2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782350

Leveraging Graph Neural Networks for MIC Prediction in Antimicrobial Resistance Studies

Zonghan Zhang[1], Ramyasri Veerapaneni[1], Moses Ayoola[2], Athish Ram Das[2], Zhiqian Chen[1]
Bindu Nanduri[2], Mahalingam Ramkumar[1]

***Abstract*—Antimicrobial resistance (AMR) poses a significant challenge in healthcare and public health, with organisms such as nontyphoidal Salmonella leading the way due to their escalating resistance to antimicrobial agents. This situation severely complicates the management and containment of diseases, highlighting the urgent need for more effective techniques to assess antimicrobial susceptibility. Conventional methods, including the broth microdilution technique for determining Minimum Inhibitory Concentrations (MICs), are time-consuming and require extensive manual effort. The advent of machine learning (ML) technologies offers a revolutionary approach to predicting MICs, thereby potentially increasing the efficacy of antimicrobial therapies. This paper explores the latest advancements in ML for MIC prediction, focusing on an innovative approach using Graph Neural Networks (GNNs), which could provide a novel insight into the correlation between gene fragment similarities and MIC values. Within this paper, we introduce the K-mer GNN, a novel GNN model designed for MIC prediction. The K-mer GNN model distinctively identifies and incorporates the similarities among k-mers, integrating these insights into GNN alongside k-mer features. This approach not only elevates the precision of MIC predictions but also sheds light on the genomic factors at the k-mer level that drive antimicrobial resistance.**

***Index Terms*—Graph Neural Networks, genomics, K-mer, Antibiotics, MIC**

I. Introduction

Antimicrobial resistance (AMR) is a growing concern in healthcare and public health [1]–[3]. The ability of pathogens like nontyphoidal Salmonella to resist antimicrobial treatments has significant implications for disease management and control. Traditional methods of determining antimicrobial susceptibility, such as broth dilution to ascertain Minimum Inhibitory Concentrations (MICs), are time-consuming and labor-intensive. Recent advances in machine learning (ML) offer a transformative approach to predicting MICs, thereby enhancing the effectiveness of antimicrobial therapies. This paper reviews the current state of ML in predicting antimicrobial MICs, focusing on nontyphoidal Salmonella, a key player in the AMR landscape.

Several studies have concentrated on employing machine learning (ML) strategies to forecast antimicrobial minimum inhibitory concentrations (MICs) and pinpoint genomic factors influencing antibiotic resistance in nontyphoidal Salmonella. These investigations span various methodologies, including the creation of portable detection systems, high-content imaging techniques, analysis of genomic features, and the application of bioinformatics, all aimed at tackling the escalating issue of antimicrobial resistance among bacterial pathogens [4]–[8]. Notably, the research conducted by [9] is distinguished by its successful application of ML to predict MICs in nontyphoidal Salmonella, establishing a foundational reference for our discussion. This is further enriched by other studies, together highlighting the potential and obstacles associated with the use of ML in this field. Graph Neural Networks (GNNs) have emerged as a significant focus in genomics research, and they are noted for their versatility across various applications. These applications range from the prediction of regulatory DNA and RNA sites to the integration of multi-omics data for purposes such as patient stratification and cancer prognosis [10]–[26]. These GNNs are utilized for groundbreaking purposes, such as analyzing thermodynamics in genome-scale metabolic networks and forecasting tumor metastasis through the integration of genomic and protein-protein interaction network data. Despite their increasing importance in the genomics domain, GNNs have yet to be explored or applied in the specific context of predicting MICs. This gap indicates a promising new direction for future investigations into antimicrobial resistance.

In the subsequent sections, we introduce a novel model, namely K-mer GNN, that leverages GNNs for predicting MICs, showcasing a significant advancement over traditional ML methodologies. GNNs, with their unique ability to model relational data, present a novel framework for analyzing complex interactions within microbial genomic data. This approach not only enhances the accuracy of MIC predictions but also offers more profound insights into the genomic factors influencing antimicrobial resistance.

II. Methods

A. Preliminary: Graph Convolutional Networks (GCN)

Graph Convolutional Networks (GCN), as the majority of GNNs, have revolutionized how we process and analyze data structured in graphs. GCNs operate by applying convolutional processes to graph-structured data, involving the aggregation of information from a node's neighbors to capture the graph's topological features effectively. The convolution in GCNs is mathematically represented as:

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

where $H^{(l)} \in R^{N\times F}$ is the feature matrix at layer l and F is a tunable hidden dimension of the neural network. $\hat{A}$ is the

[1]Computer Science and Engineering, Mississippi State University. {zz239, rv376}@msstate.edu, zchen@cse.msstate.edu, ramkumar@cse.msstate.edu
[2]Comparative Biomedical Science, College of Veterinary Medicine, Mississippi State University. {mba185,ar2903}@msstate.edu, BNanduri@cvm.msstate.edu

adjacency matrix with added self-connections ($\hat{A} = A + D$), which often represents the similarity or proximity among data instances. D is the degree matrix, $W^{(l)}$ the layer's weight matrix, and σ a non-linear activation function. H typically represents the normal features of data points (or nodes), while $\hat{A}$ denotes pairwise similarities among nodes. The fundamental capability of GCN lies in their ability to merge non-Euclidean connectivity, represented by $\hat{A}$, with Euclidean feature data, denoted as H, for conventional machine learning tasks, as shown in Equation 1. The predictive capability of GCNs is realized in the final output layer, which can be represented as:

$$Z = \text{softmax}\left(H^{(L)}W^{(L)}\right) \tag{2}$$

where Z is an intermediate representation, $H^{(L)}$ is the feature matrix at the last layer L, and $W^{(L)}$ is the weight matrix of the final layer. The softmax function is typically used in classification tasks to convert the output into probability distributions. This equation encapsulates the process where GCNs leverage the encoded features and relationships to make predictions, such as node classification, in graph-structured data. Note that this paper will perform a revised regression task, so we remove softmax in Equation :

$$Z = H^{(L)}W^{(L)} \tag{3}$$

It is essential to highlight that a graph convolutional layer is frequently combined with multiple perceptron (MLP) layers. In our investigation, we will experiment with configurations incorporating either one or two graph convolutional layers along with three MLP layers, a setup that is widely recognized as effective for GCN models.

B. Kmer-GNN: A Revised GCN Model

The Kmer-GNN model represents a new modification of GCN, specifically tailored for genomic data analysis in the context of antimicrobial resistance prediction. This section delves into the intricacies of the Kmer-GNN model, highlighting its novel aspects, particularly the graph structure and the node feature representation.

Architecture of Kmer-GNN. In Figure 1, the initial step involves extracting a selection of top K-mers to form a pool. This is followed by the derivation of their similarity graph and existence. Subsequently, the Kmer-GNN applies graph convolution, effectively integrating the similarity graph with the binary node characteristics. The Kmer-GNN architecture is further enhanced with multiple layers of graph convolution, which are essential for aggregating neighbor node information. Following this, the Kmer-GNN utilizes a series of fully connected layers that aid in the further refinement and processing of information, culminating in the prediction of MIC values.

Graph Structure in Kmer-GNN: Kmer-GNN utilizes a unique graph structure where nodes represent k-mers, and edges are established based on sequence similarity of k-mers. This structure is critical for capturing the complex relationships inherent in genomic sequences. Specifically, the

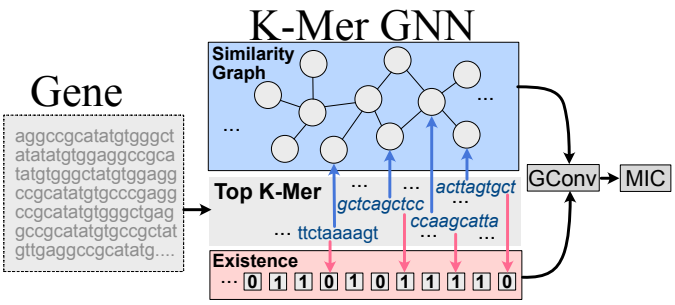

Fig. 1. K-Mer Graph Neural Networks

similarity graph in Kmer-GNN is constructed using sequence similarity such as cosine similarity and longest common sequence. This graph representation facilitates the extraction of meaningful patterns from K-mer sequences, a crucial step in predicting antimicrobial resistance. The construction of the similarity graph is mathematically represented as:

$$A_{ij} = \Phi(k_i, k_j) \tag{4}$$

where A_{ij} represents the edge weight between K-mers k_i and k_j. All similarities are normalized to the range [0, 1]. So, $A \in [0,1]^{N\times N}$ where N denote the total number of kmers. To further decrease the computational burden, a similarity threshold of 0.5 is established. Values below 0.5 are considered as 0, while those above 0.5 are regarded as 1. Finally, we have $A \in \{0,1\}^{N\times N}$, which means that each entry in A is either 0 or 1. If A is large, the matrix multiplication can result in a prohibitive computational cost. To address this issue, feature selection techniques can be applied to select the most informative 10% of all 10-mers, utilizing their feature importance scores obtained from a Random Forest algorithm. This approach will result in the identification of 1351 (i.e., N) 10-mers within our dataset.

Node Feature Representation. In the Kmer-GNN framework, the attributes of each node are encoded in a binary manner, reflecting the presence (1) or absence (0) of specific K-mers within the microbial genome. This binary encoding is fundamental for delineating the genomic landscape in the context of antimicrobial resistance analysis. For every node within the network, a binary feature is assigned, denoting the presence or lack thereof of the corresponding K-mer in the genome. This method provides a succinct yet comprehensive depiction of genomic information. The binary presence or absence of a K-mer is mathematically represented as follows:

$$H_i = [0 \text{ or } 1] \tag{5}$$

In this context, H_i represents the feature vector for node i, where the numerical values indicate the presence of specific K-mers within the pool of N K-mers, thus $H \in \{0,1\}^N$. Following this, Equations 4 and 5 are applied to incorporate these values into Equations 1 and 3 together. Note that the output of 3 is a vector Y. To enhance the outcomes, an attention layer is incorporated after the graph convolutional layers.

III. Datasets, Preprocessing, and Implementation

Salmonella MIC Prediction Model. The prediction model for Minimum Inhibitory Concentration (MIC) of *Salmonella*—encompassing 4,500 genomes, along with its software and comprehensive documentation—is publicly accessible at https://github.com/PATRIC3/mic_prediction [9]. The model's performance is quantitatively assessed with an accuracy metric defined by a ±1 two-fold dilution within a confidence interval of 95%. The analyzed strains were isolated from either raw retail meat and poultry or directly from livestock animals post-slaughter. Antimicrobial susceptibility was determined through broth microdilution techniques using the Sensititre system by Thermo Scientific, specifically for the antibiotic combination of trimethoprim-sulfamethoxazole (COT), in accordance with the protocols of FDA and USDA's NARMS laboratories. Whole-genome sequencing (WGS) was conducted using Illumina's HiSeq and MiSeq platforms, following standardized procedures.

Baseline Models. To highlight its relative efficacy, our approach was evaluated by conducting comparative analyses with well-established models, including Linear Regression (LR), Random Forest Regressor (RF), Support Vector Regressor (SVR), and XGBoost (XGB). These baseline models were developed utilizing Sklearn [27]. Additionally, the robustness and sensitivity of the proposed K-mer-based GNN were evaluated across various configurations, encompassing different graph similarity metrics (including Hamming distance [28], Levenshtein distance [29], and Needleman-Wunsch [30]), a range of k-fold cross-validation schemes (from 3 to 7 folds), and multiple hidden layer dimensions (16, 24, 32, and 48).

Training and Evaluation of Kmer-GNN. The training process for Kmer-GNN encompasses the utilization of a dataset comprising microbial genomes and their respective MIC values. Evaluation of the model's performance is predicated on its precision in predicting these MIC values. This is achieved by fine-tuning the selection of hyperparameters and implementing measures to curtail overfitting. The entire model configuration, along with its training and testing phases, is conducted using PyTorch Geometric [31]. The full code necessary to reproduce our study results has been made available at https://github.com/aquastar/kmer-gnn.

IV. Results and Analysis

Figure 2 provides a detailed comparison of the performance of various machine learning models, including GCN (Kmer-GNN), Logistic Regression (LR), Random Forest (RF), Support Vector Regression (SVR), and XGBoost (XGB), across three distinct sequence similarity metrics: Hamming Distance, Levenshtein Distance, and Needleman-Wunsch similarity. The performance of each model under different similarity metrics is depicted by bars, with error bars indicating the standard deviation. While RF, SVR, and XGB demonstrate high accuracies, the GCN model enhances performance by approximately 4-5%. Figure 3 depicts the efficacy of the same ensemble of machine learning models (GCN, LR, RF, SVR, and XGB) across various k-fold validation schemes, including 3-fold, 5-fold, and

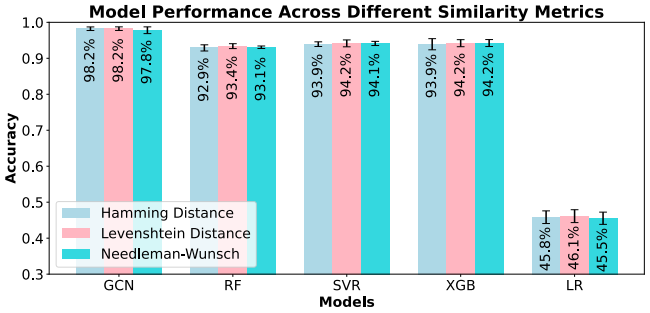

Fig. 2. Comparative Analysis of Machine Learning Models' Accuracy Across Different *Sequence Similarity Metrics*.

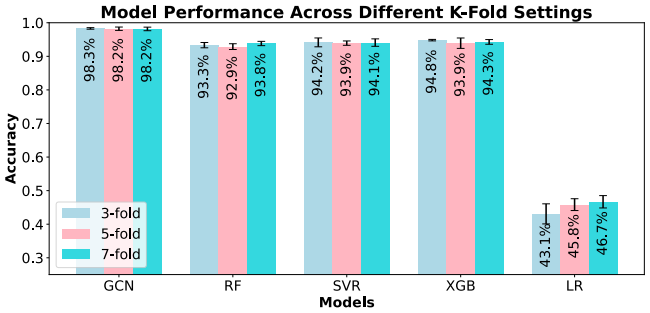

Fig. 3. Comparative Analysis of Machine Learning Models' Accuracy Across *K-Fold Settings*.

7-fold. This figure offers critical insights into the impact of the k-fold validation strategy on model performance, underscoring the significance of selecting an appropriate validation approach in the evaluation of models. Notably, there is no substantial difference in performance across different fold sizes, with the exception of Linear Regression (LR). RF, SVR, and XGB maintain high accuracy levels, yet GCN surpasses all baseline models by 4-5%. Figure 4 examines the performance of the machine learning models over a range of hidden dimension settings: 16, 24, 36, and 48. Consistently, the GCN model outperforms the baseline models by 4-5% across these settings.

Throughout these analyses, as demonstrated in all three

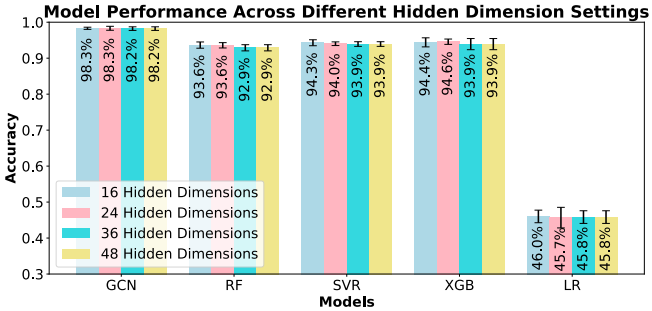

Fig. 4. Comparative Analysis of Machine Learning Models' Accuracy Across *Hidden Dimension Settings*

figures, the GCN model (Kmer-GNN) consistently exhibits superior performance with minimal variation in accuracy when compared to other models such as LR, RF, SVR, and XGB. Consequently, the GCN model is affirmed as a highly effective and adaptable model, outshining its counterparts under diverse experimental conditions and settings, thereby confirming its efficacy for tasks related to sequence similarity and genomic data analysis.

V. Conclusion

The paper has presented a comprehensive overview of the application of machine learning, specifically Graph Neural Networks (GNNs), in predicting antimicrobial minimum inhibitory concentrations (MICs) and resisting genomic determinants for nontyphoidal Salmonella. We described the underlying mechanisms, architecture, and evaluation of Kmer-GNN, a novel GNN model tailored for analyzing genomic sequences, and demonstrated its effectiveness across various experimental conditions compared to traditional machine learning models. The results clearly establish the superiority of GNNs, exemplified through Kmer-GNN, in identifying complex patterns in genomic data critical for accurate MIC prediction and understanding antimicrobial resistance. By offering enhanced predictive capabilities and deeper insights into genetic factors influencing pathogen behavior, the integration of graph-based deep learning methodologies promises to transform susceptibility testing protocols and antibiotic therapeutic strategies.

Acknowledgments

This research was supported by USDA-ARS NACA Projects 58-0200-0-002 and 58-6064-3-017, and was also partially supported by NSF IIS Award No. 2153369.

References

[1] Jay K Varma, Kåre Mølbak, Timothy J Barrett, James L Beebe, Timothy F Jones, Therese Rabatsky-Ehr, Kirk E Smith, Duc J Vugia, Hwa-Gan H Chang, and Frederick J Angulo. Antimicrobial-resistant nontyphoidal salmonella is associated with excess bloodstream infections and hospitalizations. *The Journal of infectious diseases*, 191(4):554–561, 2005.

[2] Jay K Varma, Katherine D Greene, Jessa Ovitt, Timothy J Barrett, Felicita Medalla, and Frederick J Angulo. Hospitalization and antimicrobial resistance in salmonella outbreaks, 1984–2002. *Emerging Infectious Diseases*, 11(6):943, 2005.

[3] Frederick J Angulo and Kåre Mølbak. Human health consequences of antimicrobial drug—resistant salmonella and other foodborne pathogens. *Clinical infectious diseases*, 41(11):1613–1620, 2005.

[4] S Ranganathan T Gojobori AM Khan, H Singh. Editorial on the research topic accelerating innovation to meet biological challenges: The role of bioinformatics. *Frontiers in Genetics*, 2024.

[5] K Jurica I Nastasijevic, F Proscia. Tracking antimicrobial resistance along the meat chain: One health context. *Food Reviews International*, 2023.

[6] S Sridhar S Reece O Lunguya S Baker, TA Tran. Combining machine learning with high-content imaging to infer ciprofloxacin susceptibility in clinical isolates of salmonella typhimurium. 2023.

[7] P Vohra A Chalka, TJ Dallman. The advantage of intergenic regions as genomic features for machine-learning-based host attribution of salmonella typhimurium from the usa. *Microbial Genomics*, 2023.

[8] YW Wu MR Yang, SF Su. Using bacterial pan-genome-based feature selection approach to improve the prediction of minimum inhibitory concentration (mic). *Frontiers in Genetics*, 2023.

[9] Marcus Nguyen, S Wesley Long, Patrick F McDermott, Randall J Olsen, Robert Olson, Rick L Stevens, Gregory H Tyson, Shaohua Zhao, and James J Davis. Using machine learning to predict antimicrobial mics and associated genomic features for nontyphoidal salmonella. *Journal of clinical microbiology*, 57(2):10–1128, 2019.

[10] M Blanchette D Lim, C Baek. Graphylo: A deep learning approach for predicting regulatory dna and rna sites from whole-genome multiple alignments. *iScience*, 2024.

[11] M Mutarelli A Facchiano, D Heider. Artificial intelligence and bioinformatics applications for omics and multi-omics studies. *Frontiers in Genetics*, 2024.

[12] D Huang W Zheng Z Dai W Fan, C Ding. Unraveling principles of thermodynamics for genome-scale metabolic networks using graph neural networks. *bioRxiv*, 2024.

[13] B Haibe-Kains A Goldenberg S Ma, AGX Zeng. Integrate any omics: Towards genome-wide data integration for patient stratification. *arXiv preprint arXiv:2401.07937*, 2024.

[14] Y Yao X Luo, X Chen. Mvmsgat: Integrating multiview, multi-scale graph convolutional networks with biological prior knowledge for predicting bladder cancer response to ... *Applied Sciences*, 2024.

[15] BI Kim-SK Woo H Ju, K Kim. Graph neural network model for prediction of non-small cell lung cancer lymph node metastasis using protein–protein interaction network and 18f-fdg ... *International Journal of Molecular Sciences*, 2024.

[16] X Wang-WV Li H Li T Li, K Qian. scbig for representation learning of single-cell gene expression data based on bipartite graph embedding. *NAR Genomics and Bioinformatics*, 2024.

[17] A Giovannini NS D'Souza, H Wang. Fusing modalities by multiplexed graph neural networks for outcome prediction from medical data and beyond. *Medical Image Analysis*, 2024.

[18] Z Hao-X Lei P Liang Q Chang Y Liu, Y Shao. Cuproptosis gene-related, neural network-based prognosis prediction and drug-target prediction for kirc. *Cancer Medicine*, 2023.

[19] L Maddalena-M Piccirillo M Giordano, E Falbo. Untangling the context-specificity of essential genes by means of machine learning: A constructive experience. *Biomolecules*, 2023.

[20] MS Vijaya S Devipriya. Graph convolutional neural network for ic50 prediction model with drug smiles graphs and gene expressions of amyotrophic lateral sclerosis. 2024.

[21] B Gökbağ-Y Chen L Cheng Y Shen, K Fan. Multi-layer encoder prediction of gene combination effect (mle-genecombo) using crispr-cas9 double knockout data. 2024.

[22] M Xu X Feng, Y Li. Multi-satellite cooperative scheduling method for large-scale tasks based on hybrid graph neural network and metaheuristic algorithm. *Advanced Engineering Informatics*, 2024.

[23] LC Shen-H Yan J Song S Li, Y Liu. Gmfgrn: a matrix factorization and graph neural network approach for gene regulatory network inference. *Briefings in Bioinformatics*, 2024.

[24] X Li-L Zhang DS Cao M Li T Liu, Z Fang. Assembling spatial clustering framework for heterogeneous spatial transcriptomics data with graphdeep. *Bioinformatics*, 2024.

[25] Y Li-F Lu Z Wang J Gao, Y Xu. Comprehensive exploration of multi-modal and multi-branch imaging markers for autism diagnosis and interpretation: insights from an advanced deep learning model. *Cerebral Cortex*, 2024.

[26] Y Yao X Luo, X Chen. Mvmsgat: Integrating multiview, multi-scale graph convolutional networks with biological prior knowledge for predicting bladder cancer response to immunotherapy. *Applied Sciences*, 2024.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[28] Bill Waggener and William N Waggener. *Pulse code modulation techniques*. Springer Science & Business Media, 1995.

[29] Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

[30] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[31] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.