Leveraging Machine Learning and Zinc-Aware Docking to Discover Natural Inhibitors of EptA Resistance Enzymes

Mason Cheng
masonc3000@gmail.com
Saratoga High School
Saratoga, California, USA

Mariame Diabate, PhD.
diabatem@stanford.edu
Stanford University
Stanford, California, USA

Abstract

Antibiotic resistance in multidrug-resistant (MDR) bacteria poses a critical threat to global health, contributing to millions of deaths annually. Polymyxins, a class of antibiotics, have been reintroduced as a last-resort treatment against Gram-negative MDR infections, but resistance has spread, rendering them ineffective. In light of this problem, we employed machine learning and zinc-aware docking to discover natural inhibitors of EptA resistance enzymes, a type of enzyme conferring resistance to polymyxins. We curated a training dataset of 2,088 compounds derived from known EptA inhibitors and their structural analogs, with activity labels generated from docking scores, to train a random forest (RF) model. The model achieved 93% accuracy and was applied to screen 5,000 compounds from the COCONUT natural product database, predicting 166 active compounds, of which the top 10 candidates were validated through high-accuracy docking. The top candidates demonstrated strong binding affinity to EptA's zinc-dependent active site (docking scores: -6.384 to -8.129 kcal/mol) and formed consistent hydrogen bonds with catalytic residues, warranting further evaluation through in vitro and in vivo validation.

Keywords

Bacterial resistance, Machine learning, Colistin, EptA, Polymyxin resistance enzymes, Zinc-aware molecular docking, Antibiotic resistance, Polymyxin antibiotics, Zinc-dependent metalloenzymes, Structure-based virtual screening, Gram-negative bacterial infections

ACM Reference Format:
Mason Cheng and Mariame Diabate, PhD.. 2026. Leveraging Machine Learning and Zinc-Aware Docking to Discover Natural Inhibitors of EptA Resistance Enzymes. In *Proceedings of International Journal of Secondary Computing and Applications Research (IJSCAR VOL. 3, ISSUE 1).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.5281/zenodo.18432165

1 Introduction

Multi-drug-resistant (MDR) Gram-negative bacteria pose an urgent clinical challenge due to the lack of treatment options and absence of new, effective antibiotics [6]. These infections caused 1.27 million deaths worldwide in 2019 and are projected to increase to 10 million deaths annually in 2050, posing a significant burden on healthcare systems worldwide [13]. Polymyxins, a class of antibiotics that includes colistin, have been reintroduced as last-line antibiotics for MDR infections. Despite this reintroduction, resistance has emerged rapidly and now threatens to render even these last-resort agents ineffective [10]. One of the primary resistance mechanisms involves the modification of lipid A in the bacterial outer membrane. Polymyxins rely on electrostatic interactions with negatively charged phosphate groups on lipid A. Phosphoethanolamine transferases (PEA transferases), including chromosomally encoded EptA and plasmid-mediated MCR variants, catalyze the transfer of phosphoethanolamine to lipid A, decreasing the negative charge and reducing polymyxin binding [7, 21]. The catalytic mechanism of EptA depends on a Zn^{2+} cofactor, with conserved residues coordinating the active site and enabling phosphoethanolamine transfer [16]. This zinc-dependent mechanism makes the metal coordination site an attractive target for structure-based inhibitor design [17].

EptA has been validated as a drug discovery target for restoring polymyxin activity. Inhibition of EptA can re-sensitize resistant pathogens to colistin, as demonstrated with compounds such as valnemulin, an EptA inhibitor [20]. Small molecules that disrupt EptA function have also been reported [11]. However, only a limited number of EptA inhibitors have been reported, and there remains a critical need to expand chemical diversity and scaffolds, identify inhibitors with favorable drug-like properties, and develop compounds suitable for clinical translation [11].

To address this gap, computational approaches offer a promising strategy for discovering novel EptA inhibitors. Natural products represent a valuable source of structural diversity and have been widely used in antibiotic discovery, making them a rational choice for screening efforts [17]. Artificial intelligence and computational drug discovery tools accelerate this process by enabling rapid, cost-effective screening of millions of compounds; a scale impractical for traditional experimental methods [1]. Machine learning models, combined with structure-based docking, enable the rapid evaluation of natural product libraries while accounting for structural and mechanistic features, such as zinc cofactor coordination. Random Forest models specifically allow for small training datasets and reduce overfitting, making them the best choice for this study.

This study integrates a random forest machine learning model with zinc-aware molecular docking to systematically identify natural product inhibitors of EptA, prioritizing candidates that are predicted to have strong binding affinity and direct interactions with the catalytic zinc site.

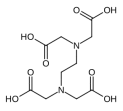
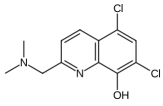
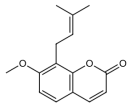
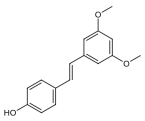
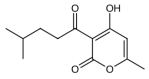
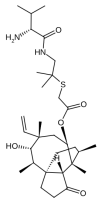
This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
IJSCAR VOL. 3, ISSUE 1
© 2026 The Author(s). Published by IJSCAR.

2 Methods

2.1 Dataset Curation: Inhibitor Selection and Analog Generation

EptA inhibitors with experimentally validated activity were identified from published literature. A total of 6 inhibitors were selected as seed compounds for analog generation (Table 1).

Table 1: A table showcasing the original 6 experimentally determined inhibitors to EptA. The 2D structures of each inhibitor are shown as well as their SMILES strings. [20], [4], [8], [23], [19], [22]

Inhibitor	Structure	SMILES
EDTA		C(CN(CC(=O)O)CC(=O)O)N(CC(=O)O)CC(=O)O
PBT2		CN(C)CC1=NC2=C(C=C1)C(=CC(=C2O)Cl)Cl
Osthole		CC(=CCC1=C(C=CC2=C1OC(=O)C=C2)OC)C
Pterostilbene		COC1=CC(=CC(=C1)/C=C/C2=CC=C(C=C2)O)OC
Pogostone		CC1=CC(=C(C(=O)O1)C(=O)CCC(C)C)O
Valnemulin		C[C@@H]1CC[C@@]23CCC(=O)[C@H]2[C@@]1([C@@H](C[C@@]([C@H]([C@@H]3C)O)(C)C=C)OC(=O)CSC(C)(C)CNC(=O)[C@@H](C(C)C)N)C

Metal-coordinating moieties present in some inhibitors were excluded to simplify docking simulations and reduce computational complexity. Simplified Molecular Input Line Entry System (SMILES) strings were obtained from the PubChem database or generated using chemical drawing software.

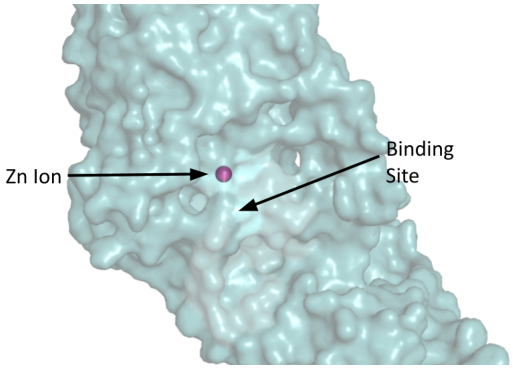


Figure 1: A surface view of EptA. Zn ion and binding site are labeled, and the Zn ion is colored for visual effect

To expand structural diversity while maintaining similarity to known inhibitors, analogs for each parent compound were retrieved from PubChem using an automated pipeline. For each of the 6 parent inhibitors, up to 1,200 compounds with Tanimoto similarity (a metric used to compare the similarity between molecules based on their fingerprints) of 0.8 and 1,200 compounds with Tanimoto similarity of 0.9 were initially retrieved using the PubChem API, yielding up to 2,400 candidate analogs per inhibitor. These candidates were then filtered using Lipinski's Rule of Five criteria, a set of chemical property guidelines for predicting whether a chemical compound's drug-likeness and favorable oral bioavailability. To reduce computational burden, a maximum of 1,000 compounds from each similarity threshold (0.8 and 0.9) were retained for further processing. The remaining compounds were subjected to Butina clustering (distance threshold = 0.3). Butina clustering is a method to cluster compounds with similar tanimoto similarities, maximizing structural diversity and reducing redundancy. Finally, a maximum of 100 analogs were selected for each parent inhibitor. This process yielded 513 analogs across all 6 inhibitors (100 analogs for 5 inhibitors but only 13 analogs for Valnemulin due to extensive filtering by Lipinski criteria). A dataset of 1,200 inactive compounds was generated using the same retrieval pipeline, but with lower Tanimoto similarities of 0.4–0.5 (rather than 0.8–0.9 used for active analogs). These compounds were labeled as inactive due to their structural dissimilarity from the known inhibitors. This approach was chosen over random sampling to ensure controlled structural dissimilarity, thereby improving the model's ability to distinguish active from inactive compounds based on relevant chemical features.

2.2 Molecular Docking for Label Generation

To generate activity labels for machine learning classification, all structural analogs were docked to the EptA active site using AutoDock Vina with the AutoDock4Zn forcefield [14]. The crystal structure of EptA was retrieved from the Protein Data Bank (ID: 5FGN). The structure was prepared using PyMOL by removing water molecules and co-crystallized ligands (Figure 1) [15]. Hydrogen atoms were

added using the Reduce software [18]. To properly model zinc-ligand interactions, the crystallographic zinc ion was replaced with a TZ pseudoatom, a specialized atom type in AutoDock4Zn that accounts for tetrahedral coordination geometry and polarization effects. The TZ pseudoatom was positioned to coordinate with HIS453, the primary zinc-coordinating residue in the active site.

All 513 analogs (derived from 6 parent inhibitors) were docked to the prepared EptA structure. Ligand SMILES strings were first converted to structure data file (SDF) format. Three-dimensional conformers were then generated and energy-minimized using MolScrub [14]. A total of 976 3D conformers were generated from the 513 input structures (average 1.9 conformers per compound) to account for stereoisomers and tautomeric forms. Ligands were converted to PDBQT format for docking using Meeko [14]. A cubic docking grid (60 × 60 × 60 points, 0.375 Å spacing) was centered on the Zn ion. Grid-based affinity maps were generated for all ligand atom types using AutoGrid4 [14]. Molecular docking was performed using AutoDock Vina with the AutoDock4Zn forcefield and an exhaustiveness parameter of 10.

Compounds with docking scores ≤ -6 kcal/mol were labeled as active, while those with scores > -6.0 kcal/mol were labeled as inactive. This threshold was selected based on the docking scores of experimentally validated inhibitors, most of which scored between -5.666 and -7.264 kcal/mol. One of experimentally validated inhibitors, EDTA, scored an outlier of -4.089 kcal/mol, which is likely due to the absence of its metal-coordinating moieties. Duplicate compounds (identified by identical SMILES strings) were removed, retaining only the conformer with the best docking score. After duplicate removal, the final docking dataset comprised 547 active compounds (docking score ≤ -6 kcal/mol) and 341 inactive compounds (score > -6 kcal/mol), for a total of 888 labeled structures. The top-scoring poses for each ligand were retained for subsequent interaction analysis and feature extraction.

Table 2: Model performance charts for each model generated using Scikit-learn. These charts were generated using a testing set held out from model training, which consisted of 30% of each training set.

Classification Report (6:18)				
	Prec.	Rec.	F1	Supp.
Class 0	0.71	0.83	0.77	6
Class 1	0.00	0.00	0.00	2
Accuracy			0.62	8
Macro Avg	0.36	0.42	0.38	8
Weighted Avg	0.54	0.62	0.58	8

Classification Report (60:180)				
	Prec.	Rec.	F1	Supp.
Class 0	0.96	0.94	0.95	54
Class 1	0.84	0.89	0.86	18
Accuracy			0.93	72
Macro Avg	0.90	0.92	0.91	72
Weighted Avg	0.93	0.93	0.93	72

Classification Report (547:1541)				
	Prec.	Rec.	F1	Supp.
Class 0	0.95	0.96	0.95	463
Class 1	0.88	0.86	0.87	164
Accuracy			0.93	627
Macro Avg	0.92	0.91	0.91	627
Weighted Avg	0.93	0.93	0.93	627

2.3 Random Forest Classification Model

To address class imbalance and prevent overfitting to the active class, the 1,200 previously generated inactive compounds (Tanimoto similarity of 0.4–0.5) were added to the docked dataset of 547 active and 341 inactive compounds, resulting in a combined dataset of 547 active and 1,541 inactive compounds (approximately 1:3 ratio). A Random Forest (RF) classifier was selected for several reasons: (1) RF models perform well on small to medium-sized datasets without requiring extensive hyperparameter tuning, (2) they are less prone to overfitting than deep neural networks when training data is limited, (3) they provide feature importance rankings that aid in interpretability, and (4) they can capture the nonlinear relationships between molecular descriptors and activity [2]. The Random Forest model was trained with 200 trees, a maximum depth of 10, and default scikit-learn parameters for other hyperparameters.

Molecular descriptors were calculated from compound SMILES strings using RDKit [12]. Features included: molecular weight, octanol-water partition coefficient (logP), hydrogen bond donor and acceptor counts, topological polar surface area (TPSA), number of rotatable bonds, aromatic ring count, and a 1024-bit Morgan fingerprint (radius = 2). Morgan fingerprints are a bit vector that encodes the presence of specific substructures, determined by a defined radius around each atom. These features capture both physicochemical properties and structural information relevant to drug-likeness and binding potential. The dataset was randomly split into training (70%, n=1461) and test (30%, n=627) sets with stratified sampling to preserve the active to inactive ratio. A random seed of 44 was used to ensure reproducibility.

The trained Random Forest model achieved an overall accuracy of 93% on the held-out test set (Table 2). Class-specific performance metrics revealed slightly lower performance for the minority (active) class: precision was 88% for active compounds versus 95% for inactive compounds, while recall was 86% for active compounds versus 96% for inactive compounds. Feature importance analysis revealed that molecular weight, logP, and rotatable bonds were the strongest predictors of activity among physicochemical descriptors, while Morgan fingerprint bits 578 and 694 (corresponding to specific structural motifs) contributed most to classification performance. The F1-scores were 87% and 95% for active and inactive classes, respectively. The area under the receiver operating characteristic curve (AUC-ROC) was 0.97, indicating strong discriminative ability (Figure 2).

The modest reduction in active compound recall (86%) reflects the challenge of identifying true positives within an imbalanced dataset,

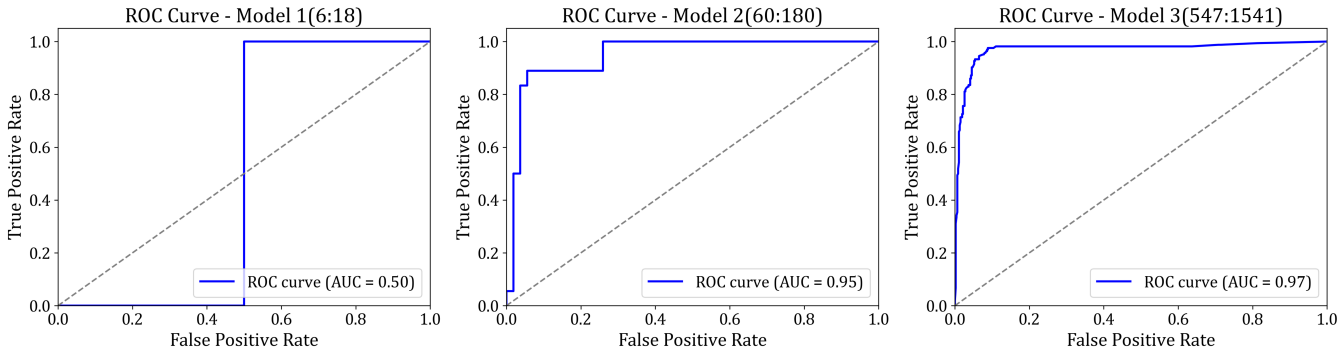


Figure 2: ROC curves of the three Random Forest models tested. The dotted line represents the random baseline. The random baseline shows the results of randomly guessing. Models with a higher area under the curve (AUC) are more capable of separating positive and negative examples correctly.

but remains acceptable for prioritizing candidates for experimental validation.

2.4 Model-Based Screening and Validation Docking

To identify novel EptA inhibitors, the trained Random Forest model was applied to screen the COCONUT (COlleCtion of Open Natural prodUcTs) database, which contains 695,119 natural product structures [3]. The entire database was filtered using Lipinski's Rule of Five, from which 5,000 compounds were randomly selected to create a computationally tractable screening set. Molecular descriptors and Morgan fingerprints were calculated for each compound using the same feature extraction protocol as described above. The Random Forest model predicted 166 compounds (3.3% of the screened set) as active based on probability scores above the classification threshold. These candidates were ranked by their prediction confidence scores to prioritize the most promising inhibitors for experimental validation. The model's prediction probability ranged from 94.5% to 50.1% for these active compounds.

This two-stage approach (machine learning pre-screening followed by high-accuracy docking validation) enabled efficient prioritization of the most promising candidates from a large chemical library. To validate the model's predictions and assess binding modes, the top 10 ranked compounds were selected for rigorous docking analysis (Table 3).

These compounds were subjected to molecular docking using the same AutoDock Vina protocol described previously (AutoDock4Zn forcefield, 60 × 60 × 60 grid centered on the Zn ion), but with increased exhaustiveness (32 vs. 10) to ensure thorough conformational sampling and more accurate binding pose prediction. The top-scoring pose for each compound was retained for binding mode analysis.

3 Results

3.1 Random Forest Model Performance and Dataset Optimization

To optimize model performance while minimizing computational cost, the effect of training dataset size on classification accuracy was systematically evaluated. Three training datasets with varying sizes were constructed, each maintaining a 1:3 active:inactive ratio: 6:18, 60:180, and 547:1,541 compounds. The smallest dataset (6:18), comprising the six parent inhibitors and three inactive analogs per inhibitor, achieved only 62% accuracy (based on a testing set created from 30% of the dataset), demonstrating the limitations of training on minimal data. Expanding the dataset to 60:180 (six parent inhibitors, 9 active analogs per inhibitor, and 30 inactive compounds per inhibitor) substantially improved accuracy to 93%, indicating that moderate dataset expansion was sufficient to achieve robust classification performance. The largest dataset (547:1,541) incorporated all compounds from the docking-based labeling procedure (described in Methods) plus 1,200 additional inactive compounds to maintain the 1:3 ratio. Notably, this dataset achieved the same 93% accuracy as the 60:180 set, suggesting that the model had reached optimal performance and that additional training data beyond $\tilde{2}40$ compounds provided diminishing returns in predictive accuracy.

To further assess model discrimination ability beyond accuracy alone, receiver operating characteristic (ROC) curves were generated for each dataset (Figure 2). ROC curves are created by finding the true positive rate (TPR) and false positive rate (FPR) at different probability threshold values, and the area under the ROC curve (AUC-ROC) can be used to compare two similarly balanced datasets. AUC-ROC values improved with dataset size: the 6:18 dataset achieved an AUC of 0.5, the 60:180 dataset reached 0.95, and the 547:1,541 dataset attained an AUC of 0.97. The AUC-ROC values for both larger datasets exceeded 0.90, indicating excellent discriminative performance and confirming the model's ability to reliably distinguish active from inactive compounds. The similarity in AUC between the two larger datasets further supports the conclusion that model performance plateaued beyond approximately 240 training compounds.

Although the 60:180 and 547:1,541 datasets achieved equivalent accuracy, the larger 547:1,541 dataset was selected for virtual screening to maximize chemical diversity and provide more robust representation of the active/inactive chemical space. The inclusion of structurally diverse inactive compounds was expected to improve the model's generalization to novel natural product scaffolds. Performance evaluation of the final Random Forest classifier on the

Ligand	SMILES	Affinity (kcal/mol)	Zn distance (Å)	Coordinating Group	Key Residue H-bond Interactions
ligand 1	CC(=O)O[C@@H]1c2c(ccc3ccc(=O)oc23)OC(C)(C)[C@@H]1OC(=O)CC(C)C	−7.345	3.84	aromatic ring	GLU114 (2), ASN106, THR110
ligand 2	COc1cc(CCc2ccc(O)cc2)cc(OC)c1	−6.384	4.58	aromatic ring	LYS328, GLU114, HIS465
ligand 3	COc1cc(O)cc(CCc2cc(C)c(OC)c(O)c2)c1	−7.228	2.66	methoxyl	LYS328, GLU114, GLY384, THR280
ligand 4	C/C=C(\C)C(=O)OC(C)(C)[C@@H]1Cc2cc3ccc(=O)oc3cc2O1	−7.262	2.97	methoxyl	LYS328, HIS465
ligand 5	CC(C)=CC(=O)O[C@@H]1c2cc3ccc(=O)oc3cc2O[C@H]1C(C)(C)O	−7.677	4.26	carbonyl	LYS328, GLU114, HIS465, THR280
ligand 6	COc1c2c(c(CC=C(C)C)c3oc(=O)ccc13)OC(C)(C)C=C2	−6.892	3.03	methoxyl	LYS328, GLU114
ligand 7	CC(C)=CC(=O)O[C@@H]1c2c(ccc3ccc(=O)oc23)OC(C)(C)[C@@H]1OC(=O)CC(C)C	−7.606	4.03	aromatic ring	GLU114 (2), ASN106, THR110
ligand 8	Cc1cc(C)c(NC(=O)COc2ccc3c(C)cc(=O)oc3c2)c(C)c1	−7.612	3.35	carbonyl	HIS465 (2), HIS453, THR280 (2)
ligand 9 (i1)	CC(=O)O[C@H]1[C@H](OC[C@H](C)C(C)=O)c2c(ccc3ccc(=O)oc23)OC1(C)C	−8.129	3.36	carbonyl	THR280 (2), ASN106, HIS465 (2)
ligand 9 (i2)	CC(=O)O[C@@H]1[C@H](OCC(C)=C(C)O)c2c(ccc3ccc(=O)oc23)OC1(C)C	−7.972	3.04	hydroxyl	THR280 (2), HIS465, VAL107, MET103
ligand 9 (i3)	C=C(O)[C@@H](C)CO[C@@H]1c2c(ccc3ccc(=O)oc23)OC(C)(C)[C@@H]1OC(C)=O	−7.739	3.07	hydroxyl	THR280 (2), HIS465
ligand 10 (i1)	CCCC(=O)c1c(O)c(CC=C(C)C)c(O)c2c([C@H](CC)OC(C)=O)cc(=O)oc12	−6.799	3.55	carbonyl	HIS465, ASN106, GLY384, GLU114
ligand 10 (i2)	CCC=C(O)c1c(O)c(CC=C(C)C)c(O)c2c([C@H](CC)OC(C)=O)cc(=O)oc12	−6.760	3.88	carbonyl	HIS465, ASN106, THR280, GLU114 (2), GLY384
ligand 10 (i3)	CCCC(=O)c1c2oc(O)cc([C@H](CC)OC(C)=O)c-2c(=O)c(CC=C(C)C)c1O	−7.154	2.34	hydroxyl	THR280, ALA281, ASN106 (2), LYS328, GLU114, HIS465
ligand 10 (i4)	CCC=C(O)c1c2oc(O)cc([C@H](CC)OC(C)=O)c-2c(=O)c(CC=C(C)C)c1O	−6.825	3.68	methyl	LYS328 (2), SER325, HIS383, THR110, ASN106
ligand 10 (i5)	CCCC(=O)c1c2oc(O)cc([C@H](CC)OC(C)=O)c-2c(O)c(CC=C(C)C)c1=O	−6.951	3.52	methyl	LYS328 (2), SER325, HIS465, ASN106, THR110, GLU114

Table 3: The docking results of the 10 ten natural products ranked by Random Forest confidence scores. Docking affinity was found through docking using the AutoDock4Zn forcefield run with an exhaustiveness of 32. Zn distance was measured from the ligand atom closest to the central Zn ion.

held-out test set (30% of data) yielded an overall accuracy of 93% (Table 2). Precision was 88% for active compounds and 95% for inactive compounds, while recall was 86% for active compounds and 96% for inactive compounds. The corresponding F1-scores were 87% for active compounds and 95% for inactive compounds, confirming decently balanced performance despite the class imbalance. Precision was 88% for active compounds and 95% for inactive compounds, while recall was 86% for active compounds and 96% for inactive compounds (Table 2). The modest reduction in active compound recall reflects the challenge of identifying minority class samples in an imbalanced dataset but remains acceptable for screening applications with subsequent validation steps.

Bit	Active Frequency	Inactive Frequency	Number of Compounds with Bit	Original Substructures SMILES
Bit_578	0.7514	0.3206	1222	CCCCC;CO;c=C(O)CCC;cO;cc(c)N(CC)CC
Bit_694	0.5210	0.1843	53	C=c(c)c;CC(C)CC[NH+];CCC;CCC=C(C)C;CCCC(N)=O
Bit_875	0.7642	0.3686	1444	C=C(O)CC(=c)O;CC(=O)NC(C)C;CNCCN;C[N@@H+](C)CC(=O)O;C[N@H+](C)CC(=O)O;ccc;ccc(cc)C(C)(C)C
Bit_580	0.5430	0.2291	679	C=C(C)C=CC;C=C(C)c(cc)c(c)O;CC(C)(C)O;CC(C)[NH3+];CC(CN)N(C)C;CCCC(C)(C)C;CCCC(C)(C)O;cc(C)c(C)c(=O)o;cc(c)CC(C)N;ccc(CC)c(c)O;coc
Bit_656	0.6088	0.3251	18	C=C(C)C=Cc;C=CC(c(c)c)C(O)O;CC=CC(C)O;CS[As](C)C;C[Si]
Bit_700	0.1517	0.0247	124	C[NH2+]C;cC(O)=CC(C)C;cc(c)C(OC)C(C)O;cc(c)n;ccc(C)c(c)O
Bit_356	0.9945	0.7657	157	CC(C)(C)CC(C)(C)O;CC(C)O;CCCC(C)N;CCN(C(=O)[O-])C(=O)[O-];C[C@@H](C)O;C[C@H](C)O;cC=NC;cc(C)o;ccc(Oc)c(c)N
Bit_147	0.0183	0.2440	295	C=CC;CC(=O)O;CC(c)(C)CCN;CC=CCC;COC;C[NH+](C)CC(=O)[O-];cC(C)(C)CCN
Bit_444	0.3163	0.1058	347	CCCCO;CNCCN;c=CCCC;cc(=O)o;cc(C)c(CC)c(c)c;cc(O)c(C)c(c)O;cc(c)OC(C)=O;cccc(C)n

Table 4: Summary of Morgan fingerprint bits and their associated substructures. Active and inactive frequencies are reported alongside the number of compounds to each bit.

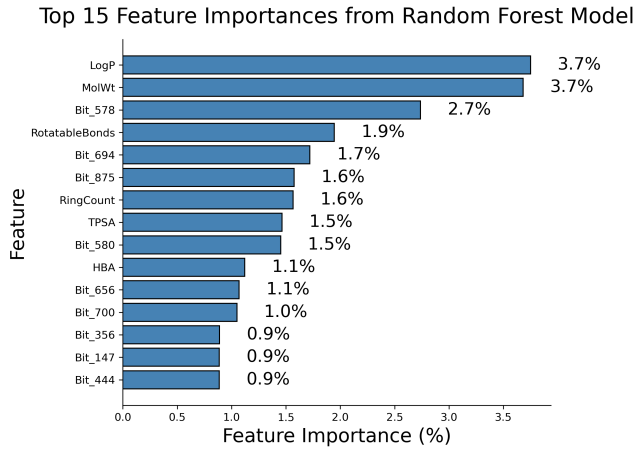


Figure 3: The top 15 feature importances from the 547:1541 Random Forest model trained using the full training dataset. Feature importance was normalized to 100%. Bits are part of the 1024-Bit Morgan Fingerprint feature (numbered Bit_0 to Bit_1023).

Among physicochemical descriptors, logP, molecular weight, and rotatable bonds were the most informative, collectively accounting for 3.75% of the model's predictive power (Figure 3).

Additionally, specific Morgan fingerprint bits (578 and 694) contributed significantly to classification, capturing structural features associated with EptA inhibition (Figure 3). This pattern indicates that both molecular fingerprints and drug-like properties are critical for distinguishing EptA inhibitors from inactive compounds. Based on these results, the 547:1,541 model was deployed to screen the COCONUT natural products database for novel EptA inhibitor candidates.

To provide chemical interpretability of these features, we analyzed the substructures associated with the most predictive Morgan fingerprint bits (Table 4). Bit 578 (active frequency: 75.1% vs. inactive: 32.1%) encodes hydroxyl groups, enols, and tertiary amines that may participate in zinc coordination or hydrogen bonding. Bit 694 (active: 52.1% vs. inactive: 18.4%) captures nitrogen-containing groups. Bit 147 serves as a negative predictor (active: 1.8% vs. inactive: 24.4%), while Bit 356 (active: 99.5%) appears to capture a core scaffold requirement. These associations link model predictions to specific chemical features relevant to EptA inhibition.

Following model validation, the trained Random Forest classifier was applied to screen 5,000 Lipinski-compliant compounds randomly sampled from the COCONUT natural products database. The model predicted 166 compounds (3.3% of the screened set) as active based on classification probability scores above 0.5. To prioritize compounds most likely to exhibit genuine EptA inhibitory activity, the top 10 candidates with the highest confidence scores (ranging from 0.85 to 0.98) were selected for validation through high-exhaustiveness molecular docking and detailed binding mode analysis (Table 3 lists compound identifiers and confidence scores). High-accuracy docking of the 10 top-ranked compounds confirmed strong predicted binding affinity across the majority of candidates. Docking scores ranged from -6.384 to -8.129 kcal/mol, with eight of the ten compounds achieving scores $\leq$ -7 kcal/mol (Table 3). For comparison, the known EptA inhibitor valnemulin yielded a docking score of -7.264 kcal/mol under identical conditions, indicating that the top natural product candidates exhibit comparable or superior predicted binding affinity. Compounds L9 and L10 (with multiple isomers/conformers analyzed for each) emerged as the

most promising hits, with L9 isomer 1 achieving the strongest binding score of -8.129 kcal/mol, significantly higher than previous EptA inhibitors. These results validate the Random Forest model's ability to successfully identify potent EptA binders from a large chemical library and demonstrate that structurally diverse natural products can achieve effective active site engagement.

3.2 Binding Mode Analysis of Top-Ranked Inhibitors

To validate the model's predictions and characterize key binding interactions, the top 10 compounds ranked by prediction confidence were subjected to high-exhaustiveness docking and detailed interaction analysis. Visual inspection of the top-ranked docking poses revealed consistent binding modes across all candidates. All 10 compounds docked to the same binding pocket adjacent to the catalytic Zn^{2+} ion, confirming that the model successfully identified compounds targeting the same enzymatic active site. To assess potential zinc coordination, the distance from the Zn^{2+} ion to the nearest ligand heteroatom was measured using each ligand's binding pose (Table 3). Zn-ligand distances ranged from 2.34 to 4.58 Å, spanning both direct coordination geometry (< 3.5 Å) and extended hydrogen bonding networks (> 3.5 Å). Typical Zn^{2+} coordination bonds with range from 2.2-2.6 Å for direct interactions with elements such as oxygen or sulfur [9]. Compounds with distances in the range of 2.3-3.5 Å likely engage in direct or near-direct coordination with the metal center, while those with distances > 3.5 Å are stabilized primarily through other interactions with surrounding residues or solvent-mediated contacts. A clear structure-affinity relationship emerged: compounds with Zn distances of 3.0-3.5 Å exhibited the strongest binding affinities (docking scores ≤ -7.5 kcal/mol), while those with distances > 3.5 Å showed progressively weaker binding (scores: ≥ -7 kcal/mol) (Figure 4). Interestingly, compounds clustered into two binding mode categories: (1) direct Zn coordination mode (3.0-3.5 Å) with carbonyl or hydroxyl groups oriented toward the metal, and (2) peripheral binding mode (4.0-4.5 Å) stabilized by hydrogen bonds with catalytic residues. Both modes produced favorable binding scores, suggesting multiple viable strategies for EptA inhibition.

Detailed analysis of individual binding poses revealed that zinc-coordinating functional groups were key determinants of binding affinity (Table 3). Compound L9 (isomer 1) exhibited the strongest binding affinity (docking score: -8.129 kcal/mol), with a carbonyl oxygen positioned at an optimal 3.36 Å from the Zn^{2+} ion, consistent with favorable electrostatic interaction geometry. Despite also presenting carbonyl groups toward the metal center, L10 isomers 1 and 2 showed weaker binding (distances: 3.55 and 3.88 Å, respectively; scores: -6.799 and -6.76 kcal/mol), likely due to suboptimal coordination geometry outside the 3.0-3.5 Å favorable range. L9 isomers 2 and 3 also demonstrated strong binding (scores: -7.972 and -7.739 kcal/mol), mediated by hydroxyl groups positioned at 3.04 and 3.07 Å, respectively, from the zinc ion, indicating that both carbonyl and hydroxyl moieties can serve as effective Zn-coordinating pharmacophores. L10 isomer 3 exhibited a high but slightly reduced binding affinity (score: -7.154 kcal/mol) despite hydroxyl-Zn coordination, with a distance of 2.34 Å, which is below the optimal 3.0-3.5 Å range. This reduced affinity may result from

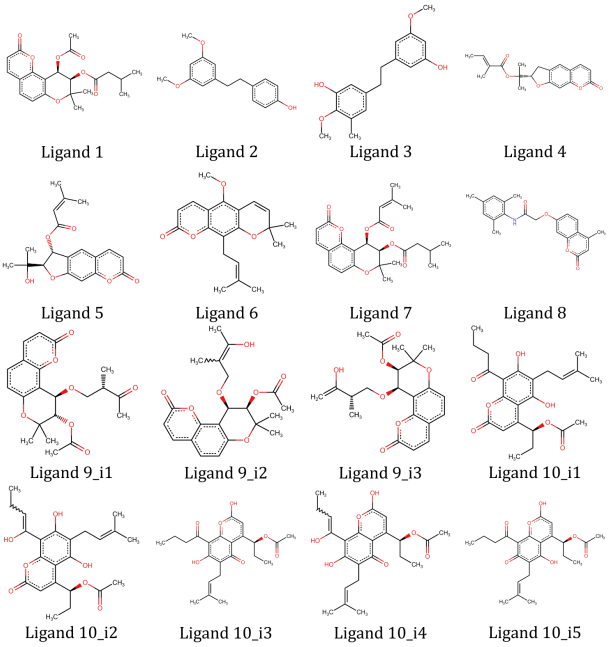


Figure 4: Diagrams of the 2D structures of the top 10 predicted inhibitors. The labels on each structure correspond with the labels in Table 3.

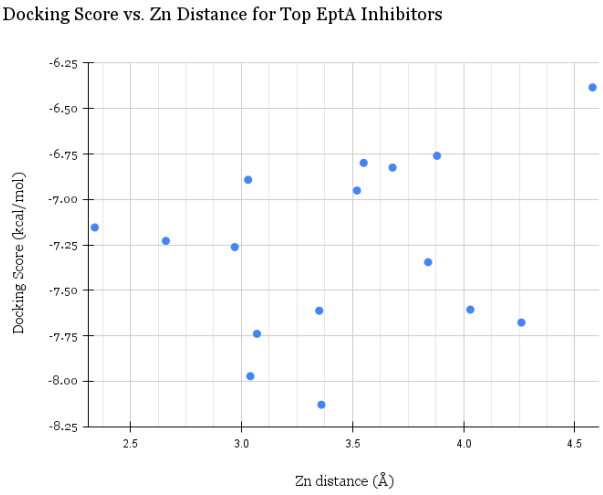


Figure 5: A graph of the Docking Score vs. Zn distance of the top 10 predicted inhibitors. Each point represents one of the inhibitors, and the lower docking scores indicate stronger binding affinity

unfavorable steric clashes or orbital overlap at very short distances, highlighting the importance of precise coordination geometry.

Beyond zinc coordination, additional stabilizing interactions were observed across the top candidates. Hydrogen bonding networks with catalytic residues such as ASN106, THR110, GLU114,

THR280, SER325, LYS328, HIS383, GLY384, HIS453, HIS465 were present in all inhibitors. Compounds with hydrogen bonding with residues spaced throughout the binding pocket exhibited high binding affinities, such as ligand 10 (i3) and ligand 3. Additionally, compounds with hydrogen bonds in the direction of the Zn ion exhibited the highest binding affinities, such as ligand8, ligand 9 (i1), and ligand 9 (i2) (scores: -7.612, -8.129, -7.972). The most common hydrogen bonding residues were GLU114 and HIS465, which formed hydrogen bonds with 10 and 11 out of 16 inhibitors, respectively. Some compounds such as ligand 9 (i2) and ligand 10 (i3) formed hydrophobic contacts with MET103, VAL107, and ALA281, further contributing to binding stability. Notably, the most potent inhibitors (L9 variants) combined optimal Zn coordination distance with complementary interactions to THR280, HIS465, and ASN106, suggesting a multi-point binding mechanism. Comparison to the known inhibitor Valnemulin revealed similarities in binding mode. The top natural product candidates showed comparable or superior docking scores (natural products: -6.384 to -8.129 kcal/mol vs. Valnemulin: -7.264 kcal/mol). Valnemulin also formed hydrogen bonds with similar key active site residues, such as GLU114 and SER325. This suggests that natural product scaffolds can achieve effective EptA inhibition through similar mechanisms.

4 Discussion

This study identified novel natural product inhibitors of EptA through combined machine learning and structure-based virtual screening. The top-ranked compounds represent promising candidates for EptA inhibition based on both their high model confidence scores (0.76-0.95) and favorable docking profiles (binding energies: -6.384 to -8.129 kcal/mol)(Table 3). Several of these molecules displayed structural features, including multiple aromatic rings, hydrogen bond donors and acceptors, and heteroatoms positioned to coordinate with the zinc ion in the catalytic site, consistent with previously reported inhibitors, such as valnemulin and other small molecules that disrupt EptA-mediated lipid A modification. Importantly, the identified natural products represent chemically distinct scaffolds from known synthetic inhibitors, potentially offering advantages such as reduced toxicity, improved bioavailability, or novel mechanisms of action; benefits commonly associated with natural product-derived therapeutics.

Detailed binding mode analysis revealed that the top-ranked compounds adopted consistent binding poses within the EptA active site, with all 10 candidates docking to the same zinc-proximal cavity. Key stabilizing interactions included direct or near-direct coordination with the catalytic Zn^{2+} ion through carbonyl and hydroxyl groups (distances: 2.3-4.6 Å) and hydrogen bonding networks with catalytic residues such as HIS453, GLU114, ASN106, THR280, and THR110. Notably, compounds with optimal Zn coordination distances (3.0-3.5 Å) exhibited the strongest binding affinities, highlighting the importance of precise metal-ligand geometry. The convergence of these interaction patterns across structurally diverse scaffolds validates the Random Forest model's ability to identify compounds based on functionally relevant pharmacophoric features rather than superficial structural similarity. This demonstrates that the machine learning approach successfully captured the key determinants of EptA binding.

Analysis of the most predictive Morgan fingerprint bits revealed chemical features consistent with EptA's catalytic mechanism (Table 4). Bits encoding hydroxyl groups, enols, and tertiary amines were enriched in active compounds. These functional groups can serve as electron donors in zinc coordination or form hydrogen bonds with active-site residues such as GLU114 and HIS465. The enrichment of nitrogen-containing moieties aligns with known zinc-binding pharmacophores in metalloenzyme inhibitors. Conversely, substructures associated with Bit 147 were depleted in active compounds, suggesting these chemical features interfere with binding. The near-universal presence of Bit 356 in active compounds indicates a conserved scaffold requirement. These findings support the model's ability to learn chemically meaningful patterns rather than arbitrary correlations.

While these computational findings are promising, several important limitations must be considered when interpreting the results. First, this study employed entirely computational methods, and the identified compounds remain experimentally unvalidated. While computational screening provides efficient prioritization, it cannot replace empirical testing. Docking scores and machine learning predictions provide useful prioritization but cannot fully capture dynamic biological factors such as membrane permeability, metabolic stability, or potential off-target effects. Additional factors not captured by these methods include protein flexibility, solvent effects, entropic contributions to binding, and the influence of the cellular environment on compound efficacy.

Second, the training dataset relied on docking-derived activity labels rather than experimental IC_{50} values, which may introduce systematic bias if the scoring function consistently over- or underestimates binding affinity for certain chemical classes. Although the AutoDock4Zn forcefield improves accuracy for zinc-dependent enzymes compared to standard forcefields, it remains a computational approximation that may not fully capture quantum mechanical effects or polarization in metal-ligand interactions. Orthogonal methods, such as molecular dynamics simulations or ligand-based pharmacophore modeling, could provide independent validation of the top candidates. Additionally, model performance for active compounds was slightly lower than for inactive compounds, indicating a potential underrepresentation of certain structural classes in the training data. This class imbalance (1:3 active:inactive ratio) may have reduced the model's sensitivity to rare but potentially important structural features present in highly active compounds. Future iterations could benefit from balanced datasets or alternative sampling strategies such as SMOTE (Synthetic Minority Over-sampling Technique). A systematic comparison of feature types, such as Morgan fingerprints alone versus physicochemical descriptors, could also optimize the precision-recall trade-off for the active class.

Third, the study screened only 5,000 compounds from the COCONUT database (<1% of the full library), meaning potentially superior candidates may remain unidentified. A comprehensive screen of the entire database would be valuable but was beyond the computational scope of this initial study. The Random Forest model can process all 695,000 compounds in minutes, and GPU-accelerated tools such as AutoDock-GPU would enable validation of a larger candidate set in future work.

Several experimental and computational approaches could build upon these findings. Most immediately, in vitro enzymatic inhibition assays using purified recombinant EptA would confirm predicted inhibitory activity and establish IC_{50} values for the top candidates. Subsequently, whole-cell assays in polymyxin-resistant bacterial strains (e.g., E. coli, K. pneumoniae, or A. baumannii expressing EptA or MCR variants) could assess whether these compounds restore colistin susceptibility and determine minimum inhibitory concentrations (MICs). For lead optimization, co-crystal structures of EptA bound to the most potent inhibitors would provide atomic-resolution insight into binding modes, enabling structure-guided medicinal chemistry to improve potency, selectivity, and drug-like properties.

ADMET (absorption, distribution, metabolism, excretion, toxicity) profiling is essential for translational development. Natural products often present challenges, including poor aqueous solubility, limited membrane permeability, and metabolic instability. Computational tools such as SwissADME, pkCSM, or ProTox-II can predict these properties and identify potential toxicity liabilities before experimental testing. Synthetic accessibility scores would further prioritize compounds based on the feasibility of obtaining sufficient quantities. Several of our top hits are naturally occurring compounds that may be available from commercial sources or have established isolation protocols.

To improve the computational pipeline, expanding the training dataset with experimentally validated actives and inactives would reduce docking-dependent bias and enhance model generalization. Incorporating additional machine learning architectures (e.g., graph neural networks) or ensemble methods could further improve prediction accuracy [5]. GNNs can learn directly from molecular graphs and may better represent complex substructures and three-dimensional conformational information, though the interpretability of fingerprint-based features was valuable for this study. Finally, expanding virtual screening to the full COCONUT database (~695,000 compounds) or other natural product libraries could identify additional candidates with diverse scaffolds and mechanisms.

5 Conclusions

This study successfully employed an integrated computational approach combining Random Forest machine learning and zinc-aware molecular docking to identify natural product–derived inhibitors of EptA, a key enzyme mediating polymyxin resistance in multidrug-resistant Gram-negative bacteria. A systematically optimized Random Forest classifier achieved 93% accuracy and AUC-ROC > 0.90 and was used to screen 5,000 COCONUT natural products, yielding 166 predicted actives from which the top 10 underwent high-exhaustiveness docking. These candidates showed strong binding affinities (-6.384 to -8.129 kcal/mol) and consistent binding modes with optimal zinc coordination and key catalytic interactions, demonstrating the effectiveness of integrating machine learning with structure-based methods for rapidly and cost-effectively prioritizing diverse natural products. Importantly, the identified compounds represent chemically distinct scaffolds from existing synthetic EptA inhibitors, potentially offering advantages such as novel mechanisms of action, reduced toxicity, or improved pharmacological properties, and by targeting EptA, this approach offers a viable strategy to restore polymyxin efficacy against multidrug-resistant pathogens.

Experimental validation now represents the crucial next step to advance these predictions toward therapeutic development. In vitro enzymatic assays with recombinant EptA will establish IC_{50} values and confirm inhibitory activity, followed by whole-cell susceptibility testing in polymyxin-resistant clinical isolates to determine whether these compounds restore colistin sensitivity through MIC measurements. Co-crystal structures will validate binding modes and guide structure-based optimization, while parallel ADMET profiling and medicinal chemistry efforts will address liabilities in absorption, metabolism, and toxicity. Ultimately, this workflow will advance the most promising candidates toward preclinical development, and beyond EptA inhibition, the integrated pipeline can be adapted to other metalloenzyme targets and resistance mechanisms, providing a generalizable strategy for accelerating antibacterial drug discovery in the face of escalating antimicrobial resistance.

References

[1] Branda, F., Scarpa, F.: Implications of artificial intelligence in addressing antimicrobial resistance: Innovations, global challenges, and healthcare's future. Antibiotics p. 502 (2024). https://doi.org/10.3390/antibiotics13060502
[2] Breiman, L.: Random forests. Machine Learning pp. 5–32 (2001). https://doi.org/10.1023/A:1010933404324
[3] Chandrasekhar, V., Rajan, K., Kanakam, S.R.S., Sharma, N., Weißenborn, V., Schaub, J., Steinbeck, C.: Coconut 2.0: a comprehensive overhaul and curation of the collection of open natural products database. Nucleic Acids Research pp. D634–D643 (2025). https://doi.org/10.1093/nar/gkae1063
[4] Cui, X.D., Zhang, J.K., Sun, Y.W., Yan, F.B., Zhao, J.F., He, D.D., Pan, Y.S., Yuan, L., Zhai, Y.J., Hu, G.Z.: Synergistic antibacterial activity of baicalin and edta in combination with colistin against colistin-resistant salmonella. Poultry Science p. 102346 (2023). https://doi.org/10.1016/j.psj.2022.102346
[5] El-Behery, H., Attia, A.F., El-Fishawy, N., Torkey, H.: An ensemble-based drug-target interaction prediction approach using multiple feature information with data balancing. Journal of Biological Engineering p. 21 (2022). https://doi.org/10.1186/s13036-022-00296-7
[6] El-Sayed Ahmed, M.A.E.G., Zhong, L.L., Shen, C., Yang, Y., Doi, Y., Tian, G.B.: Colistin and its role in the era of antibiotic resistance: an extended review (2000–2019). Emerging Microbes & Infections pp. 868–885 (2020). https://doi.org/10.1080/22221751.2020.1754133
[7] Huang, J., Zhu, Y., Han, M.L., Li, M., Song, J., Velkov, T., Li, C., Li, J.: Comparative analysis of phosphoethanolamine transferases involved in polymyxin resistance across 10 clinically relevant gram-negative bacteria. International Journal of Antimicrobial Agents pp. 586–593 (2018). https://doi.org/10.1016/j.ijantimicag.2017.12.016
[8] Jen, F.E.C., El-Deeb, I.M., Zalucki, Y.M., Edwards, J.L., Walker, M.J., von Itzstein, M., Jennings, M.P.: A drug candidate for alzheimer's and huntington's disease, pbt2, can be repurposed to render neisseria gonorrhoeae susceptible to natural cationic antimicrobial peptides. Journal of Antimicrobial Chemotherapy pp. 2850–2853 (2021). https://doi.org/10.1093/jac/dkab291
[9] Laitaoja, M., Valjakka, J., Jänis, J.: Zinc coordination spheres in protein structures. Inorganic Chemistry pp. 10983–10991 (2013)
[10] Mlynarcik, P., Kolar, M.: Molecular mechanisms of polymyxin resistance and detection of mcr genes. Biomedical Papers pp. 28–38 (2019). https://doi.org/10.5507/bp.2018.070
[11] Mullally, C., Stubbs, K.A., Thai, V.C., Anandan, A., Bartley, S., Scanlon, M.J., Jarvis, G.A., John, C.M., Lim, K.Y.L., Sullivan, C.M., Sarkar-Tyson, M., Vrielink, A., Kahler, C.M.: Novel small molecules that increase the susceptibility of neisseria gonorrhoeae to cationic antimicrobial peptides by inhibiting lipid a phosphoethanolamine transferase. Journal of Antimicrobial Chemotherapy pp. 2441–2447 (2022). https://doi.org/10.1093/jac/dkac204
[12] RDKit Community: Rdkit: Open-source cheminformatics (release 2025.09.1) (2025). https://doi.org/10.5281/zenodo.17232453
[13] Salam, M.A., Al-Amin, M.Y., Salam, M.T., Pawar, J.S., Akhter, N., Rabaan, A.A., Alqumber, M.A.A.: Antimicrobial resistance: A growing serious threat for global public health. Healthcare p. 1946 (2023). https://doi.org/10.3390/healthcare11131946
[14] Santos-Martins, D., Forli, S., Ramos, M.J., Olson, A.J.: Autodock4 zn: An improved autodock force field for small-molecule docking to zinc metalloproteins. Journal of Chemical Information and Modeling pp. 2371–2379 (2014). https://doi.org/10.1021/ci500209e

[15] Schrödinger, LLC: The pymol molecular graphics system (version 3.0)
[16] Suardíaz, R., Lythell, E., Hinchliffe, P., van der Kamp, M., Spencer, J., Fey, N., Mulholland, A.J.: Catalytic mechanism of the colistin resistance protein mcr-1. Organic & Biomolecular Chemistry pp. 3813–3819 (2021). https://doi.org/10.1039/D0OB02566F
[17] Thai, V.C., Stubbs, K.A., Sarkar-Tyson, M., Kahler, C.M.: Phosphoethanolamine transferases as drug discovery targets for therapeutic treatment of multidrug resistant pathogenic gram-negative bacteria. Antibiotics p. 1382 (2023). https://doi.org/10.3390/antibiotics12091382
[18] Word, J.M.: Reduce (version 3.16)
[19] Xie, S., Li, L., Zhan, B., Shen, X., Deng, X., Tan, W., Fang, T.: Pogostone enhances the antibacterial activity of colistin against mcr-1-positive bacteria by inhibiting the biological function of mcr-1. Molecules p. 2819 (2022). https://doi.org/10.3390/molecules27092819
[20] Xu, C., Zhang, Y., Ma, L., Zhang, G., Li, C., Zhang, C., Li, Y., Zeng, X., Li, Y., Dong, N.: Valnemulin restores colistin sensitivity against multidrug-resistant gram-negative pathogens. Communications Biology p. 1122 (2024). https://doi.org/10.1038/s42003-024-06805-2
[21] Xu, Y., Wei, W., Lei, S., Lin, J., Srinivas, S., Feng, Y.: An evolutionarily conserved mechanism for intrinsic and transferable polymyxin resistance. mBio (2018). https://doi.org/10.1128/mBio.02317-17
[22] Zhou, Y., Liu, S., Wang, T., Li, H., Tang, S., Wang, J., Wang, Y., Deng, X.: Pterostilbene, a potential mcr-1 inhibitor that enhances the efficacy of polymyxin b. Antimicrobial Agents and Chemotherapy (2018). https://doi.org/10.1128/AAC.02146-17
[23] Zhou, Y., Wang, J., Guo, Y., Liu, X., Liu, S., Niu, X., Wang, Y., Deng, X.: Discovery of a potential mcr-1 inhibitor that reverses polymyxin activity against clinical mcr-1-positive enterobacteriaceae. Journal of Infection pp. 364–372 (2019). https://doi.org/10.1016/j.jinf.2019.03.004

Received 23 December 2025; Accepted 30 January 2026