 0.5217 for E. coli and 0.3363 for A. baumannii, which are competitive but still inferior to the proposed MFAGCN. Notably, the MFGCN model, a variant of MFAGCN without the attention mechanism, attained AUPRC values of 0.5333 for E. coli and 0.3574 for A. baumannii Fig. 9b shows the AUPRC change curve of MFAGCN with and without the attention mechanism over the first 100 epochs. The decrease in performance compared to MFAGCN underscores the importance of the attention mechanism in enhancing model performance by allowing it to prioritize significant molecular features.

The superior performance of MFAGCN is attributed to its integration of multimodal feature fusion and attention mechanisms. By combining multiple molecular fingerprints (MACCS, PubChem, and ECFP), molecular graph representations, and functional group features, MFAGCN captures a comprehensive spectrum of molecular characteristics. This multimodal approach enables the model to leverage diverse sources of information, enhancing its ability to identify complex patterns associated with antibacterial activity. The attention mechanism further boosts MFAGCN's performance by dynamically assigning weights to different molecular substructures, allowing the model to prioritize the most indicative features. This selective focus not only improves prediction accuracy but also enhances the model's interpretability. The significant decline in performance of MFGCN in ablation experiments underscores the critical role of the attention mechanism in effectively modeling interactions within molecular graphs.

Prediction of molecular similarity to known antibiotics structure analysis

To avoid rediscovering known antibiotics, it is essential to conduct a structural similarity analysis between predicted molecules and known antibiotics after model prediction. In this work, the Tanimoto coefficient is utilized to calculate similarity. The Tanimoto coefficient compares the number of identical positions in molecular fingerprints to compute the similarity between molecules. The Tanimoto coefficient between molecule A and molecule B can be calculated using the molecular fingerprints through the following formula:

$$T(A,B) = \frac{c}{a+b-c}$$

In this formula, 'a' represents the number of occurrences of 1 in molecule A, 'b' represents the number of occurrences of 1 in molecule B, and 'c' represents the number of occurrences of 1 in both molecules A and B. For instance, if the molecular fingerprint of molecule A is 010101 and that of molecule B is 100,100, then a = 3, b = 2, and since there is only one common position, c = 1. Therefore, the Tanimoto coefficient between A and B is calculated as $1/(3+2-1) = 0.25$. After model prediction, priority is given to experimentally validating molecules with high prediction scores and low similarity to known antibiotic structures. This approach maximizes the chances of discovering new antibiotic structures while minimizing the likelihood of rediscovering known antibiotics[34–38]. Figure 10 visualizes the structural similarity between the active molecules in the test set and the active molecules in the training set using t-SNE.

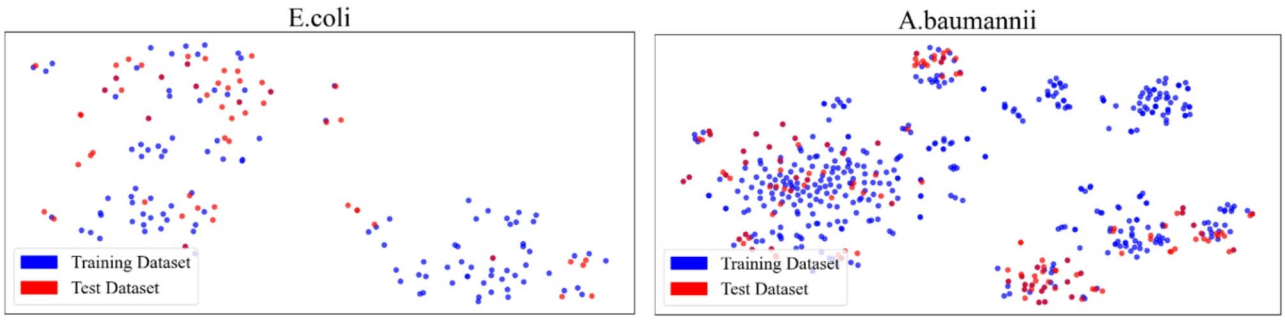

Fig. 10. t-SNE visualization of structural similarities between the active molecules in the training dataset and the active molecules in test dataset for E. coli and A. baumannii.

Conclusion

In response to the challenges of long discovery cycles, high investment costs, and the difficulty in identifying novel structural antibiotics using traditional methods, we proposed a method based on the MFAGCN model to predict the antibacterial properties of molecules. This approach leverages a graph convolutional neural network architecture, incorporating a combination of three molecular fingerprints—MACCS, PubChem, and ECFP—along with molecular graphs as input data, following a comparative analysis that demonstrated the effectiveness of this combination. Additionally, the model considers molecular functional group features and incorporates an attention mechanism to enhance predictive accuracy. Experimental results show that the MFAGCN model outperforms other existing models in predicting molecular antibacterial properties according to the AUPRC standard. Notably, the consistency in functional group distribution between the training and test sets further validates the reliability of the model's predictions.

Due to limited capabilities, this work has many limitations. The binary classification method employed in this study did not fully utilize the inhibitory rates of molecules in the dataset. It may be more beneficial to consider treating this as a regression problem rather than a classification problem, although this would require more molecular data for training. From the perspective of molecular representation, although this study has characterized molecules using molecular fingerprints, molecular graphs, and molecular functional groups, there are still many important molecular features that have not been captured, such as quantum relationships involving the electronic distribution and energy level structures of molecules. These uncaptured features may contain crucial information that significantly impacts antibacterial activity. Therefore, future research should consider incorporating more advanced molecular representation methods. Secondly, the current dataset used in this study lacks comprehensiveness in terms of sample size and diversity, especially with a severe imbalance between positive and negative samples, thereby limiting the model's generalization ability and prediction accuracy. Future research should focus on collecting larger and more diverse molecular datasets to balance the ratio of positive and negative samples. Additionally, future studies should integrate toxicity prediction into the existing prediction framework by adopting multi-objective regression or classification methods to balance antibacterial efficacy and safety. Furthermore, advanced modeling techniques such as ensemble learning can be employed to further optimize predictive performance.

Data availability

Source codes of our method are available at https://github.com/MFAGCN/MFAGCN. The E.coli dataset was downloaded from https://www.cell.com/cell/fulltext/S0092-8674(20)30102-1. The A. baumannii dataset was downloaded from https://www.nature.com/articles/s41589-023-01349-8.

Received: 9 May 2024; Accepted: 18 February 2025
Published online: 24 February 2025

References

1. Wouters, O. J., McKee, M. & Luyten, J. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *JAMA* **323**, 844–853 (2020).
2. Mullard, A. New drugs cost US$2.6 billion to develop. *Nat. Rev. Drug Discov.* **13**, 877 (2014).
3. CDC. Antibiotic Resistance Threats in the United States, 2019. Technical Report. US Department of Health and Human Services, CDC, (2019).
4. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702 (2020).
5. Melo, M. C. R. & Maasch, J. R. M. A. Fuente-Nunez, C. Accelerating antibiotic discovery through artificial intelligence. *Commun. Biol.* **4**, 1 (2021). de la.
6. Deng, J. et al. Artificial intelligence in drug discovery: Applications and techniques. *Briefings Bioinf.* **23**, bbab430 (2022).
7. Farghali, H., Kutinová Canová, N. & Arora, M. The potential applications of artificial intelligence in drug discovery and development. *Physiol. Res.* **70**, S715–S722 (2021).
8. Qureshi, R. et al. AI in drug discovery and its clinical relevance. *Heliyon* **9**, e17575 (2023).
9. Han, R. et al. Revolutionizing medicinal chemistry: The application of artificial intelligence (AI) in early drug discovery. *Pharmaceuticals* **16**, 1259 (2023).
10. Guan, S. & Fu, N. Class imbalance learning with bayesian optimization applied in drug discovery. *Sci. Rep.* **12**, 2069 (2022).
11. Tkatchenko, A. Machine learning for chemical discovery. *Nat. Commun.* **11**, 4125 (2020).

12. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
13. Xu, L. et al. Molecular property prediction by combining LSTM and GAT. *Biomolecules* **13**, 503 (2023).
14. Murcia-Soler, M. et al. Artificial neural networks and linear discriminant analysis: A valuable combination in the selection of new antibacterial compounds. *J. Chem. Inf. Model.* **44**, 1031–1041 (2004).
15. Wang, L. Le.et al. Discovering new agents active against methicillin-resistant Staphylococcus aureus with ligand-based approaches. *J. Chem. Inf. Model.* **54**, 3186–3197 (2014).
16. Ahmad, W. et al. Attention-Based graph neural network for molecular solubility prediction. *ACS Omega*. **8**, 3236–3244 (2023).
17. Chen, S. et al. MD-GNN: A mechanism-data-driven graph neural network for molecular properties prediction and new material discovery. *J. Mol. Graph Model.* **123**, 108506 (2023).
18. Zang, X. et al. Hierarchical molecular graph Self-Supervised learning for property prediction. *Commun. Chem.* **6**, 34 (2023).
19. Tian, Y. et al. Predicting molecular properties based on the interpretable graph neural network with multistep focus mechanism. *Briefings Bioinf.* **24**, bbac534 (2023).
20. Zhang, Y. et al. Attention is all you need: Utilizing attention in AI-enabled drug discovery. *Briefings Bioinf.* **25**, bbad467 (2024).
21. Carpenter, K. A. et al. Deep learning and virtual drug screening. *Future Med. Chem.* **10**, 2557–2567 (2018).
22. Liu, G. et al. Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. *Nat. Chem. Biol.* **19**, 1342–1350 (2023).
23. Mongia, M. et al. An interpretable machine learning approach to identify mechanism of action of antibiotics. *Sci. Rep.* **11**, 14229 (2022).
24. Cheng, Z. et al. Drug-target interaction prediction using multi-head self-attention and graph attention network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 2208–2218 (2022).
25. Zhu, W. et al. HiGNN: A hierarchical informative graph neural network for molecular property prediction equipped with feature-wise attention. *J. Chem. Inf. Model.* **63**, 43–55 (2023).
26. Wang, X. et al. A complete graph-based approach with multi-task learning for predicting synergistic drug combinations. *Bioinformatics* **39**, btad351 (2023).
27. Cheng, X. et al. iADRGSE: A graph-embedding and self-attention encoding for identifying adverse drug reaction in the earlier phase of drug development. *Int. J. Mol. Sci.* **23**, 16216 (2022).
28. Su, X. et al. Attention-based knowledge graph representation learning for predicting drug-drug interactions. *Briefings Bioinf.* **23**, bbac140 (2022).
29. Wen, N. et al. A fingerprints based molecular property prediction method using the BERT model. *J. Chem.* **14**, 71 (2022).
30. Zhang, J. et al. SMG-BERT: Integrating stereoscopic information and chemical representation for molecular property prediction. *Front. Mol. Biosci.* **10**, 1216765 (2023).
31. Deng, D. et al. Describe molecules by a heterogeneous graph neural network with transformer-like attention for supervised property predictions. *ACS Omega*. **7**, 3713–3721 (2022).
32. Jiang, J. et al. TranGRU: focusing on both the local and global information of molecules for molecular property prediction. *Appl. Intell. (Dordr).* **53**, 15246–15260 (2023).
33. Wang, T. et al. AttenSyn: an attention-based deep graph neural network for anticancer synergistic drug combination prediction. *J. Chem. Inf. Model.* **64**, 2854–2862 (2023).
34. Xu, J. et al. Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Briefings Bioinf.* **22**, bbab083 (2021).
35. Rentzsch, R. et al. Predicting bacterial virulence factors—evaluation of machine learning and negative data strategies. *Briefings Bioinf.* **21**, 1596–1608 (2020).
36. Rzycki, M. et al. Molecular guidelines for promising antimicrobial agents. *Sci. Rep.* **4**, 55418 (2024).
37. Miethke, M. et al. Towards the sustainable discovery and development of new antibiotics. *Nat. Rev. Chem.* **5**, 00313 (2021).
38. Polton, D. J. Installation and operational experiences with MACCS (molecular access system). *Online Rev.* **6**, 235–242 (1982).
39. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
40. Zheng, S. et al. Machine learning–enabled virtual screening indicates the anti-tuberculosis activity of aldoxorubicin and Quarfloxin with verification by molecular docking, molecular dynamics simulations, and biological evaluations. *Briefings Bioinf.* **26**, 1 (2025).

Acknowledgements

This work is supported by Special Fund to Promote High Quality Development of the Marine and Fisheries Industry of Fujian Province under Grant FJHYF-ZH-2023-09; in part by Science and Technology Program of Quanzhou under Grant 2023GZ6; in part by the Natural Science Foundation of Fujian Province under Grant 2022J01499; in part by the STS Project of CAS and Fujian Province under Grant 2024T3038, 2024T3060 and 2024T3062.

Author contributions

SY and BL contributed to conception and design of the work. SY and BZ organized the database. SY and BL performed the statistical analysis. SY wrote the first draft of the manuscript. SY wrote sections of the manuscript. BL played a guiding role in the work and was responsible for ensuring that the descriptions were accurate. All authors contributed to manuscript revision, read, and approved the submitted version.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.L.