Supporting Information : Mechanistic Principles of Antimicrobial Peptides Uncovered by Charge Density Based Machine Learning

Hrushikesh Malshikare,[†,¶] U. Deva Priyakumar,[‡] Prathit Chatterjee,[*,‡] and Durba Sengupta[*,†,¶]

[†]*Physical and Materials Chemistry Division, CSIR-National Chemical Laboratory, Dr. Homi Bhabha Road, Pune 411008, India*

[‡]*Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad 500032, India*

[¶]*Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India*

E-mail: prathit.chatterjee@iiit.ac.in; durba.sengupta@ncl.res.in

Supplementary Methods

Data Collection and Curation

Our research began with a rigorous collection and curation process to create a specialized dataset focused on antimicrobial peptides (AMPs). We organized the dataset into three distinct groups based on peptide charge density, with the goal of exploring the impact of specific biochemical properties on antimicrobial activity and identifying potential variations associated with peptide charge density. Charge density was calculated as the ratio of net charge to peptide length, where length refers to the number of amino acids. Based on the computed charge density values, peptides were divided into three categories: low charge density (-0.3 to 0.1), moderate charge density (0.1 to 0.25), and high charge density (0.25 to 0.75) as depicted in supplementary figure 2c. These divisions allowed us to systematically study the impact of increasing electrostatic character on the peptides, physicochemical properties and predictive features.

For each of the three AMP datasets, an equal number of non-AMP sequences were selected with matching charge density distributions, respectively. This ensured that comparisons were not biased by inherent disparities between positive and negative samples. Balanced datasets were thus created for each class, allowing for robust machine learning based classification and feature analysis across different peptide property regimes. The complete machine learning pipeline is illustrated in figure 1a.

Positive Dataset Creation:

To create our model, we compiled antimicrobial peptides (AMPs) from multiple publicly available databases, including dbAMP 2.0,[1] the Data Repository of Antimicrobial Peptides (DRAMP),[2] and the Database of Antimicrobial Activity and Structure of Peptides (DBAASP v3).[3] To ensure high-quality and non-redundant data, duplicate sequences across databases were removed. We retained only sequences composed entirely of standard amino acids by excluding those containing non-standard residues, and only peptides with experimentally validated activity data were included. For dbAMP, sequences labeled as "Validated" and containing target activity values (e.g., MIC or IC_{50}) below 128 μM were selected using regular expressions to extract numeric concentration values. For DRAMP, peptides were included only if cytotoxicity data were available and the IC_{50}/CC_{50} values were less than or equal to 200 μM. Sequences labeled as non-cytotoxic were excluded. For DBAASP, peptides were filtered based on standard activity measure types such as MIC, IC_{50}, EC_{50}, MIC_{90}, LC_{50}, and MIC_{50}. For the DBAASP dataset specifically, peptides were further filtered using predefined activity thresholds for each experimental measure type (e.g., MIC < 125 μM, $IC_{50} < 125$ μM, $EC_{50} < 50$ μM, $MIC_{90} < 100$ μM, $LC_{50} < 40$ μM, 50% cell death < 80 μM, $MIC_{50} < 110$ μM). Each activity type was processed separately, retaining only sequences meeting its respective cutoff, and the filtered subsets were then combined to generate a high-confidence set of experimentally active peptides.

Negative Dataset Extraction:

In the absence of a comprehensive public repository of experimentally validated non-antibacterial peptides (non-AMPs), we constructed a negative dataset by querying the UniProt database.[4] To avoid potential antimicrobial activity, entries containing any of the following keywords were excluded: antifungal, secretory, excreted, effector, virucidal, antiseptic, microbicidal, biocidal, anti-fungal, anti-biotic, anti-microbial, etc. To ensure a fair comparison with the AMP dataset, non-AMP sequences were also categorized based on charge density into three groups: high (0.25-0.75), moderate (0.10-0.25), and low (-0.30-0.10), using the same thresholds defined for AMPs.

Feature Extraction

1. Structural Features: We used ESMFold[5] to predict 3D structures for all peptide sequences and generated PDB files. Using Bio.PDB's[6] PDBParser and DSSP, we computed structural descriptors including mean relative solvent accessibility (Mean_RSA), backbone dihedral angles (Mean_Phi, Mean_Psi), and secondary structure content-percentages of helices, sheets, and coils. Additional structure-based descriptors such as radius of gyration (RoG), solvent-accessible surface area (SASA), and compactness were obtained using mdtraj.

2. Physicochemical Properties: Using the `peptides` package,[7] we calculated 12 key physicochemical descriptors such as residue frequencies, aliphatic index, Boman index, net charge, isoelectric point, and hydrophobic moment. We also included Kidera factors, which are ten principal components derived from 188 physicochemical and biochemical properties of amino acids, providing a compact yet rich representation of sequence behavior.

3. Cheminformatic Descriptors: With RDKit,[8] cheminformatic features were extracted including topological polar surface area (TPSA), LogP (a measure of lipophilicity), and heavy atom countcommonly used descriptors in molecular modeling and drug discovery.

4. Sequence-Based Pseudo Composition: We utilized the PyBioMed[9] library, pseudoAAC module to calculate Amphiphilic Pseudo Amino Acid Composition (APPACC) features, which incorporate both amino acid composition and sequence-order information, capturing local and global sequence characteristics relevant to biological activity.

Each peptide sequence was ultimately represented by a total of 57 features across these four categories as depicted in figure 1b.

Model Training and Evaluation

In this study, we utilized the open-source machine learning library *Scikit-learn* to develop predictive models for antimicrobial peptides. Several supervised classification algorithms were explored, including Random Forest (RF),[10] XGB,[11] Decision Tree (DT),[12] Naive Bayes (NB),[13]Catboost[14] . To ensure reliable performance estimation and reduce the risk of overfitting, we implemented ten-fold cross-validation. Model evaluation was conducted using multiple metrics, including precision, sensitivity (recall), and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics provide a comprehensive assessment of both the model's discriminative ability and its robustness in classifying AMPs versus non-AMPs.

Accuracy measures the overall proportion of correctly classified samples among all samples. It is defined as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. Accuracy provides a straightforward measure of model performance.

Precision measures the proportion of correctly predicted positive samples among all samples predicted as positive. It is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where TP is the number of true positives and FP is the number of false positives.

Sensitivity (Recall) measures the proportion of actual positive samples that are correctly identified by the model. It is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where FN is the number of false negatives.

Area Under the Receiver Operating Characteristic Curve (AUC-ROC) quantifies the overall ability of the model to discriminate between positive and negative classes across all classification thresholds. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR), defined as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

The AUC score ranges from 0.5 (no discriminative ability) to 1.0 (perfect discrimination).

Redundancy Reduction Using CD-HIT

To examine the potential influence of sequence redundancy and subtle sequence-level information leakage, we applied CD-HIT (Cluster Database at High Identity with Tolerance) to each charge-density subgroup using a 95% sequence identity threshold. CD-HIT clusters highly similar peptide sequences and retains a single representative sequence from each cluster, thereby removing families of near-identical peptides that may arise from evolutionary relationships or design heuristics. This procedure ensures that no two peptides in the filtered datasets share more than 95% sequence identity.

Application of CD-HIT resulted in a notable reduction in dataset size across all charge-density subgroups, reflecting the removal of closely related sequences. Machine learning models were then retrained on the resulting non-redundant datasets using the same training, validation, and evaluation protocols as in the primary analysis. For the low charge-density subgroup, the post CD-HIT model achieved a training accuracy of 0.92 and a test accuracy of 0.90, with a precision of 0.91, recall of 0.85, and ROC AUC of 0.95. The intermediate charge-density subgroup showed a training accuracy of 0.92 and a test accuracy of 0.89, with a precision of 0.90, recall of 0.87, and ROC AUC of 0.95. For the high charge-density subgroup, the corresponding values were a training accuracy of 0.92, test accuracy of 0.91, precision of 0.91, recall of 0.89, and ROC AUC of 0.97. The consistency of performance metrics before and after CD-HIT filtering indicates that the predictive behavior of the models does not rely on redundant or highly similar sequences.

External Dataset Validation

To further assess the generalizability of the model, we performed an independent external validation using newly reported antimicrobial peptides from the DBAASP database. Only peptides reported in 2025 were considered to ensure that these sequences were entirely absent from the curated datasets used for model development and interpretation. After data cleaning, this external set contained 111 AMPs. To maintain class balance, 111 non-AMP sequences were randomly selected from the UniProt negative dataset.

Because the newly reported AMPs are predominantly characterized by high charge/length values, this external benchmark effectively represents the high charge-density subgroup. The trained model was able to correctly predict these AMPs and the accuracy on the external validation set was found to be 0.98. In order to identify the SHAP values on the external benchmark, we retrained the XGBoost model on the newly reported AMP dataset using the same hyperparameters as the original high charge-density model and evaluated its performance.The model demonstrated a good performance on the external dataset, achieving a training accuracy of 0.94 and an external test accuracy of 0.89. The corresponding precision was 0.86, sensitivity (recall) was 0.90, and the ROC AUC was 0.93. These results indicate that the learned decision patterns generalize well to newly reported AMPs and are not restricted to the specific composition of the original dataset.

Feature Importance Analysis

We used SHAP (SHapley Additive exPlanations)[15] to evaluate the contribution of each feature to model predictions. SHAP values quantify how much each feature shifts the prediction from a baseline, with the sign indicating the direction of influence. For every instance in the training dataset, SHAP values were computed for all features. SHAP values provide a consistent and locally accurate measure of how individual features influence the output of a trained model, thereby facilitating mechanistically grounded interpretation of machine learning outcomes. The absolute values were taken to focus on the magnitude of influence rather than direction. These were then averaged across all instances to obtain a global importance score for each feature. Features with higher mean absolute SHAP values were considered more influential. The method provides an interpretable, model-agnostic measure of feature impact.

Feature Ablation Analysis Based on SHAP Values

To directly evaluate the robustness of the SHAP-identified descriptors, we performed a feature ablation analysis in which XGBoost models were retrained using only the top three and top five features ranked by mean SHAP values within each charge-density subgroup.

For the low charge-density subgroup, the model achieved an accuracy of 0.78, precision of 0.78, recall of 0.83, and ROC AUC of 0.87 when restricted to the top three features. Performance improved to an accuracy of 0.85, precision of 0.85, recall of 0.87, and ROC AUC of 0.92 when the top five features were included. In the intermediate subgroup, the top three features yielded an accuracy of 0.77, precision of 0.84, recall of 0.70, and ROC AUC of 0.86, while the top five features increased performance to an accuracy of 0.81, precision of 0.85, recall of 0.76, and ROC AUC of 0.90. Similarly, in the high charge-density subgroup, models trained using the top three features achieved an accuracy of 0.84, precision of 0.87, recall of 0.80, and ROC AUC of 0.92, which further improved to an accuracy of 0.86, precision of 0.88, recall of 0.85, and ROC AUC of 0.93 when the top five features were used.

These results demonstrate that the most influential features identified by SHAP retain substantial predictive power even when considered in isolation, supporting their relevance across distinct electrostatic regimes. Based on this ablation analysis, we report the top five features in the main manuscript (Fig. 4), as this set provides a balanced representation of the dominant physicochemical descriptors while preserving the strong predictive capacity observed in the reduced-feature models.

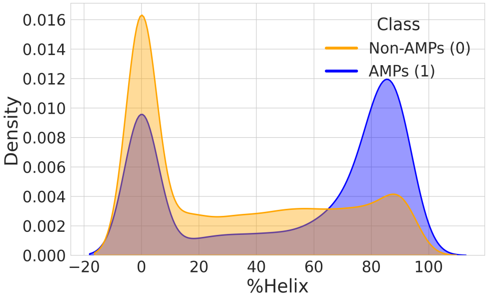

Figure S1: Distributions of helical content in non-AMPs (orange, Class 0) and AMPs (blue, Class 1) represented as kernel density estimates.

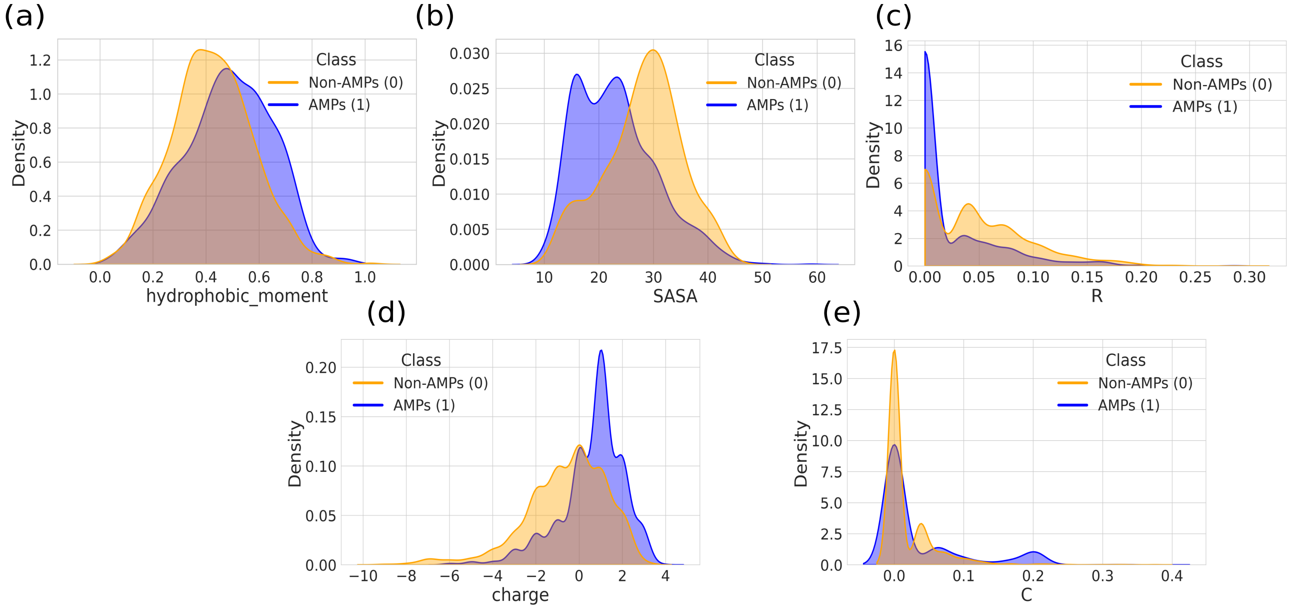

Figure S2: Distributions of top features, Hydrophobic moment , SASA, Arginine frequency, Charge and Cysteine frequency for low charge-density peptides. Non-AMPs (orange, Class 0) and AMPs (blue, Class 1) are shown as kernel density estimates.

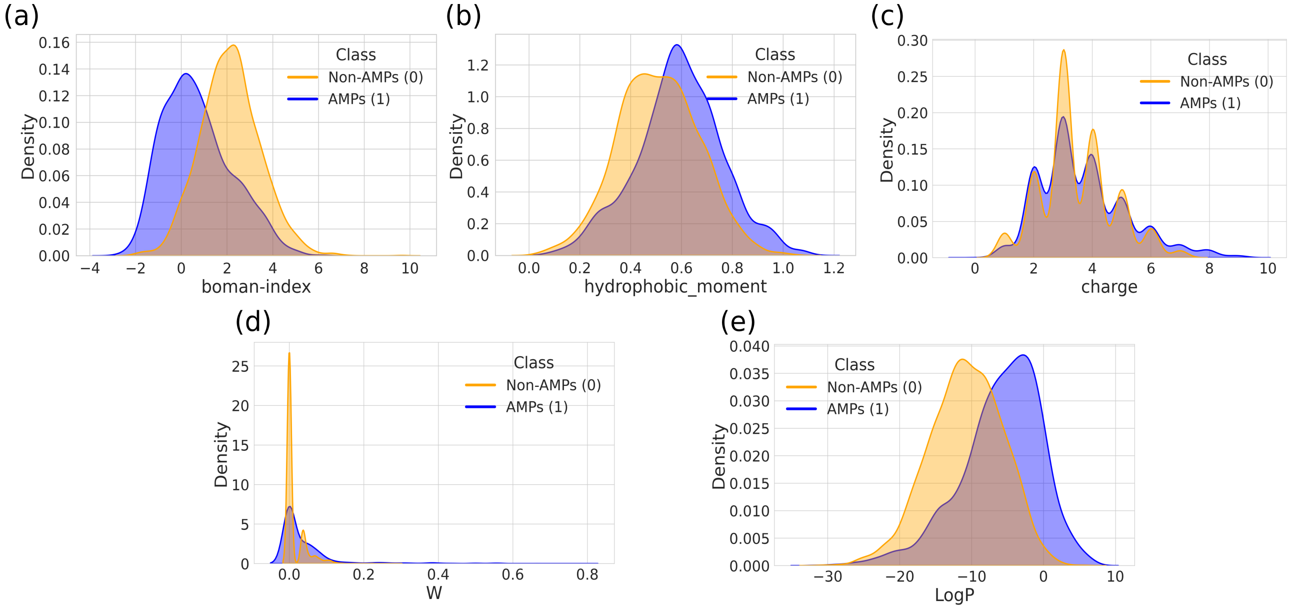

Figure S3: Distributions of top features, Boman index, Hydrophobic moment, charge, Tryptophan frequency and LogP for intermediate charge-density peptides. Non-AMPs (orange, Class 0) and AMPs (blue, Class 1) are shown as kernel density estimates.

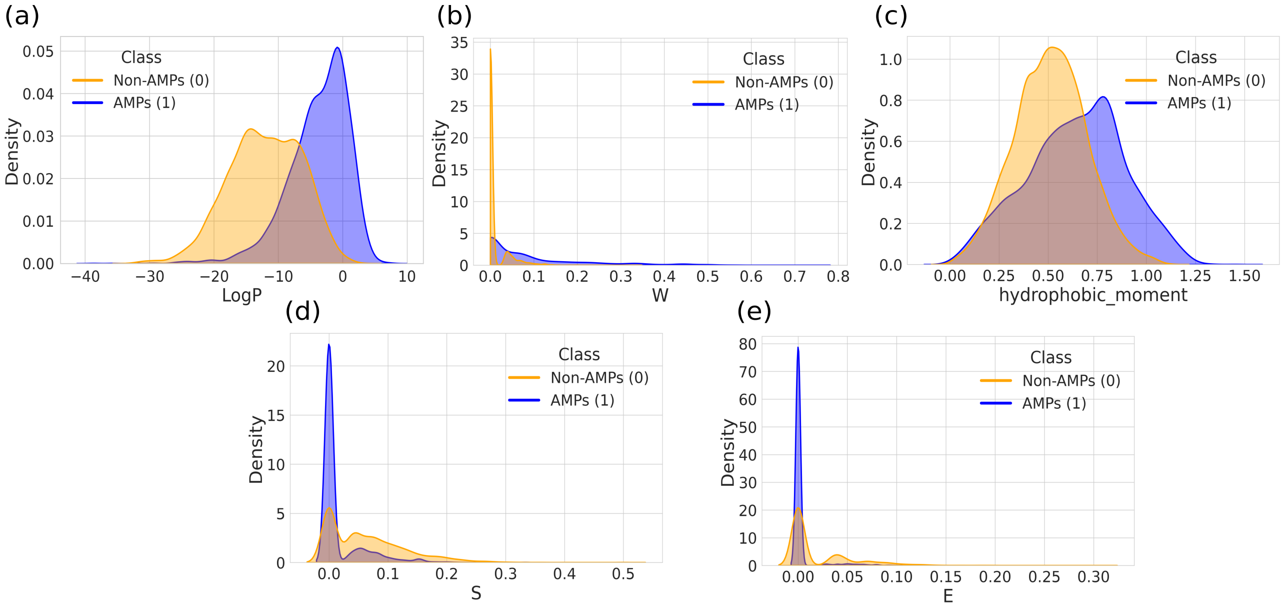

Figure S4: Distributions of top features, LogP, Tryptophan frequency, Hydrophobic moment, Serine frequency and Glutamic acid frequency for high charge-density peptides. Non-AMPs (orange, Class 0) and AMPs (blue, Class 1) are shown as kernel density estimates .

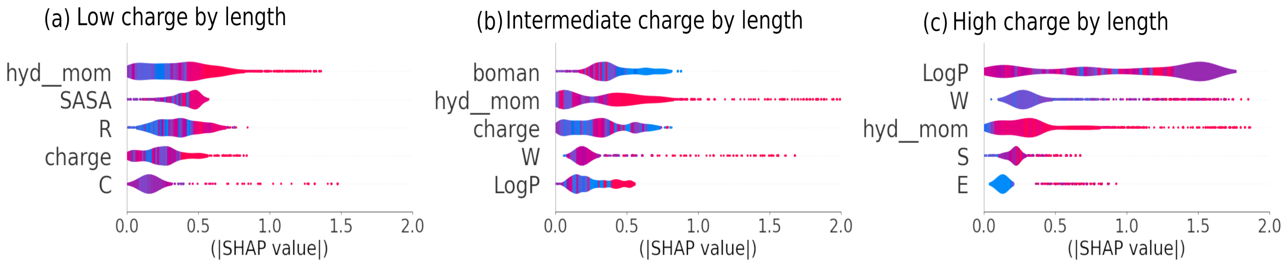

Figure S5: Absolute SHAP value distributions for the top five features in each charge/length category. Panels show (a) the low charge/length subgroup, (b) the intermediate charge/length subgroup, and (c) the high charge/length subgroup .

Table 1: Performance of five ML models across Low, Intermediate, and High charge-by-length datasets.

	Recall			Precision			Accuracy			ROC AUC		
Model	Low	Int	High	Low	Int	High	Low	Int	High	Low	Int	High
XGB	0.87	0.84	0.90	0.92	0.92	0.92	0.90	0.88	0.91	0.95	0.95	0.97
RF	0.74	0.75	0.77	0.82	0.83	0.86	0.79	0.79	0.82	0.88	0.88	0.91
CB	0.84	0.82	0.88	0.88	0.89	0.92	0.86	0.85	0.90	0.93	0.93	0.96
DT	0.78	0.81	0.70	0.82	0.73	0.86	0.81	0.74	0.79	0.84	0.79	0.861
NB	0.80	0.72	0.81	0.80	0.82	0.86	0.80	0.78	0.84	0.87	0.86	0.91

References

(1) Jhong, J.-H.; Yao, L.; Pang, Y.; Li, Z.; Chung, C.-R.; Wang, R.; Li, S.; Li, W.; Luo, M.; Ma, R., et al. dbAMP 2.0: updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data. *Nucleic acids research* **2022**, *50*, D460–D470.

(2) Shi, G.; Kang, X.; Dong, F.; Liu, Y.; Zhu, N.; Hu, Y.; Xu, H.; Lao, X.; Zheng, H. DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides. *Nucleic acids research* **2022**, *50*, D488–D496.

(3) Pirtskhalava, M.; Amstrong, A. A.; Grigolava, M.; Chubinidze, M.; Alimbarashvili, E.; Vishnepolsky, B.; Gabrielian, A.; Rosenthal, A.; Hurt, D. E.; Tartakovsky, M. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic acids research* **2021**, *49*, D288–D297.

(4) Consortium, U. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research* **2019**, *47*, D506–D515.

(5) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123–1130.

(6) Cock, P. J.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B., et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422.

(7) Osorio, D.; Rondón-Villarreal, P.; Torres, R. Peptides: a package for data mining of antimicrobial peptides. **2015**,

(8) Landrum, G. Rdkit documentation. *Release* **2013**, *1*, 4.

(9) Dong, J.; Yao, Z.-J.; Zhang, L.; Luo, F.; Lin, Q.; Lu, A.-P.; Chen, A. F.; Cao, D.-S. PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. *Journal of cheminformatics* **2018**, *10*, 16.

(10) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.

(11) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016; pp 785–794.

(12) Quinlan, J. R. Induction of decision trees. *Machine learning* **1986**, *1*, 81–106.