LEVERAGING TRANSFER LEARNING AND MULTIMODAL FOUNDATION MODELS FOR ANTIBIOTIC DISCOVERY AGAINST DATA-SCARCE *ESCHERICHIA COLI* STRAINS

Sugitha Janarthanan[1]*, Gen Zhou[1]*, Yan Yi Li[2], Zihao Jing [1], Pingzhao Hu[1,2]†
[1]Western University, London, ON, Canada
[2] University of Toronto, Toronto, ON, Canada

ABSTRACT

Antibacterial resistance is a growing global crisis, complicating the treatment of bacterial infections and bacteria-implicated diseases while increasing healthcare costs and mortality. As such, there is a pressing need for the development of novel antibiotics, but traditional drug discovery methods are costly and slow. Turning to artificial intelligence (AI) and deep learning (DL) models allows us to combat these issues, but for bacterial strains with limited experimental data for DL model training, the benefits of AI are limited. Recently, we developed CL-MFAP, an unsupervised contrastive learning (CL)-based multimodal foundation (MF) model specifically tailored for discovering small molecules with potential antibiotic properties (AP), which has shown great success in antibiotic screening. Also, Deep Message Passing Neural Networks (D-MPNN) are graph neural networks designed for molecular analysis and widely used for antibiotic screening. To combat the issue of experimental data scarcity, we propose a novel pipeline that combines these complementary architectures with transfer learning: both models are first trained on a larger, more general antibacterial dataset, and their learned embeddings are then used to train strain-specific classifiers. This approach enables effective prediction even for bacterial strains with limited data. The pipeline also incorporates extensive virtual screening of almost 11 million commercially available compounds and downstream property prediction analysis to prioritize candidates before experimental validation, significantly reducing resource requirements. By identifying potential novel antibiotic compounds for Adherent-Invasive *Escherichia coli* LF82 (AIEC LF82), we demonstrate our pipeline's potential for effective antibiotic discovery in data-scarce scenarios.

1 INTRODUCTION

Antibacterial resistance is one of the top global public health threats, posing an issue with the effectiveness of existing treatments for bacterial infections and bacteria-associated diseases. Naghavi et al. (2025) found that between 1990 and 2021, more than one million people died from drug-resistant infections each year, and this could increase to nearly 2 million by 2050. As resistance mechanisms evolve and spread, the discovery of new antibiotics has stagnated, creating an urgent need for alternative strategies to identify novel antibiotic compounds.

Unfortunately, traditional drug discovery methods are slow and costly, making them unfavorable solutions to the ever-evolving issue of antibacterial resistance. Artificial intelligence (AI) can provide a much more effective solution by speeding up the rate at which potential antibiotic compounds are identified, and saving on resources by only requiring experimental testing on a smaller, more defined, number of potential compounds. Deep learning (DL) models, particularly those leveraging multimodal representation learning, can bridge molecular structure data with biological activity,

*Equally Contributed
†Contact Author: phu49@uwo.ca

improving the generalizability and interpretability of biological datasets. Foundation models for biological data have been introduced as a powerful tool for the field, showing great success in drug discovery and molecular property prediction (Guo et al., 2025).

Previously, we developed a contrastive learning-based multimodal foundation model specifically tailored for discovering small molecules with potential antibiotic properties for antibiotic screening, named CL-MFAP (Zhou et al., 2025). It employs unsupervised contrastive learning across three molecular representations—SMILES, Morgan fingerprints, and molecular graphs—to extract meaningful embeddings for antibiotic prediction. This model integrates three distinct encoders: (1) a transformer-based encoder with rotary position embedding for SMILES processing, (2) a novel bi-level routing attention transformer for molecular graphs, and (3) a multilayer perceptron for Morgan fingerprints. By employing multimodal representation learning, CL-MFAP enhances predictive accuracy and outperforms other DL models in antibiotic screening (Zhou et al., 2025). Although CL-MFAP has been used for antibiotic screening of bacterial strains with relatively large datasets, its performance has not yet been harnessed for data-scarce bacterial strains. In addition, Deep Message Passing Neural Networks (D-MPNNs) have demonstrated success in drug discovery by leveraging a graph-based architecture that captures intricate molecular relationships through iterative message passing. Notable applications include the work of Wong et al. (2024), who used a D-MPNN to identify a novel class of antibiotics effective against methicillin-resistant *Staphylococcus aureus*, and Stokes et al. (2020) who used a D-MPNN to discover Halicin, a structurally unique antibiotic with potent activity against *Mycobacterium tuberculosis* and other pathogens.

However, for bacterial strains with limited experimental data, AI-driven approaches still struggle, with reduced predictive reliability. Data scarcity for bacterial strains stems from the high costs and time-intensive nature of traditional laboratory testing methods, combined with the rapid emergence of new resistant strains that outpaces the field's ability to thoroughly characterize them through experimental studies. To address the challenge of data scarcity in antibiotic discovery, transfer learning offers a solution. The model is first trained on a general dataset with abundant labeled examples which enables it to learn generalizable features. Once pretrained, the model is then fine-tuned on a smaller, more specific dataset corresponding to the target strain with limited experimental data. This fine-tuning process adjusts the model to focus on the particular nuances of the under-explored strain, significantly enhancing its predictive power even when fewer labeled examples are available. By transferring knowledge from a broader context, transfer learning not only reduces the need for extensive data collection but also accelerates the development of reliable predictions, making it a highly effective strategy for combating challenges in drug discovery and enabling more efficient identification of potential antimicrobial agents(Cai et al., 2020).

In this study, we focus on Adherent-Invasive *Escherichia coli* LF82 (AIEC LF82), a strain associated with Crohn's disease for which experimental data is limited (Darfeuille-Michaud et al., 1998) (Glasser et al., 2001). We present a transfer learning pipeline that integrates both CL-MFAP and D-MPNN to extract meaningful molecular embeddings. By fine-tuning pretrained models on broader datasets and applying classification techniques, our approach enables accurate antibacterial activity prediction for AIEC LF82 despite data limitations. Furthermore, our pipeline evaluates over 10 million compounds through virtual screening and downstream property prediction analysis, systematically identifying the most promising candidates for experimental validation. This strategy significantly enhances the efficiency of antibiotic discovery by reducing laboratory time and resource requirements, accelerating the development of new treatments against understudied bacterial strains.

2 METHODS

The overall workflow is presented in Figure 1. All experiments were conducted on a computing node equipped with 8× NVIDIA A100-SXM4-80GB GPUs (CUDA 12.2) and 5 CPU threads.

2.1 DATASETS AND PRE-PROCESSING

This study utilizes three distinct datasets. First, the larger training dataset consists of *Escherichia coli* (*E. coli)* Minimum Inhibitory Concentration (MIC) data from ChEMBL (Gaulton et al., 2011). MIC is the lowest concentration of an antimicrobial agent that prevents the visible growth of a microorganism and reflects the effectiveness of different compounds against a specific pathogen

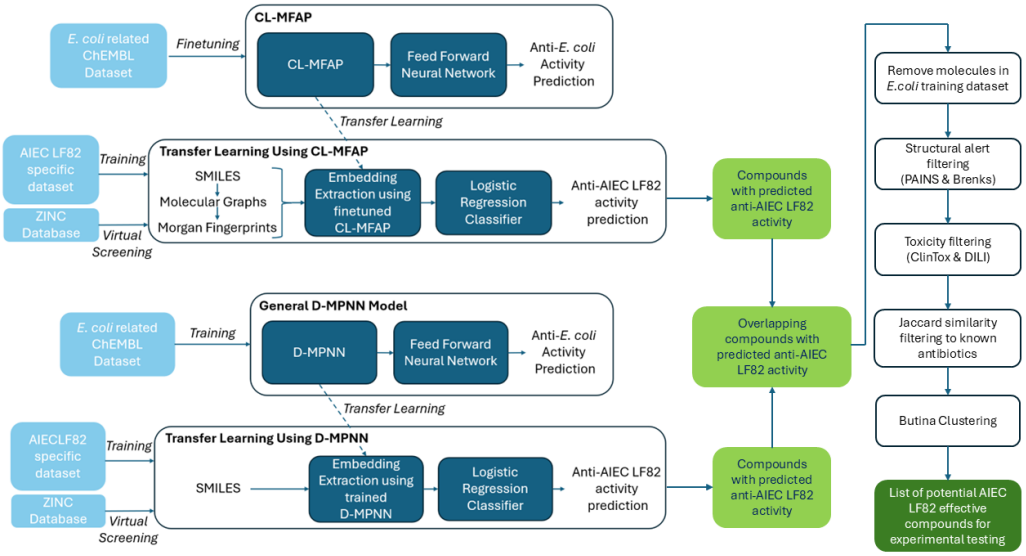

Figure 1: Proposed workflow.

(Kowalska-Krochmal & Dudek-Wicher, 2021). The dataset contains 24,602 compounds (13,391 positive (1) and 11,211 negative (0)) after preprocessing (de-duplication, unit standardization) and binarization at a 16 μg/mL threshold (positive (1) if MIC $\leq$ 16 μg/mL).

Second, the AIEC LF82-specific dataset includes antimicrobial resistance profiles for 29 antibiotics, with 16 classified susceptible, 1 classified intermediate, and 12 classified resistant, according to CLSI guidelines (Martinez-Medina et al., 2020). Both susceptible and intermediate antibiotic compounds were labeled positive (1) and resistant compounds were labeled negative (0).

Lastly, for virtual screening, a subset of the ZINC20 database was used. ZINC20 is a database of commercially available compounds used for virtual screening and drug discovery, containing millions of purchasable chemical compounds with their 3D structures and properties (Irwin et al., 2020). The library was filtered for purchasable compounds with drug-like properties (molecular weight 250 to 500 Da, logP -1 to 5) and 3D structure availability to facilitate future molecular docking, bringing the total virtual screening library to 10,846,709 compounds.

2.2 CL-MFAP FINETUNING

CL-MFAP is a novel foundation model that processes compounds through three parallel pathways: a transformer (12 layers, 8 attention heads) with rotary positional embedding for SMILES sequences, a multilayer perceptron for Morgan fingerprints (radius 2, 2048 bit length), and a transformer-based (12 layers, 16 attention heads) graph encoder with bi-level routing attention (BRA) for molecular graphs. The BRA mechanism uniquely processes molecular graphs by first identifying important structural regions at a window-to-window level before performing detailed pixel-to-pixel attention, effectively prioritizing functionally relevant molecular features. The model then employs contrastive learning using NT-Xent loss to align these three molecular representations, pulling similar pairs closer and pushing dissimilar pairs apart in the embedding space (Zhou et al., 2025).

CL-MFAP (version 1.0.0) was fine-tuned on the larger ChEMBL *E. coli* dataset that was split into 80-10-10 for training, testing, and validation, respectively, using a scaffold split. Scaffold splitting ensures robust generalization by preventing data leakage from structurally similar compounds across train, validation, and test sets, offering a greater challenge for learning algorithms than the random split. The model was finetuned for 25 epochs with a learning rate of 1e-4, weight decay of 1e-3, and batch size of 48. It also employed weighted random sampling for class imbalance and Adam optimizer for optimization. Model performance was evaluated using Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), and accuracy.

2.3 D-MPNN PRE-TRAINING

In a Directed Message Passing Neural Network (D-MPNN), atoms are represented as vertices and bonds as edges. The D-MPNN employs directed message passing with normalized aggregation to generate atomic embeddings, which were subsequently averaged into molecular embeddings and augmented with additional RDKit-derived features.

A D-MPNN was trained using the Chemprop package (version 2.1.2) (Heid et al., 2024) on the large ChEMBL *E. coli* dataset with the same split as CL-MFAP (80-10-10 for training, testing, and validation, respectively using a scaffold split). Hyperparameter optimization was performed using Optuna (Akiba et al., 2019) to refine model performance. The architecture is composed of 3 layers, message dimension of 600, and a feedforward network with 2 layers with hidden dimension of 1600, using PreLU activation, norm aggregation, and no dropout. The model was trained for 50 epochs with batch size 16, using a cyclic learning rate schedule that warmed up from 0.00061 to 0.00164 over 8 epochs before decaying to 0.00021. The loss function employed is binary cross-entropy (BCE). Model performance was evaluated using AUROC, AUPRC, and accuracy.

2.4 TRANSFER LEARNING WITH CLASSIFIERS

To make the models specific to AIEC LF82, transfer learning was employed separately for each model. Each model is leveraged to generate embeddings for the AIEC LF82 dataset, which were used to train and evaluate three classifiers: Random Forest (RF), Multilayer Perceptron (MLP), and Ridge Logistic Regression (Ridge LR). RF employs an ensemble of 100 decision trees, each trained on a bootstrap sample of the data, with final predictions made by majority voting for classification and class probabilities obtained by averaging the trees' predicted probabilities. MLP employs a feed-forward neural network with one hidden layer of 100 neurons using ReLU activation, followed by an output layer with sigmoid activation for binary classification, trained using the Adam optimizer. Ridge LR employs a linear classifier that learns a weight vector to transform input features through a logistic function (sigmoid), producing probabilities between 0 and 1 for binary classification with L2 (Ridge) regularization. All classifiers employed 10 fold cross validation and for CLMFAP, classifiers were trained for optimization on classification threshold using F1 score on precision-recall curve. Classifier performance was assessed using AUROC, AUPRC, accuracy, and F1 score. Classifiers were trained and tested separately for each model pipeline using scikit-learn (v1.5.2) (Pedregosa et al., 2011), selecting the best-performing one for each pipeline.

2.5 VIRTUAL SCREENING

The CL-MFAP and D-MPNN pipelines were independently applied for virtual screening on the ZINC dataset (10,846,709 compounds) to identify compounds predicted to be active against AIEC LF82 with a probability of $\geq$0.95. Compounds predicted as active by both models were selected, and any present in the *E. coli*-specific ChEMBL dataset were excluded. The remaining compounds were considered the final set of predicted active compounds against AIEC LF82 and were subjected to downstream property prediction for further analysis.

2.6 DOWNSTREAM PROPERTY PREDICTION TESTS

Structural Alerts Filtering. Structural alerts are specific chemical substructures or functional groups that are associated with undesirable properties. They serve as red flags in drug discovery to help filter out potentially harmful compounds early in the development process. Two common filters for structural alerts are PAINS (Pan-Assay Interference Compounds) and BRENK. PAINS are structural patterns that identify compounds which frequently appear as false positives in screening assays due to non-specific interactions rather than genuine target affinity (Baell & Holloway, 2010). These substructures are associated with compounds exhibiting activity across multiple targets, often leading to misleading results. BRENK alerts identify undesirable structural features which are associated with poor pharmacokinetics, toxicity, or chemical instability (Brenk et al., 2008). Filtering out compounds containing these structural alerts ensures a higher-quality selection for drug discovery. Therefore, the predicted active compounds were initially screened to remove any PAINS or BRENK compounds using RDKit (version 2024.9.4) (rdk).

Toxicity Filtering. The subsequent step involved eliminating compounds predicted to have toxic properties. Toxicity assessment was conducted using ADMET-AI (version 1.3.1), a state-of-the-art machine learning platform recognized for its high accuracy and rapid ADMET(Absorption, Distribution, Metabolism, Excretion, Toxicity) predictions (Swanson et al., 2024). The evaluation focused on two critical toxicity categories: Clinical Toxicity (ClinTox) and Drug-Induced Liver Injury (DILI). ClinTox predicts the likelihood of a compound causing clinical toxicity (Gayvert et al., 2016), while DILI assesses its potential for liver damage (Xu et al., 2015). Only compounds with a predicted toxicity probability of less than 20% for both ClinTox and DILI were retained for further analysis.

Jaccard Similarity Scores. To identify potential compounds with MIC activity against AIEC LF82 that are structurally distinct from existing antibiotics, MAP4C (MinHashed Atom-Pair Chiral) fingerprints (radius=2, bit length=2048) were calculated for all predicted active compounds at this stage, as well as for AIEC LF82 active antibiotic compounds in the training dataset. MAP4C fingerprints are an extension of MAP4 fingerprints that encode molecular structures by capturing atom-pair relationships while incorporating chirality, using MinHashing to generate compact and efficient representations for molecular similarity comparisons (Orsi & Reymond, 2024). They were chosen for their ability to accurately distinguish stereoisomers within complex molecular structures and their robustness across a wide range of molecular sizes. These fingerprints were then used to compute the Jaccard similarity between the predicted active compounds and the active antibiotics in the AIEC LF82 training dataset, representing structural similarities. Jaccard similarity (also known as Tanimoto similarity) is a commonly used metric in cheminformatics for assessing compound similarity (Bajusz et al., 2015), that computes the fraction of features in common between two compounds relative to the total number of features present in either compound. The Jaccard similarity between two compound fingerprints A and B is given by equation 1.

$$T(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Compounds with a Jaccard similarity score that meet a defined threshold are classified as part of the same group. To avoid experimentally testing compounds that are very structurally similar to existing antibiotic compounds, only predicted active compounds with a Jaccard similarity score of less than 0.3 to all active AIEC LF82 antibiotics in the training set were retained for further consideration.

2.7 Clustering of Similar Compounds

Butina clustering is a fast, non-hierarchical clustering algorithm used in cheminformatics to group similar compounds based on a pairwise similarity threshold (Butina, 1999). Compounds were first encoded as MAP4C fingerprints which were then used to calculate Jaccard similarity scores and generate a Jaccard distance matrix. Similar to before, MAP4C fingerprints were chosen at this step due to their ability to accurately distinguish stereoisomers within complex molecular structures and their robustness across a wide range of molecular sizes. After, Butina clustering is used to iteratively select compounds with the most neighbors as cluster centroids and group similar compounds around them. To determine the optimal Jaccard distance threshold cutoff (maximum Jaccard distance allowed between two fingerprints for them to be considered part of the same cluster), cluster size distributions (number of compounds per cluster versus cluster index) were plotted across different cutoff values from 0 to 1. The cutoff that produces the smoothest distribution with not too many singletons (clusters with 1 compound) was chosen. The intra-cluster Jaccard similarity was also calculated and visualized to ensure proper clustering. From this clustering, a subset of 100 diverse compounds for experimental testing were selected by first choosing the cluster centers, then iterating through clusters and selecting compounds based on their similarity to the cluster centers. From each cluster, up to 3 additional compounds are chosen if the cluster is large ($>$10), or 1 additional compound if the cluster is smaller, with the selection process prioritizing the most similar compounds within each cluster until 100 compounds are selected. By including additional compounds from the biggest clusters, their activity can be analyzed in experimental testing to note whether compounds from the same cluster show similar activity - if so, this may indicate a structural-activity relationship that can be further investigated. After Butina clustering, the Bemis-Murcko scaffolds of the 100 compounds were calculated and analyzed to ensure the chemical diversity of the experimental testing set (Bemis & Murcko, 1996).

By clustering the predicted active compounds, experimental testing can be streamlined, as representative compounds from each cluster can be selected for initial experimental testing. For compounds with promising antibacterial activity in the experimental stage phase, all compounds within their clusters can be tested. This strategy optimizes the experimental workflow and enhances the likelihood of identifying effective antibacterial agents.

3 RESULTS

CL-MFAP Finetuning and Classifier Training. When CL-MFAP was finetuned on *E.coli* ChEMBL dataset, the performance metrics were: ROC-AUC = 0.8645, PRC-AUC = 0.8617, and accuracy = 0.7692. When transfer learning was applied to the three classifiers with optimal threshold, the performance metrics are shown in Table 1. Both Ridge LR and MLP performed equally as well but as Ridge LR requires less computational overhead, it was chosen for the final pipeline.

Metric	Ridge LR	RF	MLP
Accuracy	0.8000 ± 0.3583	0.6667 ± 0.3143	0.8000 ± 0.3583
F1 Score	0.7800 ± 0.4158	0.7000 ± 0.3143	0.7800 ± 0.4158
ROC-AUC	0.8000 ± 0.3496	0.6000 ± 0.3162	0.8000 ± 0.3496
PRC-AUC	0.8208 ± 0.3299	0.7375 ± 0.2953	0.8208 ± 0.3299

Table 1: Performance metrics of CL-MFAP transfer learning-based classifiers.

D-MPNN Pre-training and Classifier Training. When D-MPNN was pre-trained on *E.coli* ChEMBL dataset, the performance metrics were: ROC-AUC = 0.8448, PRC-AUC = 0.8505, and accuracy = 0.7439. When transfer learning was applied to the three classifiers, performance metrics are shown in Table 2. The Ridge LR classifier performed best and was used in the final pipeline.

Metric	Ridge LR	RF	MLP
Accuracy	0.7167 ± 0.2727	0.6833 ± 0.2540	0.6833 ± 0.2540
F1 Score	0.7167 ± 0.3148	0.7467 ± 0.1951	0.6967 ± 0.3008
ROC-AUC	0.8000 ± 0.3496	0.7250 ± 0.3426	0.8000 ± 0.3496
PRC-AUC	0.8833 ± 0.2388	0.8542 ± 0.2326	0.8833 ± 0.2388

Table 2: Performance metrics of D-MPNN transfer learning-based classifiers.

Virtual Screening. Virtual screening was performed using both models. 2,223 overlapping compounds were identified with a predicted probability of 0.95 or more across both models. After filtering to remove compounds that were also present in the *E. coli* training dataset, 2,217 compounds remained for downstream property filtering.

Downstream Property Prediction. Within the predicted active compounds, 74 PAINS were first filtered out, and 725 BRENK compounds were subsequently filtered out. During toxicity filtering, 1113 compounds were filtered out, resulting in 305 compounds remaining. The large number of compounds filtered at this step can be attributed to the strict toxicity cutoffs imposed. However, these strict cutoffs were enforced as toxic compounds require no further testing. After calculating Jaccard similarity scores between compounds at this stage and active antibiotics against AIEC LF82 in the training set, all were found to be structurally diverse ($<$0.3) and thus, no compounds were filtered out at this step. The number of compounds filtered out and remaining at each step after visual screening are summarized in Table 3.

Clustering of Similar Compounds. Butina clustering was conducted with a Jaccard distance threshold cutoff of 0.6 as this results in a good balance between the number of clusters and cluster sizes. There are not many singletons and the cluster sizes don't have an extreme but smooth distribution (Figure 2). Butina clustering resulted in 74 clusters, with 31 only having one compound (singletons), 12 with over five compounds, and 2 with over twenty-five compounds. The largest cluster has 37 compounds. As the target experimental set is 100 compounds, the 74 clusters' centroids were chosen plus an additional 26 from the top clusters.

By analyzing the intra-cluster Jaccard similarity of the top 15 clusters, it is noted that no clusters show extremely low similarities, and the mean for them all is above 0.4, showing good similarity

Filtering Step	Number of Compounds Filtered	Remaining Number of Compounds
Removal of Overlapping Compounds in *E. coli* Training Set	6	2217
PAINS (Pan-Assay Interference Compounds) Filtering	74	2143
BRENK Structural Alert Filtering	725	1418
Toxicity Filtering	1113	305
Filtering Based on Jaccard Similarity to Known Antibiotics	0	305

Table 3: Downstream filtering of virtual screening predicted active compounds against AIEC LF82.

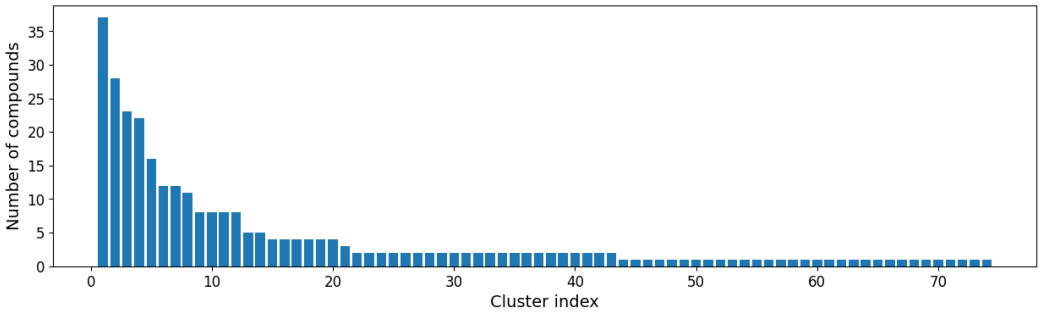

Figure 2: Number of compounds per cluster after Butina clustering using Jaccard distance threshold cutoff of 0.6.

between compounds within each cluster (Figure 3). As such, representative compounds can be confidently picked from each cluster for initial experimental testing.

The calculation of Bemis-Murcko scaffolds for the chosen 100 compounds revealed 82 scaffolds, of which 73 were singleton scaffolds, and the scaffold diversity was 0.82. This shows exceptional structural diversity. As shown in Figure 4A, the cumulative percentage of compounds rises steadily with the number of scaffolds, indicating consistent structural diversity. As shown in Figure 4B, the most frequent scaffold only appears in 4 compounds, and the 2nd and 3rd most common scaffolds in 3 compounds each. This is an ideal distribution as no single scaffold dominates. As such, the subset chosen for experimental testing shows great structural diversity.

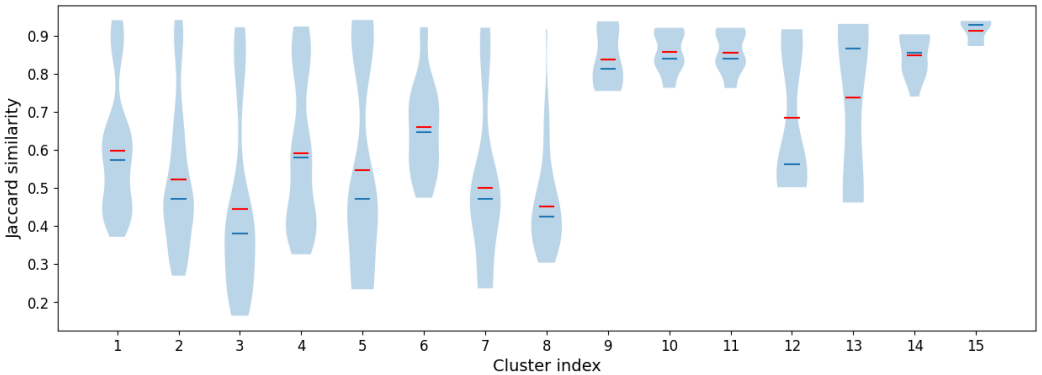

Figure 3: Jaccard similarity between compounds within their respective clusters for the 15 largest clusters. The red and blue lines indicate the mean and median of Jaccard similarity between compounds within each cluster, respectively.

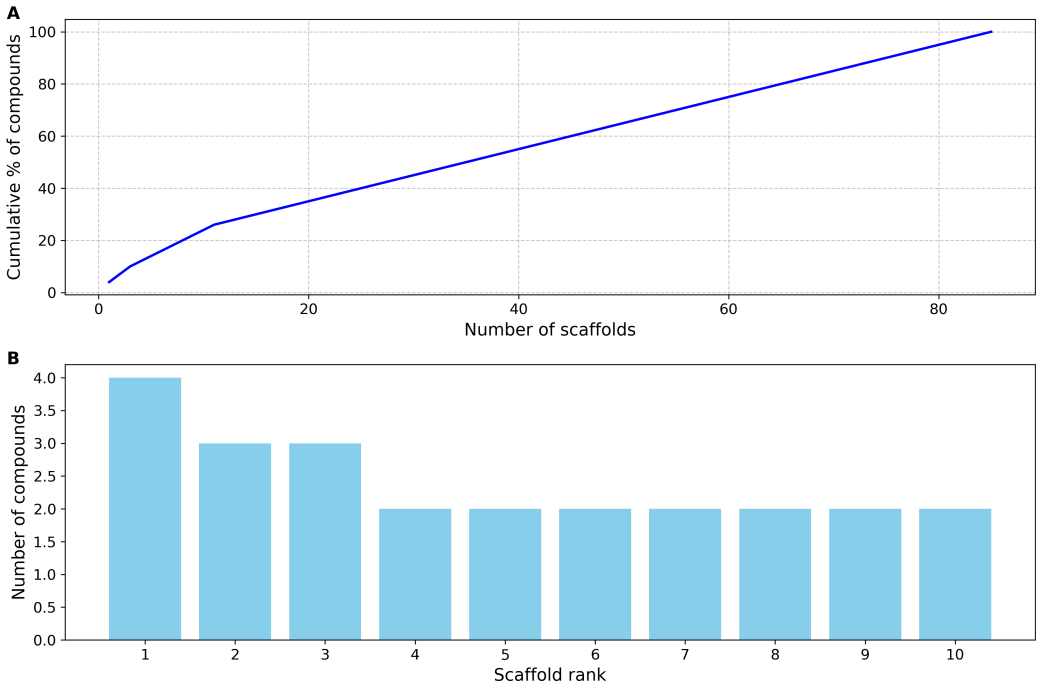

Figure 4: A) Distribution of Bemis-Murcko scaffolds across the final subset of 100 compounds ready for experimental testing. B) Ten most frequent Bemis-Murcko scaffolds identified, ranked by occurrence in the final subset of 100 compounds ready for experimental testing.

4 Conclusion

In this study, we present a comprehensive pipeline designed to identify potential antibiotic compounds, particularly when experimental data for a specific bacterial strain is limited. By leveraging advanced computational methods, including foundation models for biological data and multimodal representation learning, we successfully generated a refined list of compounds exhibiting promising activity against AIEC LF82, along with favorable molecular properties such as optimal pharmacokinetic profiles, low toxicity potential, and structural diversity. The use of high-dimensional molecular fingerprints and clustering techniques like Butina clustering allowed us to effectively navigate the complex chemical space, identify distinct clusters of compounds, and prioritize those with the highest likelihood of efficacy at the experimental testing phase.

The next phase of this work will focus on translating these computational predictions into experimental validation. This will be achieved through a close collaboration with an experimental laboratory, where the identified compounds will undergo rigorous testing. Multiple tests will be performed such as MIC assays to quantify the antimicrobial potency, time-kill assays to assess bactericidal activity, and in vivo studies to evaluate the therapeutic potential and safety of the compounds in animal models. We also plan to explore further experimental assays to investigate the mechanism of action of the most promising candidates, such as biofilm inhibition, resistance profiling, and host-pathogen interaction studies. Simultaneously, we will be performing more rigorous statistical testing (e.g. ablation studies, sensitivity analysis) to further validate the robustness of our pipeline.

The approach of utilizing state-of-the-art DL models with transfer learning aims to accelerate the drug discovery process and increase the chances of identifying novel antibiotic compounds for data-scarce strains. The integrated approach of extensive computational predictions with experimental validation aims to further this work and provide a more in depth proof of concept for this pipeline's application for other bacterial strains.

MEANINGFULNESS STATEMENT

To us, a meaningful representation of life captures the complexity, diversity, and interconnectedness of human experiences by identifying essential patterns, structures, or functions within biological, cognitive, or social systems. In scientific discovery, this means uncovering insights that deepen our understanding of human biology. Our work contributes to this by exploring novel interactions between external stimuli (compounds) and the human body, specifically in combating bacterial strains overrepresented in humans. By elucidating these interactions, we enhance our understanding of the body's complexities and accelerate antibiotic discovery, ultimately fostering a more comprehensive representation of human health and its intricate biological dynamics.

ACKNOWLEDGMENTS

This research was supported in part by the Canadian Institutes of Health Research (PLL185683, PJT 190272), the Canada Research Chairs Tier II Program (CRC-2021-00482), the Natural Sciences and Engineering Research Council of Canada (RGPIN-2021-04072) and the Canada Foundation for Innovation (CFI) John R. Evans Leaders Fund (JELF) program (grant #43481).

REFERENCES

Rdkit: Open-source cheminformatics. URL `https://www.rdkit.org`.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Jonathan B. Baell and Georgina A. Holloway. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740, April 2010. ISSN 0022-2623, 1520-4804. doi: 10.1021/jm901137j.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20, December 2015. ISSN 1758-2946. doi: 10.1186/s13321-015-0069-3.

Guy W. Bemis and Mark A. Murcko. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, January 1996. ISSN 0022-2623. doi: 10.1021/jm9602928.

Ruth Brenk, Alessandro Schipani, Daniel James, Agata Krasowski, IanHugh Gilbert, Julie Frearson, and PaulGraham Wyatt. Lessons Learnt from Assembling Screening Libraries for Drug Discovery for Neglected Diseases. *ChemMedChem*, 3(3):435–444, March 2008. ISSN 1860-7179, 1860-7187. doi: 10.1002/cmdc.200700139.

Darko Butina. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, July 1999. ISSN 0095-2338. doi: 10.1021/ci9803381.

Chenjing Cai, Shiwei Wang, Youjun Xu, Weilin Zhang, Ke Tang, Qi Ouyang, Luhua Lai, and Jianfeng Pei. Transfer learning for drug discovery. *Journal of Medicinal Chemistry*, 63(16):8683–8694, 2020. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.9b02147.

Arlette Darfeuille-Michaud, Christel Neut, Nicolas Barnich, Emmanuel Lederman, Patrick Di Martino, Pierre Desreumaux, Luc Gambiez, Bernard Joly, Antoine Cortot, and Jean-Frédéric Colombel. Presence of adherent Escherichia coli strains in ileal mucosa of patients with Crohn's disease. *Gastroenterology*, 115(6):1405–1413, December 1998. ISSN 0016-5085. doi: 10.1016/S0016-5085(98)70019-8.

Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40:D1100–D1107, 2011. doi: 10.1093/nar/gkr777.

Kaitlyn M. Gayvert, Neel S. Madhukar, and Olivier Elemento. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. *Cell Chemical Biology*, 23(10):1294–1301, October 2016. ISSN 24519456. doi: 10.1016/j.chembiol.2016.07.023.

Anne-Lise Glasser, Jerome Boudeau, Nicolas Barnich, Marie-Helene Perruchot, Jean-Frederic Colombel, and Arlette Darfeuille-Michaud. Adherent Invasive *Escherichia coli* Strains from Patients with Crohn's Disease Survive and Replicate within Macrophages without Inducing Host Cell Death. *Infection and Immunity*, 69(9):5529–5537, September 2001. ISSN 0019-9567, 1098-5522. doi: 10.1128/IAI.69.9.5529-5537.2001.

Fei Guo, Renchu Guan, Yaohang Li, Qi Liu, Xiaowo Wang, Can Yang, and Jianxin Wang. Foundation models in bioinformatics. *National Science Review*, pp. nwaf028, 2025. ISSN 2095-5138, 2053-714X. doi: 10.1093/nsr/nwaf028.

Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: A Machine Learning Package for Chemical Property Prediction. *Journal of Chemical Information and Modeling*, 64 (1):9–17, January 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c01250.

John J. Irwin, Ling Tang, Jared Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, James Mayfield, and Roger A. Sayle. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 2020. doi: 10.1021/acs.jcim.0c00675.

Beata Kowalska-Krochmal and Ruth Dudek-Wicher. The minimum inhibitory concentration of antibiotics: Methods, interpretation, clinical relevance. *Pathogens*, 10(2):165, 2021. ISSN 2076-0817. doi: 10.3390/pathogens10020165.

Margarita Martinez-Medina, Francesco Strozzi, Belén Ruiz Del Castillo, Natalia Serrano-Morillas, Nuria Ferrer Bustins, and Luis Martínez-Martínez. Antimicrobial Resistance Profiles of Adherent Invasive Escherichia coli Show Increased Resistance to -Lactams. *Antibiotics*, 9(5):251, May 2020. ISSN 2079-6382. doi: 10.3390/antibiotics9050251.

Mohsen Naghavi, Stein Emil Vollset, Kevin S Ikuta, Lucien R Swetschinski, Authia P Gray, and Wool. Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. *The Lancet*, 404(10459):1199–1226, 2025. ISSN 01406736. doi: 10.1016/S0140-6736(24)01867-1.

Markus Orsi and Jean-Louis Reymond. One chiral fingerprint to find them all. *Journal of Cheminformatics*, 16(1):53, May 2024. ISSN 1758-2946. doi: 10.1186/s13321-024-00849-6.

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.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A Deep Learning Approach to Antibiotic Discovery. *Cell*, 180(4):688–702.e13, February 2020. ISSN 00928674. doi: 10.1016/j.cell.2020.01.021.

Kyle Swanson, Parker Walther, Jeremy Leitz, Souhrid Mukherjee, Joseph C Wu, Rabindra V Shivnaraine, and James Zou. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics*, 40(7):btae416, July 2024. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btae416.

Felix Wong, Erica J. Zheng, Jacqueline A. Valeri, Nina M. Donghia, Melis N. Anahtar, Satotaka Omori, Alicia Li, Andres Cubillos-Ruiz, Aarti Krishnan, Wengong Jin, Abigail L. Manson, Jens Friedrichs, Ralf Helbig, Behnoush Hajian, Dawid K. Fiejtek, Florence F. Wagner, Holly H. Soutter, Ashlee M. Earl, Jonathan M. Stokes, Lars D. Renner, and James J. Collins. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, 626(7997):177–185, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06887-8.

Youjun Xu, Ziwei Dai, Fangjin Chen, Shuaishi Gao, Jianfeng Pei, and Luhua Lai. Deep Learning for Drug-Induced Liver Injury. *Journal of Chemical Information and Modeling*, 55(10):2085–2093, October 2015. ISSN 1549-9596. doi: 10.1021/acs.jcim.5b00238.

Gen Zhou, Sugitha Janarthanan, Yutong Lu, and Pingzhao Hu. CL-MFAP: A contrastive learning-based multimodal foundation model for molecular property prediction and antibiotic screening. In *The Thirteenth International Conference on Learning Representations*, 2025.