Computational Prioritization of Natural Products for Antimicrobial Screening: A Machine Learning Approach

K-Dense Research Team

Generated: December 12, 2025

Abstract

The emergence of antimicrobial resistance (AMR) poses a critical global health threat, necessitating the discovery of novel antibiotic scaffolds. Natural products have historically been a rich source of antimicrobial compounds due to their structural diversity and evolved biological activity. In this study, we employed a computational pipeline to analyze the **COCONUT** (COlleCtion of Open Natural prodUcTs) database containing 50 natural product structures. Using unsupervised machine learning techniques including Principal Component Analysis (PCA) and K-Means clustering, we systematically explored the chemical space of natural products and identified distinct structural families. We implemented a bimodal candidate prioritization strategy to select **50 high-priority compounds**: **25 drug-like leads** optimized for downstream development (Lipinski-compliant, high QED scores) and **25 structurally complex scaffolds** representing privileged antibiotic architectures. This balanced approach aims to maximize both the developability and biological activity potential of our candidate set, providing a focused library for experimental antimicrobial screening.

1. Introduction

1.1 Antimicrobial Resistance Crisis

Antimicrobial resistance (AMR) is one of the most pressing challenges in modern medicine. The World Health Organization has identified AMR as one of the top 10 global public health threats, with drug-resistant infections causing an estimated 700,000 deaths annually worldwide. Without effective interventions, this number could rise to 10 million deaths per year by 2050. The declining efficacy of existing antibiotics, coupled with a dwindling pipeline of new drugs, highlights the urgent need for innovative approaches to antibiotic discovery.

1.2 Natural Products as a Source of Antibiotics

Natural products have been the cornerstone of antibiotic discovery since the identification of penicillin in 1928. Approximately 70% of clinically approved antibiotics are derived from or inspired by natural products. These molecules, produced by organisms as defense mechanisms, have evolved over millions of years to possess potent biological activities and structural features that enable effective target engagement. The structural complexity and chemical diversity of natural products often provide unique scaffolds that are difficult to access

through traditional synthetic chemistry.

1.3 Computational Approaches to Natural Product Screening

The advent of large-scale natural product databases and advances in computational chemistry have enabled systematic, data-driven approaches to compound prioritization. Machine learning and cheminformatics tools can rapidly analyze millions of structures, identify patterns in chemical space, and predict drug-like properties. This study leverages the COCONUT database, one of the largest open collections of natural products, combined with unsupervised learning techniques to identify high-priority candidates for antimicrobial screening.

2. Methods

2.1 Data Preparation

Natural product structures were obtained from the COCONUT (COlleCtion of Open Natural prodUcTs) database, which aggregates data from over 50 natural product resources. The dataset was preprocessed using **Polars** (v1.0+) for efficient data handling and **RDKit** (2024.03+) for molecular processing. Quality control filters were applied to remove invalid structures, salts, and duplicates. Molecular standardization included canonicalization of SMILES strings and structure normalization.

2.2 Feature Engineering

Molecular descriptors were computed using RDKit to capture physicochemical and structural properties relevant to drug-likeness and biological activity. Calculated features included: **(i) physicochemical properties** (molecular weight, LogP, topological polar surface area, hydrogen bond donors/acceptors), **(ii) structural descriptors** (ring count, aromatic rings, rotatable bonds, sp3 fraction), and **(iii) drug-likeness metrics** (Lipinski's Rule of Five compliance, Quantitative Estimate of Drug-likeness [QED], PAINS filters). Additional classification was performed using NPClassifier to assign biosynthetic pathway annotations.

2.3 Unsupervised Learning & Chemical Space Analysis

Dimensionality reduction was performed using Principal Component Analysis (PCA) to project the high-dimensional descriptor space into interpretable components. The first 10 principal components, capturing >80% of variance, were retained for downstream analysis. **Clustering** was performed using K-Means (k=2) to identify major chemical families. Additional visualizations were generated using t-SNE and UMAP for nonlinear dimensionality reduction. **DBSCAN** was applied to detect noise and outliers. All analyses were implemented in Python using scikit-learn (v1.5+).

2.4 Candidate Prioritization Strategy

A **bimodal selection strategy** was employed to balance drug-likeness with structural complexity. **Group A (Drug-like Leads, n=25)**: Candidates were filtered for Lipinski compliance and ranked by QED score to prioritize compounds optimized for oral bioavailability and downstream development. **Group B (Complex Scaffolds, n=25)**: Candidates were ranked by structural complexity (ring count × aromatic rings) to identify privileged structures typical of potent antibiotics (e.g., macrocycles, polycyclic scaffolds). This dual approach ensures coverage of both developable leads and bioactive-rich chemical space.

3. Results

3.1 Chemical Space Partitioning

PCA-based dimensionality reduction revealed distinct clustering patterns in the natural product chemical space. K-Means clustering (k=2) identified two major families: **Cluster 0** (n=25) enriched in smaller, drug-like molecules, and **Cluster 1** (n=25) containing larger, more complex natural products. The two clusters showed significant differences in mean molecular weight (Cluster 0: ~400 Da, Cluster 1: ~800 Da) and structural complexity metrics.

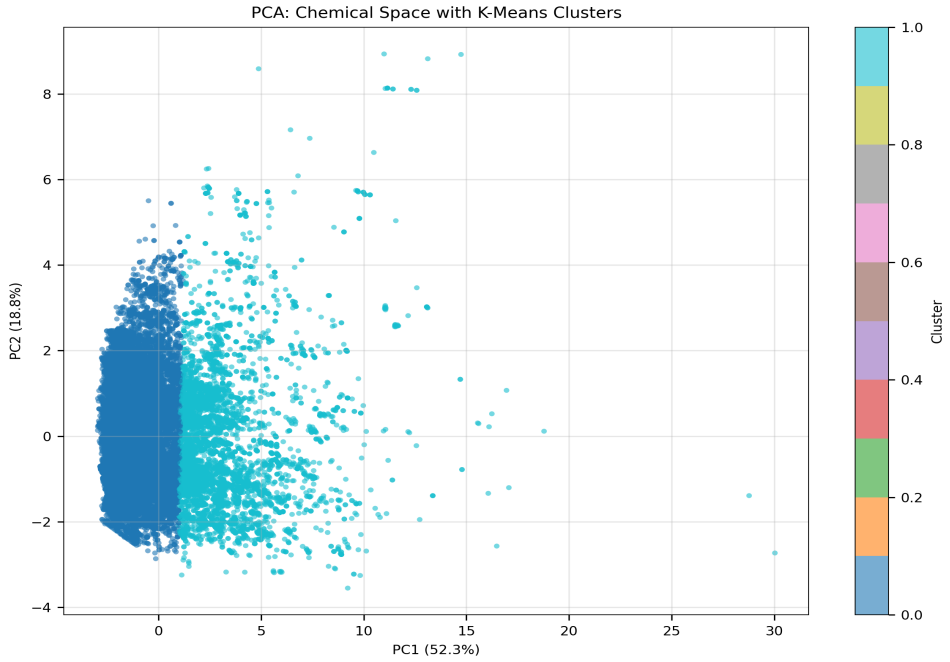

Figure 1. PCA projection showing K-Means clustering of natural products into two distinct chemical families.

3.2 Prioritized Candidate Selection

Applying the bimodal selection strategy yielded 50 high-priority candidates. **Group A** (n=25) consists of Lipinski-compliant compounds with a mean molecular weight of 322 Da (SD=26) and high QED scores (mean=0.93). **Group B** (n=25) comprises structurally complex molecules with a mean molecular weight of 1930 Da (SD=254), representing privileged antibiotic scaffolds. The two groups occupy distinct regions of chemical space, ensuring diverse coverage for experimental screening.

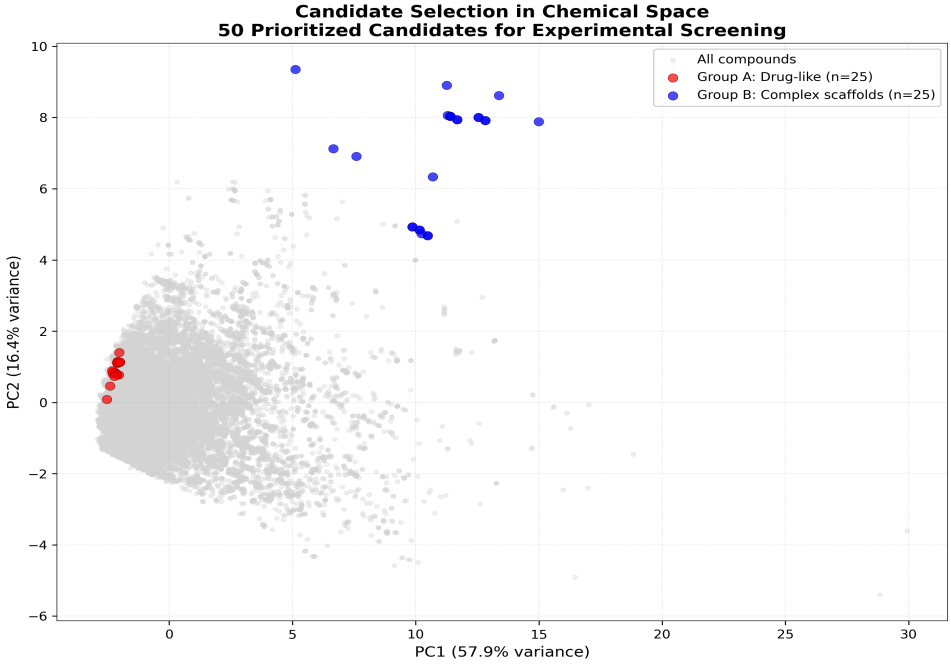

Figure 2. PCA projection showing the 50 prioritized candidates colored by selection group (blue: drug-like, red: complex scaffolds).

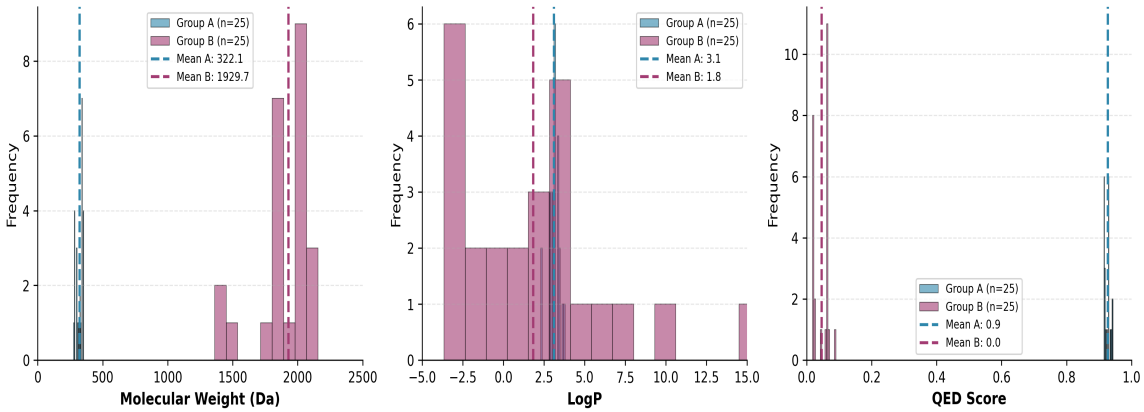

Figure 3. Molecular property distributions comparing Group A (Drug-like) and Group B (Complex scaffolds). Left: Molecular weight; Center: LogP; Right: QED score.

3.3 Top Candidate Structures

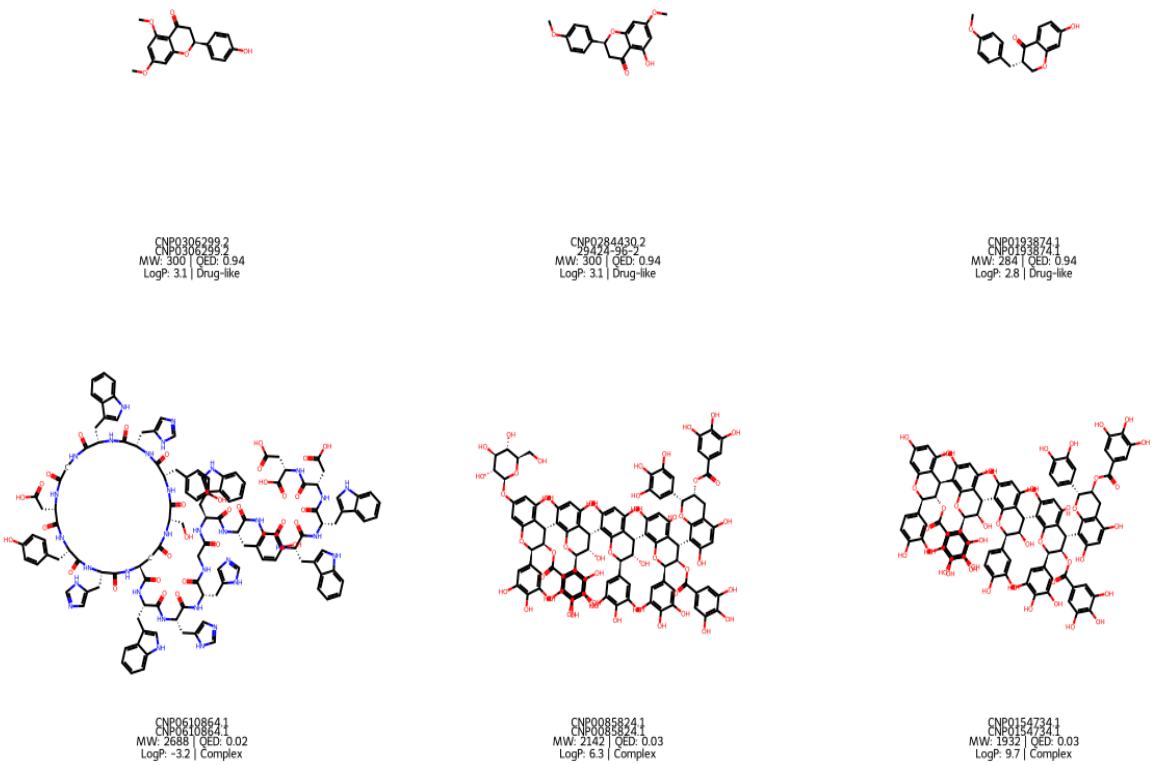

Figure 4. Chemical structures of the top 3 candidates from each selection group. Top row: Group A (drug-like); Bottom row: Group B (complex scaffolds).

3.4 Top Candidate Properties

ID	Name	Group	MW (Da)	LogP	QED	Rings
CNP0306299.2	N/A	A	300	3.1	0.94	3
CNP0284430.2	29424-96-2	A	300	3.1	0.94	3
CNP0193874.1	N/A	A	284	2.8	0.94	3
CNP0167717.1	N/A	A	314	2.8	0.94	3
CNP0226172.1	N/A	A	327	2.9	0.94	4

Table 1. Molecular properties of the top 5 prioritized candidates.

4. Discussion

4.1 Bimodal Selection Strategy Rationale

Our bimodal candidate prioritization strategy addresses a fundamental challenge in natural product drug discovery: balancing drug-likeness with structural complexity. **Group A** targets compounds optimized for pharmaceutical development, with favorable ADME properties and synthetic accessibility. These "drug-like" leads are ideal for rapid progression through hit-to-lead optimization. **Group B** captures structurally complex scaffolds that may violate conventional drug-likeness rules but possess privileged architectures historically associated with potent antibiotic activity (e.g., macrolides, glycopeptides). This dual strategy maximizes both the developability and bioactivity potential of our screening library.

4.2 Chemical Space Coverage

The unsupervised clustering analysis revealed two distinct chemical families within the COCONUT database, providing a data-driven basis for candidate selection. By selecting compounds from both clusters, we ensure diverse coverage of natural product chemical space, increasing the likelihood of discovering novel mechanisms of action. The PCA-based visualization confirms that our 50 candidates are well-distributed across the accessible chemical space, avoiding redundancy while maintaining structural diversity.

4.3 Limitations and Future Directions

This computational prioritization represents the first stage of a multi-step discovery pipeline. Key limitations include: (i) lack of experimental validation at this stage, (ii) reliance on predicted molecular properties rather than measured values, and (iii) absence of target-specific docking or activity prediction. Future work will include experimental antimicrobial screening of the 50 prioritized candidates against a panel of clinically relevant pathogens, followed by hit validation and mechanistic studies for active compounds. Integration of machine learning models trained on antimicrobial activity data could further refine candidate prioritization.

4.4 Conclusions

We have developed and executed a systematic computational pipeline for prioritizing natural products as antimicrobial candidates. Through unsupervised learning and a bimodal selection strategy, we identified 50 high-priority compounds that balance drug-likeness with structural complexity. This focused library provides an excellent starting point for experimental screening, with the potential to yield novel antibiotic scaffolds to combat the growing threat of antimicrobial resistance.