Pharmazie

Pharmazie **2026**; 81(11): 55687
https://doi.org/10.31083/Pharmazie55687

Research Article

ATC Classification and Scaffold-Split pMIC Prediction for Antibacterial Prioritization

Yevhen Pruhlo[1], Ivan Iurchenko[2,*]

[1]Department of Pharmacy, Bukovinian State Medical University, 58002 Chernivtsi, Ukraine
[2]College of Pharmacy, University of Manitoba, Winnipeg, MB R3E 0T5, Canada
*Correspondence: ivan.iurchenko@umanitoba.ca (Ivan Iurchenko)
Academic Editor: Walter E. Müller
Submitted: 7 July 2026 Revised: 14 October 2026 Accepted: 21 October 2026 Published: 13 November 2026

Abstract

Background and Aim: Antimicrobial resistance requires reproducible computational tools that distinguish antibacterial-drug class membership from measured antibacterial potency. We developed a descriptor-based workflow for Anatomical Therapeutic Chemical classification (ATC)-based prioritization, exploratory molecular generation, and pMIC prediction. **Investigations:** ChEMBL 33 structures were encoded using Mordred descriptors. Fourteen classifiers were used to distinguish 248 systemic antibacterial drugs (ATC J01/J04A) from 2540 non-anti-infective reference drugs. Models were assessed by random, independent, and Bemis–Murcko scaffold splits; An exploratory character-level recurrent neural network was used for molecular generation. Separately, eight regressors modeled *Staphylococcus aureus* pMIC for 33,505 compounds under a scaffold split. **Results:** Random forest achieved a Matthews correlation coefficient (MCC) of 0.857 on the random-split test set and 0.775 on the independent validation set. For unseen scaffolds, the MCC was 0.746; after removal of six β-lactam ring descriptors, the MCC was 0.738. Across 25 repeated scaffold splits, the random forest test MCC was 0.774 (95% confidence interval (CI) 0.752–0.795), supporting the consistency of scaffold-based performance across partitions. Char-RNN validity was 10.8%; the generated molecules were novel relative to the training set (23/25 were also absent from ChEMBL) and synthesizable. Novelty is reported relative to the training set rather than as absolute chemical-space novelty. XGBoost achieved a scaffold-held-out pMIC R^2 = 0.510, RMSE = 0.743 log units, and a Spearman ρ = 0.695. Across 10 repeated scaffold splits, XGBoost R^2 was 0.505 (95% CI 0.472–0.537), matching the single-split value. On two independent external tests—a 2019–2022 temporal split and 5090 compounds new in the later ChEMBL 37 release — absolute accuracy was lower (R^2 = 0.240 and 0.286) while rank correlation was retained (Spearman ρ = 0.502 and 0.566). These results are consistent with the model being prioritized rather than being used for accurate potency-prediction. **Conclusions:** The ATC classifier prioritizes structures resembling approved antibacterial drugs rather than establishing measured antibacterial activity. The pMIC model provides scaffold-aware estimates of S. aureus potency; both components require assessment of the applicability domain and experimental validation.

Keywords: anti-bacterial agents; microbial sensitivity tests; quantitative structure-activity relationship; machine learning; *Staphylococcus aureus*; drug discovery; bacterial drug resistance

1. Introduction

Antimicrobial resistance (AMR) is among the most pressing threats to global health: bacterial AMR was directly attributable to an estimated 1.27 million deaths in 2019 [1], with a cumulative toll approaching 39 million projected for 2025–2050 [2]. The WHO priority-pathogen list places Gram-negative (ESKAPE) organisms and rifampicin-resistant *Mycobacterium tuberculosis* in its critical tier [3]. Yet the antibacterial pipeline remains sparse and conventional discovery is slow, costly and high-attrition [4], motivating computational approaches that prioritize promising chemistry before committing scarce experimental resources.

Machine learning (ML) has emerged as a practical accelerator across this pipeline. A landmark demonstration trained a deep neural network on phenotypic growth-inhibition data and, by screening a large repurposing and virtual library, identified halicin, a compound structurally divergent from conventional antibiotics with broad-spectrum activity and *in vivo* efficacy [5]. The same strategy, retrained on a focused screen, later yielded abaucin, a narrow-spectrum agent against *Acinetobacter baumannii* that acts through lipoprotein trafficking (LolE) [6]. More recent work has pushed to ultra-large scale: a graph-neural-network virtual screen of more than one billion synthetically accessible compounds reported substantial enrichment over the high-throughput screen used to train it and recovered scaffolds dissimilar to known antibiotics [7]. Because publicly available antibacterial data are scarce, transfer learning, using pre-training on large auxiliary molecular datasets and fine-tuning on limited antibacterial labels, has been used to identify sub-micromolar broad-spectrum hits against ESKAPE pathogens [8]. These studies establish that ML can explore chemical space and prioritize candidates for experimental follow-up.

Copyright: © 2026 The Author(s). Published by IMR Press.
This is an open access article under the CC BY 4.0 license.

Publisher's Note: IMR Press stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Three complementary ML tasks recur across this literature, and it is useful to keep them distinct. (i) Structure-based classification predicts whether a molecule resembles a defined antibacterial class. Graph neural networks can be advantageous when training data are abundant, while classical learners such as Random Forest remain competitive and more interpretable for smaller or specialized chemical spaces. A recurring difficulty is class imbalance: antibacterials are a small minority of any drug-like library, so naive sampling can inflate apparent performance. Careful resampling and construction of the negative set are therefore necessary. Equally important is the train/test split: because close analogues are common, random splits can leak structural information, and scaffold-based splitting, which groups molecules by their Bemis-Murcko frameworks [9] and was popularized as a hard generalization benchmark by MoleculeNet [10], is increasingly used for more stringent estimates of generalization. Notably, most classifiers in this area define the positive label from measured activity (MIC or growth inhibition), rather than from a regulatory or therapeutic class annotation. (ii) Generative design proposes new structures using variational autoencoders, chemical language models, or recurrent networks. Character-level RNNs can be trained as molecular language models and fine-tuned on small active sets to enrich generated libraries toward a target [11], while continuous-latent-space autoencoders enable optimization over chemical space [12]. A central tension is that optimizing structural novelty often lowers initial potency, and novelty and activity are commonly reconciled through iterative derivative synthesis; molecular validity also varies substantially between architectures, and validity alone does not establish synthetic accessibility or utility. (iii) Potency (MIC) regression predicts a continuous activity value. Reported accuracy varies with dataset homogeneity and the complexity of bacterial defence mechanisms.

Despite this progress, several methodological considerations remain unevenly addressed and motivate the present work. First, reproducibility: many pipelines rely on proprietary data or under-specified protocols, whereas open databases (e.g., ChEMBL), fixed seeds and containerized environments allow independent re-execution. Second, interpretability: feature-level analysis can connect model behaviour to recognisable pharmacophoric chemistry, which is valuable both for trust and for hypothesis generation. Third, honest evaluation: performance must be read against the split strategy and the class definition used, since a therapeutic-class label answers a different question (“does this resemble an approved antibacterial drug?”) than a *measured* potency endpoint (“how active is it?”). Conflating the two overstates what a classifier has learned, particularly when a single dominant scaffold (e.g., the β-lactam ring) drives separability.

Here, we present an interpretable and reproducible machine-learning framework using Mordred descriptors and ChEMBL 33 data. Task A compares fourteen classification algorithms for distinguishing systemic antibacterial drugs (ATC J01/J04A) from non-anti-infective reference compounds. Model performance is assessed using random, independent, and Bemis–Murcko scaffold splits, together with applicability-domain analysis. An exploratory character-level recurrent neural network generates candidate structures, which are ranked by classifier scores, with novelty assessed relative to the training set. Task B evaluates eight regression algorithms for predicting *Staphylococcus aureus* pMIC values using scaffold-based validation. This framework extends a workflow previously applied to an antidiabetic target [13].

The study extends conventional model evaluation through repeated scaffold splits, β-lactam sensitivity analyses, and comparison with a molecular-similarity baseline. For pMIC regression, temporal validation and testing against compounds newly represented in ChEMBL 37 provide additional assessments of predictive reliability. These analyses aim to establish the strengths and limitations of each approach and their potential utility in early-stage antibacterial compound prioritization.

2. Materials and Methods

This study defines two complementary tasks that should not be conflated. Task A (classification) predicts whether a molecule belongs to the therapeutic class of approved antibacterial drugs, using an ATC-based label (J01/J04A). Task B (regression) predicts a molecule’s measured potency (pMIC against *Staphylococcus aureus*) from experimental bioactivity data. Task A asks “does this resemble an antibiotic drug?”; Task B asks “how potent is it?”. The complete workflow for both tasks—dataset construction, descriptor generation, cleaning, splitting, balancing, model training, and validation—is summarized in Fig. 1.

2.1 Dataset Assembly (ChEMBL 33, ATC-Based Labelling)

Data were taken from the ChEMBL 33 SQLite release (https://doi.org/10.6019/CHEMBL.database.33) [14]. External validation was performed using ChEMBL version 37 (https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_37/). ATC annotations were obtained by joining molecule_atc_classification (field level5) to molecule_dictionary and canonical SMILES in compound_structures (via molregno). The positive class (label = 1) comprised molecules whose ATC level-5 code began with J01 (antibacterials for systemic use) or J04A (antituberculars). The negative class (label = 0) comprised all other ATC-classified molecules with a valid SMILES, excluding compounds assigned to selected anti-infective ATC categoriesto minimize the potential inclusion of antibacterial agents in the negative dataset. This yielded 248 antibacterial and 2540 non-antibacterial compounds (Table 1).

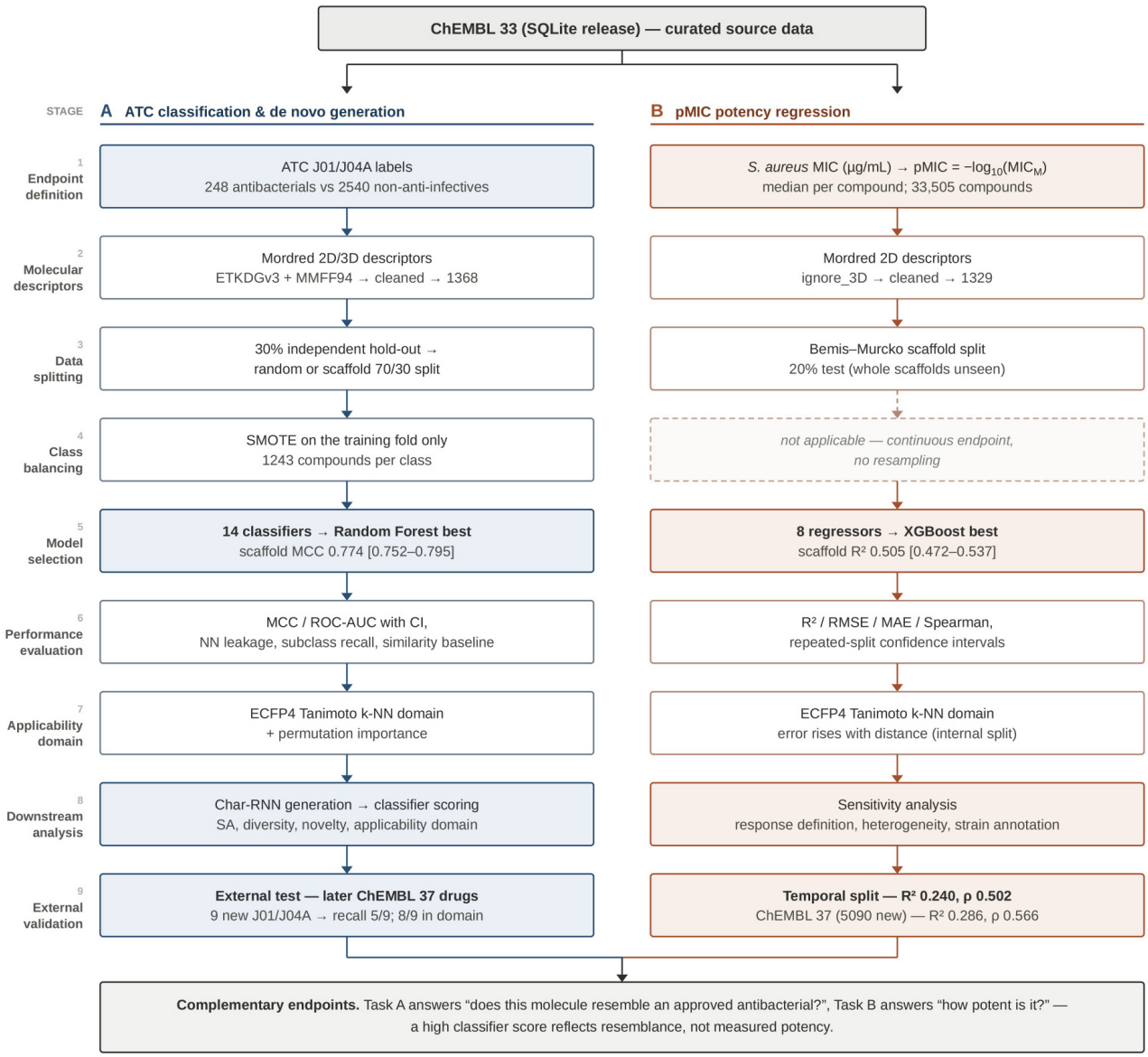


Fig. 1. Reproducible two-task workflow. Task A (left): ATC-based classification of 248 antibacterials vs 2540 non-anti-infectives from Mordred descriptors, with random/independent/Bemis–Murcko-scaffold splits, SMOTE on the training fold only, fourteen classifiers, applicability-domain and importance analysis, and exploratory Char-RNN generation. Task B (right): S. aureus pMIC regression on 33,505 compounds under a scaffold split with eight regressors and an ECFP4 Tanimoto k-NN applicability domain. The nine-stage pipeline runs from source data through endpoint definition, descriptors, splitting, balancing, model selection, evaluation, applicability domain, downstream analysis and external validation (ChEMBL 37 and temporal split). ATC, Anatomical Therapeutic Chemical classification; SMOTE, Synthetic Minority Over-sampling Technique; Char-RNN, Character-level Recurrent Neural Network.

Each compound enters once (one row per molregno/canonical SMILES), so there are no duplicate structures to distribute across the training, test or validation splits. The ATC J01/J04A label was chosen because it is a curated, standardized therapeutic-class annotation that is reproducible and free of the assay-to-assay noise of measured MIC values, making it a well-defined target for a class-recognition task. The resulting model therefore identifies chemistry resembling approved systemic antibacterials—it prioritizes known-antibacterial-like structures rather than discovering mechanistically novel agents, which is the complementary role of measured-potency modelling (Task B) and experimental screening. Compared with a single similarity search against known antibiotics, the classifier learns a multi-descriptor decision surface across a chemically heterogeneous positive class, is paired with an applicability domain that flags extrapolative inputs, and yields interpretable feature-level structure–class associations (a quantitative comparison against a tuned similarity-search baseline is given in Section 3.2). Its main limitation is that the nega-

tive class comprises other ATC-classified drugs (all anti-infectives excluded) rather than compounds with experimentally demonstrated antibacterial inactivity; some negatives may thus possess unrecorded antibacterial activity, a limitation the pMIC regression on measured data partly offsets. The positive class is chemically heterogeneous, including β-lactams, aminoglycosides, macrolides, fluoroquinolones, tetracyclines, glycopeptides, sulfonamides, and polymyxins (MW 138–2393 g mol^{-1}); the most populated J01 subclasses were J01MA (n = 22), J01CA (n = 19), J01DD (n = 18), and J01FA (n = 15).

Table 1. Dataset composition for ATC-based classification.

Set	Definition	N
Positive (y = 1)	ATC J01 + J04A	248
Negative (y = 0)	other ATC drugs, all anti-infectives excluded	2540
Total		2788

2.2 Descriptor Generation (Mordred 2D/3D)

Descriptors were computed with Mordred 1.2.0 [15] (ignore_3D = False; 1826 raw features). Three-dimensional geometries were generated with RDKit [16] using ETKDGv3 embedding (seed 42, 200 iterations) and MMFF94 optimization (≤500 iterations), a reproducible, license-free substitute for the original Gaussian/PM6 protocol. Stereochemistry present in the ChEMBL canonical SMILES was retained during descriptor generation (Mordred computes both 2D descriptors that ignore stereochemistry and 3D descriptors sensitive to the embedded geometry); Bemis–Murcko scaffolds, by contrast, were extracted with chirality removed (RDKit MurckoScaffoldSmiles, includeChirality = False), so stereoisomeric analogues share a scaffold and are kept together on the same side of every scaffold split. Molecules with more than 120 heavy atoms or failing to embed (30/248 positives and 97/2540 negatives; chiefly glyco- or polypeptide antibiotics) used 2D features only. A sensitivity analysis (Table 2) confirms that this mixed treatment does not bias the classifier and that a fully-2D model performs comparably. ClassyFire and ChEBI ontological descriptors were not used.

2.3 Cleaning, Splitting and Balancing (Task A)

Descriptors were coerced to numeric values, ±∞ values were set to NaN, descriptors missing in more than 20% of molecules were removed, remaining values were median-imputed, and zero-variance columns were removed, leaving 1368 descriptors. Descriptor missingness/-variance filtering and median imputation were computed on the full descriptor matrix before splitting. Because this procedure uses information from held-out compounds, it may introduce test-set information leakage even though their labels were not used. Its impact on predictive performance has not been quantified and remains a limitation of the present analysis. SMOTE and model fitting were restricted to training folds. With a fixed seed (random_state = 42), a stratified 30% independent validation set was hidden first (74 active/763 inactive); the remainder was split 70/30 into training and test sets (test: 52 active/534 inactive). SMOTE (k_neighbors = 5, random_state = 42) was applied to the training fold only [17], yielding 1243 molecules in each class. Because SMOTE interpolates in the high-dimensional descriptor space, its synthetic vectors do not necessarily correspond to chemically valid molecules; it is therefore used only to rebalance the training fold and never enters test, validation or cross-validation folds (confirmed by the SMOTE-inside-CV design of Section 2.4). The full classification dataset contains 1466 unique Bemis–Murcko scaffolds (158 among the 248 actives, 1322 among the 2540 inactives). Under the scaffold split (Section 2.4), the training and test sets share no scaffold (1151 vs 315 scaffolds; zero overlap).

2.4 Classification Models and Evaluation (Task A)

Fourteen scikit-learn [18] and XGBoost [19] classifiers were trained. Scale-sensitive learners were wrapped in a scaler: Logistic Regression, Linear SVM, k-NN, SGD, Lasso, Elastic Net, and Ridge used StandardScaler; Multinomial Naive Bayes used MinMaxScaler. Tree-based models (Decision Tree, Random Forest, Gradient Boosting, and XGBoost), Bernoulli Naive Bayes, and Linear Discriminant Analysis used raw features. All models used random_state = 42 where applicable (Logistic Regression: solver = 'liblinear', max_iter = 2000). To avoid confusion with continuous regressors, the penalized linear *classifiers* are labelled "Lasso (L1-logistic)", "Ridge classifier" and "Elastic Net" in the tables; these are, respectively, an L1-penalized logistic-regression classifier, a RidgeClassifier, and an elastic-net logistic SGDClassifier (log-loss)—none is a continuous regression model. Where predict_proba was unavailable, decision_function scores were used for ROC-AUC without probability calibration. Models were assessed by 5-fold stratified CV on the training fold and by the held-out test and validation sets (accuracy, precision, recall, F1, ROC-AUC, balanced accuracy, and MCC). No hyperparameter tuning was performed for the fourteen-model comparison by design; the aim was a like-for-like baseline across algorithms, not a state-of-the-art single score. The appropriateness of this choice for the two principal models was verified post hoc by nested cross-validation (Table 3). To probe what the classifier exploits, two robustness analyses were added (Section 3.2): a Bemis–Murcko scaffold split (whole scaffolds held out) replacing the random split, and a β-lactam ablation removing the six four-membered-ring descriptors (n4Ring, n4ARing, n4HRing, n4AHRing, Xch-4d, and Xch-4dv) before retraining. To quantify the stability of scaffold-split performance rather than rely on a single partition, the Bemis–Murcko scaffold

Table 2. 2D/3D descriptor sensitivity: Random Forest and XGBoost test MCC (mean, 95% CI) over 25 repeated scaffold splits.

Scenario	n positive	Random forest MCC (95% CI)	XGBoost MCC (95% CI)
A. Baseline (2D/3D, all)	248	0.774 (0.752–0.795)	0.756 (0.732–0.780)
B. Exclude 2D-only compounds	218	0.735 (0.700–0.769)	0.760 (0.732–0.787)
C. Fully-2D model (all)	248	0.765 (0.739–0.792)	0.760 (0.735–0.785)

A, baseline mixed 2D/3D on all compounds; B, 2D-only compounds excluded; C, fully-2D model (3D descriptors dropped for all). MCC, Matthews correlation coefficient.

split was additionally repeated over 25 independent random scaffold-group partitions (seeds 0–24); for each held-out set MCC, ROC-AUC, balanced accuracy, and F1 were recomputed, and we report the mean with a 95% confidence interval (Student t-interval across repeats) together with the standard deviation and range. To assess residual analogue leakage, a nearest-neighbour analysis was added: for every held-out test compound the maximum ECFP4 (Morgan radius 2, 2048-bit) Tanimoto similarity to any training compound was computed, and the distributions were compared between the scaffold split and a random split of the same set. As a reference point for the learned model, a nearest-antibiotic similarity-search baseline was also evaluated over the same 25 scaffold splits: each test compound was scored by its maximum ECFP4 Tanimoto similarity to the training-set antibacterials and classified with a decision threshold tuned on the training fold only (leave-one-out for training positives), so the baseline never sees test labels. As an external validation of the classifier, the ChEMBL 33-trained Random Forest was also applied to the J01/J04A drugs that are new in the later ChEMBL 37 ATC release (InChIKey absent from the ChEMBL 33 positives), and its recall on these unseen approved antibacterials was recorded. To test whether classification is driven by β-lactam chemistry, two further analyses used the exact ATC subclass of each positive (from the ATC level-3/4 annotation): (i) a leave-β-lactam-compounds-out experiment in which all 96 β-lactam positives (ATC J01C penicillins and J01D cephalosporins/carbapenems/monobactams) were removed and the classifier retrained over the same 25 repeated scaffold splits; and (ii) a per-subclass recall analysis in which the held-out sensitivity of Random Forest was aggregated across the repeated scaffold splits for each antibacterial subclass. Finally, to test whether the default configurations are appropriate and whether the model ranking is stable under optimization, Random Forest and XGBoost were evaluated by nested cross-validation (outer StratifiedKFold, inner RandomizedSearchCV over standard hyperparameter ranges, MCC scoring), with SMOTE placed inside the cross-validation pipeline so that synthetic samples never entered a validation fold. Default and tuned models were scored on the identical outer folds and were then both refit within the same SMOTE pipeline and evaluated on the main held-out test set, so that the default-versus-tuned comparison is paired throughout. A 2D/3D descriptor sensitivity analysis compared three scenarios over the repeated scaffold splits: (A) the baseline mixed 2D/3D descriptors on all compounds; (B) exclusion of the 127 compounds that used 2D features only (geom_tag ≠ 3D); and (C) a fully-2D model with all Mordred 3D descriptors dropped for every compound. The 2D-only compounds were also characterised by molecular weight and ATC subclass.

2.5 Feature-Frequency and Importance Analysis

Each descriptor was reduced to a presence/absence (non-zero) indicator and tested between classes with Pearson's χ^2 (2×2 contingency table); *p*-values were BH-FDR corrected ($q < 0.05$). Because Mordred descriptors are numerous and strongly intercorrelated, this binarisation discards magnitude information and the FDR procedure (which assumes limited dependence) is only approximate; the χ^2 results are used as a qualitative enrichment screen, not formal inference. Model-intrinsic importances were taken from Random Forest, Gradient Boosting and XGBoost. Because impurity-based importance can be inflated for correlated descriptors and does not establish causality, model-agnostic permutation importance was additionally computed for the Random Forest classifier on the held-out test set (30 permutation repeats, ROC-AUC and balanced-accuracy scoring; scikit-learn). Permutation importance is itself diluted across groups of correlated descriptors, so all importances are interpreted as statistical predictors / structural correlates rather than mechanistic determinants.

2.6 Applicability Domain (Task A)

The applicability domain (AD) was defined in ECFP4 fingerprint space as the distance to the *k* nearest training neighbours: for each compound the distance is one minus the mean Tanimoto similarity (Morgan radius 2, 2048-bit) to its five nearest training molecules, and the AD cutoff is the 95th percentile of the training leave-one-out distance distribution (here a mean top-5 Tanimoto ≥0.20). This fingerprint-based, chemically interpretable domain replaces the earlier Mahalanobis-distance definition; unlike a single ellipsoid in the singular ~1400-descriptor space it needs no dimensionality reduction, and it discriminates reliable from unreliable predictions more sharply (Section 3.6). The same k-NN Tanimoto domain is used for the regression task (Section 2.8).

Table 3. Hyperparameter optimization: nested-CV MCC (3 outer folds, SMOTE inside the CV pipeline) and main-split held-out test MCC, default vs tuned, for the two principal models.

Model	Nested-CV MCC (default)	Nested-CV MCC (tuned)	Test MCC (default)	Test MCC (tuned)
Random forest	0.843 ± 0.054	0.835 ± 0.054	0.845	0.836
XGBoost	0.819 ± 0.077	0.834 ± 0.052	0.846	0.868

Selected hyperparameters (main split): Random Forest — 600 trees, max_depth 10, max_features sqrt, min_samples_leaf 2, min_samples_split 5; XGBoost — 400 trees, max_depth 9, learning_rate 0.03, subsample 1.0, colsample_bytree 1.0.

Both test-MCC columns come from the same SMOTE-in-pipeline refit, which resamples the descriptor matrix as a numeric array; the default column is therefore the matched baseline for this comparison and is not interchangeable with the corresponding value in Table 7, where SMOTE was applied to the typed descriptor table.

2.7 Generative Model (Char-RNN, Exploratory)

This branch is a deliberately small proof of concept, not a competitive generative model. It is presented as an exploratory demonstration; a scaled generative component with transfer learning from a large SMILES corpus is planned as separate work and is beyond the scope of the present study. A two-layer LSTM (256 hidden units, dropout 0.3, seed 42) was trained on the 248 antibacterial SMILES augmented by stochastic non-canonical enumeration (20 variants/molecule → 4960 strings), using Adam (learning rate 1×10^{-3}), cross-entropy loss, batch 64, sequence length 80, gradient clip 5, for 45 epochs (CPU). Molecules were produced by autoregressive top-k sampling (k = 4) from seven prime strings (C, N, O, CC, c1, C(, [C@); no temperature scaling beyond top-k truncation. A molecule was counted as "novel" if it was a valid canonical SMILES with ≥6 heavy atoms not exactly matching any training molecule—novelty defined *relative to the training set only*, not against external databases. Survivors were scored by the trained Random Forest, Gradient Boosting, XGBoost and Linear SVM and ranked by mean predicted probability.

2.8 Potency Regression (Task B: pMIC vs S. aureus)

MIC values against S. aureus were extracted from ChEMBL activities (standard_type = 'MIC', standard_units = 'ug.mL^{-1}', standard_relation = '=', standard_value > 0), joined to assays (assay_organism = '*Staphylococcus aureus*'). Salts were reduced to the largest organic fragment (RDKit FragmentParent) and MW computed with RDKit. Records were then aggregated to one row per canonical parent SMILES using the median pMIC, so that replicate and conflicting measurements of the same structure are collapsed to a single value; stereoisomers, which have distinct canonical SMILES, were retained as separate compounds. Because one value per compound is modelled and the split is by Bemis–Murcko scaffold, identical structures cannot appear in both the training and test sets. MIC was converted to molar units and expressed as **pMIC = $-\log_{10}$(MIC_M)**, where **MIC_M (mol L^{-1}) = MIC (μg mL^{-1}) × 10^{-3} / MW (g mol^{-1})**.

The median pMIC was taken per compound. The corpus is highly heterogeneous—the 90 916 MIC records come from 11,998 distinct ChEMBL assays across many laboratories, S. aureus strains (MRSA/MSSA; reference strains such as ATCC 25923/29213 and clinical isolates) and media, and are organism-level phenotypic assays (confidence_score = 1 for essentially all records); the ChEMBL confidence score grades how well an assay maps to its molecular target, from 0 (unassigned) through 1 (organism-level, non-molecular) to 9 (a single well-defined protein target). A value of 1 therefore denotes whole-organism phenotypic growth-inhibition assays with no assigned molecular target—appropriate for phenotypic MIC data, but signalling target-level heterogeneity. Strain/protocol are not modelled, imposing an intrinsic ceiling on R^2. Only 100 ≤ MW ≤ 900 g mol^{-1} was retained, giving 33 505 modelled compounds (pMIC = 4.79 ± 1.10). Physically implausible records—pMIC <2 or >9, i.e., molar MIC outside ≈ 13 mM–1 nM (208 and 34 records, 0.7% total)—were flagged as likely erroneous ChEMBL entries; their removal is reported as a sensitivity check in Section 3.7 (negligible effect). Features were Mordred 2D descriptors (ignore_3D=True; 1615 raw → 1329 after cleaning). Compounds were split by Bemis–Murcko scaffold (RDKit MurckoScaffoldSmiles, seed 42): whole scaffolds formed a 20% test set (6707 molecules; 26 798 for training), so test chemotypes are unseen. The dataset spans 10,692 unique Bemis–Murcko scaffolds; the training and test sets share none (8330 vs 2362 scaffolds; zero overlap). Eight regressors were evaluated—Ridge (α = 10), Lasso (α = 0.01), Elastic Net (α = 0.01, l1_ratio = 0.5), k-NN (k = 5), RBF-Nyström (500 components) + Ridge, Random Forest (300 trees), Gradient Boosting (default), XGBoost (400 trees, learning rate 0.05, depth 6, subsample 0.8)—each linear/kernel/distance model preceded by StandardScaler. Metrics were R^2, RMSE, MAE and Spearman ρ. The two leading regressors were hyperparameter-tuned: XGBoost by nested cross-validation over three scaffold outer folds, and both XGBoost and Random Forest by RandomizedSearchCV (12 iterations, inner 3-fold CV, R^2 scoring) on the main scaffold split, comparing default and tuned held-out R^2 and recording the selected hyperparameters (Table 4). The ap-

Table 4. Hyperparameter optimization of the pMIC regressors (RandomizedSearchCV, 12 iterations, R^2 scoring).

Model	Default test R^2	Tuned test R^2	Nested-CV default R^2	Nested-CV tuned R^2	Selected hyperparameters
XGBoost	0.510	0.531	0.485 ± 0.024	0.515 ± 0.026	600 trees, depth 8, lr 0.05, subsample 0.7, colsample 0.7
Random Forest	0.501	0.498	—	—	200 trees, min_samples_leaf 1, max_features 0.5, unlimited depth

XGBoost is additionally assessed by nested cross-validation over three scaffold outer folds; both models are tuned on the main scaffold split and compared default-vs-tuned on the held-out test set. Tuning yields a small gain for XGBoost and none for Random Forest.

plicability domain used the same ECFP4 Tanimoto k-NN definition as the classifier (Section 2.6): compounds whose mean top-5 Tanimoto similarity to the training set fell below the 95th-percentile training cutoff (here 0.44) were classified as outside the domain. To assess the stability of the scaffold-split estimates, the scaffold split was additionally repeated over 10 independent random scaffold-group partitions (seeds 0–9); for each held-out set R^2, RMSE, MAE and Spearman ρ were recomputed and are reported as the mean with a 95% Student t-interval (Table 5). The repeated analysis covers the linear, kernel, distance and XGBoost regressors; Random Forest and Gradient Boosting were omitted from it because their exact refit over 10 repeats at this scale was computationally prohibitive, and XGBoost used the equivalent histogram tree method. The same nearest-neighbour maximum-Tanimoto analysis described in Section 2.4 was applied to the scaffold-held-out regression test set. Because the molar transformation adds a $+\log_{10}(\mathrm{MW})$ term to the response while MW-related descriptors are also predictors, a response-definition sensitivity analysis compared the molar pMIC with a mass-based response ($-\log_{10}$ of MIC in μg mL^{-1}, which contains no MW term) under identical scaffold splits, and quantified the variance contributed by the $\log_{10}(\mathrm{MW})$ term and how well descriptors alone recover $\log_{10}(\mathrm{MW})$. To validate the applicability domain, the scaffold-held-out test compounds were binned into deciles of their mean top-5 ECFP4 Tanimoto similarity to the training set, and RMSE, MAE and R^2 were computed per decile together with the Spearman correlation between Tanimoto distance and absolute error. As a temporal (quasi-external) validation, compounds were partitioned by the publication year of their earliest S. aureus MIC record (ChEMBL 33 docs.year): the model was trained only on compounds first reported up to 2018 and tested on those first reported in 2019–2022, so the test set is held out in time and dominated by later, mostly novel chemistry. As an independent external test against a later database release, the ChEMBL 33-trained model was additionally applied to the 5090 compounds whose *S. aureus* MIC is new in ChEMBL 37 (InChIKey absent from ChEMBL 33); Mordred 2D descriptors were computed for these compounds and predictions compared with their measured median pMIC.

2.9 Software and Reproducibility

Python 3.10 with RDKit 2024.03.1 [16], Mordred 1.2.0, scikit-learn 1.6.1, XGBoost 2.1.4, imbalanced-learn 0.14.2 [20], statsmodels 0.14.4, and PyTorch (CPU) [21] were used. Random seeds were fixed at 42 for the primary analyses, unless otherwise specified. For repeated scaffold-split evaluations, seeds 0–24 were used for classification (Task A), and seeds 0–9 were used for pMIC regression (Task B), as described in Sections 2.4 and 2.8, respectively. Seven executable notebooks (including a ChEMBL 37 external-validation notebook and a per-compound external prediction-table export), helper scripts, descriptor matrices, trained models, metric tables, and figures were prepared with a Dockerfile for reproducibility.

3. Results

3.1 Descriptor-Frequency Signature of Antibacterial Compounds

A presence/absence χ^2 analysis with BH-FDR correction identified the descriptors most strongly separating antibacterial (y = 1) from non-antibacterial (y = 0) drugs (Table 6; all $q \ll 10^{-20}$). The strongest discriminators encode small strained rings: four-membered ring counts (n4Ring, n4ARing, n4HRing, n4AHRing) occur in ~40% of antibacterials but ≤0.9% of other drugs—a direct fingerprint of the β-lactam class. Antibacterials are further enriched in sulfur (nS), acidic groups (nAcid), hydroxyls (SsOH), secondary amines (SssNH) and polar surface (PEOE_VSA12), reflecting the polar, heteroatom-rich character required to reach bacterial targets.

3.2 Development, Evaluation and Application of Classification Models

Because the antibacterial class is the minority (approximately 9% of the test set), model quality is interpreted primarily with MCC and balanced accuracy. Tree ensembles gave the strongest performance on the random-split held-out test set (Table 7): Random Forest led (MCC = 0.857; ROC-AUC = 0.975), with Gradient Boosting (MCC = 0.836) and XGBoost (MCC = 0.826) close behind. Once wrapped in a StandardScaler, the scale-sensitive linear models were competitive (Logistic Regression MCC = 0.831; Ridge = 0.739). These random-split results should be interpreted as analogue-aware rather than prospective

Table 5. Stability of scaffold-split pMIC regression over 10 repeated scaffold-group partitions.

Model	R^2 (mean)	R^2 95% CI	RMSE (mean)	RMSE 95% CI	MAE (mean)	MAE 95% CI	Spearman ρ (mean)	Spearman 95% CI
XGBoost	0.505	0.472–0.537	0.771	0.748–0.795	0.588	0.575–0.600	0.704	0.687–0.721
k-NN	0.464	0.426–0.502	0.802	0.778–0.827	0.589	0.576–0.602	0.689	0.671–0.707
SVR (RBF/Nyström)	0.367	0.336–0.398	0.872	0.854–0.890	0.676	0.667–0.686	0.599	0.578–0.620
Elastic Net	0.316	0.288–0.344	0.907	0.893–0.921	0.706	0.697–0.714	0.565	0.541–0.590
Lasso	0.305	0.279–0.331	0.914	0.901–0.928	0.713	0.705–0.721	0.555	0.530–0.579
Ridge	0.299	0.236–0.363	0.916	0.880–0.952	0.694	0.682–0.705	0.595	0.575–0.614

Values are mean (95% Student t-interval) of the held-out metric across repeats. Random Forest and Gradient Boosting are represented by their full-configuration single-split values in Table 20; their exact refit over 10 repeats on the 33,505 × 1329 matrix was computationally prohibitive, and XGBoost (the best model) uses the equivalent histogram tree method here.

Table 6. Top Mordred descriptors discriminating antibacterial from non-antibacterial drugs (presence frequency; chi-square with BH-FDR).

Descriptor	Freq. y = 1	Freq. y = 0	Δ	χ^2	q (FDR)	Interpretation
Xch-4d / Xch-4dv	0.460	0.033	0.426	619	2.5×10^{-134}	4-membered-ring path (β-lactam)
n4AHRing / n4HRing	0.399	0.005	0.394	910	4.5×10^{-197}	4-membered heterocycle (β-lactam)
n4Ring / n4ARing	0.399	0.009	0.390	814	1.8×10^{-176}	4-membered ring count
PEOE_VSA12	0.694	0.321	0.373	137	6.7×10^{-30}	polar surface area
nS	0.524	0.187	0.337	151	7.3×10^{-33}	sulfur atom count
nAcid	0.552	0.216	0.336	138	5.5×10^{-30}	acidic group count
SsOH / NsOH	0.802	0.467	0.336	102	2.8×10^{-22}	hydroxyl groups
SssNH	0.681	0.348	0.333	107	2.9×10^{-23}	secondary amine (–NH–)

Δ = frequency difference (y = 1 - y = 0).

scaffold-level estimates. The CV-F1 column was computed with SMOTE applied inside each cross-validation fold rather than to the whole training set beforehand; this removes the fold-to-fold leakage of synthetic neighbours that had previously inflated CV-F1 to ~0.99, and the corrected values (e.g., Random Forest 0.813, XGBoost 0.820) now agree with the independent held-out F1 in the same row, providing additional evidence of model stability while reducing oversampling-related cross-validation leakage.

The confusion matrix (Table 8) shows Random Forest produced only 3 false positives out of 534 negatives while recovering 42/52 actives; Logistic Regression gave a balanced 44 true positives at 8 false positives.

On the hidden validation set (74 active/763 inactive), which took no part in training or balancing (Table 9), Random Forest retained MCC = 0.775. The test-to-validation decrease is consistent with some generalization within the same ChEMBL/ATC sampling frame, although it does not eliminate the possibility of analogue-level optimism under random splitting. Random Forest was selected for downstream ranking analyses.

What does the classifier learn? The strong random-split performance must be read with care. As shown in Section 3.1, the most discriminative descriptors encode the four-membered β-lactam ring (n4Ring/Xch-4d: 40% of actives versus 0.5% of negatives), which dominates the Gradient Boosting model (Section 3.5). This raises the concern that the model is mainly a β-lactam detector and that the random split can place close analogues on both sides. We tested both concerns directly (Table 10): a Bemis-Murcko scaffold split (unseen test chemotypes) and a β-lactam ablation that removed the six four-membered-ring descriptors, as well as both settings combined.

Two findings emerge. First, the signal is not limited to the six explicit β-lactam descriptors: removing them did not reduce random-split performance (Random Forest 0.857 to 0.860; Gradient Boosting 0.836 to 0.862), indicating that correlated structural features, including heteroatoms, polar surface, and sulfur-related descriptors, also contribute. Second, analogue-level optimism is present but bounded. Moving to a scaffold split lowered MCC (Random Forest 0.857 to 0.746; XGBoost 0.826 to 0.703), confirming that random splitting is optimistic for close analogues; nonetheless, performance remained substantial on held-out scaffolds. In the hardest setting, scaffold split plus β-lactam ablation, Random Forest and XGBoost retained MCC values of approximately 0.74–0.75. Thus, within this dataset, the model captures structural correlates of ATC antibacterial class beyond direct scaffold memorization and the six explicit β-lactam descriptors. The scaffold-split estimate, rather than the random-split value, is the appropriate primary estimate for prospective use. Because a single scaffold partition may be unrepresentative, the scaffold split was repeated over 25 independent scaffold-group partitions (Table 11). Held-out performance was stable: Random Forest gave a mean test MCC of 0.774 (95% CI 0.752–

Table 7. Classification performance on the held-out random-split test set for all 14 models (sorted by MCC); CV-F1 = 5-fold cross-validation F1 (mean ± SD) with SMOTE applied inside each fold (imblearn pipeline), so no synthetic sample crosses folds.

Model	Accuracy	Precision	Recall	F1	ROC-AUC	bal.acc	MCC	CV-F1
Random Forest	0.978	0.933	0.808	0.866	0.975	0.901	0.857	0.813 ± 0.049
Gradient Boosting	0.974	0.894	0.808	0.849	0.979	0.899	0.836	0.792 ± 0.047
Logistic Regression	0.973	0.846	0.846	0.846	0.978	0.916	0.831	0.745 ± 0.049
XGBoost	0.973	0.875	0.808	0.840	0.976	0.898	0.826	0.820 ± 0.053
Linear SVM	0.964	0.763	0.865	0.811	0.977	0.920	0.793	0.716 ± 0.057
Elastic Net	0.959	0.726	0.865	0.790	0.975	0.917	0.771	0.673 ± 0.118
SGD Classifier	0.959	0.759	0.789	0.774	0.971	0.882	0.751	0.658 ± 0.089
Ridge classifier	0.954	0.705	0.827	0.761	0.959	0.897	0.739	0.681 ± 0.024
Decision Tree	0.935	0.606	0.769	0.678	0.860	0.860	0.648	0.634 ± 0.069
k-NN	0.882	0.429	0.981	0.597	0.959	0.927	0.603	0.517 ± 0.036
Lasso (L1-logistic)	0.896	0.455	0.865	0.596	0.937	0.882	0.580	0.503 ± 0.016
LDA	0.882	0.414	0.789	0.543	0.868	0.840	0.516	0.419 ± 0.031
Multinomial NB	0.887	0.421	0.712	0.529	0.885	0.808	0.490	0.529 ± 0.052
Bernoulli NB	0.846	0.344	0.808	0.483	0.899	0.829	0.461	0.493 ± 0.053

Table 8. Random-split test-set confusion matrices for all 14 models (TN / FP / FN / TP).

Model	TN	FP	FN	TP
Random Forest	531	3	10	42
Gradient Boosting	529	5	10	42
XGBoost	528	6	10	42
Logistic Regression	526	8	8	44
Linear SVM	520	14	7	45
Elastic Net	517	17	7	45
Ridge classifier	516	18	9	43
SGD Classifier	521	13	11	41
Decision Tree	508	26	12	40
Multinomial NB	483	51	15	37
LDA	476	58	11	41
Lasso (L1-logistic)	480	54	7	45
k-NN	466	68	1	51
Bernoulli NB	454	80	10	42

Table 9. Independent hidden-validation performance for all 14 models (sorted by MCC).

Model	F1	Recall	Precision	bal.acc	MCC
Random Forest	0.781	0.676	0.926	0.835	0.775
XGBoost	0.773	0.689	0.879	0.840	0.760
Gradient Boosting	0.767	0.689	0.864	0.839	0.753
SGD Classifier	0.710	0.662	0.766	0.821	0.686
Logistic Regression	0.701	0.730	0.675	0.848	0.672
Linear SVM	0.693	0.703	0.684	0.836	0.663
Elastic Net	0.693	0.703	0.684	0.836	0.663
Ridge classifier	0.647	0.730	0.581	0.839	0.613
Decision Tree	0.602	0.676	0.544	0.810	0.563
k-NN	0.564	0.892	0.413	0.884	0.555
Lasso (L1-logistic)	0.509	0.757	0.384	0.819	0.478
LDA	0.485	0.784	0.352	0.822	0.459
Multinomial NB	0.492	0.608	0.413	0.762	0.442
Bernoulli NB	0.475	0.703	0.359	0.790	0.436

0.795; SD 0.052; range 0.675–0.872), with XGBoost 0.756 (0.732–0.780) and Gradient Boosting 0.728 (0.700–0.755), and ROC-AUC remained ≈ 0.92–0.93 for all tree ensembles. The single-split Random Forest value reported above (0.746) lies at the lower edge of this interval, confirming it is a conservative rather than optimistic estimate of scaffold-level generalization.

A nearest-neighbour analysis confirms that the scaffold split removes most analogue leakage (Table 12, Fig. 2). For the scaffold-held-out test set the median maximum ECFP4 Tanimoto to the training set was 0.34 (mean 0.38), and only 4.7% of test compounds had a close analogue (maximum Tanimoto >0.7) in training. Under a random split of the same set the median maximum Tanimoto rose to 0.51 and the fraction with a close training analogue rose to 22.6%. The higher structural proximity under random splitting is consistent with its more optimistic MCC (0.857 vs 0.746), and confirms that the scaffold-split estimate better reflects generalization to structurally distinct compounds.

Table 10. Robustness of test-set MCC under the evaluation protocols and β-lactam ablation (top five models).

Model	S1	S2	S3	S4
Random Forest	0.857	0.746	0.860	0.738
Gradient Boosting	0.836	0.741	0.862	0.723
XGBoost	0.826	0.703	0.828	0.749
Logistic Regression	0.831	0.558	0.831	0.558
Linear SVM	0.793	0.665	0.781	0.665

S1, random split/full features; S2, scaffold split/-full features; S3, random split/β-lactam descriptors removed; S4, scaffold split/β-lactam descriptors removed.

Table 11. Stability of scaffold-split classification over 25 repeated scaffold-group partitions (top five models).

Model	Test MCC (mean)	MCC 95% CI	SD	Range	ROC-AUC (95% CI)	Bal. acc. (95% CI)	F1 (95% CI)
Random Forest	0.774	0.752–0.795	0.052	0.675–0.872	0.924 (0.910–0.938)	0.835 (0.818–0.853)	0.777 (0.753–0.801)
XGBoost	0.756	0.732–0.780	0.059	0.638–0.850	0.925 (0.910–0.940)	0.838 (0.820–0.856)	0.765 (0.739–0.791)
Gradient Boosting	0.728	0.700–0.755	0.067	0.588–0.838	0.929 (0.917–0.941)	0.843 (0.825–0.860)	0.745 (0.718–0.772)
Linear SVM	0.666	0.641–0.691	0.060	0.544–0.771	0.915 (0.901–0.929)	0.836 (0.821–0.851)	0.691 (0.666–0.716)
Logistic Regression	0.626	0.586–0.667	0.099	0.367–0.735	0.908 (0.893–0.922)	0.831 (0.813–0.849)	0.653 (0.612–0.694)

Values are mean across repeats; a 95% Student t-interval across repeats is given for every metric (for MCC as a dedicated column, for ROC-AUC, balanced accuracy and F1 in parentheses), and SD/range describe partition-to-partition variability of test MCC.

Table 12. Nearest-neighbour analysis: maximum ECFP4 (Morgan r = 2, 2048-bit) Tanimoto similarity of each held-out test compound to the training set.

Split	n test	Median max Tc	Mean max Tc	Q1–Q3	% with max Tc >0.7	% with max Tc ≤0.4
Classification, scaffold split	836	0.34	0.38	0.25–0.51	4.7%	59.1%
Classification, random split	837	0.51	0.52	0.35–0.69	22.6%	33.6%
Regression (pMIC), scaffold split	6707	0.68	0.66	0.55–0.78	45.8%	8.5%

n test is the number of held-out compounds, not of scaffold groups: a 30% hold-out of the full 2788-compound classification set under each scheme (stratified random, n = 837; whole-scaffold Bemis–Murcko, n = 836 — intact scaffold groups make the partition only approximately 30%), and a 20% scaffold hold-out of the 33,505 pMIC compounds (n = 6707).

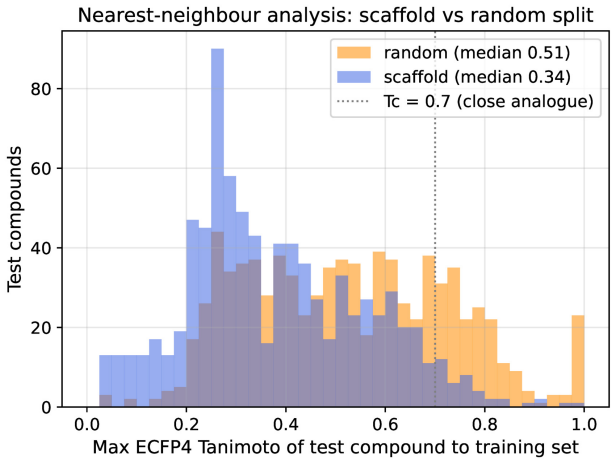


Fig. 2. Distribution of the maximum ECFP4 Tanimoto similarity of each held-out test compound to the training set, for the classification scaffold and random splits. Dashed line, Tc = 0.7.

The scaffold-size distribution of the classification dataset (Fig. 3) shows that most Bemis–Murcko scaffolds are singletons, which is why whole-scaffold hold-out removes analogue pairs rather than merely thinning them.

Does the classifier add anything over conventional similarity searching against known antibiotics? We compared it with a nearest-antibiotic baseline that labels a compound positive when its maximum ECFP4 Tanimoto similarity to any training-set antibacterial exceeds a threshold tuned on the training fold only, evaluated over the identical 25 repeated scaffold splits. The baseline reached a mean test MCC of 0.778 (95% CI 0.756–0.801), statistically indistinguishable from Random Forest (0.774, 0.752–0.795). This is an informative result: on this ATC endpoint, class recognition is largely a similarity-detection problem, and the multi-descriptor model buys no large accuracy gain over a single similarity search. Its advantages are therefore operational rather than accuracy-based—it needs no curated reference set of known antibiotics at inference, returns a continuous per-compound model score rather than one nearest-neighbour distance, yields interpretable descriptor-level structure–class associations (Section 3.5), and is coupled to an applicability domain that flags extrapolative inputs (Section 3.6). It also reinforces the paper's central caveat: because a plain similarity search performs comparably, the classifier is best understood as prioritizing antibacterial-like chemistry rather than as evidence of mechanistic activity.

To give the classifier a genuinely external check—complementing the repeated scaffold splits, which reuse the ChEMBL 33/ATC set—we applied it to approved or investigational antibacterials added to the ATC vocabulary only in the later ChEMBL 37 release. Nine J01/J04A drugs were new relative to the ChEMBL 33 positives (the curated ATC annotation grows slowly, which limits the sam-

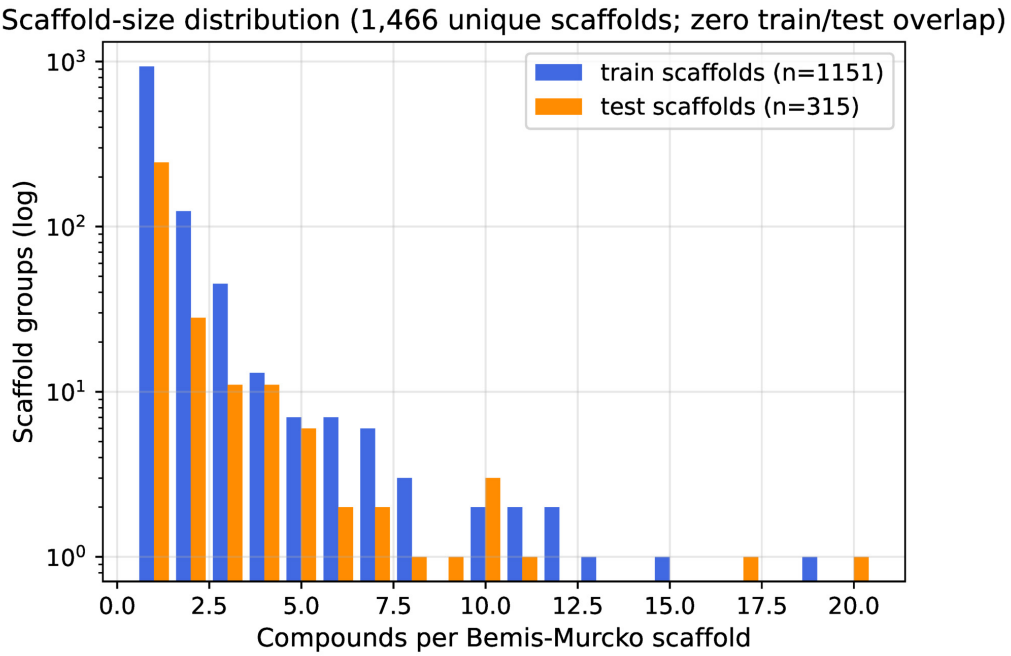


Fig. 3. Bemis–Murcko scaffold-size distribution for the classification dataset (1466 unique scaffolds), split by whether the scaffold falls in the training (1151) or held-out test (315) set under the seed-42 scaffold split; the two sets share no scaffold.

ple size). The ChEMBL 33-trained Random Forest recognised 5 of the 9 as antibacterial (recall 0.56), with 8 of 9 inside the Tanimoto applicability domain; the four misses were structurally distinctive or under-represented chemotypes (an antitubercular, two fluoroquinolones and one other-class agent), consistent with the uneven per-subclass recall of Table 13. The sample is very small, so the estimate is imprecise (exact 95% Clopper–Pearson interval for 5/9 is 0.21–0.86); within that limitation the result is consistent with the classifier extending to genuinely unseen approved antibacterials, and it gives Task A the external check that the potency model receives in Section 3.7.

Table 13. Per-subclass recall (sensitivity) of the Random Forest classifier, aggregated over held-out test appearances across the 25 repeated scaffold splits.

Antibacterial subclass (ATC)	n positives	Recall
β-lactams (J01C/J01D)	96	0.95
Aminoglycosides (J01G)	14	0.93
Tetracyclines (J01A)	15	0.86
Macrolides/lincosamides (J01F)	19	0.85
Quinolones (J01M)	30	0.71
Sulfonamides/trimethoprim (J01E)	23	0.55
Glycopeptides/polymyxins etc. (J01X)	25	0.23
Antituberculars (J04A)	22	0.17
Amphenicols (J01B)	4	0.00

Is the classifier merely a β-lactam detector? To answer this we used the exact ATC subclass of each antibacterial. β-Lactams (ATC J01C/J01D) are the largest subclass (96/248, 38.7% of positives), consistent with the four-membered-ring enrichment in Section 3.1. Removing all 96 β-lactam compounds and retraining over the 25 repeated scaffold splits lowered but did not collapse performance (Table 14): Random Forest test MCC fell from 0.774 (95% CI 0.752–0.795) to 0.685 (0.648–0.722), a drop of 0.089, with ROC-AUC 0.924 → 0.908; XGBoost behaved similarly (0.756 → 0.689). The classifier therefore retains substantial discriminative power on non-β-lactam antibacterials and is not simply a β-lactam detector. The per-subclass recall (Table 13) reinforces this: Random Forest recovers β-lactams best (0.95) but also aminoglycosides (0.93), tetracyclines (0.86), macrolides/lincosamides (0.85) and quinolones (0.71) well, while structurally distinctive or under-represented classes are recognised poorly—sulfonamides/trimethoprim (0.55), glycopeptides/polymyxins (0.23), antituberculars (0.17) and amphenicols (0.00, n = 4). The antibacterial signal is thus distributed across several major families rather than confined to β-lactams, but coverage is uneven and weakest for the most distinctive chemotypes.

Does the ranking survive hyperparameter optimization? Nested cross-validation (Table 3) shows that the default configurations are already near-optimal for the two principal models. For Random Forest, tuning did not improve performance (nested-CV MCC 0.843 ± 0.054 default vs 0.835 ± 0.054 tuned); for XGBoost, tuning gave a small gain (0.819 ± 0.077 → 0.834 ± 0.052). On the main held-out test set the same pattern held (Random Forest 0.845 → 0.836; XGBoost 0.846 → 0.868). All values lie within

Table 14. Leave-β-lactam-compounds-out experiment: test MCC over 25 repeated scaffold splits, full dataset vs all 96 β-lactam positives (ATC J01C/J01D) removed.

Model	MCC full (95% CI)	MCC without β-lactams (95% CI)	Δ MCC	ROC-AUC full	ROC-AUC without β-lactams
Random Forest	0.774 (0.752–0.795)	0.685 (0.648–0.722)	−0.089	0.924	0.908
XGBoost	0.756 (0.732–0.780)	0.689 (0.650–0.729)	−0.067	0.925	0.911

one standard deviation of each other, so the model ranking is stable and Random Forest with default settings remains a well-justified choice for the downstream analyses, while XGBoost is competitive after tuning. This supports the deliberate use of default configurations for the fourteen-model baseline.

Does mixing 2D and 3D descriptor procedures bias the results? 127 compounds (30 positive, 97 negative) used 2D features only because they failed 3D embedding or exceeded the heavy-atom limit. These compounds are structurally distinctive — large molecules (median MW 1145 vs 329 g mol^{-1} for 3D-embedded compounds), predominantly macrolides/lincosamides (14), glycopeptides/polymyxins (8) and other large natural-product antibiotics. However, their different treatment does not bias the model (Table 2): excluding them changed Random Forest test MCC only from 0.774 to 0.735 (XGBoost 0.756 → 0.760), and a fully-2D model with all Mordred 3D descriptors removed performed comparably (Random Forest 0.765, XGBoost 0.760); all values overlap within their 95% confidence intervals. Only 53 of the 213 Mordred 3D descriptors survived the missingness/variance filtering (precisely because the 2D-only compounds lack them), so the 3D block contributes little and the mixed 2D/3D procedure introduces no material bias.

3.3 Generation of New Molecules From SMILES (Exploratory)

The 248 antibacterial SMILES were augmented to 4960 non-canonical strings and used to train a two-layer LSTM (256 units, 45 epochs; loss 1.33 → 0.79). Top-k sampling (k = 4) from seven primes produced 268 candidates, with explicit metrics validity 10.8% (29/268), uniqueness 100%, training-set novelty 86% (25/29). The low validity—far below modern generators (>90%)—reflects the tiny 248-molecule corpus and minimal model, which is why this branch is framed as exploratory; "novelty" is relative to the training set only (Section 2.7), not a claim of absolute novelty. Rather than arbitrary fragments, the generator recombined scaffolds from its 248 training molecules, reproducing the major antibacterial pharmacophores (Fig. 4): penicillin β-lactam cores, a cephalosporin with an aminothiazole–methoxyimino side chain (the defining motif of cefotaxime/ceftriaxone), aminoglycoside polyols, sulfanilamides and diaminopyrimidines (trimethoprim-like).

3.4 Selection of Generated Molecules Using Classification Models

Each generated molecule was scored by the four best classifiers and ranked by mean model score (Table 15). All ten top candidates had mean scores of at least 0.87; the leading structures are well-formed β-lactams and a cefotaxime-like cephalosporin, with close agreement across models. These scores indicate similarity to the ATC-positive training class and should not be interpreted as calibrated probabilities of measured antibacterial activity. In particular, the high classifier scores of these generated molecules are evidence of antibacterial-like chemistry, not of antibacterial potency: potency can only be estimated from measured activity, which is the separate role of the pMIC model (Task B, Section 3.7). This resemblance-versus-potency distinction is maintained throughout — abstract, Results, Discussion, Conclusions, and the Table 15 caption.

Table 15. Top 10 generated molecules ranked by mean classifier score.

#	Mean	RF	GB	XGB	SVM	Scaffold class
1	0.972	1.00	0.99	1.00	0.90	penicillin (β-lactam)
2	0.968	0.95	0.99	1.00	0.93	penicillin (β-lactam)
3	0.956	1.00	1.00	1.00	0.83	cephalosporin (aminothiazole–oxime)
4	0.948	0.84	0.99	1.00	0.96	β-lactam derivative
5	0.946	0.96	0.99	1.00	0.83	β-lactam / heterocycle
6	0.929	0.91	0.99	1.00	0.82	sulfonamide/heterocycle
7	0.917	0.83	0.95	1.00	0.89	aminopyrimidine
8	0.912	0.92	0.98	1.00	0.75	heteroaromatic
9	0.874	0.81	0.96	0.90	0.83	diaminopyrimidine
10	0.871	0.78	0.99	1.00	0.72	heteroaromatic

Scores reflect similarity to the ATC-positive training class and are not calibrated activity probabilities.

Leading SMILES: (1) CC1(C)S[C@@H]2[C@H](NC(=O)C(=O)Nc3ccccc3)C(=O)N2[C@H]1C(=O)O; (2) COc1ccc(C(=O)C(=O)N[C@@H]2C(=O)N3[C@@H]2SC(C)(C)[C@@H]3C(=O)O)cc1; (3) CO/N=C(\C(=O)N[C@@H]1C(=O)N2[C@@H]1SCC(C=O)[C@@H]2C(=O)O)c1csc(N)n1. These are close analogues of known drugs rather than new chemotypes; candidate 3 reproduces the cefotaxime/ceftriaxone side chain. Their novelty is relative to the training set only. Before experimental prioritization, candidates should be checked against external structure, reaction, and commercial databases,

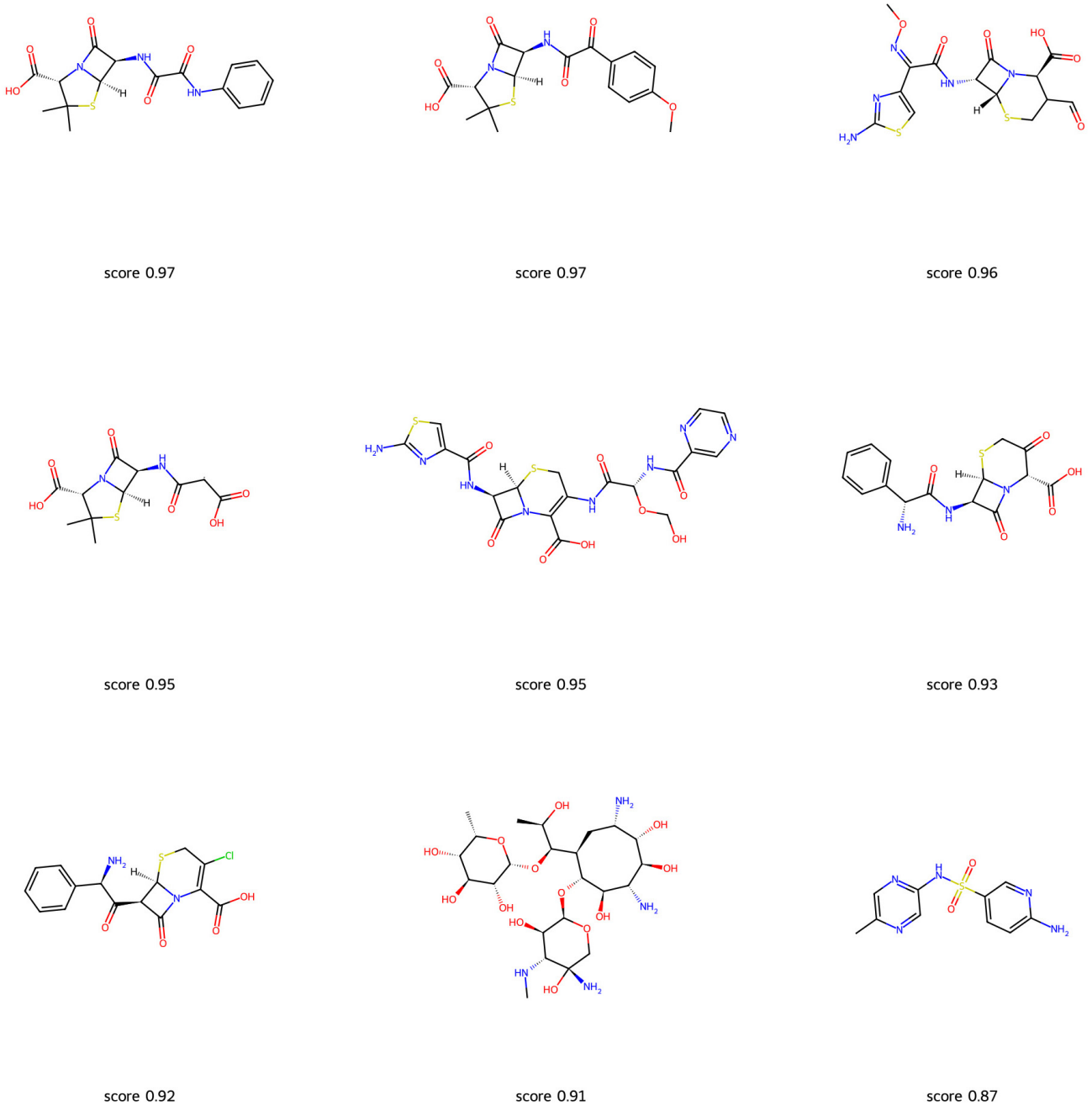


Fig. 4. Representative molecules generated by the exploratory Char-RNN, showing recognizable antibacterial chemotypes recombined from the 248-molecule training set.

screened for synthesis feasibility and liabilities, assessed for applicability-domain membership, and then evaluated in appropriate MIC assays. Consensus scores are hypothesis-generating outputs, not standalone selection decisions.

A quantitative characterisation of the 25 valid generated molecules supports their use as exploratory, synthesizable hypotheses. All are readily synthesizable (RDKit synthetic-accessibility score median 3.4, range 1.8–5.9, all ≤6 — comparable to or simpler than the training antibacterials, median 3.9), internally diverse (mean pairwise ECFP4 Tanimoto 0.14, i.e., diversity 0.86), and all training-set-novel (no exact match to any training molecule). Checked against ChEMBL 33 by InChIKey, 23 of the 25 are absent from the database entirely (only 2 have an exact match), so the set is largely externally novel and not merely a re-derivation of catalogued structures. Their maximum ECFP4 Tanimoto to the 248 known antibacterials is moderate (median 0.44, range 0.15–0.67; none above 0.7): the top-ranked candidates reproduce recognisable β-lactam/cephalosporin motifs (Table 15), but the set as a whole consists of novel recombinations of training-

Table 16. Top eight Random Forest descriptor importances for ATC-based classification.

Rank	Descriptor	Importance	Meaning
1	AATS6se	0.024	electronegativity autocorrelation (lag 6)
2	TopoPSA	0.017	topological polar surface area
3	FNSA5	0.016	charged partial surface area
4	AATS5se	0.014	electronegativity autocorrelation (lag 5)
5	AATS7pe	0.013	polarisability autocorrelation (lag 7)
6	nHetero	0.011	number of heteroatoms
7	SdssC	0.011	E-state of =C< carbons
8	AATS7are	0.011	Allred–Rochow autocorrelation (lag 7)

derived pharmacophores rather than near-duplicates of known drugs. Twenty-three of the 25 also fall inside the classifier applicability domain (Section 3.6). Together these metrics confirm the branch is a genuine, if small, proof-of-concept: it produces synthesizable, diverse, training-set-novel antibacterial-like structures, with the low validity (10.8%) remaining its principal limitation.

3.5 Biochemical and Pharmacophoric Patterns From Feature Importance

Random Forest distributes importance across many weak signals (Table 16, Fig. 5): the top descriptors are electronegativity autocorrelations, **topological polar surface area (TopoPSA)** and heteroatom count (nHetero)—the hallmark of molecules that must cross bacterial envelopes. The model-agnostic permutation importance (Fig. 5, Table 17) ranks the same families highest—a sulfur E-state (SdS), Sanderson-electronegativity-weighted autocorrelations (AATS/ATSC*se) and TopoPSA—confirming the signal is not an artefact of impurity-importance bias. Gradient Boosting concentrates on the β-lactam ring path (Xch-4dv/Xch-4d) and the primary-amine E-state (SsNH2) (Table 18). AATS6se and AATS7are recur in both ensembles.

Table 17. Random Forest permutation importance on the held-out test set (mean ROC-AUC decrease over 30 repeats), with the impurity-based rank for comparison; top eight.

Descriptor	Permutation importance (ΔROC-AUC)	Impurity rank
SdS	0.0037	—
AATS7se	0.0018	—
GATS6i	0.0012	—
TopoPSA	0.0012	2
FNSA5	0.0011	3
PEOE_VSA6	0.0010	—
ATSC1se	0.0010	—
AATS5se	0.0009	4

Because impurity-based importance can be inflated for correlated descriptors, we additionally computed model-agnostic permutation importance for Random Forest on the held-out test set (Table 17). The two rankings broadly agree on polar-surface and heteroatom-related autocorrelation descriptors—TopoPSA and FNSA5 rank in the top five by both measures—but they also diverge: AATS6se, the top impurity descriptor, falls well down the permutation ranking, and nHetero has near-zero permutation importance, indicating their apparent individual importance is largely shared with correlated descriptors. The permutation importances are small in absolute terms (top descriptor ≈ 0.004 ROC-AUC drop), which is the expected signature of a highly intercorrelated descriptor space in which no single variable is individually decisive; the four-membered-ring β-lactam descriptors do not appear among the top permutation features for the same reason (their information is redundant across the group), consistent with the ablation result in Section 3.2. These descriptors should therefore be read as statistical predictors and structural correlates of the ATC-antibacterial class—reflecting the polar, heteroatom-rich character of molecules that engage bacterial targets—rather than as mechanistic determinants of activity.

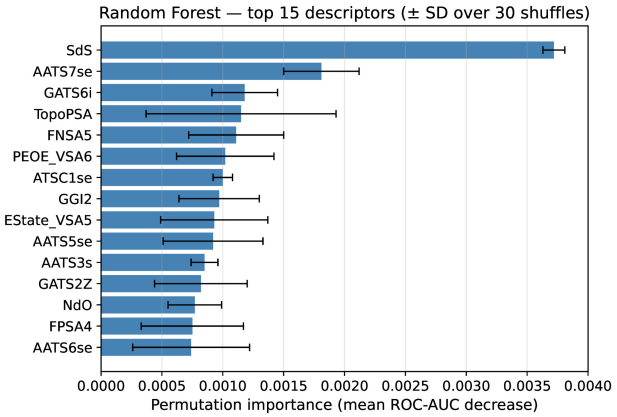


Fig. 5. Top 15 Random Forest descriptors for ATC-based classification. Bars show permutation importance (mean decrease in ROC-AUC over 30 shuffles) with ± SD error bars; permutation importance is used because it is model-agnostic and less biased toward high-cardinality descriptors than impurity importance (Table 17).

Table 18. Top eight Gradient Boosting descriptor importances for ATC-based classification.

Rank	Descriptor	Importance	Meaning
1	Xch-4dv	0.157	4-membered-ring chi path (β-lactam)
2	SsNH2	0.112	primary-amine E-state
3	AATS6se	0.065	electronegativity autocorrelation
4	Xch-4d	0.057	4-membered-ring chi path
5	JGI9	0.044	topological charge index (order 9)
6	EState_VSA10	0.044	E-state × VSA surface bin
7	AATS7are	0.037	Allred–Rochow autocorrelation
8	nG12AHRing	0.030	fused-ring descriptor

3.6 Applicability Domain and Prediction Reliability

The applicability domain (AD) was defined in ECFP4 fingerprint space as the distance to the five nearest training neighbours (Section 2.6; a compound is in-domain when its mean top-5 Tanimoto to training is ≥0.20, the 95th-percentile cutoff). Of the hidden-validation molecules, 792/837 (94.6%) fell inside the AD (Fig. 6); the 5% outside are the more structurally distinct validation compounds, so this fingerprint domain already discriminates more than the earlier Mahalanobis definition, which had placed 836/837 inside. Its operational role is to flag extrapolative external or generated inputs. Applied to the 25 Char-RNN-generated molecules, 23/25 fell inside the AD (mean top-5 Tanimoto 0.37)—two of the more novel structures are correctly flagged as out-of-domain, whereas the earlier Mahalanobis AD (distorted by a single high-variance descriptor) had placed all 25 inside. This is consistent with the generator recombining training-derived pharmacophores while remaining moderately distinct from known antibacterials (ECFP4 median 0.44, none above 0.7) and training-set-novel (Section 3.4). For the regression task, prediction error demonstrably increases with this same Tanimoto distance from the training set (Table 19, Fig. 7), so the domain discriminates reliable from unreliable predictions.

Table 19. pMIC prediction error by Tanimoto distance to the training set: scaffold-held-out test compounds binned into deciles of mean top-5 ECFP4 Tanimoto similarity to their nearest training neighbours (XGBoost).

Similarity decile	n	Mean top-5 Tanimoto	RMSE (log)	MAE (log)	R^2
1 (most similar)	671	0.83	0.60	0.46	0.66
2	671	0.76	0.63	0.49	0.60
3	670	0.71	0.68	0.53	0.60
4	671	0.67	0.69	0.54	0.56
5	670	0.62	0.69	0.53	0.52
6	671	0.57	0.80	0.62	0.41
7	671	0.52	0.76	0.58	0.33
8	669	0.46	0.80	0.61	0.34
9	672	0.40	0.84	0.68	0.20
10 (least similar)	671	0.32	0.88	0.69	0.08

Error rises and R^2 falls monotonically as similarity decreases.

3.7 From Therapeutic-Class Label to Measured Potency: pMIC Regression

Beyond the ATC label, measured potency against *Staphylococcus aureus* was modelled for 33,505 compounds, with pMIC = $-\log_{10}$(MIC_M) and MIC_M (mol L^{-1}) = MIC (μg mL^{-1}) × 10^{-3} / MW (g mol^{-1}) (pMIC = 4.79 ± 1.10; Fig. 8A). Eight regressors using Mordred 2D descriptors were evaluated under a Bemis-Murcko scaffold split so that test chemotypes were unseen. XGBoost performed best (R^2 = 0.510, RMSE = 0.743 log units, Spearman ρ = 0.695), narrowly ahead of Random Forest (Table 20). To confirm this is not an artefact of one partition, the scaffold split was repeated over 10 independent scaffold-group partitions (Table 5): XGBoost gave a mean R^2 of 0.505 (95% CI 0.472–0.537; SD 0.045), RMSE 0.771 (0.748–0.795) log units and Spearman ρ 0.704 (0.687–0.721), so the single-split values are representative and stable. Rankings were preserved, with XGBoost and k-NN the leading regressors across repeats. The nearest-neighbour analysis (Table 12, Fig. 2) shows, however, that scaffold splitting is less effective at separating chemotypes for this dataset than for classification: the median maximum ECFP4 Tanimoto of scaffold-held-out test compounds to the training set was 0.68, and 45.8% still had a close training analogue (maximum Tanimoto >0.7). This reflects the dense, congeneric sampling of the large S. aureus MIC corpus, in which many distinct Bemis–Murcko scaffolds are nonetheless structurally similar; consequently, the reported R^2 partly benefits from near-neighbour information, and truly extrapolative prediction would be harder. Stricter cluster- or similarity-based splitting is a useful direction for future work. The RMSE corresponds to an approximate 5.5-fold multiplicative error scale; it is not a per-compound prediction interval. The parity plot (Fig. 8B) shows central agreement but also regression toward the mean, especially at the potency extremes. A Tanimoto k-NN applicability domain (Section 2.6) retained 5260/6707 test molecules, with R^2 rising to 0.563 for in-domain molecules. To verify that the applicability domain is meaningful, we tested whether prediction error grows with distance from the training set (Table 19, Fig. 7). Prediction error generally increased with increasing distance from the training set, although the relationship was not strictly mono-

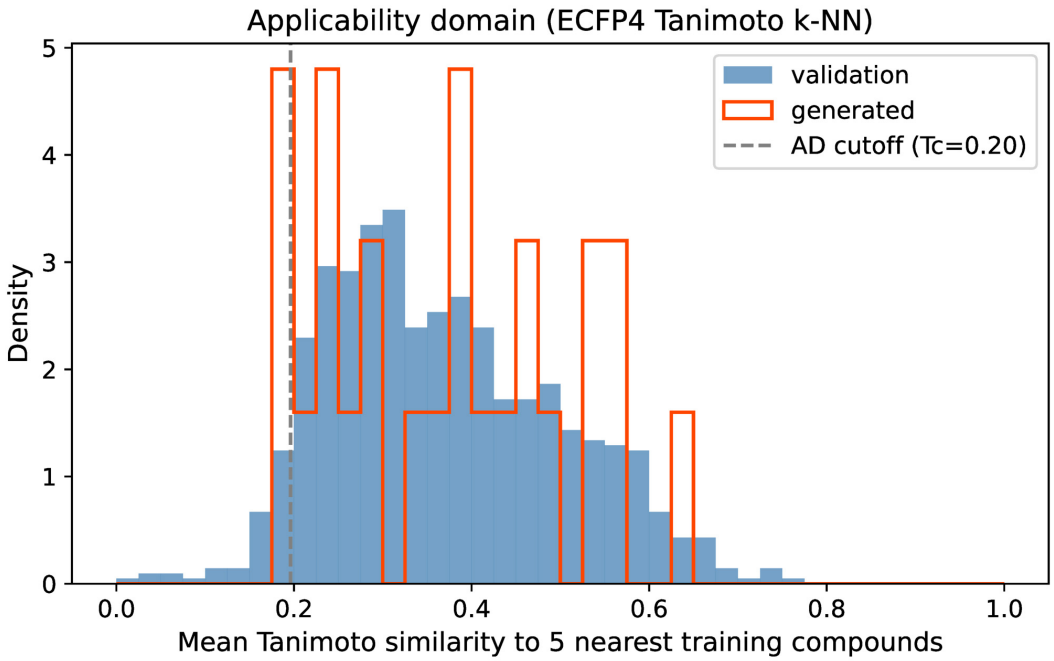


Fig. 6. Applicability domain of the classifier in ECFP4 fingerprint space. Distribution of the mean Tanimoto similarity of each compound to its 5 nearest training molecules, for the hidden-validation set (filled) and the Char-RNN-generated molecules (outline); dashed line = 95th-percentile AD cutoff (mean top-5 Tanimoto ≈ 0.20). 792/837 validation and 23/25 generated molecules are in-domain.

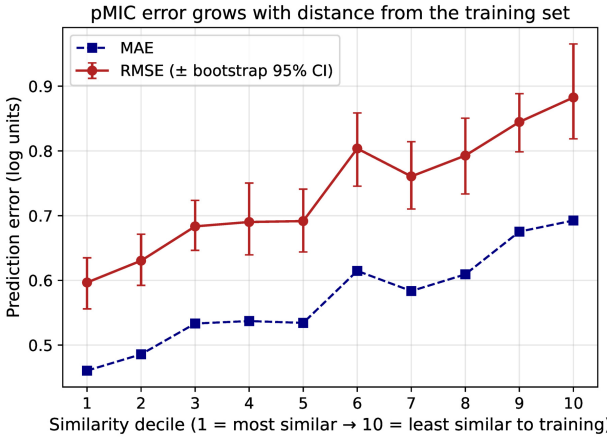


Fig. 7. pMIC prediction error versus Tanimoto distance from the training set. Test compounds are binned into deciles of mean top-5 ECFP4 Tanimoto similarity to their nearest training neighbours (decile 1 = most similar); RMSE (points, with bootstrap 95% CI whiskers) and MAE generally increased as similarity decreased, although fluctuations were observed between adjacent deciles.

tonic across similarity deciles. When the scaffold-held-out compounds were grouped into deciles based on their mean top-5 Tanimoto similarity to the training set, RMSE rose from 0.60 log units in the most-similar decile to 0.88 in the least-similar, MAE from 0.46 to 0.69, and per-decile R^2 fell from 0.66 to 0.08 (Spearman between distance and absolute error = 0.14 over 6707 compounds). Split at the 95th-percentile Tanimoto AD cutoff, in-domain compounds (5260) had R^2 0.563, RMSE 0.709, whereas out-of-domain compounds (1447)—the chemically most distinct—degraded to R^2 0.157, RMSE 0.855. This fingerprint domain separates reliable from unreliable predictions much more sharply than the earlier descriptor-space k-NN domain (which gave in-/out-of-domain R^2 0.528/0.345), and it therefore correctly flags where predictions become unreliable.

Table 20. pMIC-regression performance on the scaffold-held-out test set (6707 compounds).

Model	R^2	RMSE	MAE	Spearman ρ
XGBoost	0.510	0.743	0.573	0.695
Random Forest	0.501	0.750	0.576	0.693
k-NN	0.463	0.778	0.581	0.676
Gradient Boosting	0.363	0.847	0.668	0.584
Ridge	0.356	0.852	0.666	0.596
SVR (RBF/Nyström)	0.343	0.861	0.675	0.574
Elastic Net	0.317	0.877	0.693	0.559
Lasso	0.300	0.888	0.704	0.544

Data heterogeneity and robustness. These results must be read against the corpus: the 90,916 raw MIC measurements come from 11,998 distinct ChEMBL assays across

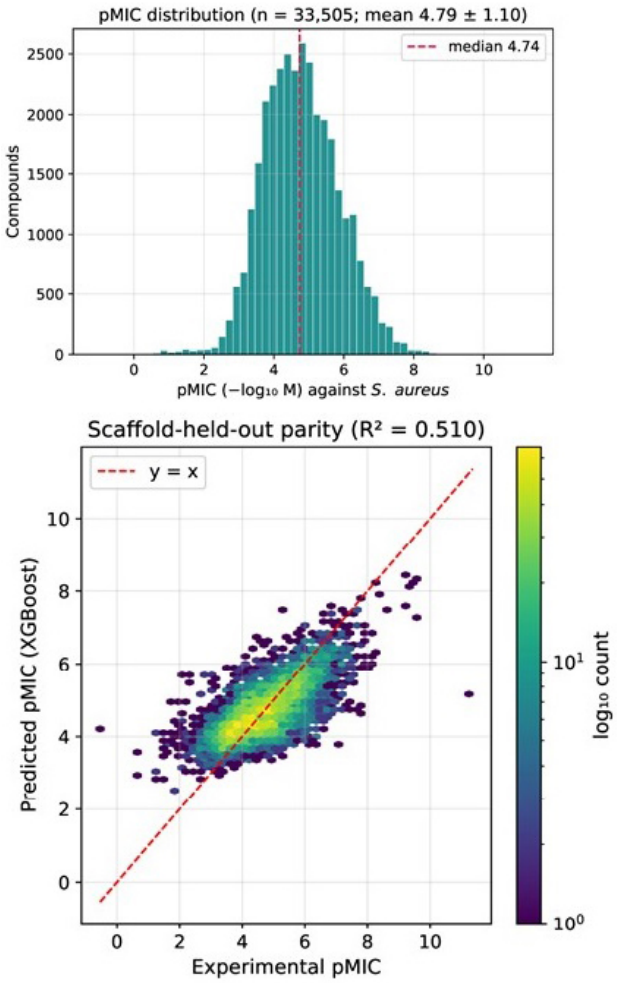


Fig. 8. Potency regression. (Top panel) Distribution of pMIC against *Staphylococcus aureus* across 33,505 ChEMBL compounds. (Bottom panel) Experimental versus predicted pMIC for XGBoost on the scaffold-held-out test set ($R^2 = 0.510$). The panel is now shown as a hexbin density plot with a dashed identity (y = x) line; the colour encodes compound count (log scale). Predictions concentrate around the identity line, with visible regression toward the mean at the potency extremes.

multiple laboratories, *Staphylococcus aureus* strains (MRSA/MSSA; ATCC reference strains and clinical isolates), and media, at low target confidence (confidence_score = 1 for essentially all records). Strain and protocol were not modelled, which places a practical ceiling on R^2. Quantifying this heterogeneity directly: the 90,916 records map to 37,160 measured compounds and 11,998 assays, and 43.1% of compounds carry more than one MIC measurement (mean 2.4, maximum ~3000). Where replicates exist, however, the within-compound scatter is modest—the median per-compound range of pMIC is 0.30 log units (≈2-fold in MIC) and the median SD is 0.21 log units, although 20.5% of replicated compounds still span more than one log unit (10-fold). Conflicting measurements were reduced to the per-compound median (Section 2.8), which is robust to these outliers; because a single value per compound is modelled and the split is by scaffold, no compound appears in both the training and test sets. Regarding strain, 25,977 of the 90,916 records (28.6%), drawn from 3925 of the 11,998 assays (32.7%), carry an explicit MRSA / methicillin-resistant annotation, and 6889 records (7.6%) from 774 assays (6.5%) an MSSA / methicillin-susceptible annotation, while 58,050 records (63.9%) are unlabelled; a strain-stratified analysis is therefore feasible for the annotated subset. Evaluating the scaffold-held-out predictions against strain-specific target pMIC (median over MRSA-only or MSSA-only records) shows the global model serves both phenotypes comparably: on test compounds carrying an MRSA annotation (n = 2160) R^2 was 0.329 (RMSE 0.752, Spearman ρ 0.569), and on MSSA-annotated test compounds (n = 691) R^2 was 0.276 (RMSE 0.790, ρ 0.611) — similar absolute error and ranking for both, with no systematic advantage for either strain. This supports modelling *S. aureus* potency with a single global model rather than separate strain-specific models; the lower R^2 relative to the all-strain test (0.510) reflects the smaller, noisier strain-specific target sets rather than a resistance-phenotype effect.

Is the reported performance an artefact of the molar transformation? Because $\text{pMIC} = -\log_{10}(\text{MIC_M}) = -\log_{10}(\text{MIC in } \mu\text{g mL}^{-1}) + 3 + \log_{10}(\text{MW})$, the molar response explicitly contains a $+\log_{10}(\text{MW})$ term, and XGBoost recovers $\log_{10}(\text{MW})$ from the descriptors essentially perfectly ($R^2 = 1.000$ on unseen scaffolds). However, this term is a negligible part of the response: $\log_{10}(\text{MW})$ has an SD of only 0.135 over the 100–900 g mol^{-1} window and accounts for just 1.5% of the pMIC variance. Consequently, re-modelling a mass-based response ($-\log_{10}$ of MIC in μg mL^{-1}, which contains no MW term) barely changed performance (XGBoost R^2 0.462, 95% CI 0.405–0.519, vs 0.493, 0.445–0.542, for molar pMIC over 5 repeated scaffold splits; identical RMSE ≈ 0.77). The predictive signal is therefore robust to the response definition and reflects genuine structure–potency information rather than molecular-weight prediction.

How does the model perform prospectively, on genuinely later data? As a temporal (quasi-external) validation, we trained only on compounds whose earliest S. aureus MIC record predates 2019 (28,453 compounds) and tested on those first reported in 2019–2022 (5052 compounds), of which 90% carry scaffolds unseen in training. Performance dropped relative to the scaffold split: XGBoost gave R^2 = 0.240, RMSE = 0.875 log units and MAE = 0.689, but the rank correlation was retained at Spearman ρ = 0.502. This is expected—temporal validation compounds a hard scaffold-generalization test with prospective conditions (later chemistry, laboratories and protocols)—and it quantifies honestly that absolute predictive accuracy on external, time-split data is lower than the internal scaffold-split estimate

($R^2 \approx 0.51$), while the model still ranks potency usefully ($\rho \approx 0.5$), supporting its intended prioritization role rather than precise potency prediction.

We further validated against a genuinely independent later database release. The model, trained on ChEMBL 33, was applied to the 5090 compounds whose *S. aureus* MIC is new in ChEMBL 37 (absent from the ChEMBL 33 training set by InChIKey; a dedicated notebook reproduces this analysis). Performance closely matched the temporal split: $R^2 = 0.286$, RMSE = 0.879 log units, MAE = 0.691 and Spearman $\rho = 0.566$, with 43% of the new compounds inside the Tanimoto applicability domain—the low in-domain fraction reflecting how structurally distinct the later-release compounds are. Applying the identical physical filter ($2 \leq$ pMIC ≤ 9; Section 2.8) to this external set removes only 15 records and leaves the external R^2 essentially unchanged (0.286 → 0.282), confirming the filter behaves consistently across the internal and external data. Critically, the clean internal error-versus-distance relationship (Table 19) did not transfer to this genuinely external release (Table 21): across the same Tanimoto-similarity deciles neither RMSE nor MAE rose monotonically as similarity fell, and the overall Spearman correlation between distance and absolute error was ≈ 0 (−0.01 over the 5090 compounds). Split at the 95th-percentile cutoff, in-domain compounds showed a higher R^2 (0.294 vs 0.179 out-of-domain) but also a higher RMSE (0.920 vs 0.847)—so the apparent in-domain R^2 advantage reflects the larger pMIC variance of the in-domain subset, not smaller absolute error. The applicability domain, calibrated on ChEMBL 33, thus flags reliability well within the source distribution (Table 19) but is a weak per-compound error predictor on structurally and temporally shifted external chemistry; this is an honest limitation that argues for recalibrating the domain on the deployment distribution rather than a claim that the domain transfers unchanged. The agreement of two independent external tests—a within-release time split and a genuinely later ChEMBL 37 release—gives a consistent, honest picture: on external, previously unseen compounds the model's absolute accuracy is modest ($R^2 \approx 0.24$–0.29) but its potency ranking is preserved ($\rho \approx 0.50$–0.57), which is what its prioritization use requires.

Excluding the 0.7% of physically implausible records (pMIC <2 or >9; Section 2.8) as a sensitivity check lowered the best-model R^2 only slightly (0.510 to 0.492). In this large, multi-source and noisy dataset evaluated under a strict scaffold split, $R^2 \approx 0.5$ and RMSE ≈ 0.74 log units represent a useful prioritization signal, but not an assay-replacement endpoint.

In contrast to the internal split (Table 19), error does not rise monotonically as similarity to the ChEMBL 33 training set decreases (Spearman between distance and absolute error ≈ 0), so the ChEMBL 33-calibrated applicability domain is a poor per-compound error predictor on shifted external chemistry.

Table 21. External test of the same analysis on the genuinely later ChEMBL 37 release (5090 compounds new vs ChEMBL 33; XGBoost).

Similarity decile	n	Range of mean top-5 Tanimoto	RMSE (log)	MAE (log)	R^2
1 (most similar)	508	0.67–1.00	0.84	0.65	0.34
2	510	0.61–0.67	1.04	0.83	0.11
3	509	0.54–0.61	0.97	0.79	0.24
4	509	0.48–0.54	0.82	0.63	0.26
5	509	0.44–0.48	0.79	0.63	0.26
6	509	0.40–0.44	0.76	0.61	0.19
7	509	0.37–0.40	0.78	0.64	0.25
8	509	0.34–0.37	0.85	0.67	0.20
9	508	0.30–0.34	0.87	0.68	0.12
10 (least similar)	510	0.19–0.30	1.03	0.77	0.10

Hyperparameter tuning of the regressors gave only a modest improvement, confirming that the default configurations are near-optimal (Table 4). For XGBoost, nested cross-validation over scaffold outer folds raised R^2 from 0.485 to 0.515 (Δ +0.031), and on the main scaffold split the tuned model reached $R^2 = 0.531$ versus 0.510 for the default; Random Forest did not benefit (0.498 vs 0.501). The selected XGBoost hyperparameters (600 trees, depth 8, learning rate 0.05, subsample 0.7, colsample_bytree 0.7) are close to the default configuration, so all other results are reported with the default models.

The potency-driving descriptors (Table 22), including lipophilic surface (SlogP_VSA2), information content (IC2), aromatic-ring autocorrelations (ATSC0are, nARing), and a fluorine E-state (SsF, consistent with fluoroquinolones), differ from the class-defining descriptors. This indicates that recognizing ATC antibacterial class membership and estimating measured potency depend on different structural correlates. The data-extraction workflow can be adapted to other pathogens after pathogen-specific curation, model development, and independent validation; the present potency results apply only to *Staphylococcus aureus*.

Table 22. Top descriptors for pMIC prediction from Random Forest and XGBoost.

Random Forest	Imp.	XGBoost	Imp.
SlogP_VSA2	0.067	ATSC0are	0.040
IC2	0.041	nARing	0.020
ATSC0are	0.027	IC2	0.019
SpDiam_A	0.020	SlogP_VSA2	0.017
ATSC0pe	0.011	SsF	0.012
SlogP_VSA3	0.010	ETA_eta	0.012

Table 23. Positioning relative to representative antibacterial machine-learning studies.

Study	Task	Data	Validation	Reported metric
Stokes et al. 2020 [5]	*E. coli* growth-inhibition classification	2335 compounds	external repurposing/ virtual screen	Not compared quantitatively
Liu et al. 2023 [6]	*A. baumannii* classification	7684 compounds	prospective screen	hit-rate enrichment
Scalia et al. 2026 [7]	GNN virtual screen	$>10^9$ enumerated	prospective	enrichment over HTS
García-Ortegón et al. 2025 [8]	transfer-learning hit identification	ESKAPE screens	prospective	sub-μM hits
This work — Task A	ATC therapeutic-class classification	2788 compounds	25 repeated scaffold splits	MCC 0.774 (0.752–0.795)
This work — Task B	*S. aureus* pMIC regression	33,505 compounds	scaffold + 2 external tests	R^2 0.51 int / 0.29 ext; ρ 0.70 / 0.57

Tasks, datasets and validation regimes differ, so metrics are not directly comparable; the table gives context rather than a ranking.

4. Discussion

4.1 Interpretation and Limitations

The classifier learns an ATC-based therapeutic-class definition of "antibacterial drug"; the negative class may contain compounds with unrecorded or latent antibacterial activity, a limitation that the pMIC regression partly addresses by using measured activity. The classifier signal is enriched in four-membered β-lactam descriptors (40% of actives versus 0.5% of negatives; Sections 3.1–3.2). However, ablation and scaffold-split analyses showed that predictive performance remained robust after removing six descriptors, with MCC values around 0.75 on held-out scaffolds. Removing the β-lactam *compounds* entirely (Table 14) lowered Random Forest MCC only from 0.774 to 0.685, and the per-subclass recall (Table 13) shows the model also recognises aminoglycosides, tetracyclines, macrolides and quinolones, so the antibacterial signal is distributed across families rather than confined to β-lactams; coverage is nonetheless uneven and weakest for structurally distinctive classes such as glycopeptides, polymyxins and antituberculars, which the global model should flag through the applicability domain. The scaffold-split performance estimate, rather than the random-split result, is therefore more appropriate for assessing prospective generalization. Random splitting can place close analogues on both sides and does not exclude analogue-level leakage. Repeated scaffold-aware validation over 25 partitions confirmed the stability of the held-out estimate (Table 11), and the class-holdout and nearest-neighbour similarity analyses provide further evidence that the reported performance is not driven solely by close analogues. Antibacterials are exceptionally heterogeneous, so a global model averages over unrelated chemotypes and the applicability domain should accompany external predictions. The MIC corpus aggregates 11,998 assays across multiple laboratories, strains, and protocols, caps achievable R^2, and pertains to *Staphylococcus aureus* only. Three-dimensional geometries used ETKDG/MMFF rather than quantum optimization. Hyperparameter tuning of both the classifiers (Table 3) and the regressors (Table 4) was performed and gave only marginal gains over the default configurations, which are therefore retained throughout. The models are best used as *in silico* pre-screening tools for library prioritization, candidate ranking, potency estimation, and AD-guided selection for experimental validation; they complement, rather than replace, expert judgment and experiments.

4.2 Novelty and Positioning

Unlike previous studies primarily focused on predicting microbial growth inhibition or MIC values, the present work examines two distinct endpoints: ATC-defined antibacterial drug classification and experimentally measured antibacterial potency. The former characterizes molecular features associated with established antibacterial drugs, whereas the latter estimates activity against *S. aureus*. Their separate evaluation allows the interpretation of classification results without assuming that structural resemblance necessarily indicates antibacterial activity.

An important methodological contribution is the comprehensive assessment of model generalization beyond conventional random splitting. Bemis–Murcko scaffold partitioning, repeated scaffold-group validation, nearest-neighbour similarity analysis, and β-lactam sensitivity analyses were used to evaluate robustness across structurally distinct compounds. Comparison with a tuned molecular-similarity baseline showed comparable classification performance, suggesting that the descriptor-based model offers advantages primarily in interpretability and applicability-domain assessment rather than predictive accuracy.

The potency-regression analysis further incorporates temporal validation and an independent external dataset from ChEMBL 37. Although absolute predictive accuracy declined on these datasets, moderate rank correlations were retained. These results illustrate the challenges of modelling heterogeneous antibacterial susceptibility data and suggest that the regression model may be more useful for

preliminary compound ranking than for precise potency estimation.

Overall, the novelty of this work lies in its evaluation strategy rather than the development of new machine-learning architectures. By incorporating complementary prediction endpoints, structural sensitivity analyses, and external validation, the study provides a more comprehensive assessment of model reliability and practical limitations. Table 23 (Ref. [5,6,7,8]) positions these methodological contributions relative to representative antibacterial machine-learning studies. Given differences in datasets, prediction tasks, and validation strategies, the comparison is intended to provide context rather than a direct performance ranking.

5. Conclusions

This study establishes a reproducible machine-learning framework that distinguishes recognition of established antibacterial drug classes from prediction of experimentally measured antibacterial potency. Its principal methodological contribution is the integration of scaffold-aware evaluation, chemical-class sensitivity analyses, molecular-similarity benchmarking, and external validation within a transparent compound-prioritization workflow. The results demonstrate that ATC-based classification can identify antibacterial-like molecular structures, while scaffold-based pMIC regression provides moderate predictive performance for *Staphylococcus aureus*, with reduced accuracy on external datasets. These findings support the use of both approaches as complementary tools for early-stage compound prioritization rather than substitutes for experimental antibacterial susceptibility testing. Future studies should focus on prospective MIC validation, broader pathogen coverage, and further assessment of generated molecular candidates.

Abbreviations

ATC, Anatomical Therapeutic Chemical classification; MIC, minimum inhibitory concentration; MCC, Matthews correlation coefficient; AD, applicability domain; SMOTE, Synthetic Minority Over-sampling Technique; FDR, false discovery rate; PSA, polar surface area; CV, cross-validation; MW, molecular weight.

Availability of Data and Materials

All primary data are publicly available: the study was built from ChEMBL 33 (https://doi.org/10.6019/CHEMBL.database.33) and externally validated against ChEMBL 37 (https://doi.org/10.6019/CHEMBL.database.37). The seven executable notebooks (dataset assembly, descriptor computation, classification, Char-RNN generation, pMIC regression, ChEMBL 37 external validation, and the per-compound external prediction table), helper scripts, cleaned Mordred descriptor matrices, trained models, metric tables and figures, together with a Dockerfile pinning the software environment, are openly available in the project repository at https://github.com/pruhlo/abx-ml. No other datasets were generated or analysed during this study. The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Author Contributions

YP: Conceptualization, methodology, software, data curation, formal analysis, investigation, validation, visualization, and writing – original draft preparation. II: Methodology, supervision, validation, interpretation of results, and writing – review and editing. Both authors have read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

The authors would like to thank all the peer reviewers for their valuable opinions and suggestions.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet (London, England). 2022; 399: 629–655. https://doi.org/10.1016/S0140-6736(21)02724-0

[2] GBD 2021 Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance 1990-2021: a systematic analysis with forecasts to 2050. Lancet (London, England). 2024; 404: 1199–1226. https://doi.org/10.1016/S0140-6736(24)01867-1

[3] 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. World Health Organization: Geneva. 2024. Available at: https://www.who.int/publications/i/item/9789240093461 (Accessed: 21 October 2026).

[4] Theuretzbacher U, Outterson K, Engel A, Karlén A. The global preclinical antibacterial pipeline. Nature Reviews. Microbiology. 2020; 18: 275–285. https://doi.org/10.1038/s41579-019-0288-0

[5] Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A Deep Learning Approach to Antibiotic Discovery. Cell. 2020; 180: 688–702.e13. https://doi.org/10.1016/j.cell.2020.01.021

[6] Liu G, Catacutan DB, Rathod K, Swanson K, Jin W, Mohammed JC, et al. Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. Nature Chemical Biology. 2023; 19: 1342–1350. https://doi.org/10.1038/s41589-023-01349-8

[7] Scalia G, Rutherford ST, Lu Z, Buchholz KR, Skelton N, Chuang K, et al. Deep-learning-based virtual screening of antibacterial

compounds. Nature Biotechnology. 2026; 44: 1154–1167. https://doi.org/10.1038/s41587-025-02814-6

[8] García-Ortegón M, Seal S, Geddes E, Littler JL, Guy CS, Whiteside J, et al. Transfer learning enables discovery of sub-micromolar antibacterials for ESKAPE pathogens from ultra-large chemical spaces. Chemical Science. 2025; 16: 21518–21533. https://doi.org/10.1039/d5sc03055b

[9] Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. Journal of Medicinal Chemistry. 1996; 39: 2887–2893. https://doi.org/10.1021/jm9602928

[10] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chemical Science. 2017; 9: 513–530. https://doi.org/10.1039/c7sc02664a

[11] Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Central Science. 2018; 4: 120–131. https://doi.org/10.1021/acscentsci.7b00512

[12] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Central Science. 2018; 4: 268–276. https://doi.org/10.1021/acscentsci.7b00572

[13] Pruhlo Y, Iurchenko I, Tomenko A. Modeling of New Agents with Potential Antidiabetic Activity Based on Machine Learning Algorithms. AppliedChem. 2025; 5: 30. https://doi.org/10.3390/appliedchem5040030

[14] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Research. 2019; 47: D930–D940. https://doi.org/10.1093/nar/gky1075

[15] Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. Journal of Cheminformatics. 2018; 10: 4. https://doi.org/10.1186/s13321-018-0258-y

[16] RDKit. RDKit: open-source cheminformatics software. 2024. Available at: https://www.rdkit.org (Accessed: 21 October 2026).

[17] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002; 16: 321–357. https://doi.org/10.1613/jair.953

[18] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011; 12: 2825–2830. Available at: https://www.jmlr.org/papers/v12/pedregosa11a.html (Accessed: 21 October 2026).

[19] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery: New York. 2016.

[20] Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research. 2017; 18: 1–5.

[21] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems. 2019; 32: 8024–8035.