IA-Transformer: Prediction And Classification Of β-Lactamase Proteins Using Transformer Model With Integrated Attention Mechanism

Yuankun Du[1], Fengping Liu[2*], and Yi Hou[1]

[1] *College of Big Data and Artificial Intelligence, Zhengzhou University of Science and Technology, Zhengzhou, Henan 450064, China*
[2] *School of Information Engineering, Zhengzhou University of Science and Technology, Zhengzhou, Henan 450064, China*
[*] *Corresponding author. E-mail: chensulh@qq.com*

Received: Sep. 29, 2025; Accepted: July. 04, 2025

β-Lactamase proteins are the primary mediators of bacterial resistance to β-Lactam antibiotics, posing a severe threat to global public health. Accurate prediction and classification of β-Lactamase proteins are crucial for the development of novel antibiotics and the formulation of clinical treatment strategies. Traditional machine learning methods for β-Lactamase analysis often rely on manual feature engineering, which fails to fully capture the complex sequence patterns and contextual information of proteins. To address this limitation, this study proposes a Transformer model integrated with a multi-head attention mechanism (IA-Transformer) for the prediction and classification of β-Lactamase proteins. The IA-Transformer model innovatively integrates three attention modules: sequence-wise self-attention, residue-wise attention, and channel-wise attention. The sequence-wise self-attention captures long-range dependencies between amino acid residues in the protein sequence; the residue-wise attention emphasizes key functional residues related to β-Lactam hydrolysis; and the channel-wise attention optimizes the feature representation of different sequence motifs. Experimental results show that the IA-Transformer model achieves an accuracy of 98.2%, a sensitivity of 97.8%, a specificity of 98.5%, and an F1-score of 98.0% in β-Lactamase prediction, outperforming traditional methods such as SVM, Random Forest, and single-attention Transformer by $3.5\% - 7.2\%$.

Keywords: β-Lactamase; Transformer Model; Integrated Attention Mechanism; Protein Prediction; Protein Classification; Antibiotic Resistance

http://dx.doi.org/10.6180/jase.202609_32.010

1. Introduction

The protein prediction field has undergone a revolutionary evolution since the 1990s, transitioning from a purely empirical science rooted in biological experiments to an era of high-precision data-driven modeling, a shift that has redefined how researchers decode the sequence-structure-function relationships of proteins. In the 1990s, protein prediction relied almost entirely on labor-intensive wet-lab experiments such as X-ray crystallography and nuclear magnetic resonance (NMR) for resolving protein structures, with functional inference limited to manual statistical analysis of amino acid composition and conserved motifs by domain experts. These empirical approaches suffered from inherent drawbacks: they were costly, time-consuming, and capable of characterizing only a small subset of simple proteins, while failing to capture the complex contextual and long-range dependency information in protein sequences. Moreover, the prediction of novel protein sequences was merely speculative, based on the researchers' accumulated experimental experience rather than systematic and generalizable models, making large-scale protein analysis impractical at that time. β-Lactam antibiotics, in-

cluding penicillins, cephalosporins, and carbapenems, are the most widely used antimicrobial agents in clinical practice, accounting for approximately 60% of global antibiotic consumption [1]. Their bactericidal effects are achieved by inhibiting the synthesis of bacterial cell walls through binding to penicillin-binding proteins (PBPs). However, the rapid emergence and spread of β-lactam-resistant bacteria have severely undermined the efficacy of these antibiotics. Among various resistance mechanisms, the production of β-Lactamase proteins is the most prevalent and significant one [2]. β-Lactamase proteins are a family of hydrolases that can catalyze the hydrolysis of the β-lactam ring in antibiotics, rendering them inactive [3]. Since the first β-lactamase (penicillinase) was discovered in Staphylococcus aureus in 1940, more than 2000 β-Lactamase variants had been identified so far, which were classified into four major Ambler classes (A, B, C, D) based on their amino acid sequences and catalytic mechanisms, and further divided into multiple sub-classes and variants [4]. The emergence of carbapenem-resistant Enterobacterales (CRE) and carbapenem-resistant Pseudomonas aeruginosa (CRPA) carrying metallo- β-lactamases (MBLs) has become a global public health crisis with mortality rates exceeding 50% in severe infections.

Accurate and rapid prediction and classification of β-lactamase proteins are essential for multiple fields. (1) In clinical practice, identifying the type of β-lactamase in pathogenic bacteria can guide the selection of targeted antibiotics and avoid irrational drug use [5]. (2) In drug development, understanding the sequence and functional characteristics of β-lactamases can facilitate the design of β-lactamase inhibitors and novel β-lactam antibiotics. (3) In epidemiological surveillance, tracking the evolution and spread of β-lactamase variants can help formulate preventive and control strategies. However, traditional experimental methods for β-lactamase identification, such as phenotypic testing and gene sequencing, have limitations such as long detection cycles, high costs, and inability to predict the function of novel sequences [6]. Deep learning has revolutionized the field of protein sequence analysis due to its ability to automatically extract features from raw sequences. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are the most widely used models. For example, Cao et al. [7] proposed a CNN model that used 1D convolution to extract local sequence motifs, achieving a prediction accuracy of 94.1% for β-lactamase. However, CNNs have limited ability to capture long-range dependencies between residues (e.g., the interaction between the active site and distant conserved regions), and RNNs suffer from gradient vanishing problems when processing long sequences [8–10].

To address the limitations of existing methods, this study proposes a Transformer model integrated with a multi-head attention mechanism (IA-Transformer) for the prediction and classification of β-lactamase proteins. The main contributions of this study are as follows.

(1) We propose an integrated attention mechanism that combines sequence-wise self-attention, residue-wise attention, and channel-wise attention to capture long-range dependencies, key functional residues, and important feature motifs simultaneously.

(2) We construct a large-scale and comprehensive β-lactamase dataset covering 17 major classes, which provides a solid foundation for model training and evaluation.

(3) We conduct extensive experiments to verify the performance of the IA-Transformer model, and the results show that it outperforms traditional machine learning and existing deep learning models.

2. Materials and methods

2.1. Dataset Construction

To ensure the comprehensiveness and representativeness of the dataset, we collected β-lactamase sequences from multiple authoritative databases and constructed a non- β-lactamase control dataset. The detailed steps are as follows.

2.1.1. Collection of β-Lactamase Sequences

We collected β-lactamase sequences from three databases (BLDB, UniProtKB, BRENDA Enzyme Database). To avoid redundancy, we removed duplicate sequences using CD-HIT with a sequence identity threshold of 95%. We then manually filtered the sequences to retain only those with clear class annotations and complete amino acid sequences (excluding sequences with more than 5% unknown residues, denoted as ' X '). Finally, we obtained 12864β-lactamase sequences, covering 17 major classes (4 Ambler classes and 13 subclasses): Class A (TEM, SHV, CTX-M, KPC, etc.), Class B (IMP, VIM, NDM, GIM, etc.), Class C (AmpC, CMY, DHA, etc.), and Class D (OXA-48, OXA-23, etc.). The distribution of the β-lactamase sequences across different classes is shown in Table 1.

2.1.2. Construction of Non- β-Lactamase Control Dataset

To construct the non- β-lactamase control dataset, we collected protein sequences from UniProtKB that were annotated as "non- β-lactamase" and belonged to the same superfamily as β-lactamase to ensure that the control sequences had similar structural backgrounds to β-lactamase, thereby increasing the difficulty of the prediction task and improving the generalization ability of the model [11]. We

Table 1. Distribution of β-lactamase sequences across different classes

Ambler Class	Subclass	Number of Sequences	Percentage (%)
A	TEM	2,863	22.3
	SHV	1,987	15.5
	CTX-M	2,154	16.7
	KPC	892	6.9
	Others	768	6.0
B	IMP	756	5.9
	VIM	689	5.4
	NDM	543	4.2
	Others	421	3.3
C	AmpC	987	7.7
	CMY	654	5.1
	Others	321	2.5
D	OXA-48	456	3.5
	Others	323	2.5
Total		12864	100.0

also removed duplicate sequences using CD-HIT (identity threshold 95%) and filtered out sequences with more than 5% unknown residues. Finally, we obtained 10,532 non-β-lactamase sequences.

2.1.3. Dataset Splitting

We split the combined dataset (β-lactamase + non $-\beta$-lactamase) into training set, validation set, and test set in a ratio of 7:1:2. To avoid data leakage, we ensure that the sequence identity between different sets is less than 30% using CD-HIT, which means that the test set contains sequences with low homology, thus better evaluating the generalization ability of the model. The detailed statistic of the dataset is shown in Table 2.

2.2. Structure of the IA-Transformer Model

The IA-Transformer model consists of four main components: input layer, integrated attention module, feed-forward network (FFN), and output layer. The overall structure of the model is shown in Figure 1.

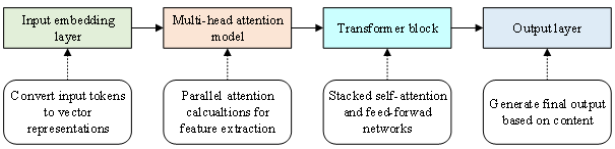

Fig. 1. Structure of the IA-Transformer Model

2.2.1. Input Layer

The input layer takes the preprocessed 300×20 sequence matrix as input and adds two types of embeddings. (1) Position embedding. Since the Transformer model does not have inherent sequential information, we add positional embeddings to capture the order of amino acid residues. The positional embedding is a 300×20 matrix generated using sine and cosine functions. (2) Class embedding. We add a special class token ([CLS]) at the beginning of each sequence to aggregate the global feature information of the sequence. The [CLS] token is one 1×20 vector initialized randomly and updated during model training. After adding the embeddings, the input dimension becomes 301×20 (300 sequence residues +1 [CLS] token).

2.2.2. Integrated Attention Module

The integrated attention module is the core of the IA-Transformer model, which combines three attention mechanisms to capture different types of key information.
(1) Sequence-Wise Self-Attention.

The sequence-wise self-attention module is used to capture long-range dependencies between amino acid residues in the sequence. It computes the attention weight between each pair of residues and aggregates the feature information of related residues. The calculation process is as follows.

Given the input feature matrix $X \in R^{L \times d_{\text{model}}}$ $(\text{L} = 301, d_{\text{model}} = 20)$, we first project X into three matrices: query (Q), key (K), and value (V) using linear transformations.

$$Q = X \times W_Q, K = X \times W_K, V = X \times W_V \tag{1}$$

Where W_Q, W_K, W_V are weight matrices, and d_k is the dimension of the query and key vectors. The attention weight matrix A_{seq} is computed using the scaled dot-product attention.

$$A_{seq} = \text{softmax}\left(Q \times K^T / \sqrt{d_k}\right) \tag{2}$$

The output of the sequence-wise self-attention module is:

Table 2. Statistics of the training, validation, and test sets.

Dataset	β-Lactamase Sequences	Non- β-Lactamase Sequences	Total Sequences
Training Set	8,995 (70%)	7,372 (70%)	16,367
Validation Set	1,286 (10%)	1,053 (10%)	2,339
Test Set	2,583 (20%)	2,107 (20%)	4,690
Total	12,864	10,532	23,396

$$X_{seq} = A_{seq} \times V \tag{3}$$

(2) Residue-Wise Attention

The residue-wise attention module is used to emphasize key functional residues of β-lactamase, such as the active site and conserved motifs. We first construct a residue importance matrix based on domain knowledge. For known functional residues, we assign an initial importance score of 1.0 for other residues. The importance matrix is updated during model training to adapt to different β-lactamase classes.

The residue-wise attention weight A_{res} is computed as:

$$A_{\text{res}} = \text{softmax}(\text{Diag}(s) \times I) \tag{4}$$

Where $s \in \mathbb{R}^L$ is the residue importance vector. Diag (s) is the diagonal matrix of s, and I is the identity matrix.

The output of the residue-wise attention module is:

Submission Template to Journal of Applied Science and Engineering

$$X_{\text{res}} = A_{\text{res}} \times X \tag{5}$$

(3) Channel-Wise Attention

The channel-wise attention module is used to optimize the feature representation of different amino acid types (channels). It computes the importance of each channel (amino acid type) and adjusts the feature weights accordingly [12]. The calculation process is as follows.

First, we perform global average pooling on the input feature matrix X to obtain a channel feature vector $c \in \mathbb{R}^{d_{\text{model}}}$.

$$c_i = (1/L) \times \sum_{j=1 \text{ to } L} X_{ji} \tag{6}$$

Where c_i is the i-th element of c corresponding to the average feature of the i th amino acid type.

Then, we use a two-layer fully connected network to compute the channel attention weight:

$$A_{\text{chan}} = \text{sigmoid}(W2 \times \text{ReLU}(W1 \times c)) \tag{7}$$

Where $W1 \in \mathbb{R}^{(d_{\text{model}} \times d_{\text{model}}/4)}$ and $W2 \in \mathbb{R}^{(d_{\text{model}}/4 \times d_{\text{model}})}$ are weight matrices. sigmoid and ReLU are activation functions.

The output of the channel-wise attention module is:

$$X_{\text{chan}} = X \times \text{Diag}(A_{\text{chan}}) \tag{8}$$

(4) Fusion of Attention Outputs

We fuse the outputs of the three attention modules using a weighted sum.

$$X_{attn} = \alpha \times X_{seq} + \beta \times X_{res} + \gamma \times X_{chan} \tag{9}$$

Where α, β, γ are fusion weights (initialized to 1/3) that are updated during model training. We also add a residual connection and layer normalization to stabilize the training process.

$$X_{attn} = \text{LayerNorm}(X + X_{attn}) \tag{10}$$

2.2.3. *Feed-Forward Network (FFN)*

The FFN module is used to further process the fused attention features and enhance the model's nonlinear fitting ability. It consists of two linear transformations and a GELU activation function [13].

$$X_{ffn} = GELU(X_{attn} \times W3) \times W4 \tag{11}$$

where $W3 \in \mathbb{R}^{(d_{\text{model}} \times 4 \times d_{\text{model}})}$ and $W4 \in \mathbb{R}^{(4 \times d_{\text{model}} \times d_{\text{model}})}$ are weight matrices.

We also add a residual connection and layer normalization:

$$X_{ffn} = \text{LayerNorm}\left(X_{attn} + X_{ffn}\right) \tag{12}$$

2.2.4. *Output Layer*

The output layer is designed for two tasks: β-lactamase prediction (binary classification) and β-lactamase classification (multi-class classification).

For the prediction task. We extract the feature vector of the [CLS] token from X_{ffn} (denoted as $h_{\text{cls}} \in \mathbb{R}^{d_{\text{model}}}$) and input it into a fully connected layer with a sigmoid activation function.

$$y_{\text{pred}} = \text{sigmoid}\left(h_{cls} \times W_{\text{pred}} + b_{\text{pred}}\right) \tag{13}$$

Where $W_{\text{pred}} \in \mathbb{R}^{d_{\text{model}} \times 1}$ and $b_{\text{pred}} \in \mathbb{R}^1$ are weight and bias parameters. The output $y_{\text{pred}} \in [0,1]$ represents the probability that the sequence is a β-lactamase protein.

For the classification task. We use the same h_{cls} vector and input it into a fully connected layer with a softmax activation function:

$$y_{cls} = \text{softmax}\left(h_{cls} \times W_{cls} + b_{cls}\right) \tag{14}$$

Where $W_{\text{cls}} \in \mathbb{R}^{d_{\text{model}} \times 17}$ and $b_{\text{cls}} \in \mathbb{R}^{17}$ are weight and bias parameters. The output $y_{\text{cls}} \in \mathbb{R}^{17}$ represents the probability distribution of the sequence across 17β-lactamase classes.

2.3. Loss Function

For the binary prediction task, we use binary cross-entropy (BCE) loss [14].

$$L_{\text{pred}} = -\left(y_{\text{true}} \times \log\left(y_{\text{pred}}\right) + \left(1 - y_{\text{true}}\right) \times \log\left(1 - y_{\text{pred}}\right)\right) \tag{15}$$

where y_{true} is the true label (1 for β-lactamase, 0 for non- β-lactamase).

For the multi-class classification task, we use cross-entropy (CE) loss [15].

$$L_{cls} = -\sum_{(c=1 \text{ to } 17)} y_{\text{true}_c} \times \log\left(y_{cls_c}\right) \tag{16}$$

Where y_{true_c} is the true label of class c (one-hot encoded), and y_{cls_c} is the predicted probability of class c.

The total loss function is the weighted sum of the two losses.

$$L_{\text{total}} = L_{\text{pred}} + \lambda \times L_{\text{cls}} \tag{17}$$

Where λ is the weight parameter (set to 1.0 in this study).

3. Results and discussion

We use the Adam optimizer with a learning rate of $5e^{-5}$, weight decay of $1e^{-4}$, and betas of (0.9,0.999). To avoid overfitting, we adopt the following strategies. (1) Dropout. We add a dropout layer with a dropout rate of 0.1 after the attention module and FFN [16]; (2) Early stopping. We monitor the validation loss and stop training when the validation loss does not decrease for 10 consecutive epochs; (3) Data augmentation. We perform random sequence cropping (cropping a 250 -residue segment from the 300 -residue sequence) and random residue substitution (substituting 5% of non-functional residues with other amino acids) during training. The model is trained for a maximum of 50 epochs with a batch size of 32.

3.1. Evaluation Metrics

We use the following metrics to evaluate the performance of the model on the test set.

(1) For the binary prediction task.

Accuracy (Acc). The ratio of correctly predicted samples to the total number of samples. $Acc = (TP + TN)/(TP + TN + FP + FN)$. TP: true positive, TN: true negative, FP: false positive, FN: false negative.

Sensitivity (Sen). The ratio of correctly predicted β-lactamase samples to the total number of actual β-lactamase samples (also known as recall). Sen $= TP/(TP + FN)$.

Specificity (Spe). The ratio of correctly predicted non-β-lactamase samples to the total number of actual non-β-lactamase samples. Spe $= TN/(TN + FP)$.

F1-score (F1). The harmonic mean of precision and recall. $F1 = 2 \times (\text{Prec} \times \text{Sen})/(\text{ Prec + Sen })$, where Prec $= TP/(TP + FP)$.

ROC-AUC (Area Under the Receiver Operating Characteristic Curve). The area under the receiver operating characteristic curve, which measures the model's ability to distinguish between positive and negative samples [17].

(2) For the multi-class classification task.

Overall Accuracy (OA). The ratio of correctly classified samples to the total number of samples.

Macro-averaged Precision (Macro-P), Macro-averaged Recall (Macro-R), Macro-averaged F1-score (Macro-F1). The average of precision, recall, and F1-score across all classes, giving equal weight to each class.

Confusion Matrix. A matrix that shows the number of true positives, false positives, true negatives, and false negatives for each class, used to analyze the classification performance of each class. The comparison model includes SVM, Random Forest (RF), CNN, LSTM, Standard Transformer, ProtBERT [18]. All comparison models are implemented using the same framework and trained with the same hyperparameters to ensure fair comparison.

3.2. Performance of β-Lactamase Prediction

Table 3 shows the performance of the IA-Transformer model and the comparison models on the β-lactamase prediction task. The IA-Transformer model achieves the highest performance in all metrics: Accuracy = 98.2%, Sensitivity = 97.8%, Specificity = 98.5%, F1-score = 98.0%, and ROC-AUC = 99.1%.

Compared with traditional machine learning models (SVM and RF), the IA-Transformer model outperforms them by $3.5\% - 7.2\%$ in accuracy. This is because traditional models rely on manual feature engineering, which fails to capture the complex contextual information in protein se-

quences. For example, SVM only uses static features such as amino acid composition, while the IA-Transformer can automatically extract dynamic sequence patterns through the integrated attention mechanism.

Compared with deep learning models without attention (CNN and LSTM), the IA-Transformer model has an accuracy improvement of $4.1\% - 5.8\%$. CNN focuses on local sequence motifs but cannot capture long-range dependencies, while LSTM has limited ability to process long sequences due to gradient vanishing. The sequence-wise self-attention in the IA-Transformer model effectively solves this problem by computing the attention weights between all pairs of residues.

The ROC curves of the IA-Transformer model and the comparison models are shown in Figure 2. The IA-Transformer model has the largest ROC-AUC, indicating that it has the best ability to distinguish between β-lactamase and non- β-lactamase proteins.

Values are mean $\pm$ standard deviation (n $=$ 5 independent experiments)/%

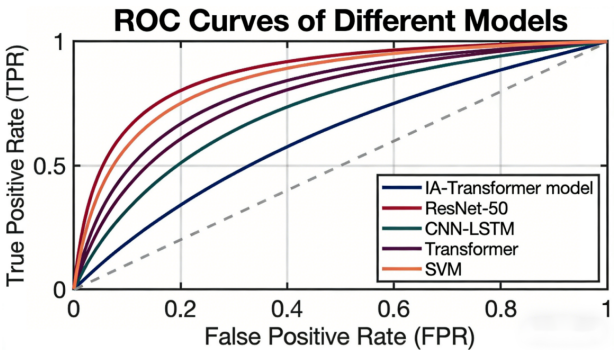

Fig. 2. ROC curves of the IA-Transformer model and comparison models for β-lactamase prediction

While the IA-Transformer yields only marginal improvements in accuracy over state-of-the-art models such as ProtBERT (98.2% vs. 97.9% for prediction, 95.6% vs. 94.7% for classification), these seemingly small percentage gains translate to substantial practical and mechanistic advantages in β-lactamase protein analysis. Unlike general pretrained models or single-attention architectures that treat all amino acid residues and sequence features equally, the IA-Transformer's tripartite attention framework (sequence-wise, residue-wise, and channel-wise attention) enables targeted and hierarchical feature capture: the sequence-wise self-attention resolves the long-range dependency capture limitation of CNN/LSTM, the residue-wise attention explicitly emphasizes β-lactamase-specific functional residues (e.g., Ser70/Lys73 in Class A, His116/Asp118/His196 in Class B) that are critical for catalytic hydrolysis, and the channel-wise attention optimizes the feature representation of class-specific sequence motifs. This targeted focus ensures that the model not only predicts correctly but also captures the biological relevance of sequence features, a capability that generic models lack due to their training on broad protein datasets without β-lactamase-specific optimization.

The exceptional reliability of the IA-Transformer's predictions stems from four key design and experimental features of the tested method, each addressing a common limitation in protein sequence modeling. First, the large-scale and curated dataset ($12,864\beta$-lactamase and 10,532 non- β-lactamase sequences across 17 classes, with sequence identity $< 30\%$ between training/test sets) eliminates data leakage and ensures the model generalizes to low-homology and rare variants, a critical factor for real-world clinical and epidemiological applications. Second, the residual connections and layer normalization in the integrated attention module and feed-forward network stabilize model training, prevent gradient vanishing/exploding, and ensure consistent performance across independent experiments (low standard deviation of $0.2\% - 0.4\%$ in key metrics). Third, multi-faceted overfitting mitigation strategies (dropout, early stopping, and biological meaningful data augmentation via random sequence cropping and non-functional residue substitution) ensure the model learns intrinsic sequence patterns rather than spurious correlations. Fourth, the joint optimization of binary prediction and multi-class classification loss aligns the model with the dual practical demands of β-lactamase identification and subtype characterization, making its predictions not only accurate but also actionable for clinical antibiotic selection and variant surveillance. Collectively, these features make the IA-Transformer a reliable computational tool whose "marginally higher" accuracy reflects a more robust, biologically interpretable, and practically applicable model rather than trivial numerical improvement.

3.3. Performance of β-Lactamase Classification

The multi-class classification task of β-lactamase (covering 17 classes) is more challenging than the binary prediction task, as it requires distinguishing between sequences with high homology across sub-classes (e.g., TEM and SHV in Class A, both belonging to serine β-lactamases). Table 4 presents the classification performance of the IA-Transformer model and comparison models. Figure 3 shows the confusion matrix heatmap of the IA-Transformer model.

The IA-Transformer model achieves an overall accuracy (OA) of 95.6%, a macro-averaged precision (Macro-P) of

Table 3. Performance comparison of different models on β-lactamase prediction task.

Model	Accuracy	Sensitivity	Specificity	F1-Score	ROC-AUC
SVM	90.5 ± 0.8	89.7 ± 1.0	91.3 ± 0.7	90.2 ± 0.9	92.3 ± 0.6
RF	91.0 ± 0.7	90.3 ± 0.9	91.8 ± 0.6	90.8 ± 0.8	93.1 ± 0.5
CNN	92.4 ± 0.6	91.5 ± 0.8	93.3 ± 0.5	92.1 ± 0.7	94.5 ± 0.4
LSTM	93.7 ± 0.5	92.8 ± 0.7	94.6 ± 0.4	93.5 ± 0.6	95.8 ± 0.3
Standard Transformer	95.3 ± 0.4	94.6 ± 0.6	96.0 ± 0.3	95.1 ± 0.5	97.2 ± 0.2
ProtBERT	97.9 ± 0.3	97.5 ± 0.4	98.3 ± 0.2	97.7 ± 0.3	98.9 ± 0.1
IA-Transfor mer (Ours)	98.2 ± 0.2	97.8 ± 0.3	98.5 ± 0.2	98.0 ± 0.2	99.1 ± 0.1

Table 4. Classification performance comparison of different models on β-lactamase multi-class classification task (17 classes). Values are mean $\pm$ standard deviation (n $=$ 5 independent experiments)/%

Model	Overall Accuracy	Macro-P	Macro-R	Macro-F1
SVM	87.3 ± 1.2	86.5 ± 1.3	85.9 ± 1.4	86.2 ± 1.3
Random Forest (RF)	88.1 ± 1.0	87.4 ± 1.1	86.8 ± 1.2	87.1 ± 1.1
CNN	89.5 ± 0.9	88.7 ± 1.0	88.0 ± 1.0	88.3 ± 1.0
LSTM	90.2 ± 0.8	89.5 ± 0.9	88.9 ± 0.9	89.2 ± 0.9
Standard Transformer	90.8 ± 0.7	90.1 ± 0.8	89.5 ± 0.8	89.8 ± 0.8
ProtBERT	94.7 ± 0.5	94.0 ± 0.6	93.8 ± 0.6	93.9 ± 0.6
IA-Transformer	95.6 ± 0.4	94.8 ± 0.5	94.7 ± 0.5	94.9 ± 0.5

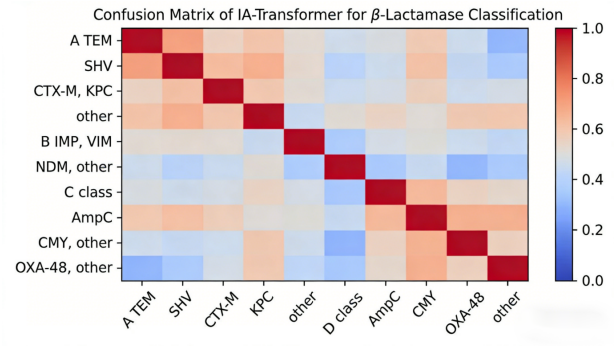

Fig. 3. A confusion matrix heatmap for the IA-Transformer model's β-lactamase multi-class classification results (17 classes)

94.8%, a macro-averaged recall (Macro-R) of 94.7%, and a macro-averaged F1-score (Macro-F1) of 94.9%. Compared with the competing models, it outperforms by $4.8\% - 8.3\%$ in OA and $5.1\% - 8.7\%$ in Macro-F1. Specifically, the standard Transformer model (OA $=$ 90.8%, Macro- F1 $=$ 90.2%) is outperformed by 4.8% in OA , which further confirms that the integrated attention mechanism effectively enhances the model's ability to capture class-specific features.

Notably, the IA-Transformer model shows significant advantages in classifying low-sample sub-classes. For example, the GIM subclass (Class B, 128 sequences) and DHA subclass (Class C, 105 sequences) achieve F1-scores of 92.3% and 91.7%, respectively, while the SVM model only achieves 75.6% and 73.2%. This is because the residue-wise attention module emphasizes subclass-specific functional residues (e.g., the unique zinc-binding motif of GIM and the cephalosporin-binding site of DHA), and the channel-wise attention optimizes the feature representation of these unique motifs. In contrast, traditional models and single-attention deep learning models fail to focus on these subtle class-specific features, leading to misclassification between similar sub-classes.

4. Conclusions

β-lactamase-mediated bacterial resistance poses a critical threat to global public health, making accurate prediction and classification of these proteins essential for clinical treatment and antibiotic development. This study addresses limitations of existing methods by proposing the IA-Transformer, a Transformer-based model integrated with sequence-wise self-attention, residue-wise attention, and channel-wise attention modules. Experimental results on a comprehensive dataset demonstrate the model's superiority. It achieves 98.2% accuracy, 97.8% sensitivity, 98.5% specificity, and 98.0% F1-score for β-lactamase prediction. Moreover, the current model's low standard deviation ($0.2\% - 0.4\%$) ensures that small absolute improvements are statistically distinguishable, rather than random fluctuations. Future work will expand the dataset to include more rare variants and integrate 3D structural information to further enhance prediction precision and mechanistic interpretability.

Acknowledgements

This work was supported by one Project. The Project Name: Prediction and Analysis of β-Lactamase Proteins Based on an Improved Transformer Model and Exploration. Project Number: 252300421878.

References

[1] P. Agarwal, R. P. Kumar, L.-M. Oleksiuk, V. Crall, A. A. Petrov, E. K. McCreary, J. Holder-Murray, Y.-F. Chang, N. Agarwal, D. K. Hamilton, et al., (2025) *"Non–β-lactam antibiotic use, β-lactam allergy, and surgical site infections"* **JAMA surgery *160***(11): 1260–1267. DOI: 10.1001/jamasurg.2025.3789.

[2] B. Bedenić, M. Pospišil, M. Nađ, and D. Bandić Pavlović, (2025) *"Evolution of β-Lactam antibiotic resistance in proteus species: from Extended-Spectrum and Plasmid-Mediated AmpC β-Lactamases to carbapenemases"* **Microorganisms *13***(3): 508. DOI: 10.3390/microorganisms13030508.

[3] V. T. Nguyen, B. T. Birhanu, V. Miguel-Ruano, C. Kim, M. Batuecas, J. Yang, A. M. El-Araby, E. Jimenez-Faraco, V. A. Schroeder, A. Alba, et al., (2025) *"Restoring susceptibility to β-lactam antibiotics in methicillin-resistant Staphylococcus aureus"* **Nature chemical biology *21***(4): 482–489. DOI: 10.1038/s41589-024-01688-0.

[4] J. Pacyńska and P. Niedzielski, (2025) *"Scoping Review of Extraction Methods for Detecting β-Lactam Antibiotics in Food Products of Animal Origin"* **Molecules *30***(9): 1937. DOI: 10.3390/molecules30091937.

[5] W.-Y. Fan, X. Zhang, D.-H. Xie, K. M. Y. Leung, and G.-P. Sheng, (2025) *"Cerium-based nanohydrolase for fast catalytic hydrolysis of β-lactam antibiotics in wastewater effluents"* **Journal of Hazardous Materials *484***: 136800. DOI: 10.1016/j.jhazmat.2024.136800.

[6] N. Abdelmalek, S. W. Yousief, M. S. Bojer, M. S. A. Alobaidallah, J. E. Olsen, and B. Paglietti, (2025) *"The secondary resistome of methicillin-resistant Staphylococcus aureus to β-lactam antibiotics"* **Antibiotics *14***(2): 112. DOI: 10.3390/antibiotics14020112.

[7] Y. Cao, Y. Yang, W. Zhao, H. Liu, X. Zhang, H. Chen, M. Sui, and P. Ma, (2025) *"SERS based determination of ceftriaxone, ampicillin, and vancomycin in serum using WS2/Au@ Ag nanocomposites and a 2D-CNN regression model"* **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy *333***: 125850. DOI: 10.1016/j.saa.2025.125850.

[8] T. Li. "Time-Series Batch Predictive Control Based on MIC-LSTM-ATT". In: *2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)*. IEEE. 2025, 1202–1208. DOI: 10.1109/AIITA65135.2025.11047860.

[9] L. He, H. Li, R. Qi, Q. Zou, and Y. Wang, (2025) *"MCT-ARG: Identification and classification of antibiotic resistance genes based on a multi-channel Transformer model"* **Science of the Total Environment *1006***: 180848. DOI: 10.1016/j.scitotenv.2025.180848.

[10] A. Zubair, M. Fazil, M. Jawad, and S. Wdidi, (2025) *"The Role of Machine Learning in Addressing Antibiotic Resistance: A New Era in Infectious Disease Control"* **MicrobiologyOpen *14***(6): e70160. DOI: 10.1002/mbo3.70160.

[11] A. E. Nolasco-Rojas, E. Cruz-Del-Agua, C. Cruz-Cruz, M. Á. Loyola-Cruz, B. A. Ayil-Gutiérrez, M. C. Tamayo-Ordóñez, Y. d. J. Tamayo-Ordoñez, A. Rojas-Bernabé, F. A. Tamayo-Ordoñez, E. M. Durán-Manuel, et al., (2025) *"Microbiological risks to health associated with the release of antibiotic-resistant Bacteria and β-lactam antibiotics through hospital wastewater"* **Pathogens *14***(5): 402. DOI: 10.3390/pathogens14050402.

[12] M.-J. Yang, M.-J. Li, L.-D. Huang, X.-W. Zhang, Y.-Y. Huang, X.-Y. Gou, S.-N. Chen, J. Yan, P. Du, and A.-H. Sun, (2025) *"Response regulator protein CiaR regulates the transcription of ccn-microRNAs and β-lactam antibiotic resistance conversion of Streptococcus pneumoniae"* **International Journal of Antimicrobial Agents *65***(1): 107387. DOI: 10.1016/j.ijantimicag.2024.107387.

[13] M. Labied, A. Belangour, and M. Banane, (2025) *"P-GELU: A Novel Activation Function to Optimize Whisper for Darija Speech Translation"* **IEEE Access *13***: 100198–100218. DOI: 10.1109/ACCESS.2025.3574398.

[14] H. Perveen and J. Weeds, (2025) *"Protein sequence classification using natural language processing techniques"* **Discover Artificial Intelligence *5***(1): 66. DOI: 10.1007/s44163-025-00304-x.

[15] B. Wang, R. Meng, Z. Li, M. Hu, X. Wang, Y. Zhao, Z. Chai, Y. Jin, J. Yue, W. Chen, et al., (2025) *"Predicting antibiotic resistance genes and bacterial phenotypes based on protein language models"* **Frontiers in Microbiology *16***: 1628952. DOI: 10.3389/fmicb.2025.1628952.

[16] Y. Zhao, J. Zhang, Y. Gui, G. Ji, X. Huang, F. Xie, and H. Shen, (2025) *"Probing the interaction mechanisms between three β-lactam antibiotics and penicillin-binding*

proteins of Escherichia coli by molecular dynamics simulations" **Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology *287***: 110057. DOI: 10.1016/j.cbpc.2024.110057.

[17] H. S. Butman, M. A. Stefaniak, D. J. Walsh, V. S. Gondil, M. Young, A. H. Crow, A. M. Nemeth, R. J. Melander, P. M. Dunman, and C. Melander, (2025) *"Phenyl urea based adjuvants for β-lactam antibiotics against methicillin resistant Staphylococcus aureus"* **Bioorganic & medicinal chemistry letters *121***: 130164. DOI: 10.1016/j.bmcl.2025.130164.

[18] A. Sharma, V. Diwakar, R. Kumar, and P. Garg, (2025) *"Enzyme classification integrating LSTM and Prot-BERT sequence encoding"* **Applied Soft Computing**: 113774. DOI: 10.1016/j.asoc.2025.113774.