RESEARCH ARTICLE
DOI: 10.5152/TrendsPharm.2025.25018

Predicting Active Antimicrobial Compounds Using Machine Learning

ABSTRACT

Background: *Acinetobacter baumannii* is a multidrug-resistant (MDR) pathogen recognized by the World Health Organization as a critical priority due to its high prevalence in hospital-acquired infections and limited treatment options.

Methods: To address the urgent need for novel therapeutics, artificial intelligence (AI)-based approaches were applied to predict compounds with potential antibacterial activity against *A. baumannii*.

Results: The used training set included 11 084 compounds with experimentally determined minimum inhibitory concentrations (MICs). The prediction set to be analyzed included 5835 structurally diverse compounds filtered from the ZINC20 database. Molecular descriptors, MACCS keys, and Morgan fingerprints were generated using RDKit, and a random forest classifier was trained using scaffold-based cross-validation to classify compounds as active (MIC < 32 μg/mL) or inactive. The model achieved AUROC values of 0.73-0.82 and average precision scores of 0.84-0.90, demonstrating strong predictive performance. Application of the trained model to the prediction dataset identified 375 compounds (6%) as potentially active, including 7 high-confidence candidates (probability > .85).

Conclusion: Scaffold analysis revealed considerable structural diversity among predicted compounds, supporting the potential for novel chemotypes. These findings highlight the utility of AI-driven drug discovery workflows for accelerating the identification of antibacterial agents targeting MDR *A. baumannii*.

Keywords: *Acinetobacter baumannii*, antimicrobial compounds, machine learning

What is already known on this topic?

- *Acinetobacter baumannii* is recognized by the World Health Organization as a critical multidrug-resistant pathogen with very limited treatment options.
- Artificial intelligence–based methods are increasingly used to accelerate the early stages of antibacterial drug discovery.
- Previous research shows that machine learning models trained on laboratory growth-inhibition data can help identify compounds with antibacterial potential, but studies focused specifically on *Acinetobacter baumannii* are still limited.

What this study adds on this topic?

- This study develops an artificial intelligence model trained on more than eleven thousand compounds with experimentally determined growth-inhibition values to predict antibacterial activity against *Acinetobacter baumannii*.
- The model shows strong predictive performance and identifies three hundred seventy-five compounds as potentially active, including seven with very high confidence.
- The structurally diverse compounds discovered in this study introduce new chemical possibilities and demonstrate the practical value of artificial intelligence-guided workflows for finding new antibacterial agents targeting *Acinetobacter baumannii*.

INTRODUCTION

Acinetobacter baumannii is a multidrug-resistant (MDR) bacterial species listed among the top priority pathogens by the World Health Organization and is a major cause of severe infections in hospital settings.[1] It can lead to serious nosocomial infections, including ventilator-associated pneumonia, bloodstream infections, and sepsis, particularly in intensive care units.[2] Multidrug-resistant *A. baumannii* infections are associated with increased mortality due to their poor response to existing antibiotic treatments, creating an urgent need for the discovery of new antibacterial agents. The growing prevalence of antibiotic resistance, coupled with the bacterium's inherent and acquired resistance mechanisms, further complicates clinical management.[3]

Conventional antibiotic discovery is time-consuming, costly, and often yields limited success. In silico approaches and artificial intelligence (AI)-based methods have gained increasing importance for the rapid and reliable identification of novel active compounds.[4] AI-assisted drug discovery workflows generally involve several key steps. Initially, chemical data are collected from large-scale databases (e.g., PubChem, ChEMBL, ZINC20), and compounds are represented using molecular fingerprints and other descriptors that capture structural and chemical features with experimental data such as biological activities.[5,6] Machine learning (ML) algorithms are then employed to predict active and inactive compounds as basic classifications or further data via regression analysis. These predictions help prioritize high-probability candidates for lead molecules for designing and biological evaluation.[7] In the context of AI-assisted ML for prediction of inhibitor molecules against bacteria, particularly a large

İbrahim Arman

Department of Molecular Biology and Genetics, Zonguldak Bülent Ecevit University Faculty of Science, Zonguldak, Türkiye

Corresponding author:
İbrahim Arman
✉ ibrahim.arman@beun.edu.tr

Received: September 22, 2025
Revision Requested: October 07, 2025
Last Revision Received: October 13, 2025
Accepted: October 20, 2025
Publication Date: November 25, 2025

Cite this article as: Arman İ. Predicting active antimicrobial compounds using machine learning. *Trends in Pharmacy* 2025, 2, 0018, doi: 10.5152/TrendsPharm.2025.25018.

dataset of biological activities (e.g., minimum inhibitory concentration [MIC]) is critical for refining AI models and guiding candidate selection.[8,9]

Recent studies have demonstrated that AI and ML-based approaches are effective in antibacterial drug discovery, including the identification of potential candidates against *A. baumannii* and other MDR pathogens.[10] By employing diverse algorithms, fingerprint types, and descriptor combinations, predictive model performance has been enhanced, yielding highly accurate results. These methods not only save time and resources during preclinical evaluation but also provide more targeted candidates for subsequent laboratory testing.[11]

In this study, ML models were applied to predict potentially active compounds against *A. baumannii* based on a large dataset of MICs, and performance comparisons were conducted across different molecular fingerprints. The primary objective was to assess the effectiveness and reliability of AI approaches in antibacterial agent discovery and to provide preliminary data for the identification of novel compounds targeting *A. baumannii*.

MATERIALS AND METHODS

Data Curation

In this study, 2 types of datasets were custom-prepared for training ML models and predicting active compounds against *A. baumannii*. Dataset #1 included SMILES data with MIC values of commercial antibiotics, FDA-approved compounds, compounds in phase studies (phase 1, 2, and 3), and finally MICs of compounds from preclinical studies that were obtained from the ChEMBL.[12,13] The Dataset#2 included compounds with unknown MIC values in the literature to make predictions of active/non-active against *A. baumannii* that were obtained from ZINC20.[14] Because ZINC20 included a high number of compounds (up to 10 M), additional filtrations were applied based on in-stock properties, presence of 3D structure, non-reactiveness, and several physicochemical properties convenient to Lipinski's Rule of 5.[15] Additionally, compounds with high toxicity were discarded from the list based on in silico analysis. The SMILES generation and toxicity filtering of the compounds in Database #2 were performed using DataWarrior v6.5.1.

Both databases were prepared as .xlsx files including SMILES data, MIC values, unit (ug/mL) and species name for Database #1, and only SMILES data for Database #2.

System Preparation

Machine learning (training) and prediction studies were performed based on python scripts via Conda on the Windows Subsystem for Linux 2.5.10 operating on Windows 11. Various sub-softwares have been integrated to run the scripts. RDKit[16] was used to process molecular structures, desalt and normalize compounds in SMILES format, and generate molecular descriptors such as the Morgan fingerprint.[17] Scikit-learn (sklearn)[18] was used in the ML stages to train a classification model (RandomForestClassifier), evaluate model performance through cross-validation (scaffold-split), and calculate metrics based on AUROC and average precision (AP). Pandas was used to read, edit, and store Excel data files, while NumPy was used for matrix operations and numerical calculations. Bemis–Murcko scaffold analysis[14] was conducted with RDKit to assess structural similarity among the predicted active compounds.

Machine Learning

The analysis was performed using 4 different custom written scripts for training the system, making predictions, and assessing the similarity of compounds against *A. baumannii*. A final script (one_click_pipeline_v4.py) was designed to perform a one-click pipeline using the scripts via conda (environment name: micml2) for analyses.

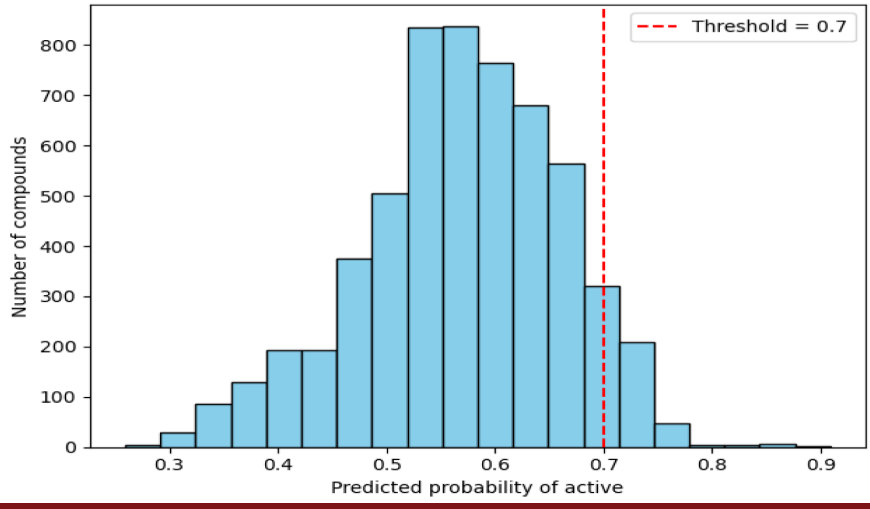

Figure 1. Probability distribution of active compounds with a 0.7 threshold from prediction analysis.

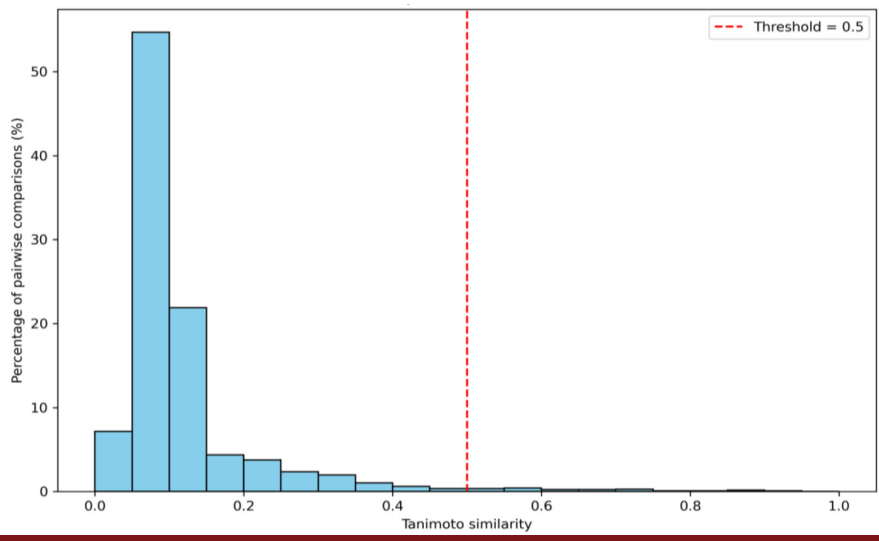

Figure 2. Distribution of pairwise Tanimoto similarities.

In the context of the scripts, each compound was represented in SMILES format, and during preprocessing steps, structures containing salt/counter ions were removed, leaving only the parent molecule. Descriptor calculations for the molecules were performed directly with the RDKit module,[16] and features included molecular weight (MW), logP, topological polar surface area (TPSA), number of rotatable bonds, hydrogen bond donors (HBD), and hydrogen bond acceptors (HBA). Additionally, Morgan fingerprints (ECFP, radius = 2, nBits = 2048) along with MACCS keys (166-bit) were calculated and combined with the descriptor matrix to represent the molecular structure.

The training script (train_pipeline_cleaned101.py) uses training data from the "training_compounds.xlsx" file. The training script included RandomForestClassifier[19] for model training, and compounds were classified as Active/Inactive based on their MIC values. The cutoff value for this classification was set as MIC < 32 µg/mL → Active, and MIC ≥ 32 µg/mL → Inactive. Model performance was assessed using a 5-fold cross-validation based on molecular scaffolds, and classification metrics such as AUROC, AP, and accuracy were calculated. AUROC ranges of 0.65-0.85 and AP ranges of 0.75-0.90 were used as reference values for QSAR and MIC prediction studies.

The prediction script (predict_preclinical101.py) was designed to predict the biological activity of new compounds from the "new_compounds.xlsx" based on the generated training model. The probability of each molecule being active is calculated using a predetermined strict threshold (0.75). The compounds are classified as active (1) or inactive (0), and the results are recorded along with the predicted probability.

Bemis–Murcko scaffold analysis was performed with an additional script (fingerprinting.py) using the prediction file (new_compounds_prediction.xlsx) to determine the skeletal structures of compounds predicted. For each compound, scaffolds were identified and the similarity distribution of active compounds was generated as a Tanimoto similarity[20] graphic.

All scripts used in the study, as well as training data, are provided in Supplementary File 1.

RESULTS

A custom-prepared training dataset (training_compounds.xlsx) was used in the study. The training set contained a total of 11 084 compounds with MIC values against *A. baumannii* obtained from the ChEMBL database (91

Table 1. The List of Predicted Highest Active Compounds (probability > .85), Along with Their Corresponding Properties

No	SMILES	Probability	ZINC Codes
1	C[C@@H]([C@H]([C@H]1C)N2C(C(O)=O)=C1S[C@@H]1C[NH2+][C@H](CNS(N)(=O)=O)C1)C2=O	0,90	ZINC390823134
2	CC(C)[C@H](C(OC[C@H](CO)OC[n]1c(N=C(N)NC2=O)c2nc1)=O)[NH3+]	0,89	ZINC11616800
3	CC(OC[C@H](CO)OC[n]1c(N=C(N)NC2=O)c2nc1)=O	0,88	ZINC22059880
4	CN(CN1[C@@H]([C@@H]2O)O[C@H](CO)[C@H]2O)C2=C1N=C(N)NC2=O	0,88	ZINC2390988
5	Cc1[nH+]cc[n]1Cc1cccc(CNC(NC(C(N)=O)C(N)=O)=O)c1	0,87	ZINC48447475
6	CN(C1)C(C(N=C(N)N2)=O)=C2N1[C@@H]([C@@H]1O)O[C@H](CO)[C@H]1O	0,86	ZINC2390988
7	NC(NCC(N(CC1)[C@@H](CS(C2)(=O)=O)[C@@H]2N1c1ncccn1)=O)=O	0,85	ZINC219712880

antibiotics, 10 993 approved and preclinical compounds) (Supplementary file 1). When multiple MIC values existed for the same compound, the geometric mean was calculated. The prediction set contained 5835 compounds with unknown MIC values, selected according to Lipinski's Rule of 5.

For model development, each compound was represented by Morgan fingerprints (ECFP, radius = 2, nBits = 2048), MACCS keys (166-bit), and molecular descriptors (MW, logP, TPSA, HBD, HBA, rotatable bonds). Scaffold-based cross-validation was used to minimize bias from structural similarity. A random forest classifier was trained with balanced class weighting, and a probability threshold of 0.75 was chosen for high-confidence predictions.

From the training, CV AUROC and AP values were found in the range of 0.73-0.82 (mean = 0.8) and 0.84-0.90 (mean = 0.89), respectively, and were at acceptable levels (AUROC 0.65-0.85, AP 0.75-0.90).

Among the predicted compounds, 375 (6%) were classified as active. Of these, 7 (0.2%) compounds were identified with a high probability of activity (probability > .85) (Figure 1). The list of predicted highest 7 active compounds (probability > .85), along with their corresponding properties, is provided in Table 1.

Scaffold analysis showed that the predicted active compounds did not adhere to a single chemical skeleton, maintaining structural diversity (Figure 2).

DISCUSSION

In this study, 2 datasets were used to predict potential *A. baumannii*-active compounds: a training dataset (17 944 compounds, with MIC values) and a prediction dataset (5835 compounds, without MIC values). Molecules were represented in SMILES format; salts and counterions were removed, leaving only the parent molecule. Molecular representations were enriched with Morgan fingerprints (radius = 2, nBits = 2048) combined with MACCS keys (166-bit). A random forest classifier was trained and cross-validated; AUROC and AP mean values were 0.8 and 0.89, respectively. In the prediction pipeline, 6% of the compounds were classified as active, and scaffold analysis demonstrated structural diversity, demonstrating the model's reliability.

Machine learning and AI-assisted methods have recently played an important role in discovering lead antibacterial compounds against resistant pathogens like *A. baumannii*. Several studies have predicted antimicrobial activity using random forest, support vector machines, and machine-learning models, and assessed molecular diversity with scaffold-based analyses.[21] Notably, Liu et al[21] discovered a new antibiotic, abaucin, using a deep-learning model trained on growth-inhibition data; the model's predictions were validated by laboratory testing. Similarly, Jia et al[22] integrated deep neural networks with gene-expression data to predict antimicrobial phenotypes in multidrug-resistant *A. baumannii* isolates. Palacios-Can et al[2] performed a Quantitative Structure-Activity/Property relationships model to perform a virtual screening on a natural product database against *Staphylococcus aureus*, *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*, and found 60 potential lead molecules. Fermandes et al[23] and Sayıner et al[24] performed Linear Regression and Artificial Neural Networks to predict active/inactive classification based on a relatively low number of compounds (n = 29) against *A. baumannii*. In contrast, this study was based on a large dataset for an active–inactive classification approach using a combination of Morgan fingerprints (radius = 2, nBits = 2048) and MACCS keys (166-bit), a random forest classifier. Scaffold analysis showed that the predicted active compounds covered structurally diverse scaffolds, and high prediction probabilities (>0.75) support the model's reliability for AI-assisted antimicrobial discovery applications. Recently, Stokes et al[10] used graph neural networks to predict antibiotic activity for an in-house activity-determined dataset including around 32k compounds against *S. aureus*.

These findings are consistent with similar approaches in the literature and indicate that ML-based active–inactive classification is an effective method for antibacterial discovery.[10,21]

This study has several limitations that the predicted active compounds were not experimentally validated, leaving their true activity uncertain. Using a binary classification oversimplifies antimicrobial effects, and reliance on a single random forest model may limit performance compared with other approaches. In addition, the lack of biological context and external validation raises concerns about generalizability, while important factors such as toxicity and pharmacokinetics were not considered.

This study highlights the potential of AI-driven ML to accelerate the discovery of novel antibacterial agents against *A. baumannii*. The developed model achieved strong predictive performance and identified structurally diverse high-confidence candidates, offering a practical pipeline to guide future experimental validation and drug development efforts.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Ethics Committee Approval: Ethical approval and patient consensus were not necessary due to the totally *in-silico* design of the study

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – I.A.; Design – I.A.; Supervision – I.A.; Resource – I.A.; Materials – I.A.; Data Collection and/or

Processing – I.A.; Analysis and/or Interpretation – I.A.; Literature Search – I.A.; Writing – I.A.; Critical Reviews – I.A.

Declaration of Interests: The author has no conflicts of interest to declare.

Funding: The author declares that this study received no financial support.

References

1. World Health Organization. *WHO bacterial priority pathogens list, 2024: Bacterial Pathogens of Public Health Importance to Guide Research, Development and Strategies to Prevent and Control Antimicrobial Resistance*. Geneva: World Health Organization; 2024. ISBN: 978-92-4-009346-1.
2. Palacios-Can FJ, Silva-Sánchez J, León-Rivera I, Tlahuext H, Pastor N, Razo-Hernández RS. Identification of a family of glycoside derivatives biologically active against Acinetobacter baumannii and other MDR bacteria using a QSPR model. *Pharmaceuticals (Basel)*. 2023;16(2):250. **[CrossRef]**
3. Geisinger E, Huo W, Hernandez-Bird J, Isberg RR. Acinetobacter baumannii: envelope determinants that control drug resistance, virulence, and surface variability. *Annu Rev Microbiol*. 2019;73:481-506. **[CrossRef]**
4. Bilal H, Khan MN, Khan S, et al. The role of artificial intelligence and machine learning in predicting and combating antimicrobial resistance. *Comput Struct Biotechnol J*. 2025;27:423-439. **[CrossRef]**
5. Wang T, Wu MB, Lin JP, Yang LR. Quantitative structure-activity relationship: promising advances in drug discovery platforms. *Expert Opin Drug Discov*. 2015;10(12):1283-1300. **[CrossRef]**
6. Muratov EN, Bajorath J, Sheridan RP, et al. QSAR without borders. *Chem Soc Rev*. 2020;49(11):3525-3564. **[CrossRef]**
7. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today*. 2015;20(3):318-331. **[CrossRef]**
8. Liu GY, Yu D, Fan MM, et al. Antimicrobial resistance crisis: could artificial intelligence be the solution? *Mil Med Res*. 2024;11(1):7. **[CrossRef]**
9. Ardila CM, González-Arroyave D, Tobón S. Machine learning for predicting antimicrobial resistance in critical and high-priority pathogens: a systematic review considering antimicrobial susceptibility tests in real-world healthcare settings. *PLoS One*. 2025;20(2):e0319460. **[CrossRef]**
10. Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. *Cell*. 2020;181(2):475-483. **[CrossRef]**
11. Bugeac CA, Ancuceanu R, Dinu M. QSAR models for active substances against Pseudomonas aeruginosa using disk-diffusion test data. *Molecules*. 2021;26(6):1734. **[CrossRef]**
12. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 2012;40(Database issue):D1100-D1107. **[CrossRef]**
13. Zdrazil B, Felix E, Hunter F, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res*. 2024;52(D1):D1180-D1192. **[CrossRef]**
14. Petrov KP, Bender A. An open-source implementation of the Scaffold Identification and Naming System (SCINS) and example applications. *J Chem Inf Model*. 2024;64(20):7905-7916. **[CrossRef]**
15. Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technol*. 2004;1(4):337-341. **[CrossRef]**
16. Landrum G. RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling. Accession. Available at: https://www.rdkit.org/RDKit_Overview.pdf.
17. Morgan HL. The generation of a unique machine description for chemical structures-A technique developed at Chemical Abstracts Service. *J Chem Doc*. 1965;5(2):107-113. **[CrossRef]**
18. Pedregosa F, Varoquaux G, Gramfort A, Michel V. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
19. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. **[CrossRef]**
20. Chung NC, Miasojedow B, Startek M, Gambin A. Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. *BMC Bioinformatics*. 2019;20(Suppl 15):644. **[CrossRef]**
21. Liu G, Catacutan DB, Rathod K, et al. Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. *Nat Chem Biol*. 2023;19(11):1342-1350. **[CrossRef]**
22. Liu J, Zhu X, Xu M, et al. A deep learning–guided discovery of abaucin, a narrow-spectrum antibacterial agent targeting Acinetobacter baumannii. *Nat Chem Biol*. 2023;19(9):1141-1149. **[CrossRef]**
23. Fernandes PO, Dias ALT, Dos Santos Júnior VS, et al. Machine learning-based virtual screening of antibacterial agents against methicillin-susceptible and resistant Staphylococcus aureus. *J Chem Inf Model*. 2024;64(6):1932-1944. **[CrossRef]**
24. Sayiner HS, Abdalrahm AAS, Basaran MA, Kovalishyn V, Kandemirli F. The quantum chemical and QSAR studies on Acinetobacter baumannii Oxphos inhibitors. *Med Chem*. 2018;14(3):253-268. **[CrossRef]**

Supplementary File 1. Includes Scripts and training dataset

https://drive.google.com/drive/folders/1ml_C6l-7ATj6SpaIyUtKs8HWGyQiWmky.