

# Sequence and Physicochemical Associations with Antimicrobial Peptide Activity against *Acinetobacter baumannii*: A Homology-Aware Regression and Explainable Machine-Learning Study

Aziza Omari<sup>1</sup>, Nurlan Sadykov<sup>1</sup>, Bakhytzhan Seksenbayev<sup>2</sup>, Yerlan Suleimenov<sup>2,3</sup>, Assylkhan Bekbayev<sup>2</sup>

<sup>1</sup>School of Pharmacy, Sharmanov National Medical University, Astana, Kazakhstan

<sup>2</sup>OLYMP Clinical Diagnostic Laboratories LLP, Astana, Kazakhstan

<sup>3</sup>QazGene LLP, Astana, Kazakhstan

Received: 2026-07-28.

Accepted: 2026-10-01.



This work is licensed under a  
Creative Commons Attribution 4.0  
International License

J Clin Med Kaz 2026; 23(5): 65-75

Corresponding author:

Aziza Omari.

E-mail: [omari.a@amu.kz](mailto:omari.a@amu.kz).

ORCID: 0000-0002-5300-5783.

## ABSTRACT

**Background:** Carbapenem-resistant *Acinetobacter baumannii* is a critical-priority pathogen, but the available database records were predominantly not annotated for carbapenem resistance. We evaluated associations between interpretable peptide features and *A. baumannii* minimum inhibitory concentration (MIC), and tested transfer across sequence-disjoint clusters.

**Methods:** From 1,740 raw records, 809 exact or reported-mean MIC observations were eligible and were aggregated into 525 conservative chemical entities. The outcome was entity-level median log<sub>2</sub> MIC in micromolar units. Twelve prespecified sequence and physicochemical features were analyzed by multiple regression with HC3 and 80%-sequence-cluster-robust uncertainty estimates. Peptides sharing at least 80% normalized global identity were kept within the same development or test partition. A median baseline, Elastic Net, Random Forest, and Extra Trees were compared by nested grouped cross-validation; the locked Extra Trees model was evaluated in 104 test entities. Descriptor, assay-metadata, structural, censoring, alignment, homology-threshold, and similarity-domain sensitivities were examined.

**Results:** Only 19 of 809 observations (2.3%; 14 entities) were explicitly labeled carbapenem-resistant, and 578 (71.4%) lacked a resistance label. Seven features had nominal HC3 P values below 0.05, including negative associations of MIC with charge, mean hydrophobicity, beta-strand hydrophobic moment, cysteine fraction, and C-terminal amidation. Cluster-robust and leave-one-cluster-out analyses reduced precision and showed that the cysteine and N-terminal acetylation findings were concentrated in particular sequence families. Extra Trees showed weak development-set transfer (pooled out-of-fold  $R^2=0.054$ ; mean outer-fold RMSE=2.174, versus 2.236 for the median baseline). In the test set, RMSE was 1.775 log<sub>2</sub> units,  $R^2$  was 0.381 (95% cluster-bootstrap CI -0.024 to 0.543), and Spearman rho was 0.579. Performance weakened when more remote families were separated and in the named-strain-only subset.

**Conclusions:** The data support hypothesis-generating associations rather than causal feature effects. Predictive transfer to unseen peptide clusters was weak during grouped cross-validation, and the more favorable single test estimate was uncertain. The model is not validated as a stand-alone prioritization method and does not support carbapenem-resistant-specific inference; standardized prospective MIC testing, toxicity and stability assessment, and external validation remain necessary.

**Keywords:** *Acinetobacter baumannii*; antimicrobial peptides; minimum inhibitory concentration; homology-aware validation; machine learning; sequence clustering.

## Introduction

Antimicrobial resistance is a major cause of mortality and health-system burden. A systematic analysis estimated 1.27 million deaths directly attributable to bacterial antimicrobial resistance in 2019, with a substantially larger number associated with resistant infections [1]. Updated global estimates confirm that the burden has continued and is expected to remain substantial through 2050 [2]. The World Health Organization classifies carbapenem-resistant *Acinetobacter baumannii* among its critical-priority pathogens for antibacterial research and development [3].

*A. baumannii* is particularly problematic in intensive-care and other healthcare settings because it combines environmental persistence, genomic plasticity, multidrug resistance, and a repertoire of virulence and envelope adaptations [4,5]. Treatment options for resistant infections are limited and can depend on agents with narrow therapeutic margins. This creates a need for antibacterial chemotypes that operate through mechanisms distinct from those of conventional small molecules.

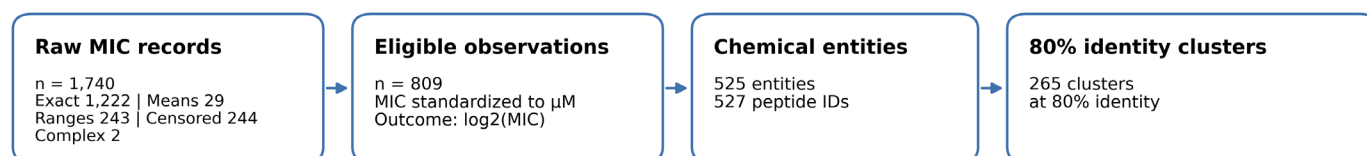
Antimicrobial peptides (AMPs) are structurally diverse innate-defense and synthetic molecules with direct antimicrobial, membrane-active, and sometimes immunomodulatory functions [6,7]. Their activity commonly depends on a balance among positive charge, hydrophobicity, amphipathicity, conformational propensity, sequence length, and terminal or side-chain modification [6-8]. These same properties can also influence solubility, proteolytic stability, hemolysis, and host-cell toxicity, so activity optimization is intrinsically multivariable.

Curated repositories such as DBAASP make species-specific sequence-activity studies possible [9], and the AbAMPdb resource reflects growing interest in peptides tested against *A. baumannii* [10]. Rationally modified helical peptides and compact de novo peptides have shown activity against resistant isolates [11,12], and few-shot learning has been coupled to experimental validation of short-peptide candidates [13]. These examples do not remove the need to distinguish measured associations from family, structure, and assay effects in heterogeneous databases.

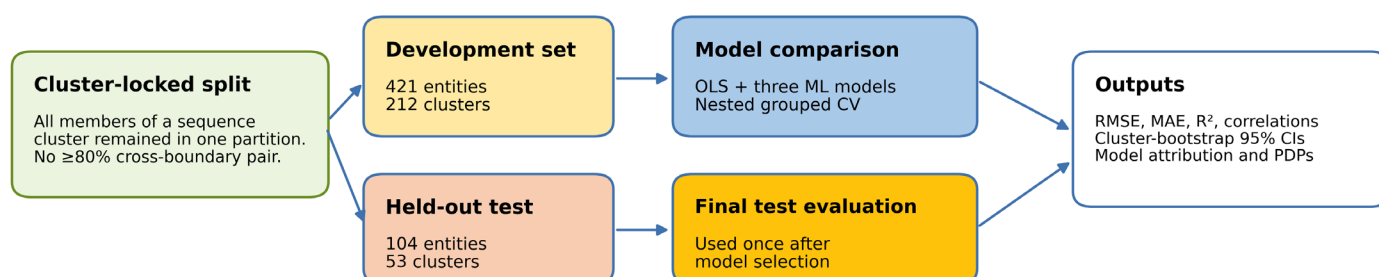
Database-derived model performance can be inflated when near-identical analogues are divided between training and test sets. This risk is especially high in AMP collections containing alanine scans, terminal modifications, and analogue series. DOME recommendations emphasize transparent provenance, independent evaluation, and leakage control [14], and AMP benchmarking studies show that dataset construction can materially change apparent performance [15]. Assay heterogeneity and incomplete strain-level metadata can introduce additional confounding that sequence-only models cannot resolve.

This database-derived, homology-aware study had three objectives: to estimate adjusted associations between 12 interpretable sequence and physicochemical features and *A. baumannii* MIC; to quantify predictive transfer to 80%-identity clusters absent from development; and to characterize how resistance labeling, assay metadata, descriptor specification, family concentration, and homology definition affect interpretation. The analysis was designed to bound uncertainty

## A Data curation and entity construction



## B Leakage-controlled model development and evaluation



## C Interpretation and robustness

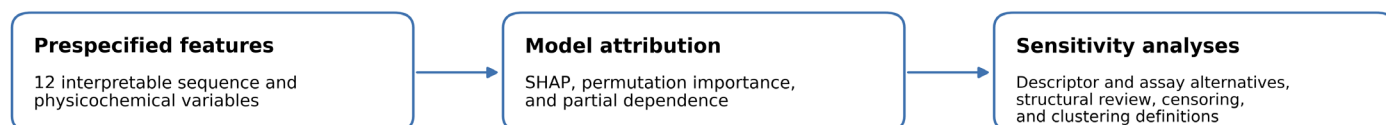


**Figure 1** – Study flow, leakage-controlled validation, and robustness framework

Raw database-derived MIC records were harmonized and aggregated into conservative chemical entities. Sequence clusters were locked within partitions and cross-validation folds. The locked model was evaluated in an untouched test set, and interpretation was accompanied by descriptor, assay, family, alignment, and homology sensitivities. Abbreviations: MIC, minimum inhibitory concentration; OLS, ordinary least squares; PDP, partial dependence plot; SHAP, Shapley additive explanations.

and domain of applicability rather than infer causal feature effects or establish a carbapenem-resistant-specific prediction tool.

## Methods

### Study design and reporting

This was a secondary, database-derived regression and machine-learning study. The primary unit of analysis was a conservative peptide chemical entity, and the continuous outcome was the median log<sub>2</sub>-transformed MIC in micromolar units. Analyses were organized according to the Data, Optimization, Model, and Evaluation domains of the DOME recommendations [14]. Figure 1 summarizes the workflow, and Supplementary Table S9 provides the reporting checklist.

### Data source, eligibility, MIC harmonization, and censoring

Activity records were obtained from a curated DBAASP-derived workbook [9]. The source contained 1,740 raw MIC records: 1,222 exact values, 29 mean-plus-standard-deviation values, 243 ranges, 244 censored values, and two complex records. Eligibility required *A. baumannii* as the target species, a valid peptide sequence and molecular weight, an exact numeric MIC or reported mean, units of micromolar or micrograms per milliliter, and successful descriptor quality control. Range, censored, and complex records were excluded rather than assigned arbitrary point estimates. Because exclusion may be informative if limit-reported values are concentrated among very weak or very strong activities, the resulting regression is conditional on numerically resolved records; the direction and magnitude of that selection cannot be recovered from the available metadata.

The eligible observation-level set comprised 809 measurements: 782 exact values and 27 reported means. Of these, 402 were originally reported in micromolar units and 407 in micrograms per milliliter. Mass concentrations were converted as MIC (μM)=MIC (μg/mL)×1,000/molecular weight (Da). Salt form and hydration state were not recorded consistently and could not be incorporated into molecular-weight conversion. The outcome was log<sub>2</sub> transformed to express error on a twofold-dilution scale. A threshold of 16 μM was used only to assess descriptive split balance. Sensitivity analyses restricted the data to values originally reported in micromolar units and to exact values.

Resistance labels and assay-condition fields were summarized before modeling. Nineteen observations were explicitly labeled carbapenem-resistant, 171 multidrug-resistant, 41 other resistance, and 578 had no resistance label; no explicit susceptible category was available. Strain labels, medium, inoculum, pH, ionic strength, salt type, incubation time, and temperature were audited for completeness (Table 2; Supplementary Table S11). Sparse and nonstandard fields precluded reliable multivariable adjustment. Instead, separate analyses were restricted to MHB/CAMHB observations, MHB/CAMHB observations with exact inocula of 5×10<sup>5</sup> or 1×10<sup>6</sup> CFU, and observations with a named strain.

### Chemical-entity construction and repeated measurements

Records were aggregated by normalized amino-acid sequence and supported terminal chemistry. Exact sequence matches with the same N- and C-terminal state were treated as

one chemical entity. Explicit cyclic peptides and four OH-CM6 variants with distinct reported cross-link patterns were retained as separate entities pending source-level structural verification. This produced 525 entities from 527 DBAASP peptide identifiers and 519 unique normalized sequences.

For each entity, the median of eligible log<sub>2</sub> MIC observations was used as the primary outcome. Median aggregation prevented frequently tested peptides from receiving disproportionate weight and reduced sensitivity to isolated extremes. The full 809-row observation-level dataset was retained for audit. Publication identifiers were not consistently resolvable from the database reference field, so a publication-level random-effects model could not be fitted.

### Sequence-similarity clustering and leakage control

Primary global identity was defined as 1 minus Levenshtein edit distance divided by the longer sequence length. This normalized edit measure was selected because it is deterministic, symmetric, gap-penalizing, and applicable to short sequences of different lengths without choosing a substitution matrix. An undirected graph linked pairs meeting the identity threshold, and connected components defined clusters; this can merge chains of pairwise-similar sequences and was therefore evaluated at 70%, 80%, and 90% thresholds. As an alignment-based sensitivity, 80% clusters were recomputed with global Needleman-Wunsch alignment (match 1, mismatch -1, gap-open -1, gap-extension -0.5), with identity defined as exact matches divided by aligned length.

All entities in an 80%-identity cluster were assigned atomically to the same partition. An 80:20 split was selected from 1,000 candidate group splits to balance the median outcome and the proportion with MIC ≤16 μM; the fixed split seed was 20260806. The development set contained 421 entities in 212 clusters and the untouched test set contained 104 entities in 53 clusters. No sequence pair with identity ≥80% crossed the final boundary. By contrast, a prior exact-sequence split contained 304 cross-boundary pairs at this threshold. Five grouped outer folds were assigned within the development set while keeping clusters intact.

### Prespecified features and descriptor sensitivities

The primary feature set contained 12 variables chosen a priori for interpretability: sequence length; net charge at pH 7.4; Eisenberg mean hydrophobicity [16]; maximum 11-residue-window hydrophobic moment at 100 degrees (alpha-helical geometry) and 160 degrees (beta-strand geometry); aliphatic index [18]; normalized Shannon sequence entropy [19]; cysteine, glycine, and proline fractions; N-terminal acetylation; and C-terminal amidation. Charge used EMBOSS-compatible pKa values and Henderson-Hasselbalch ionization with supported terminal modifications [17].

Modified-termini isoelectric point was missing for 218 of 525 entities; charge density is algebraically related to charge and length; and Boman index correlated strongly with mean hydrophobicity in development (r approximately -0.955). These variables were therefore not entered together in the prespecified model. Their biological relevance was examined in alternative models: Boman index added to or replacing mean hydrophobicity; charge density added to or replacing charge and length; and isoelectric point analyzed in complete cases and after median imputation with a missingness indicator. Variance inflation factors and model fit were reported for each specification (Supplementary Table S13).

### Adjusted regression and family-cluster sensitivity

Multiple ordinary least-squares regression was fitted in the development set with all 12 prespecified variables. Standardized coefficients compared association magnitudes; unstandardized coefficients retained the original units. HC3 heteroscedasticity-consistent standard errors and 95% confidence intervals were the primary inferential estimates [20]. A second covariance estimate treated the 80%-identity sequence cluster as the dependence unit. Diagnostics included variance inflation factors, residual-versus-fitted inspection, Breusch-Pagan and White tests, a normal Q-Q plot, the Jarque-Bera test, and Cook's distance. All P values were exploratory and were not interpreted as causal evidence.

Two focused leave-one-cluster-out analyses assessed family concentration. For N-terminal acetylation, the 80%-identity cluster containing most acetylated development entities was removed and the 12-feature model refitted. For cysteine, the cluster containing the largest number of entities with at least two cysteines was removed and the model refitted. These analyses test dependence on dominant families but cannot eliminate residual structural-class confounding.

### Machine-learning development, audit, and test evaluation

A median-prediction baseline was compared with Elastic Net [21], Random Forest [22], and Extra Trees [23]. The archived primary analysis used five outer grouped folds and three inner grouped folds, with 80%-identity clusters kept intact and RMSE as the optimization metric. Three inner folds were used to retain adequate numbers of clusters in each partition of the 421-entity development set while limiting instability and computational burden [24]. The primary archived outputs preserve fold-level performance, selected settings, and the locked final model, but the launch script containing the complete original candidate lists was not recoverable. To avoid reconstructing unverified settings or changing the test model after test outcomes were known, the primary results were retained and a separate fully specified repeat audit was performed (Supplementary Tables S16-S17).

The locked primary Extra Trees model used 160 trees, max\_features=0.5, min\_samples\_leaf=3, min\_samples\_split=2, and unrestricted depth. The repeat audit specified the complete Elastic Net grid and tree-model search spaces, used three grouped inner folds, 12 randomized tree-parameter draws per outer fold, split seed 20260806, fold seed 20260755, and analysis seed 20260717. The audit was interpreted as a reproducibility and model-selection-stability check, not as grounds to replace the locked test model.

Performance metrics were RMSE, mean absolute error (MAE), R<sup>2</sup>, Spearman rank correlation, and Pearson correlation. The held-out test set was evaluated with the locked Extra Trees model. Ninety-five percent confidence intervals were estimated from 1,000 bootstrap samples of intact test clusters [27].

### Model attribution and sensitivity analyses

For the locked model, SHAP values were calculated on held-out test predictions to describe each feature's contribution to predicted log<sub>2</sub> MIC [26]. Global mean absolute SHAP values, a SHAP distribution plot, permutation importance, and partial-dependence functions were reported. These are model attributions, not causal effects. Fold-wise out-of-fold SHAP rankings were compared to assess stability beyond the single test partition.

Prespecified data-quality analyses restricted the data to exact MIC values, original micromolar values, exact original-micromolar values, and entities not requiring structural review; another excluded entity medians below 0.05 or above 256  $\mu$ M. Regression robustness was examined after excluding Cook-influential entities. Generalization was re-estimated under 70%, 80%, and 90% grouping, and with Needleman-Wunsch clusters. Additional analyses evaluated assay-restricted subsets, alternative descriptors, cluster-robust inference, dominant-family exclusion, and absolute error across nearest-development-sequence identity strata. These analyses were treated as robustness and domain-of-applicability checks rather than independent confirmatory tests.

### Software and reproducibility

Analyses used Python 3.13.5 with NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, scikit-learn 1.8.0 [25], statsmodels 0.14.6, SHAP 0.50.0, Matplotlib 3.10.8, joblib 1.5.3, and Biopython 1.86. The analysis-ready workbook, original and repeat-audit outputs, executable scripts, fitted model, and figure-generation code are included in the submission package and deposited at Zenodo (<https://doi.org/10.5281/zenodo.22170145>).

## Results

### Dataset composition, resistance labels, and assay metadata

The 809 eligible observations represented 525 conservative chemical entities. Entity-level median MIC was 4.32  $\mu$ M (interquartile range, 1.93-12.50), 426 entities (81.1%) had median MIC  $\leq 16$   $\mu$ M, median sequence length was 17 residues, and 280 entities (53.3%) were C-terminally amidated. Eighty-two entities contained at least two cysteines, and six explicitly cyclic or otherwise cross-linked entities required manual structural review.

Outcome balance was similar across the leakage-controlled split. Median log<sub>2</sub> MIC was 2.117 in development and 2.056 in test, while the proportions with MIC  $\leq 16$   $\mu$ M were 81.2% and 80.8%, respectively (Table 1). The test set had somewhat longer sequences but remained completely disjoint from development at the primary 80% cluster definition.

Resistance status was inadequate for a susceptible-versus-resistant comparison. The 19 CRAB observations represented 14 entities; 171 observations were labeled MDR, 41 had

**Table 1** Entity-level dataset and leakage-controlled split

| Partition         | Entities | 80% identity clusters | Median MIC, $\mu$ M [IQR] | Length, residues [IQR] | Net charge [IQR] | C-terminal amidation | MIC $\leq 16$ $\mu$ M |
|-------------------|----------|-----------------------|---------------------------|------------------------|------------------|----------------------|-----------------------|
| Overall           | 525      | 265                   | 4.32 [1.93-12.50]         | 17.0 [13.0-22.0]       | 5.94 [4.94-7.00] | 280 (53.3%)          | 426 (81.1%)           |
| Development set   | 421      | 212                   | 4.34 [1.97-12.50]         | 16.0 [13.0-22.0]       | 5.79 [4.64-7.00] | 224 (53.2%)          | 342 (81.2%)           |
| Held-out test set | 104      | 53                    | 4.16 [1.64-11.61]         | 19.5 [14.0-29.0]       | 5.94 [4.94-7.94] | 56 (53.8%)           | 84 (80.8%)            |

*Note.* The descriptive activity threshold was used only to assess split balance. Lower MIC denotes greater measured activity. Abbreviations: IQR, interquartile range; MIC, minimum inhibitory concentration.

another resistance label, and 578 lacked a resistance label. No observation category explicitly identified susceptible isolates. Medium was reported for 806 observations and inoculum for 760, but named strain labels were available for 320; pH, ionic strength, salt type, incubation time, and temperature were largely or entirely unavailable (Table 2; Supplementary Table S11).

#### Adjusted associations and family sensitivity

The 12-feature model explained 28.1% of development-set variance (adjusted  $R^2$ , 26.0%). All primary-model variance inflation factors were below 3.73. Under HC3 inference, seven variables had nominal  $P$  values below 0.05 (Table 3; Figure 2): net positive charge (standardized  $\beta$  -0.305, 95% CI -0.515 to -0.095;  $P=0.004$ ), mean hydrophobicity ( $\beta$  -0.206;  $P=0.023$ ), beta-strand hydrophobic moment ( $\beta$  -0.141;  $P=0.028$ ), cysteine fraction ( $\beta$  -0.469;  $P<0.001$ ), C-terminal amidation ( $\beta$  -0.127;  $P=0.023$ ), aliphatic index ( $\beta$  0.141;  $P=0.035$ ), and N-terminal

acetylation ( $\beta$  0.083;  $P<0.001$ ). These estimates describe conditional associations in the development data.

Cluster-robust uncertainty was wider. Cysteine fraction and N-terminal acetylation retained  $P<0.001$ , C-terminal amidation had  $P=0.078$ , and the charge, hydrophobicity, beta-strand moment, and aliphatic-index  $P$  values ranged from 0.091 to 0.165. N-terminal acetylation was present in 14 entities overall and 13 development entities distributed across three development clusters; 11 of 13 were in cluster SC80\_0075. After excluding that cluster, the acetylation coefficient was 1.016 (95% HC3 CI -1.347 to 3.379;  $P=0.399$ ). Cluster SC80\_0026 contained 30 entities, 29 with at least two cysteines; after its removal, the cysteine-fraction coefficient was -4.197 (95% CI -9.296 to 0.903;  $P=0.107$ ). Thus, both findings were sensitive to family concentration.

Regression diagnostics showed heteroscedasticity (Breusch-Pagan  $P<0.001$ ; White-test  $P<0.001$ ) and non-normal residual tails (Jarque-Bera  $P<0.001$ ), supporting robust rather

**Table 2** Observation-level resistance labels and assay-metadata availability

| Domain           | Category or field           | Available n/N (%) | Interpretive consequence                          |
|------------------|-----------------------------|-------------------|---------------------------------------------------|
| Resistance label | Carbapenem-resistant (CRAB) | 19 (2.3%)         | Only 14 chemical entities; no CRAB-specific model |
| Resistance label | Multidrug-resistant (MDR)   | 171 (21.1%)       | Descriptive only                                  |
| Resistance label | Other resistance reported   | 41 (5.1%)         | Descriptive only                                  |
| Resistance label | Resistance not reported     | 578 (71.4%)       | No explicit susceptible comparison group          |
| Assay metadata   | Named strain label          | 320 (39.6%)       | 45 distinct reported labels                       |
| Assay metadata   | Growth medium               | 806 (99.6%)       | MHB 570; CAMHB 115; other media 121               |
| Assay metadata   | Inoculum (CFU)              | 760 (93.9%)       | Values and ranges were heterogeneous              |
| Assay metadata   | pH                          | 4 (0.5%)          | Insufficient for adjustment                       |
| Assay metadata   | Ionic strength              | 36 (4.4%)         | Insufficient for adjustment                       |
| Assay metadata   | Salt type                   | 0 (0.0%)          | Not reported                                      |
| Assay metadata   | Incubation time             | 0 (0.0%)          | Not reported                                      |
| Assay metadata   | Temperature                 | 0 (0.0%)          | Not reported                                      |

*Note. Resistance categories were mutually exclusive as recorded. No category explicitly identified susceptible isolates. Abbreviations: CAMHB, cation-adjusted Mueller-Hinton broth; CFU, colony-forming units; CRAB, carbapenem-resistant Acinetobacter baumannii; MDR, multidrug-resistant; MHB, Mueller-Hinton broth.*

**Table 3** Adjusted linear associations with entity-level median log2 MIC in the development set

| Feature                          | Standardized $\beta$ | 95% HC3 CI       | HC3 $P$ value | 95% 80%-cluster-robust CI | Cluster-robust $P$ value |
|----------------------------------|----------------------|------------------|---------------|---------------------------|--------------------------|
| Sequence length                  | 0.083                | -0.123 to 0.288  | 0.428         | -0.295 to 0.461           | 0.666                    |
| Net charge at pH 7.4             | -0.305               | -0.515 to -0.095 | 0.004         | -0.659 to 0.049           | 0.091                    |
| Mean hydrophobicity              | -0.206               | -0.383 to -0.028 | 0.023         | -0.488 to 0.077           | 0.153                    |
| Alpha-helical hydrophobic moment | -0.078               | -0.203 to 0.046  | 0.216         | -0.265 to 0.108           | 0.408                    |
| Beta-strand hydrophobic moment   | -0.141               | -0.266 to -0.015 | 0.028         | -0.328 to 0.047           | 0.141                    |
| Aliphatic index                  | 0.141                | 0.010 to 0.271   | 0.035         | -0.058 to 0.340           | 0.165                    |
| Normalized sequence entropy      | 0.019                | -0.110 to 0.148  | 0.768         | -0.168 to 0.206           | 0.838                    |
| Cysteine fraction                | -0.469               | -0.617 to -0.320 | <0.001        | -0.722 to -0.215          | <0.001                   |
| Glycine fraction                 | -0.015               | -0.109 to 0.079  | 0.756         | -0.136 to 0.106           | 0.809                    |
| Proline fraction                 | -0.005               | -0.091 to 0.081  | 0.909         | -0.122 to 0.112           | 0.933                    |
| N-terminal acetylation           | 0.083                | 0.034 to 0.131   | <0.001        | 0.039 to 0.126            | <0.001                   |
| C-terminal amidation             | -0.127               | -0.236 to -0.018 | 0.023         | -0.268 to 0.014           | 0.078                    |

*Note. HC3 and 80%-cluster-robust estimates address different error structures but do not remove unmeasured confounding.  $P$  values are nominal and exploratory.*

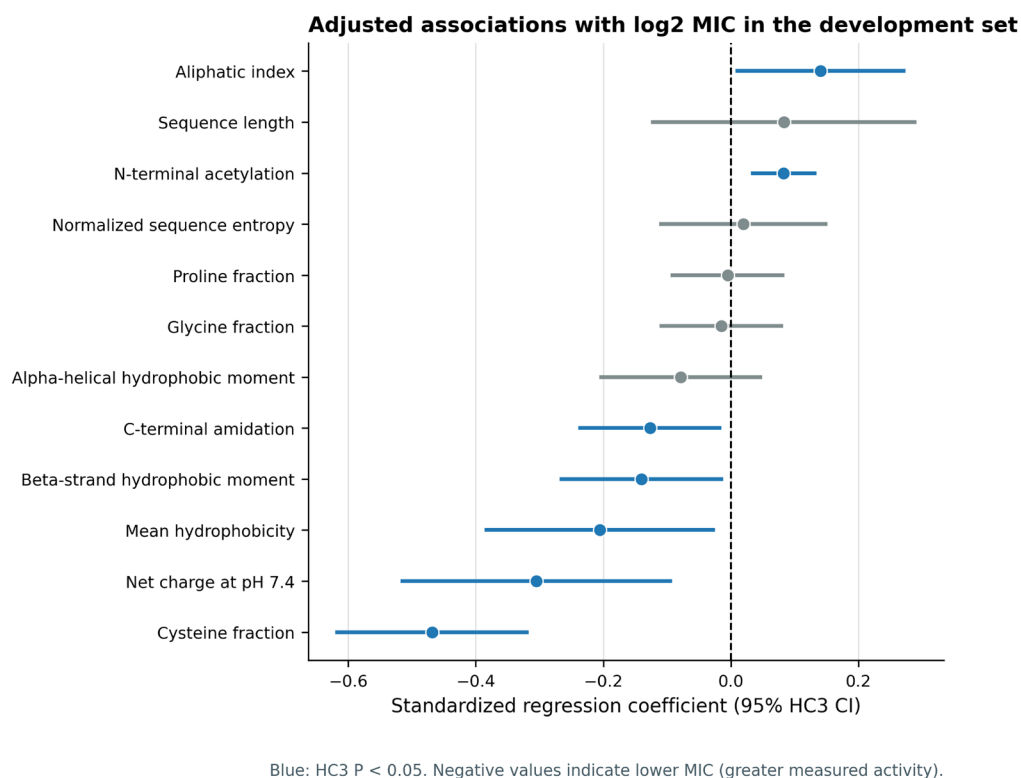


**Figure 2** – Adjusted linear associations with entity-level median log2 MIC in the development set

Points are standardized coefficients and horizontal lines are 95% HC3 confidence intervals. Blue coefficients have nominal HC3  $P < 0.05$ . Negative coefficients indicate lower MIC and therefore greater measured activity. The figure does not display the wider cluster-robust intervals reported in Table 3. Abbreviations: CI, confidence interval; HC3, heteroscedasticity-consistent covariance estimator type 3; MIC, minimum inhibitory concentration.

**Table 4** Nested grouped cross-validation performance in the development set

| Model           | Fold RMSE, mean $\pm$ SD | Fold MAE, mean $\pm$ SD | Fold $R^2$ , mean $\pm$ SD | Pooled OOF RMSE | Pooled OOF $R^2$ | Pooled OOF Spearman rho |
|-----------------|--------------------------|-------------------------|----------------------------|-----------------|------------------|-------------------------|
| Median baseline | 2.236 $\pm$ 0.768        | 1.753 $\pm$ 0.628       | -0.003 $\pm$ 0.003         | 2.369           | 0.002            | -0.014                  |
| Elastic Net     | 2.294 $\pm$ 0.763        | 1.794 $\pm$ 0.600       | -0.065 $\pm$ 0.137         | 2.420           | -0.042           | 0.143                   |
| Random Forest   | 2.180 $\pm$ 0.801        | 1.723 $\pm$ 0.674       | 0.054 $\pm$ 0.082          | 2.326           | 0.038            | 0.304                   |
| Extra Trees     | 2.174 $\pm$ 0.767        | 1.710 $\pm$ 0.619       | 0.053 $\pm$ 0.128          | 2.306           | 0.054            | 0.301                   |

Note. All folds were grouped by 80%-identity cluster. Small model differences do not support stable superiority.

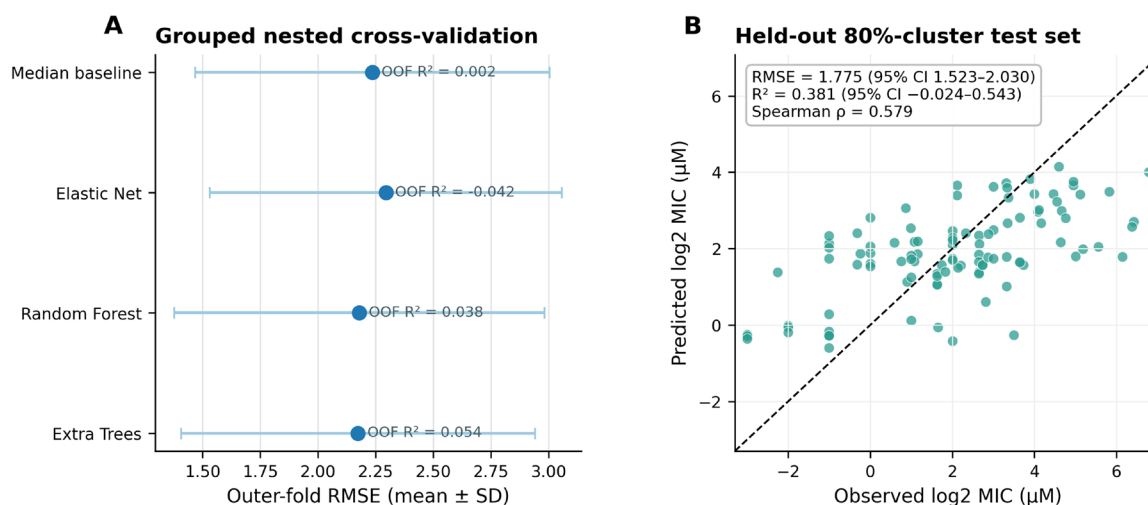


**Figure 3** – Homology-aware model performance

(A) Mean and standard deviation of RMSE across five outer grouped folds, with pooled out-of-fold  $R^2$  shown for each model. (B) Observed versus Extra Trees-predicted log2 MIC in the held-out 80%-identity-cluster test set. The dashed line denotes perfect agreement, and the annotation reports the cluster-bootstrap confidence interval for test  $R^2$ . Abbreviations: MIC, minimum inhibitory concentration; OOF, out-of-fold; RMSE, root-mean-square error.

Table 5

Locked Extra Trees performance in the untouched held-out test set

| Metric         | Estimate | 95% cluster-bootstrap CI |
|----------------|----------|--------------------------|
| RMSE           | 1.775    | 1.523 to 2.030           |
| MAE            | 1.442    | 1.204 to 1.692           |
| R <sup>2</sup> | 0.381    | -0.024 to 0.543          |
| Spearman rho   | 0.579    | 0.252 to 0.748           |
| Pearson r      | 0.632    | 0.288 to 0.772           |

Note. Confidence intervals use 1,000 intact-cluster bootstrap samples; the R<sup>2</sup> confidence interval includes zero.

than conventional standard errors. Twenty-two development entities exceeded Cook's distance 4/n. Held-out regression performance was RMSE 2.135, MAE 1.675, R<sup>2</sup> 0.105, and Spearman rho 0.300. Residual plots are shown in Supplementary Figure S1.

### Grouped model comparison and held-out performance

Nested grouped cross-validation showed negligible separation among learned models and the median baseline (Table 4; Figure 3A). The locked primary results gave mean outer-fold RMSE values of 2.174 for Extra Trees, 2.180 for Random Forest, 2.236 for the baseline, and 2.294 for Elastic Net. Pooled out-of-fold R<sup>2</sup> was 0.054 for Extra Trees and 0.038 for Random Forest. These values indicate that the learned models explained little variance when entire 80%-identity clusters were unseen during development.

Extra Trees was retained as the locked test model under the original pretest RMSE rule. On the 104-entity test set, RMSE was 1.775 log<sub>2</sub> MIC units (95% cluster-bootstrap CI 1.523-2.030), MAE was 1.442 (1.204-1.692), R<sup>2</sup> was 0.381 (-0.024 to 0.543), Spearman rho was 0.579 (0.252-0.748), and Pearson r

was 0.632 (0.288-0.772) (Table 5; Figure 3B). The R<sup>2</sup> interval included zero, so the point estimate from this single cluster composition does not establish reliable generalization.

In the fully specified repeat audit, Random Forest and Extra Trees remained effectively tied: mean outer-fold RMSE was 2.190 and 2.194, while pooled out-of-fold R<sup>2</sup> was 0.033 and 0.042, respectively. The 0.004 difference in mean RMSE selected Random Forest in that audit, whereas Extra Trees had slightly lower pooled RMSE. This instability reinforces that no learned model showed a reproducible performance advantage (Supplementary Table S17).

### Held-out model attributions

Cysteine fraction had the largest test-set mean absolute SHAP value (0.543), followed by sequence length (0.296), net charge (0.249), C-terminal amidation (0.134), aliphatic index (0.125), and beta-strand hydrophobic moment (0.112) (Figure 4). Permutation importance similarly ranked cysteine fraction first. These rankings characterize the fitted Extra Trees model and do not distinguish direct chemistry from correlated family or structural-class information.

The SHAP distribution and partial-dependence functions showed nonlinear fitted patterns (Figures 4 and 5). Higher cysteine fractions and greater charge generally shifted predictions toward lower MIC, and C-terminal amidation shifted predictions toward greater predicted activity. Sequence length showed a curved model relationship despite a non-significant adjusted linear coefficient. The plots describe conditional model behavior over the observed data range and do not establish causal or universally transferable effects.

### Data-quality, assay, and descriptor sensitivity

Held-out performance was similar in the exact-MIC analysis (RMSE 1.758; R<sup>2</sup> 0.393), after excluding structurally unverified entities (RMSE 1.804; R<sup>2</sup> 0.380), and after excluding

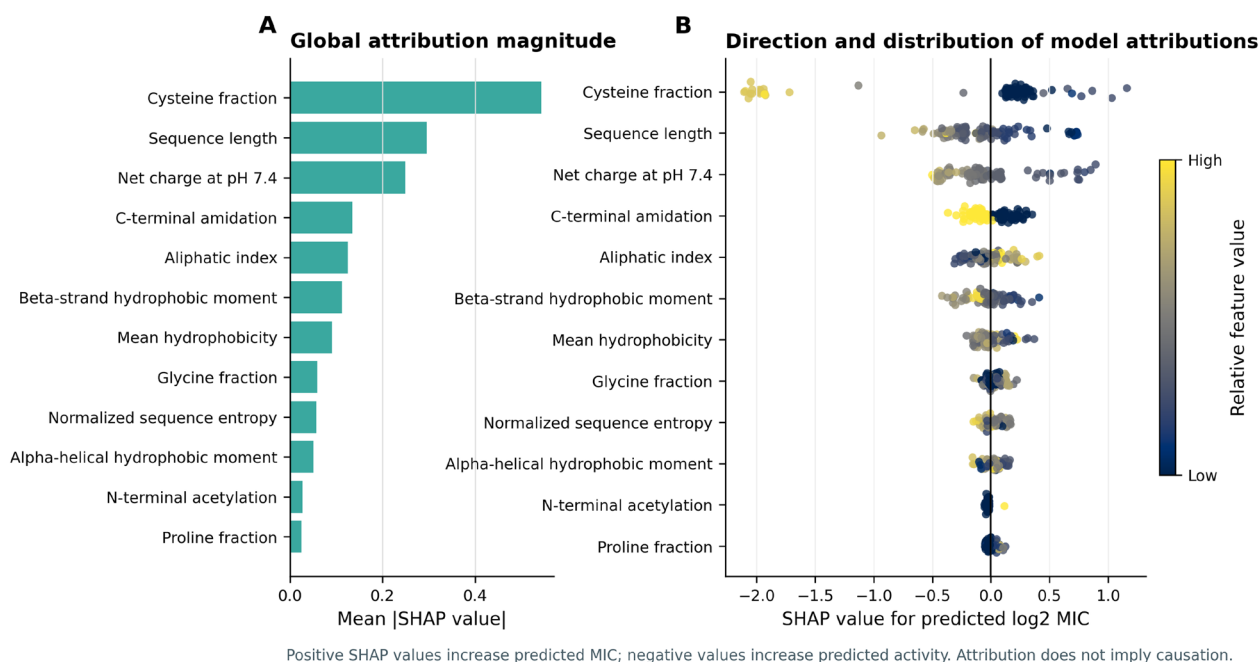
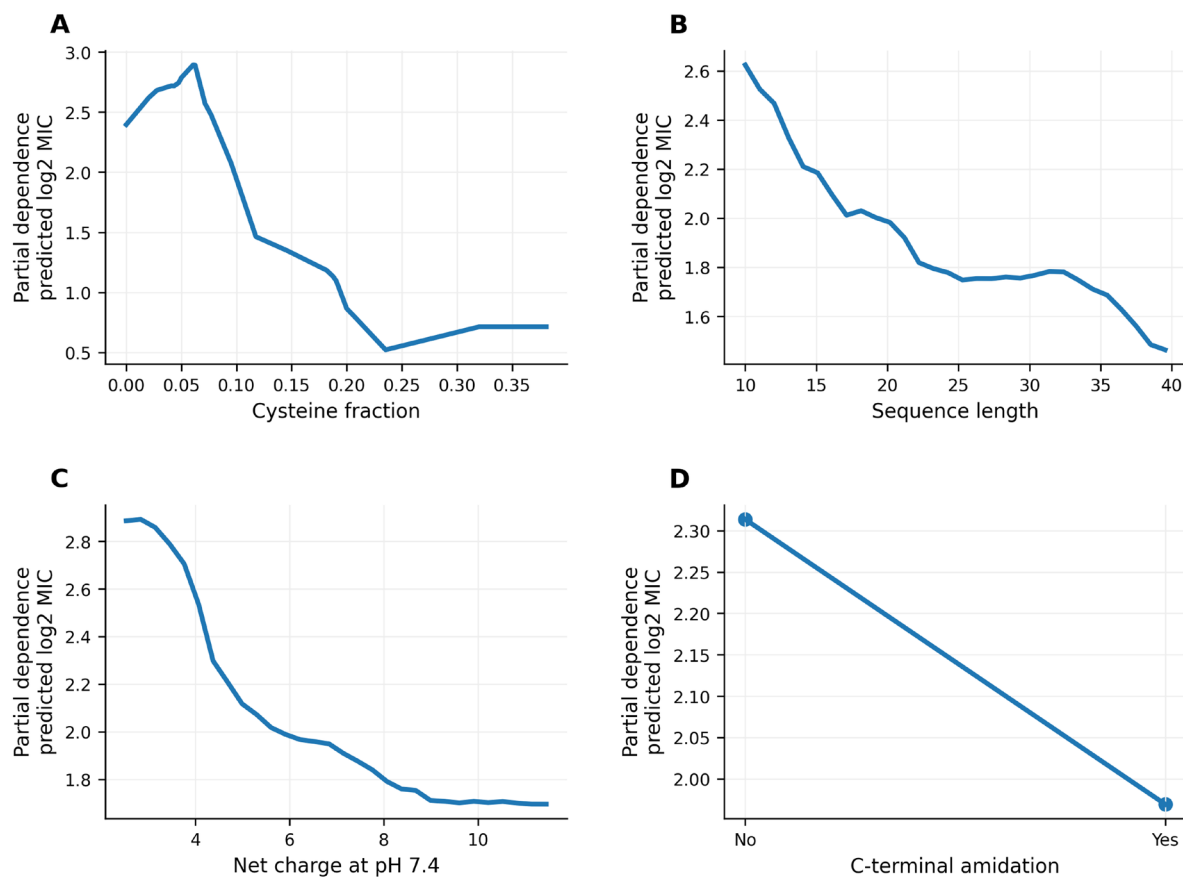


Figure 4 – Held-out model attributions for the locked Extra Trees model

(A) Mean absolute SHAP values rank global attribution magnitude. (B) SHAP distributions show direction and heterogeneity of attributions; the color scale uses monotonically varying luminance to remain interpretable in grayscale. Positive values increase predicted MIC, whereas negative values increase predicted activity. Attribution does not imply causation. Abbreviations: MIC, minimum inhibitory concentration; SHAP, Shapley additive explanations.



Curves describe the fitted model over observed feature ranges and do not establish causal effects.

**Figure 5** – Partial-dependence functions for the four leading interpretable features

Panels show model-averaged predicted log2 MIC across cysteine fraction, sequence length, net charge at pH 7.4, and C-terminal amidation. Curves describe fitted model behavior over observed feature ranges and do not establish causal effects. Abbreviation: MIC, minimum inhibitory concentration.

extreme entity medians (RMSE 1.806;  $R^2$  0.359). Restriction to values originally reported in micromolar units reduced the sample to 219 development and 71 test entities and yielded RMSE 2.100 and  $R^2$  0.058; the exact-and-original-micromolar subset yielded RMSE 1.900 and  $R^2$  0.229 (Supplementary Table S3; Supplementary Figure S3). These results show sensitivity to unit-source composition and do not quantify salt-form or hydration-state error directly.

The MHB/CAMHB-only analysis yielded RMSE 1.817 and  $R^2$  0.368 (354 development, 89 test entities). Restricting further to exact inocula of  $5 \times 10^5$  or  $1 \times 10^6$  CFU yielded RMSE 1.707 and  $R^2$  0.433 (193 development, 51 test). In contrast, the named-strain-only subset was small (176 development, 31 test) and performed poorly (RMSE 1.974;  $R^2$  -0.177; Spearman rho 0.116). These restrictions do not remove laboratory or strain confounding, but they demonstrate that conclusions depend on which assay metadata are available (Supplementary Table S14; Supplementary Figure S3).

Alternative-descriptor models did not identify a single specification that resolved the primary uncertainty. Adding Boman index to mean hydrophobicity produced severe collinearity (maximum VIF 22.17) and a Boman coefficient near zero ( $P=0.961$ ); replacing hydrophobicity with Boman index produced similar adjusted  $R^2$  (0.258) and a positive Boman coefficient ( $P=0.030$ ). Adding charge density produced maximum VIF 11.53 and  $P=0.724$ ; replacing charge and length reduced adjusted  $R^2$  to 0.231 and gave  $P=0.065$ . Modified-

termini isoelectric point was not significant in complete cases ( $P=0.247$ ) or after median imputation ( $P=0.386$ ), while the missingness indicator was borderline ( $P=0.066$ ) (Supplementary Table S13).

#### Homology threshold, alignment definition, and similarity domain

With fixed Extra Trees settings, grouped out-of-fold  $R^2$  was -0.148 at 70% identity, -0.107 at 80%, and 0.261 at 90%, with RMSE 2.540, 2.495, and 2.038, respectively (Supplementary Table S4; Supplementary Figure S2). Lower identity thresholds separate more remote families and produced worse transfer. Needleman-Wunsch and normalized-edit 80% clustering generated the same 212 development clusters (adjusted Rand index 1.000), no cross-boundary pair reached 80% Needleman-Wunsch identity, and the maximum was 0.775 (Supplementary Table S18).

No test entity had  $\geq 80\%$  identity to a development sequence by design. Mean absolute error was 1.634 for the 50 entities with nearest-development identity  $< 50\%$ , 1.419 for 31 entities at  $50\% - < 70\%$ , and 1.058 for 23 entities at  $70\% - < 80\%$  (Supplementary Table S15; Supplementary Figure S4). This descriptive gradient supports similarity-aware uncertainty reporting, but it does not establish performance for close analogues above 80% or for entirely novel structural classes.

Across the five outer folds, cysteine fraction and net charge appeared in the top three SHAP features in all folds, and

sequence length did so in four folds (Supplementary Table S5). Stable rank does not remove confounding by family composition or weak out-of-fold predictive performance.

## Discussion

This study separates adjusted database associations from predictive transfer under sequence-cluster separation. Several sequence and physicochemical variables were associated with log2 MIC in the primary HC3 model, but cluster-robust and family-exclusion analyses showed that inferential precision depended on sequence-family composition. Nested grouped cross-validation explained little out-of-fold variance, and the more favorable test point estimate had a confidence interval that included zero. The results therefore support hypothesis generation, not a validated stand-alone screening or prioritization instrument.

The inverse HC3 associations of positive charge, mean hydrophobicity, and beta-strand hydrophobic moment are compatible with established membrane-interaction hypotheses [6-8], but compatibility with mechanism is not evidence that these variables cause activity in this dataset. Peptide family, secondary structure, resistance phenotype, and assay conditions were incompletely observed and may jointly influence both descriptors and MIC. The wider 80%-cluster-robust intervals for these terms illustrate this dependence.

Cysteine fraction dominated both the linear and Extra Trees descriptions. Cysteine can enable disulfide-constrained scaffolds, but disulfide connectivity, redox state, and three-dimensional topology were not encoded. The concentration of 29 cysteine-rich entities in one 30-entity cluster and loss of statistical precision after removing that cluster show that cysteine fraction partly represented family or structural-class membership. Its prominence identifies a family-dependent hypothesis for prospective testing rather than a universal chemical rule.

Terminal chemistry was also family-sensitive. C-terminal amidation had a negative HC3 association but a cluster-robust P value of 0.078. N-terminal acetylation was rare and concentrated: 11 of 13 acetylated development entities belonged to one cluster, and the association was not precise after that cluster was removed. The positive primary coefficient should therefore not be interpreted as evidence that acetylation intrinsically reduces activity.

The alternative-descriptor analyses clarify why Boman index, charge density, and isoelectric point were not simply added to the main model. Boman index contained information when substituted for hydrophobicity, but adding both produced severe collinearity and no independent Boman contribution. Charge density added little when charge and length were retained, and replacing both reduced fit. Isoelectric point analyses were limited by 41.5% missingness and did not show an independent association; the borderline missingness indicator suggests that missingness may encode modification patterns or reporting practices. These results make the specification choice transparent but do not privilege one descriptor as causally superior.

Resistance and assay metadata restrict the biological scope. Only 19 observations were explicitly CRAB, most records lacked a resistance label, and no explicit susceptible category existed. The study therefore concerns *A. baumannii* activity broadly, not carbapenem-resistant activity specifically. Medium and inoculum restrictions were broadly similar to the primary result, whereas the named-strain subset performed poorly and

was small. Absence of temperature, incubation-time, pH, ionic-strength, and salt metadata prevented systematic adjustment, and residual laboratory and strain heterogeneity remains a major source of confounding.

The validation results are more informative than the single test point estimate alone. Exact-sequence separation would have allowed 304 cross-boundary pairs at 80% identity, whereas cluster locking removed that leakage route. Under the primary nested procedure, Extra Trees improved mean RMSE by only 0.062 log2 units relative to the median baseline and had pooled out-of-fold  $R^2$  of 0.054. The fully specified repeat audit again found a near-tie between tree models and weak pooled  $R^2$ . The test  $R^2$  of 0.381 therefore should be viewed as one uncertain realization over 53 unequal clusters, not as stable evidence of generalization.

Homology sensitivity further narrows the domain of applicability. Performance was poorest under the 70% grouping, improved under the less stringent 90% grouping, and absolute error decreased as test sequences approached 80% identity to development. Because all  $\geq 80\%$  analogues were intentionally excluded from the test set, the analysis cannot estimate performance for close analogues above that threshold. Predictions for remote families, unusual chemistries, cyclic topologies, or features outside the observed range should be treated as unvalidated hypotheses.

Recent protein-language-model and few-shot approaches have coupled computational screening with experimental validation against *A. baumannii* [13]. The present study addresses a narrower question: how much transferable information is available from 12 conventional descriptors once close sequence relatives are excluded? The answer is limited. Future work may combine sequence embeddings, explicit structural representations, standardized assay covariates, and family-aware external validation, but model development should preserve cluster-disjoint evaluation and avoid post hoc test-set selection.

Strengths include a species-specific continuous outcome, conservative entity aggregation, a locked cluster-disjoint test set, nested grouped model comparison, cluster-bootstrap uncertainty, cluster-robust regression, descriptor and alignment sensitivities, assay-restricted analyses, and held-out model attribution. The submission package includes entity- and observation-level audit trails, scripts, outputs, and publication-quality figures.

Limitations include exclusion of 489 range, censored, or complex records with potentially informative censoring; incomplete study identifiers; uncertain salt and hydration states for mass-to-molar conversion; sparse resistance and assay metadata; unresolved structural details for six cyclic or modified entities; unmodeled disulfide connectivity; and family concentration of rare modifications. Toxicity, hemolysis, serum stability, pharmacokinetics, and resistance emergence were not outcomes. The structural-exclusion, original-unit, extreme-value, assay, descriptor, and family analyses bound some sensitivities but cannot replace source-level verification or prospective experiments.

Any computational use should therefore be embedded in a broader experimental workflow. Similarity to the development set, feature-range checks, uncertainty reporting, toxicity and developability filters, and standardized broth-microdilution testing are required before candidate advancement. The current model should not independently rank or select peptides for development. Prospective studies should use predefined panels of genetically and phenotypically diverse *A. baumannii* isolates,

blinded candidate selection, standardized assay conditions, and external replication.

## Conclusion

In this heterogeneous database, net charge, hydrophobicity, beta-strand amphipathicity, cysteine content, aliphatic composition, and terminal chemistry showed adjusted associations with *A. baumannii* MIC, but several estimates were sensitive to sequence-family clustering and descriptor specification. Homology-aware cross-validation showed weak out-of-fold transfer, and the test-set  $R^2$  estimate was uncertain. Sparse resistance labels preclude CRAB-specific conclusions, while incomplete assay and structural metadata limit biological interpretation. The findings generate testable hypotheses and define sources of model uncertainty; they do not establish causal feature effects or validate stand-alone computational prioritization. Prospective standardized susceptibility, toxicity, stability, and external-validation studies are required.

## Supplementary materials

The Supplementary information includes:

- Supplementary Table S1. Prespecified feature definitions and variance inflation factors;
- Supplementary Table S2. Regression diagnostics;
- Supplementary Table S3. Predictive sensitivity analyses;
- Supplementary Table S4. Homology-threshold sensitivity in grouped out-of-fold prediction;
- Supplementary Table S5. Stability of out-of-fold SHAP ranking across five grouped outer folds;
- Supplementary Table S6. Robustness of key HC3 associations across sample restrictions;
- Supplementary Table S7. Entities requiring manual structural verification;
- Supplementary Table S8. Extreme entity-median MIC values used in the exclusion sensitivity;
- Supplementary Table S9. DOME-aligned machine-learning reporting checklist;
- Supplementary Table S10. Software environment and locked model settings;
- Supplementary Table S11. Detailed resistance and assay-metadata audit;
- Supplementary Table S12. Cluster-robust and dominant-family sensitivity of regression associations;
- Supplementary Table S13. Alternative descriptor specifications;
- Supplementary Table S14. Assay-metadata-restricted predictive analyses;
- Supplementary Table S15. Absolute prediction error by nearest-development-sequence identity;

- Supplementary Table S16. Fully specified hyperparameter search space for the repeat audit;
- Supplementary Table S17. Nested grouped cross-validation in the fully specified repeat audit;
- Supplementary Table S18. Needleman-Wunsch alignment sensitivity;
- Supplementary Figure S1. Development-set regression diagnostics;
- Supplementary Figure S2. Homology-threshold and alignment-definition sensitivity;
- Supplementary Figure S3. Held-out predictive performance across data-quality and assay-metadata restrictions;
- Supplementary Figure S4. Descriptive relationship between nearest-development-sequence identity and absolute test error.

These supplementary materials have been provided by the authors to give readers additional information about their work.

The file can be accessed using: [https://www.editorialpark.com/download/article-supp/795/05\\_Supplementary\\_Material\(2\).docx](https://www.editorialpark.com/download/article-supp/795/05_Supplementary_Material(2).docx).

**Author Contributions:** Conceptualization, B.S. and A.O.; methodology, B.S., Y.S. and N.S.; validation, Y.S. and N.S.; formal analysis, B.S. and Y.S.; investigation, B.S.; resources, A.B.; data curation, B.S.; writing – original draft preparation, B.S.; writing – review and editing, Y.S., A.B., N.S. and A.O.; visualization, B.S. and Y.S.; supervision, A.B. and A.O.; project administration, A.B. and A.O.; funding acquisition, not applicable. All authors have read and agreed to the submitted version of the manuscript.

**Disclosures:** The authors have no conflicts of interest.

**Acknowledgments:** None.

**Funding:** None.

**Data availability statement:** The analysis-ready data, executable code, sensitivity outputs, and fitted model are available at Zenodo: <https://doi.org/10.5281/zenodo.22170145>.

**Artificial Intelligence (AI) Disclosure Statement:** During preparation of this work, the authors used OpenAI ChatGPT to assist with manuscript organization, language refinement, and drafting and debugging of analytical code. All methodological decisions, statistical analyses, references, numerical results, and interpretations were independently reviewed and verified by the authors. The figures were generated deterministically from analysis outputs using Matplotlib; no generative image model was used. The authors take full responsibility for the content.

## References

1. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*. 2022;399(10325):629-655. [https://doi.org/10.1016/S0140-6736\(21\)02724-0](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*. 2024;404(10459):1199-1226. [https://doi.org/10.1016/S0140-6736\(24\)01867-1](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. Geneva: World Health Organization; 2024. ISBN 978-92-4-009346-1.
4. Harding CM, Hennon SW, Feldman MF. Uncovering the mechanisms of *Acinetobacter baumannii* virulence. *Nat Rev Microbiol*. 2018;16(2):91-102. <https://doi.org/10.1038/nrmicro.2017.148>
5. Peleg AY, Seifert H, Paterson DL. *Acinetobacter baumannii*: emergence of a successful pathogen. *Clin Microbiol Rev*. 2008;21(3):538-582. <https://doi.org/10.1128/CMR.00058-07>
6. Mookherjee N, Anderson MA, Haagsman HP, Davidson DJ. Antimicrobial host defence peptides: functions and clinical potential. *Nat Rev Drug Discov*. 2020;19(5):311-332. <https://doi.org/10.1038/s41573-019-0058-8>
7. Hancock REW, Sahl HG. Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nat Biotechnol*. 2006;24(12):1551-1557. <https://doi.org/10.1038/nbt1267>
8. Fjell CD, Hiss JA, Hancock REW, Schneider G. Designing antimicrobial peptides: form follows function. *Nat Rev Drug Discov*. 2012;11(1):37-51. <https://doi.org/10.1038/nrd3591>
9. Pirtskhalava M, Armstrong AA, Grigolava M, Chubinidze M, Alimbarashvili E, Vishnepolsky B, Gabrielian A, Rosenthal A, Hurt DE, Tartakovsky M. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res*. 2021;49(D1):D288-D297. <https://doi.org/10.1093/nar/gkaa991>
10. Anwer F, Navid A, Faiz F, Haider U, Nasir S, Farooq M, Zahra M, Bano A, Bashir HH, Ahmad M, Abbas SA, Room SE, Saeed MT, Ali A. AbAMPdb: a database of *Acinetobacter baumannii* specific antimicrobial peptides. *Database (Oxford)*. 2024;2024:baae096. <https://doi.org/10.1093/database/baae096>
11. Mant CT, Jiang Z, Gera L, Davis T, Nelson KL, Bevers S, Hodges RS. De novo designed amphipathic alpha-helical antimicrobial peptides incorporating Dab and Dap residues on the polar face to treat the Gram-negative pathogen *Acinetobacter baumannii*. *J Med Chem*. 2019;62(7):3354-3366. <https://doi.org/10.1021/acs.jmedchem.8b01785>
12. Chen SP, Chen EHL, Yang SY, Kuo PS, Jan HM, Yang TC, Hsieh MY, Lee KT, Lin CH, Chen RPY. A systematic study of the stability, safety, and efficacy of the de novo designed antimicrobial peptide PepD2 and its modified derivatives against *Acinetobacter baumannii*. *Front Microbiol*. 2021;12:678330. <https://doi.org/10.3389/fmicb.2021.678330>
13. Huang J, Zhang W, Wang A, Jiang Y, Lai Y, Xu Y, Wang C, Zhao J, Zhang P, Ji J. Discovery of antimicrobial peptides targeting *Acinetobacter baumannii* via a pre-trained and fine-tuned few-shot learning-based pipeline. *Nat Commun*. 2026;17:2475. <https://doi.org/10.1038/s41467-026-69306-2>
14. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, ELIXIR Machine Learning Focus Group, Harrow J, Psomopoulos FE, Tosatto SCE. DOME: recommendations for supervised machine learning validation in biology. *Nat Methods*. 2021;18(10):1122-1127. <https://doi.org/10.1038/s41592-021-01205-4>
15. Sidorczuk K, Gagat P, Pietluch F, Kala J, Rafacz D, Bakala L, Slowik J, Kolenda R, Rodiger S, Fingerhut LCHW, Cooke IR, Mackiewicz P, Burdukiewicz M. Benchmarks in antimicrobial peptide prediction are biased due to the selection of negative data. *Brief Bioinform*. 2022;23(5):bbac343. <https://doi.org/10.1093/bib/bbac343>
16. Eisenberg D, Weiss RM, Terwilliger TC. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*. 1982;299(5881):371-374. <https://doi.org/10.1038/299371a0>
17. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. 2000;16(6):276-277. [https://doi.org/10.1016/S0168-9525\(00\)00204-2](https://doi.org/10.1016/S0168-9525(00)00204-2)
18. Ikai A. Thermostability and aliphatic index of globular proteins. *J Biochem*. 1980;88(6):1895-1898. <https://doi.org/10.1093/oxfordjournals.jbchem.a133168>
19. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27:379-423, 623-656.
20. MacKinnon JG, White H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J Econom*. 1985;29(3):305-325. [https://doi.org/10.1016/0304-4076\(85\)90158-7](https://doi.org/10.1016/0304-4076(85)90158-7)
21. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67(2):301-320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>
22. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32. <https://doi.org/10.1023/A:1010933404324>
23. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;63:3-42. <https://doi.org/10.1007/s10994-006-6226-1>
24. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7:91. <https://doi.org/10.1186/1471-2105-7-91>
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
26. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765-4774.
27. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat*. 1979;7(1):1-26. <https://doi.org/10.1214/aos/1176344552>