scientific reports

Check for updates

OPEN

Prediction of inhibitory peptides against *E.coli* with desired MIC value

Nisha Bajiya, Nishant Kumar & Gajendra P. S. Raghava ✉

In the past, several methods have been developed for predicting antibacterial and antimicrobial peptides, but only limited attempts have been made to predict their minimum inhibitory concentration (MIC) values. In this study, we developed predictive models for MIC values of antibacterial peptides against *Escherichia coli* (*E. coli*), comprised of 3143 peptides for training and 786 peptides for validation, with experimentally determined MIC values. We found that the Composition Enhanced Transition and Distribution (CeTD) attributes significantly correlate with MIC values. Initially, we attempted to estimate MIC using BLAST similarity searches but found them inadequate. Subsequently, we employed machine learning regression models that integrated various features, including peptide composition, binary profiles and embeddings from large language models. Feature selection techniques, particularly mRMR, were utilized to refine our model inputs. Our Random Forest regressor built using default parameters achieved a correlation coefficient (R) of 0.78, R^2 of 0.59, and RMSE of 0.53 on the validation set. Our best model outperformed existing methods when benchmarked on an independent dataset of 498 anti-*E. coli* peptides. Additionally, we screened anti-*E. coli* proteins in the proteomes of three probiotic bacterial strains and created a web-based platform, "EIPpred", enabling users to design peptides with desired MIC values (https://webs.iiitd.edu.in/raghava/eippred).

Keywords Inhibitory peptides, Minimum Inhibitory concentration, Machine learning, *Escherichia coli*, Peptide design, Regression models

Abbreviations

ABP	Antibacterial peptide
R	Correlation coefficient
R^2	Coefficient of determination
MSE	Mean squared error
RMSE	Root mean squared error
MAPE	Mean absolute percentage error
MAE	Mean absolute error
MRE	Maximum residual error
BLAST	Basic local alignment search tool
mRMR	Minimum redundancy maximum relevance
RFR	Random Forest regression
CNN	Convolutional neural network
DNN	Deep Neural network

The introduction of antibiotics in the 1940s revolutionized medicine by effectively treating bacterial infections and reducing mortality rates. However, issues such as antibiotic resistance and allergic reactions quickly surfaced[1]. By the early twenty-first century, resistance had rendered many antibiotics ineffective, causing an estimated 5 million deaths annually, with projections rising to 10 million by 2050[2]. The escalating resistance of pathogens to conventional antibiotics has become a global crisis, demanding new treatment approaches. In response, alternative therapies like biologics-based treatments have garnered interest for their specificity and reduced potential for resistance development[3]. These innovative strategies offer a promising solution to the limitations of traditional antibiotics in the current era of resistance. Among the leading alternatives are peptide/protein-based therapies, including antimicrobial peptides (AMPs) and monoclonal antibodies (mAbs).

Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, Okhla Industrial Estate, Phase III (Near Govind Puri Metro Station), A-302 (R&D Block), New Delhi 110020, India. ✉email: raghava@iiitd.ac.in

AMPs are highly effective against a range of pathogens, including multidrug-resistant ones, due to their rapid action and unique mechanisms that limit resistance development. Part of the innate immune system, AMPs are generally amphipathic and contain cationic amino acids, which enable them to disrupt the negatively charged bacterial membrane[4]. They work through mechanisms such as membrane disruption (carpet formation, barrel stave formation, toroidal pore formation) or non-membrane pathways, including the inhibition of protein synthesis, DNA/RNA synthesis, metabolic processes, or cell wall formation [Fig. 1][5].

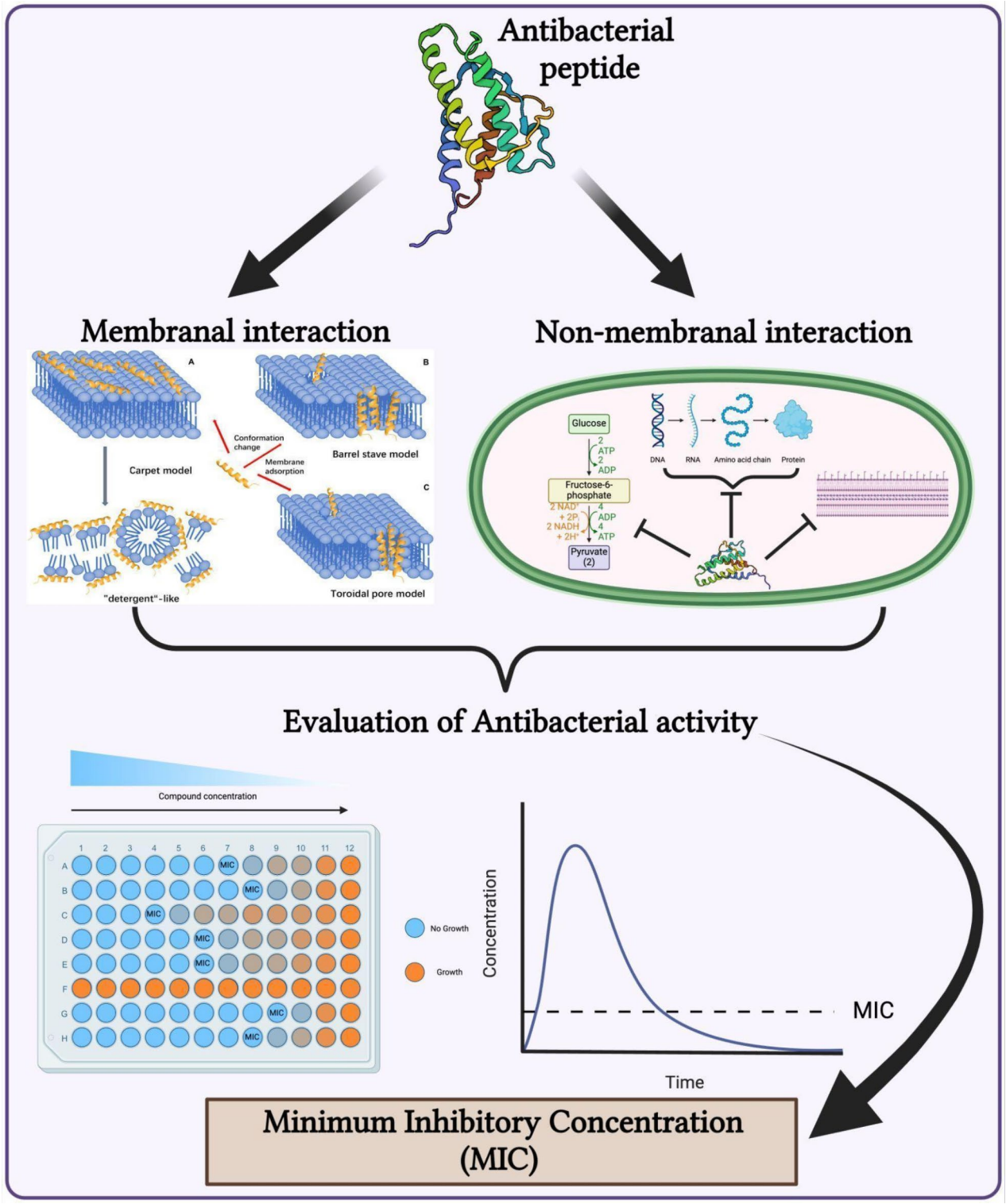

Fig. 1. Graphical representation of the working of antibacterial peptides and their activity prediction.

In the field of peptides, extensive research has led to the creation of comprehensive databases accessible to the scientific community. The first such database, the Antimicrobial Peptide Database (APD), was established in 2004 and contains detailed information on peptides with various properties, including antitumor, antiviral, antifungal, and antibacterial activities[6]. Since its inception, numerous public repositories have emerged, providing data on natural, synthetic, and predicted antimicrobial/antibacterial peptides. Notable examples include CAMPR4, DBAASP v3, DRAMP 3.0 and AntiTbPdb, which offer comprehensive information on the peptide structure, functional activities, target species, and sources[7–10]. These databases, populated with data from both experimental methods and computational resources, are invaluable for designing new drugs and advancing peptide research.

With the growth of publicly available data on AMPs, there has been a significant increase in the development of accurate prediction tools for identifying and designing AMP sequences. Examples of these methods include AI4AMP, AMPDiscover, AMPScanner vr.2, AntiBP3, AntiFP and AntiMPmod[11–16]. Similarly, species-specific predictive methods like AntiTbPred are designed to distinguish antituberculosis peptides[17]. These methods aim to predict peptides that can inhibit bacterial growth but do not assess the inhibitory potential or the amount of peptide required to inhibit bacterial growth[18]. Recent studies have attempted to predict the minimum inhibitory concentration (MIC) value of peptides against *Escherichia coli* (*E. coli*) or other bacteria[5,19–22]. To complement these existing methods, we have systematically developed an improved method for predicting the MIC value of peptides against *E. coli*. One of the major objectives of our study is to facilitate the scientific community in designing peptides with desired MIC values against *E. coli*. Therefore, we developed standalone software called EIPpred and a web-based platform to support this goal. Our standalone server can be used to scan antibacterial peptides in proteins against *E. coli* at a proteome scale. In addition to prediction, our server maps residues in peptides responsible for increasing or decreasing MIC values.

Hence, to deploy the application of our web server, we have performed a case study aimed at determining the effectiveness of our models in predicting the inhibitory peptides/regions from the protein sequences of the whole genome of the probiotic organisms. *E.coli* is a common bacterial species found in the flora of the human gut; however, certain enteropathogenic strains of *E.coli* can cause various infections and can easily develop resistance, such as gastrointestinal disorders like loss of appetite, nausea, hemorrhagic colitis (HC), diarrhoea and Inflammatory Bowel Disease (IBD)[23,24]. Hence, several probiotics in the market have been released possessing antimicrobial properties that help in the elimination of intestinal pathogenic bacteria[25,26]. These probiotics consist of non-pathogenic live microorganisms, namely Lactic Acid Bacteria (LAB) that produce bacteriocin and lactic acid to limit the proliferation of distinct or related bacteria and others such as *Bifidobacterium*, *Propionibacterium* and the fungi *Saccharomyces boulardii* are effective according to WHO and Food and Agriculture Organisation of the United Nations (FAO)[27]. *Enterococcus* belongs to the LAB bacterial species that produce a different variety of enterocin that suppress the flourishing gram-negative and gram-positive bacteria[28,29]. They provide several health benefits, including regulating the host immune system, anti-inflammatory effects, antimicrobial effects, enhancing nutrient absorptions, etc.[30,31]. For this, we have selected the three most commonly used probiotic strains of *Lactobacillus*, *Bifidobacterium*, and *Enterococcus* that inhibit the growth of *E.coli* and thus have antibacterial properties in their proteome data. By employing the standalone version of EIPpred, we scan the whole proteome of the selected strains and identify the proteins possessing maximum inhibitory activity against *E. coli*, which could be beneficial in designing novel antibacterial peptide-based drugs.

Materials and methods

Collection of dataset

Main dataset

The pre-processed dataset of 3929 peptides with MIC value against *E. coli* was obtained from paper MBC-CNN attention[32], which was originally curated from the DBAASP v3 database[8]. This dataset was divided into training and validation sets, with the training set containing 3,143 (80%) sequences and the validation set containing 786 (20%) sequences. To avoid any biases in the distribution of the training and validation datasets, we first sorted all peptides in decreasing order based on their MIC values, then added the first peptide to the validation set, the next four peptides to the training set, the sixth peptide to the validation set and so on.

Independent dataset

We have already used 20% of the data to validate our models using external validation. In addition, we have created an independent dataset of peptides with MIC values against *E. coli* to evaluate the performance of our final model and the existing methods. We extracted 498 unique experimentally validated peptides, curated from DBAASP[8], DRAMP[33] and APD3[6] databases with a count of 127, 342 and 29 peptides, respectively. We ensured none of these sequences were identical to the training and validation dataset. We removed all those 8–60 length peptides that contain unnatural amino acids, 'BJOUZX'.

Target variable

In this study, the target MIC value is represented as negative logarithmic MIC (in microMolar) against the bacteria, represented in Eq. 1. However, a single peptide can have multiple MIC values; in that scenario, the mean of all MIC values is considered the target prediction value for such multiple entries[21].

$$Target\ variable\ unit = -log_{10}\left(MIC\ in\ microMolar,\ \mu M\right) \quad (1)$$

Workflow of the study

The detailed workflow of the study is represented in Fig. 2, which depicts the progression of this study from the data curation to the final tool development and web server development.

Cross-validation

To ensure accurate error estimation, we employed a fivefold cross-validation (CV) approach, which prevents overfitting, reduces bias, and allows us to evaluate the predictive ability of our models. We applied the fivefold CV technique on the training dataset while keeping the validation dataset untouched as a standard procedure which has been extensively used in several previous studies[34–36]. This approach divides the entire training dataset into five parts: four parts for training the model and the fifth part for testing it. This process is repeated five times, with each part serving as a test set once. The overall performance is calculated as the average of these five iterations. The untouched validation dataset is then used to externally validate our models.

BLAST for similarity search

BLAST is heavily used in literature to annotate protein sequences[37–39]. We have used it to identify peptides with antibacterial activity based on the similarity of peptides with known antibacterial activity, believing that similar sequences might have the same functional properties. The blastp (protein–protein BLAST) suite of BLAST + version 2.7.1 was used to develop the similarity-based search module[40], where the query sequences were hit against the custom database built using the training set consisting of known activity of ABPs. Then, a peptide is searched against a custom database using BLAST for different E-value cutoffs. For the hits obtained after running the BLAST, the validation peptide was assigned to the particular MIC values from the database, and the performance was evaluated using various e-values.

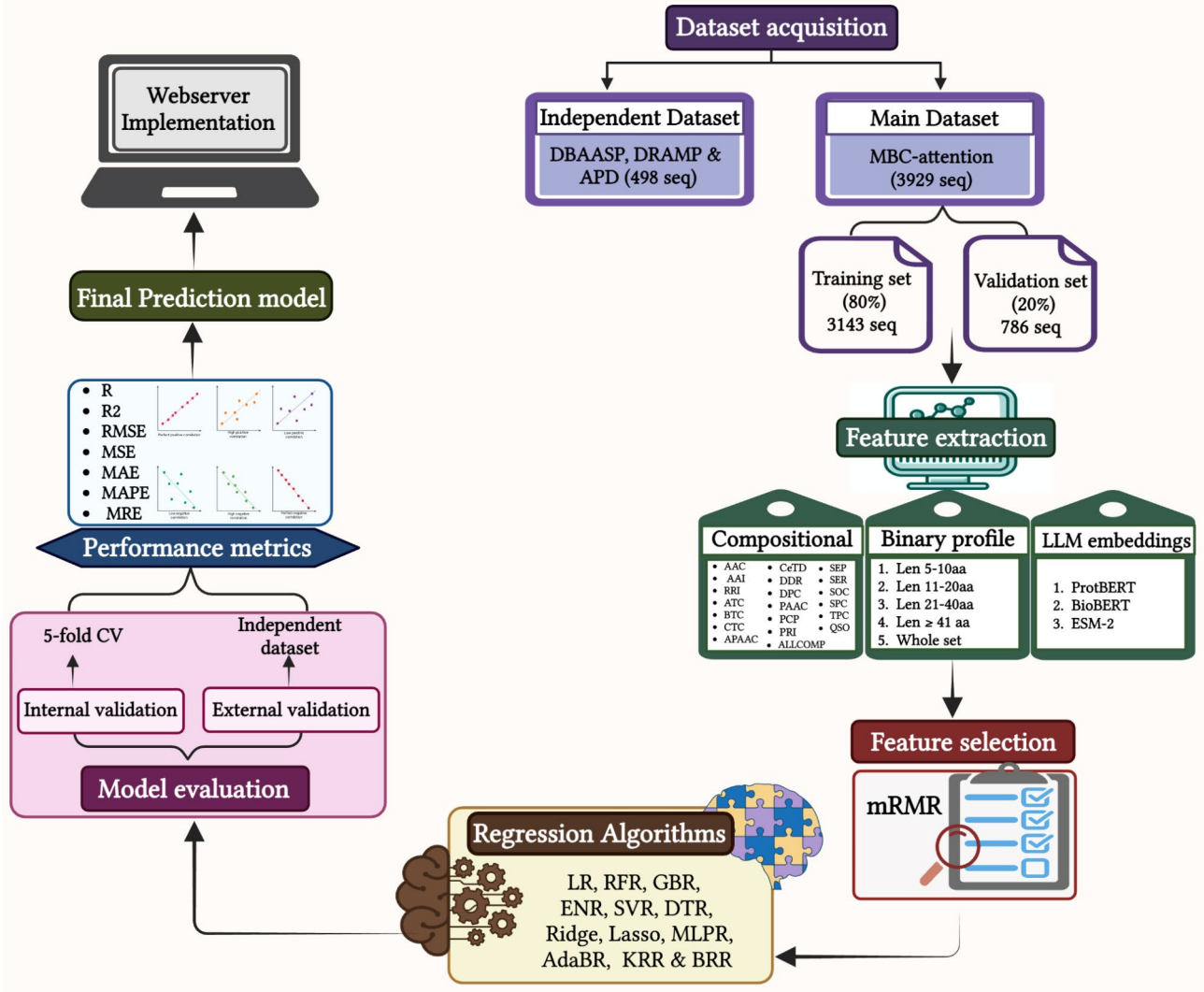

Fig. 2. The detailed workflow of the study.

Peptide features

Composition-based features

Residue information of the protein was used in the form of various compositional features for developing ML regression models. With the intent to calculate a diverse range of features from the sequences of protein or peptide, we have implemented the Pfeature package[41]. We deployed the composition-based module of Pfeature to generate more than 9000 descriptors of peptide sequences in the datasets. We have calculated nineteen types of features (AAC, DPC, RRI, DDR, SEP, SER, SPC, PRI, AAI, CTC, CeTD, PAAC, APAAC, QSO, TPC, ATC, PCP, BTC and SOC) as shown in Supplementary Table S1. The input vector of 9189 descriptors was used further for feature selection and ML purposes[42].

Binary profiles

We generated a binary profile or one hot encoding of patterns by assigning binary values to the amino acids in fixed-length patterns. First, we have divided the whole data set into four different groups according to the length of the peptides: 1st group with a length of 5-10aa, 2nd group with a length of 11-20aa, 3rd group with a length of 21-40aa and 4th group with a length of 41aa or more. For each group, we have further created fixed length vectors such as for group 5-10aa, a vector of 10 residues is formed by taking five residues from the N-terminal and five residues from the C-terminal, then generating a pattern with a vector size of 10×20 (here, 20 corresponds to each amino acid present in the pattern and 10 is the length of the pattern)[43]. Similarly, patterns of 20×20, 40×20 and 80×20 are generated for the second, third and fourth groups, respectively, as depicted in Supplementary Table S2.

Large language model (LLM) embeddings

Transformer-based language models have recently been prominent in natural language processing (NLP) and have been applied to protein sequences. These protein language models (PLMs) seek to enhance protein representation by learning contextualized representation (also called embeddings) from an abundance of protein sequence data, improving tasks involving protein functions[44].

We evaluated three pre-trained protein models (protBERT, BioBERT and ESM-2) to extract features from the peptide sequences for regressive analysis. ProtBert [45] was trained on the Big Fantastic Database (BFD), which contains over 2.3 million protein sequences and is based on the BERT algorithm[46,47]. The last attention layer produces an output of a 1024-dimensional embedding for each residue[48]. BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a model pre-trained on large-scale biomedical corpora (PubMed abstracts and PubMed Central full-text articles)[49] that generates a total of 768 embeddings and ESM-2 (Evolutionary Scale Modeling)[50] is a general-purpose PLM based on the BERT transformer architecture and trained on UniRef50 to predict masked amino acids using all the preceding and following amino acids in the sequence. In this study, we used a model called esm2_t33_650M_UR50D (termed ESM-2), which has approximately 3 billion learnable parameters. The model's output was an embedding of a feature dimension of 1280 for each amino acid[44].

Feature selection

Pfeature has generated a large number of peptide features, i.e., 9189, and all might not be relevant; some may be highly correlated with each other and can cause overfitting. Therefore, selecting features that have majorly contributed to the performance of ML and removing the highly correlated features are necessary to identify relevant features, thus reducing the complexity of models. Among various feature selection methods, we have applied the mRMR (minimum redundancy maximum relevance) method[51] that uses mutual information to compute the relevance and the redundancy among features/classes. We have used the mrmr_regression program to select the top 200, 500, 1000, 1500 and 2000 features and evaluated the ML performance on the selected features.

Regression models

Regression aims to create a model of the relationship between a certain number of features and a continuous target variable. Twelve different regression algorithms were implemented in our study using the Python Scikit-learn library[52]. Regressive algorithms used in the study are Linear regression (LR)[53], Support vector regression (SVR)[54], Ridge regression (Rigde)[55], Lasso regression [56], Gradient Boosting regression (GBR)[57], MLP regression (MLPR)[58], AdaBoost regression(AdaBR)[59], Elastic Net regression (ENR)[60], Kernel Ridge regression (KRR)[61] and Bayesian Ridge regression(BRR)[62] to determine which model could deliver the most accurate performance in the prediction of the exact value of MIC. The models are run with the default parameters.

Hybrid approach

We intended to enhance further our best-performing regressor model, built on the best feature set, by incorporating an ensemble method. This method combines the regressive power of the similarity search method with the ML regressors. Unlike a classification problem where the blast scores are added with the prediction scores, instead here, the MIC values are replaced for the obtained blast hits with the corresponding ML prediction scores. Various E-value cutoffs were used to evaluate the performance of the hybrid method.

Performance metrics

We have used the Correlation coefficient (R-score), Mean absolute percentage error (MAPE), Maximum residual error (MRE), Mean absolute error (MAE), Mean squared error (MSE), Root mean squared error (RMSE), and Coefficient of determination (R^2) as metrics to measure the performance of the regressive models as shown

in Equations from 2 to 7. Note that the lower the MAE, MSE, MAPE, MRE, or RMSE value, the better the regression models. Conversely, the higher the value of R and R^2, the better the regressive models.

$$R = \frac{\Sigma\left(x_i - \underline{x}\right)\left(y_i - \underline{y}\right)}{\sqrt{\Sigma\left(x_i - \underline{x}\right)^2 \Sigma\left(y_i - \underline{y}\right)^2}} \tag{2}$$

$$R^2 = 1 - \frac{\Sigma\left(y_i - \widehat{y_i}\right)^2}{\Sigma\left(y_i - \underline{y}\right)^2} \tag{3}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \widehat{y_i}\right| \tag{4}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{y_i}\right)^2} \tag{5}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| \times 100 \tag{6}$$

$$MRE = max\left(\left|y_i - \widehat{y_i}\right|\right) \tag{7}$$

where,y_i and x_i are the data points, N is the number of data points, and $\underline{y}$, $\widehat{y_i}$ and $\underline{x}$ indicates the mean value of y, predicted value of actual value (y), and average of x, respectively.

Results

Primary analysis

Length distribution

We have explored the datasets according to their length and found that more than 80% of the data length lies in between 11 and 40 residues for all train, validation and independent sets. The dataset is graphically represented in Fig. 3 as per length distribution.

Correlation analysis

In order to find the strength of the relationship between the features and the target MIC values, we have computed the Pearson correlation of amino acids, dipeptides, and mRMR selected 1000 features, as depicted in Fig. 4. Its value ranges from − 1 to 1, where − 1 represents a negative correlation, which means the target MIC value is less dependent, and 1 represents a positive correlation, which means the target MIC value is more dependent on that particular feature, and a value of 0 indicates no correlation.

In amino acids, residues such as R and Y show a positive correlation of 0.19 and 0.14, while residues such as E and S show significant negative correlation of − 0.16. In the case of dipeptides, among 400 dipeptide features, the top 10 positive correlated and bottom 10 negative correlated features have been selected for representation; dipeptides like CY and YR have shown the highest correlation of 0.23. In contrast, EN and NE have shown the lowest correlation of − 0.16 and − 0.15, respectively. Similarly, for the 1000 mRMR selected features, the top 10 and bottom 10 have been used for representation, where the CeTD features show a maximum positive correlation of 0.41, and PRI features show a minimum correlation of − 0.28.

Performance of similarity search method

We have also utilized the ability of BLAST to predict the MIC values of validation data for the hits obtained from the customized database. For the validation hits obtained, we have assigned them the MIC values corresponding to the database sequences. After that, their performances with respect to the actual MIC values of the validation set were calculated, as shown in Table 1. At the highest e-value, the performance increases, but at the same time, the number of sequence hits obtained by the BLAST is much less, which signifies the inefficiency of the similarity search method in predicting the MIC values of the peptides.

Performance on ML-based regressors

Compositional features

We have developed prediction models using 12 regression algorithms, including LR, RFR, GBR, ENR, SVR, DTR, Ridge, Lasso, MLPR, AdaBR, KRR, and BRR. We designed prediction models employing all different types of compositional features, as shown in Supplementary Table S1. Figure 5 shows that the RF-based regressor with default parameters using cross-validation performs best among all other regression models with the highest R^2 score and lowest RMSE value using the AAC feature; therefore, we have applied RFR on default parameters (random_state = 95) for further analysis. The detailed list of performance of RFR using compositional features is depicted in Table 2; the ALLCOMP feature outperforms other models with an R and R^2 score of 0.77 and 0.59 on

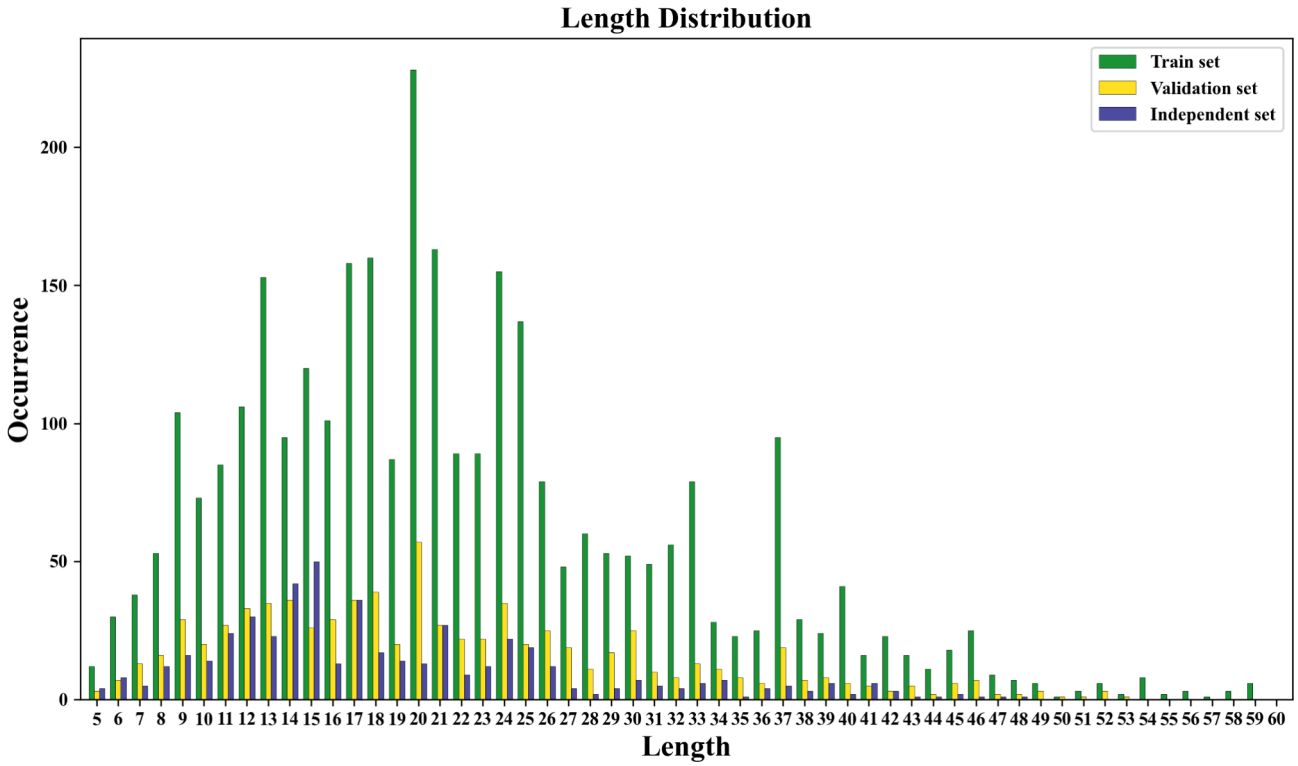

Fig. 3. The histogram of the distribution of length of antibacterial peptide for *E. coli* for training, validation and independent set.

the validation dataset with an RMSE value of 0.53. The additional performance metrics of the ALLCOMP feature are represented in Supplementary Table S3.

Binary profiles

To understand the importance of building different models for different length sets, we have developed binary profiles for different length groups and used them as input features to train and evaluate the prediction models by implementing various ML regressors. As shown in Table 3, the RFR performs extremely poor in the length group of 41 or more residues with the highest RMSE value on the validation set, while it performs better for 11–20 residues with an R^2 and RMSE value of 0.55 and 0.57 on the validation set. The detailed results of binary profiles for all length groups are depicted in Supplementary Tables S4, S5, S6, S7 and S8.

LLM embeddings

To generate the embedding using large language models, first, we have to modify the LL models for regression analysis and then pass out data to compute the embeddings. The performance across various LLM embeddings is reported in Table 4. Models that exploited ESM-2 embeddings saw the best performance with lower RMSE and high correlation scores of 0.56 and 0.76, respectively. Supplementary Tables S9, S10, and S11 show the results obtained with all three LLM embeddings.

Performance on selected features

We have implemented the mRMR feature selection technique to get useful features that help enhance the models' performances. We have also evaluated the performance of features at intervals of 100 to provide a better understanding of the relevance of each feature subset, as reported in Supplementary Table S17. Since the variation is very minimal, we have selected the top 200, 500, 1000, 1500 and 2000 features to check the effectiveness of different feature subsets. The performance of the RF-based regressor on selected features is depicted in Table 5, which shows that the top 1000 features contributed most to effective understanding by the model and got a maximum R/R^2 of 0.78/0.59 with RMSE 0.53 on the validation set. The detailed performance of all the selected feature sets are presented in Supplementary Tables S12, S13, S14, S15 and S16.

Performance of hybrid method

With the aim of improving the performance of the ML-based regressors, we have tried a hybrid approach that combines the similarity search BLAST with the non-similarity-based ML methods. Here, we first run the BLAST on the validation set and obtain the hits from the customized database. After that, for the validation hits obtained, we assigned them the MIC values as predicted by the best-performing RF-based regressor model built on mRMR-selected 1000 features and their performances with respect to the actual MIC values of the validation set were calculated, as shown in Supplementary Table S18. Our RFR model obtained an R score of 0.78 on the

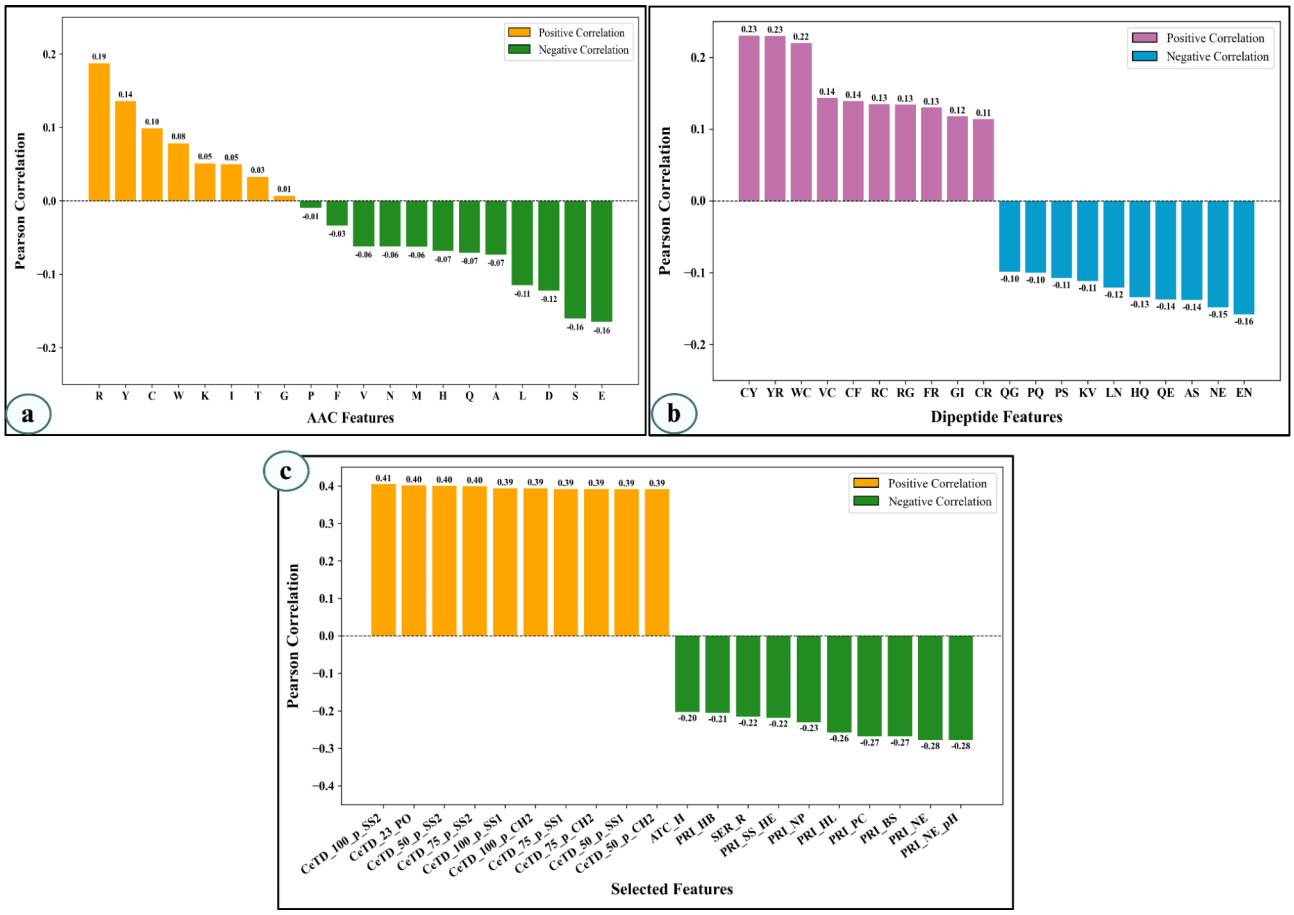

Fig. 4. The bar graph showing the Pearson correlation value of **a** AAC, **b** Dipeptide and **c** 1000 selected features with the target MIC values.

E-value	Hits (Total peptides = 786)	R	R^2	MSE	RMSE	MAPE	MAE	MRE
1.00E-20	18	0.88	0.74	0.12	0.34	1.79	0.23	0.72
1.00E-10	174	0.72	0.48	0.32	0.57	15.10	0.37	2.11
1.00E-06	334	0.72	0.49	0.36	0.60	104.60	0.41	2.20
0.0001	406	0.70	0.45	0.37	0.60	93.80	0.41	2.84
0.001	461	0.69	0.41	0.39	0.62	83.53	0.42	2.84
0.01	516	0.70	0.43	0.38	0.61	74.72	0.41	2.84
0.1	570	0.69	0.42	0.39	0.63	71.75	0.42	2.84
1	619	0.67	0.38	0.42	0.65	66.12	0.44	3.26

Table 1. Performance of BLAST on the validation set at different e-values. #R = Correlation coefficient, R^2 = Coefficient of determination, MSE = Mean squared error, RMSE = Root mean squared error, MAPE = Mean absolute percentage error, MAE = Mean absolute error, MRE = Maximum residual error.

validation set, while the hybrid method only for obtained hits shows a maximum R score of 0.92 at an e-value of 1.00E−20 while its coverage is only 18 peptides from the total 786 peptides, which is very low implies that the combined approach is not improving the performance significantly.

Benchmarking on an independent set

We have selected an RF-based regressor-based model on default parameters with the top 1000 selected features as our final model. To evaluate the performance of our model, we have utilized the independent set and compared our performance with the existing methods. The MBC-attention model[32] was built using CNN with an attention layer and tuned on various hyperparameters such as the number of layers, number of filters, dropout layers, loss function and early stopping parameters. We have used the same non-redundant dataset that is used in MBC-attention. However, the dataset splitting is performed differently. Similarly, the model architecture of

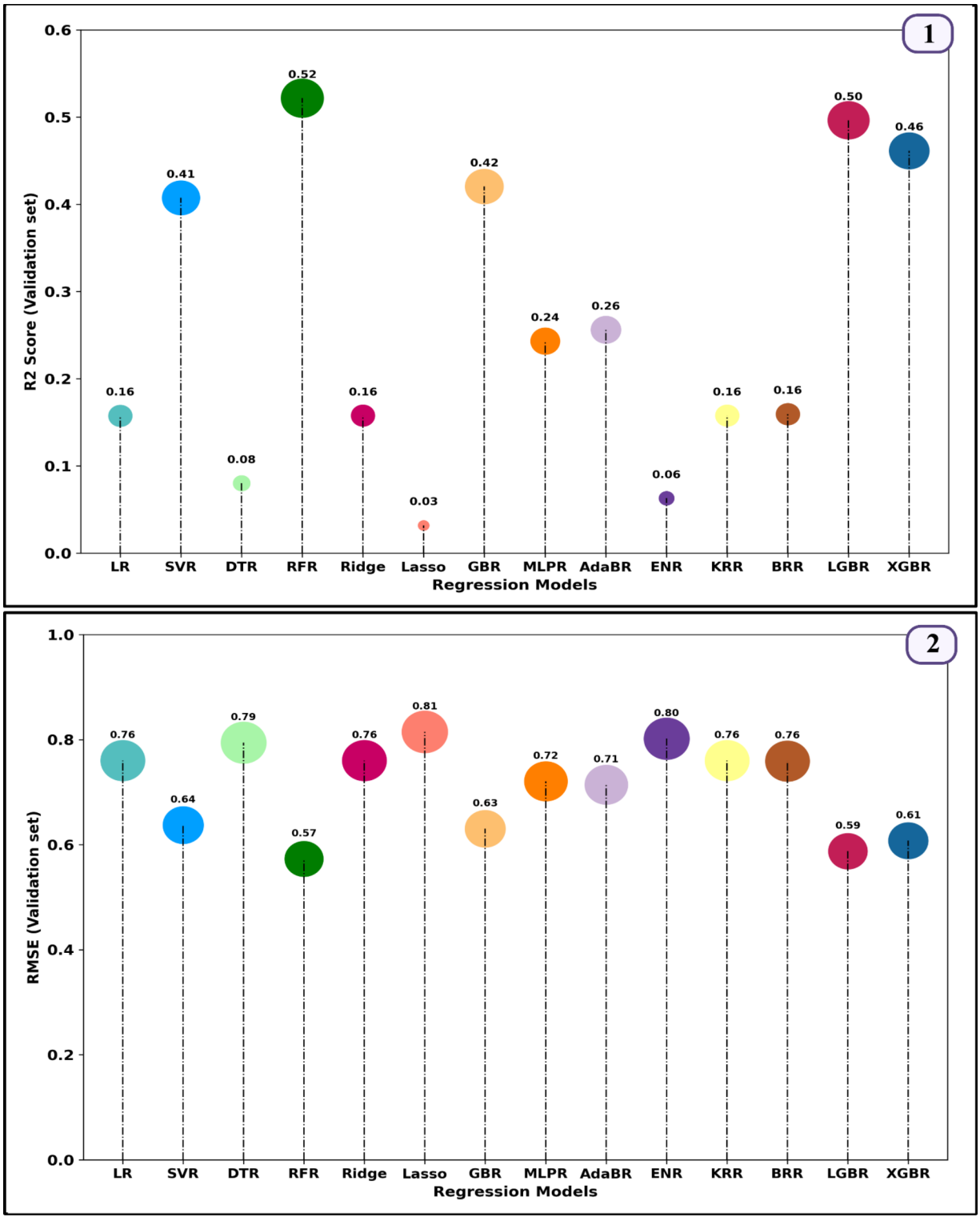

Fig. 5. Performance of ML-based regressors in terms of (1) R^2 and (2) RMSE scores developed using AAC feature on the validation set using a fivefold cross-validation strategy.

AMPActiPred is the deep neural network that first classifies the ABP or non-ABP against different bacterial groups and finally predicts the activity of those peptides in terms of MIC values[22]. The performance of different models on the independent dataset is shown in Table 6, which suggests that our model performs better on the independent set despite being built using basic computational procedures with an R^2 and RMSE value of 0.410 and 0.709, respectively.

Features	R	R^2	RMSE	R	R^2	RMSE
	Test set (CV)			Validation set		
AAC	0.67	0.45	0.61	0.73	0.52	0.57
AAI	0.64	0.41	0.63	0.68	0.46	0.61
APAAC	0.69	0.47	0.60	0.72	0.51	0.58
ATC	0.49	0.23	0.72	0.53	0.27	0.71
BTC	0.51	0.23	0.73	0.50	0.21	0.74
CTC	0.65	0.42	0.63	0.71	0.51	0.58
CeTD	0.70	0.49	0.59	0.76	0.57	0.55
DDR	0.66	0.44	0.62	0.71	0.49	0.59
DPC	0.69	0.47	0.60	0.74	0.54	0.56
PAAC	0.67	0.44	0.62	0.70	0.48	0.60
PCP	0.64	0.40	0.64	0.67	0.45	0.62
PRI	0.66	0.43	0.62	0.70	0.48	0.60
QSO	0.68	0.47	0.60	0.74	0.54	0.56
RRI	0.63	0.40	0.64	0.67	0.45	0.62
SEP	0.27	-0.07	0.85	0.30	-0.03	0.84
SER	0.68	0.46	0.61	0.73	0.52	0.57
SOC	0.28	0.00	0.82	0.32	0.04	0.81
SPC	0.63	0.40	0.64	0.67	0.44	0.62
TPC	0.67	0.45	0.61	0.72	0.52	0.57
ALLCOMP	0.72	0.52	0.57	0.77	0.59	0.53

Table 2. Performance of RF-based regressor on compositional features of test set (CV) and validation set. #R = Correlation coefficient, R^2 = Coefficient of determination, RMSE = Root mean squared error.

Length groups	R	R^2	RMSE	R	R^2	RMSE
	Test set (CV)			Validation set		
5-10aa	0.67	0.44	0.69	0.73	0.53	0.63
11-20aa	0.70	0.49	0.6	0.74	0.55	0.57
21-40aa	0.60	0.36	0.58	0.53	0.28	0.6
41 or more aa	0.55	0.21	0.67	0.48	0.21	0.69
Whole data	0.59	0.35	0.66	0.64	0.41	0.64

Table 3. Performance of RF-based regressor on Binary profiles of different length groups of test set (CV) and validation set. #R = Correlation coefficient, R^2 = Coefficient of determination, RMSE = Root mean squared error.

LLM	R	R^2	RMSE	R	R^2	RMSE
	Test set (CV)			Validation set		
ProtBERT	0.65	0.40	0.64	0.70	0.46	0.61
BioBERT	0.58	0.31	0.69	0.63	0.36	0.66
ESM-2	0.69	0.46	0.61	0.76	0.55	0.56

Table 4. Performance of RF-based regressor on LLM embeddings of test set (CV) and validation set. #R = Correlation coefficient, R^2 = Coefficient of determination, RMSE = Root mean squared error.

Web server implementation

To better serve the scientific community, we have developed a user-friendly prediction web interface named "EIPpred" (https://webs.iiitd.edu.in/raghava/eippred) and executed our best model to predict the MIC values of ABPs. The web server includes three different modules: "Predict", "Design", and "Protein scan". The module 'Predict' allows users to predict the inhibitory activity of the submitted sequence against *E.coli* in terms of MIC value (in $-\log_{10}$ μM). The 'Design' module allows users to create mutants by generating all possible mutations replacing one residue at a time in the input sequence. The result of the design module includes the mutant sequences with inhibitory activity and represents their predicted MIC values in order to better understand the importance of each residue in imparting the inhibitory activity to the peptide. The 'Protein scan' module scans the protein sequences to identify the inhibitory regions/domains based on the predicted MIC values against

Selected features	R	R^2	MSE	RMSE	MAE	MRE	R	R^2	MSE	RMSE	MAE	MRE
	Test set (CV)						Validation set					
200	0.71	0.50	0.34	0.58	0.46	2.61	0.76	0.57	0.30	0.55	0.42	2.06
500	0.72	0.52	0.33	0.57	0.44	2.47	0.77	0.58	0.29	0.53	0.41	2.03
1000	0.73	0.52	0.33	0.57	0.44	2.58	0.78	0.59	0.28	0.53	0.41	2.12
1500	0.72	0.52	0.33	0.57	0.44	2.47	0.77	0.59	0.28	0.53	0.41	2.12
2000	0.72	0.52	0.33	0.57	0.44	2.39	0.77	0.59	0.28	0.53	0.41	2.06

Table 5. Performance of RF-based regressor on different selected feature subsets on test set (CV) and validation set. #R = Correlation coefficient, R^2 = Coefficient of determination, MSE = Mean squared error, RMSE = Root mean squared error, MAE = Mean absolute error, MRE = Maximum residual error.

Model	R	R^2	MSE	RMSE	MAPE	MAE	MRE
Our models							
EIPpred (1000 selected features-RFR)	0.676	0.410	0.503	0.709	5.975	0.567	2.362
ALLCOMP (RFR)	0.673	0.401	0.511	0.715	5.968	0.572	2.335
ESM-2 (RFR)	0.611	0.326	0.575	0.758	6.571	0.591	2.944
Existing models							
MBC-attention (CNN)	0.657	0.392	0.519	0.720	7.436	0.554	2.763
AMPActiPred (DNN)	0.334	−0.071	0.914	0.956	6.298	0.585	4.258

Table 6. Comparative performance of our models and existing methods on the independent dataset. #R = Correlation coefficient, R^2 = Coefficient of determination, MSE = Mean squared error, RMSE = Root mean squared error, MAPE = Mean absolute percentage error, MAE = Mean absolute error, MRE = Maximum residual error, RFR = Random Forest Regression, CNN = Convolutional Neural Network, DNN = Deep Neural Network.

Bacterium	Total proteome (no. of proteins)	Proteins (less than 0.5 µg/ml MIC value)	Percentage of inhibitory proteins (less than 0.5 µg/ml MIC)	Inhibitory peptides (less than 0.5 µg/ml MIC value)
Lactobacillus acidophilus (La-5)	1787	621	34.751	3428
Bifidobacterium lactis (Bb-12)	1542	920	59.663	7716
Enterococcus faecalis (DSM 16431)	2630	1114	42.357	6666

Table 7. Comparative analysis of the proteome of three probiotic bacteria for the inhibitory peptide against *E.coli*.

E.coli of the desired lengths defined by the user. We also allow users to download the training, validation, and independent dataset (https://webs.iiitd.edu.in/raghava/eippred/dataset.php), which we have used in this study, available in CSV file format.

Case study: inhibitory protein scan in the proteome of probiotic species

Several studies have reported the antagonist role of probiotic single or multi strains against *E.coli*[24,25,28,31,63]. The whole genome proteome data for the probiotic strains was downloaded from NCBI; Lactobacillus acidophilus (La-5) [https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=1579&search_text=La-5], Bifidobacterium lactis (Bb-12) [https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=552531], and Enterococcus faecalis (Symbioflor 1) [https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=1261557]. For *Lactobacillus acidophilus* (La-5), we obtained a total of 1787 proteins, 1542 proteins for *Bifidobacterium lactis* (Bb-12) and 2630 proteins for *Enterococcus faecalis* (Symbioflor 1). To identify the inhibitory regions in the proteins or inhibitory proteins, we used the "Protein Scan" module of the 'EIPpred' standalone version, using the window length of 20 (i.e., length of peptides as 20 amino acids). We then converted the µM MIC values into µg/ml and made a stringent cutoff to keep those proteins and peptides with MIC values less than or equal to 0.5 µg/ml. Table 7 represents the density of the inhibitory proteins in the proteome of the three probiotic species after the cutoff, which shows that the proteome of *Bifidobacterium lactis* (Bb-12) contains nearly 60% of the inhibitory proteins, more than the other two species. The density of inhibitory peptides found in each of the three bacterial proteomes with MIC value less than 0.5 µg/m/ is depicted in Supplementary Tables S19, S21 and S23, while the list of top 20 inhibitory peptides with their predicted MIC values for each species is reported in Supplementary Tables S20, S22 and S24.

Discussion

Bacteria are well-studied organisms in terms of their structure, biology, and infection processes; nevertheless, our understanding remains incomplete due to their fast progression. The rat race between bacterial evolution and antibiotic research demands alternative options to overcome resistance. Peptide-based therapies have emerged as a promising treatment option that offers targeted approaches to inhibit bacterial survival pathways. *Escherichia coli*, a common clinical infection, frequently develops resistance to several antimicrobials, creating an enormous global issue[64]. Thus, predicting the inhibitory activity of antimicrobial peptides, specifically through MIC values, is crucial for assessing peptide-based drug effectiveness against specific bacterial species.

Since existing approaches are computationally expensive and have limited species-specificity, the activity prediction of ABPs requires improvement; hence, designing newer inhibitory peptides has become essential in the post-antibiotic era. In this study, we report a quick and efficient RF-based regressor tool aimed at improving the prediction of quantitative MIC values of ABPs against E. coli. Our model relies on a supervised ML approach based on sequence information employing various types of feature sets to predict the activity of the peptides. The main advantage of our model is its simple architecture with superior performance that allows us to learn from multiple feature types. This is particularly useful in scenarios where employing heavy computation is not possible.

The preliminary analysis of the correlation of the selected 1000 features with the target MIC values shows a strong correlation with the CeTD feature, implying that the distribution of secondary structure (SS), charge (CH) and polarity (PO) along the whole length of the sequence are important in determining inhibitory properties of the ABPs. Our results indicate that the residue repeats for physio-chemical properties, such as neutral, basic and positively charged residues, negatively correlate with the target variable. This implies that the distribution of residues capable of forming secondary structures is essential in conferring inhibitory activity.

In this study, we leverage the regressive ability of various ML algorithms for quantitative MIC prediction and enhance their predictive capabilities by incorporating different feature sets such as compositional, binary profiles, and LLM embedding. Among the features set used, the model built using all compositional features performs significantly better than individual compositional features with an R and R^2 score of 0.77 and 0.59, respectively, on the validation set. Secondly, the binary profiles with lengths from 5 to 20 residues are better at capturing the properties of the peptide to predict the MIC values quantitatively. Third and lastly, ESM-2 performs well but not better than all compositional feature sets among the LLM embedding generated from the pre-trained models.

Since the number of all compositional features (ALLCOMP) is high, we employed mRMR to select the features that greatly contributed to the performance of the model and from Table 5, we conclude that the top 1000 feature set is performing superior to other selected feature sets, thus reducing overfitting and irrelevant features. Moreover, we also perform BLAST to understand the effect of the similarity on predicting the activity of the peptides; however, our results suggest that utilizing BLAST hits doesn't significantly improve the performance of our models. Indeed, the prediction given by our ML models for the hits is far better than the hybrid method where the validation hits MIC values are replaced by the subject MICs (Supplementary Table S18).

It is important to note that by performing the comparative analysis on the independent dataset, our final RFR model built using 1000 selected features outperforms the existing methods by a small margin, achieving R and RMSE values of 0.676 and 0.709, still employing simple architecture instead of sophisticated neural networks, greatly enhancing the performance of our model. This reduces the computational cost and time needed to predict the activity of ABPs in *E. coli.* Lastly, we built a web server—EIPpred and in addition to the web server facility, GitHub (https://github.com/NishaBajiya/EIPpred), the standalone version, and the pip package (https://pypi.org/project/eippred/) of our model are also publicly available to researchers. Finally, to present the applicability of our tool, we have deployed the "Protein scan" module of the standalone version of our tool and performed the analysis on the whole-genome proteome data to identify the inhibitory proteins showing antibacterial activity against *E.coli.* Our analysis revealed that the proteome of the *Bifidobacterium lactis* (Bb-12), among others, has a vast distribution of the inhibitory proteins encompassing nearly 60% of the whole proteome.

Lastly, the application of our RFR-based tool offers a quick, accurate, and user-friendly way to design new antimicrobial peptides, marking a breakthrough in the computational prediction of peptide activity.

Conclusion

In conclusion, novel inhibitory ABPs can be designed against *E. coli* as defined by the user with their predicted MIC values by utilizing our web server, which could prove essential in the post-antibiotic era. Similarly, our model can be implemented to find the probiotic strains against *E.coli* by analysing their proteomes for inhibitory proteins/peptides. Eventually, the model needs more improvement and extends its domain to predict the activity of peptides for other pathogenic species.

Data availability

The dataset files and code used in EIPpred are available on the web server (https://webs.iiitd.edu.in/raghava/eippred). BioRxiv doi: https://doi.org/https://doi.org/10.1101/2024.07.18.604028

Received: 21 October 2024; Accepted: 13 January 2025
Published online: 08 February 2025

References

1. Daulaire, N., Bang, A., Tomson, G., Kalyango, J. N. & Cars, O. Universal access to effective antibiotics is essential for tackling antibiotic resistance. *J. Law Med. Ethics* **43**(Suppl 3), 17–21 (2015).

2. Tang, K. W. K., Millar, B. C. & Moore, J. E. Antimicrobial resistance (AMR). *Br. J. Biomed. Sci.* **80**, 11387 (2023).
3. Halawa, M., Akantibila, M., Reid, B. E. & Carabetta, V. J. Therapeutic proteins have the potential to become new weapons in the fight against antibiotic resistance. *Front. Bacteriol.* **2**, 1304444 (2023).
4. Jenssen, H., Hamill, P. & Hancock, R. E. W. Peptide antimicrobial agents. *Clin. Microbiol. Rev.* **19**, 491–511 (2006).
5. Nagarajan, D. et al. Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. *J. Biol. Chem.* **293**, 3492–3509 (2018).
6. Wang, G., Li, X. & Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **44**, D1087–D1093 (2016).
7. Gawde, U. et al. CAMPR4: A database of natural and synthetic antimicrobial peptides. *Nucleic Acids Res.* **51**, D377–D383 (2023).
8. Pirtskhalava, M. et al. DBAASP v3: Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* **49**, D288–D297 (2021).
9. Shi, G. et al. DRAMP 3.0: An enhanced comprehensive data repository of antimicrobial peptides. *Nucleic Acids Res.* **50**, D488–D496 (2022).
10. Usmani, S. S., Kumar, R., Kumar, V., Singh, S. & Raghava, G. P. S. AntiTbPdb: A knowledgebase of anti-tubercular peptides. *Database* **2018**, (2018).
11. Lin, T.-T. et al. AI4AMP: An antimicrobial peptide predictor using physicochemical property-based encoding method and deep learning. *Systems* **6**, e0029921 (2021).
12. Pinacho-Castellanos, S. A., García-Jacas, C. R., Gilson, M. K. & Brizuela, C. A. Alignment-free antimicrobial peptide predictors: improving performance by a thorough analysis of the largest available data set. *J. Chem. Inf. Model.* **61**, 3141–3157 (2021).
13. Veltri, D., Kamath, U. & Shehu, A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* **34**, 2740–2747 (2018).
14. Bajiya, N., Choudhury, S., Dhall, A. & Raghava, G. P. S. AntiBP3: A method for predicting antibacterial peptides against gram-positive/negative/variable bacteria. *Antibiotics (Basel)* **13** (2024).
15. Agrawal, P. et al. Approach for prediction of antifungal peptides. *Front. Microbiol.* **9**, 323 (2018).
16. Agrawal, P. & Raghava, G. P. S. Prediction of antimicrobial potential of a chemically modified peptide from its tertiary structure. *Front. Microbiol.* **9**, 2551 (2018).
17. Usmani, S. S., Bhalla, S. & Raghava, G. P. S. Prediction of antitubercular peptides from sequence information using ensemble classifier and hybrid features. *Front. Pharmacol.* **9**, 954 (2018).
18. Balouiri, M., Sadiki, M. & Ibnsouda, S. K. Methods for evaluating antimicrobial activity: A review. *J Pharm Anal* **6**, 71–79 (2016).
19. Witten, J. & Witten, Z. Deep learning regression model for antimicrobial peptide design. *bioRxiv* 692681 (2019) https://doi.org/10.1101/692681.
20. Yasir, M., Karim, A. M., Malik, S. K., Bajaffer, A. A. & Azhar, E. I. Prediction of antimicrobial minimal inhibitory concentrations for using machine learning models. *Saudi J. Biol. Sci.* **29**, 3687–3693 (2022).
21. Dean, S. N., Alvarez, J. A. E., Zabetakis, D., Walper, S. A. & Malanoski, A. P. PepVAE: Variational autoencoder framework for antimicrobial peptide generation and activity prediction. *Front. Microbiol.* **12**, 725727 (2021).
22. Yao, L. et al. AMPActiPred: A three-stage framework for predicting antibacterial peptides and activity levels with deep forest. *Protein Sci.* **33**, e5006 (2024).
23. LeBlanc, J. J. Implication of virulence factors in *Escherichia coil* O157: H7 pathogenesis. *Crit. Rev. Microbiol.* **29**, 277–296 (2003).
24. Poppi, L. B. et al. Effect of *Lactobacillus* sp. isolates supernatant on *Escherichia coli* O157:H7 enhances the role of organic acids production as a factor for pathogen control. *Pesqui. Vet. Bras.* **35**, 353–359 (2015).
25. Fijan, S., Šulc, D. & Steyer, A. Study of the in vitro antagonistic activity of various single-strain and multi-strain probiotics against. *Int. J. Environ. Res. Public Health* **15** (2018).
26. Roozbahani, F. et al. Characterization of antimicrobial activities BB-12 and their inhibitory effect against some foodborne pathogens. *Foodborne Pathog. Dis.* **21**, 370–377 (2024).
27. Acharjee, M. et al. antibacterial activity of commercially available probiotics on food-borne pathogens along with their synergistic effects with synthetic drugs. *Metabol Open* **14**, 100187 (2022).
28. Krawczyk, B., Wityk, P., Gałęcka, M. & Michalik, M. The Many Faces of spp.-Commensal, Probiotic and Opportunistic Pathogen. *Microorganisms* **9**, (2021).
29. Almeida-Santos, A. C., Novais, C., Peixe, L. & Freitas, A. R. spp. as a producer and target of bacteriocins: A double-edged sword in the antimicrobial resistance crisis context. *Antibiotics (Basel)* **10** (2021).
30. Hill, C. et al. Expert consensus document: The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. *Nat. Rev. Gastroenterol. Hepatol.* **11**, 506–514 (2014).
31. Chen, J., Chen, X. & Ho, C. L. Recent development of probiotic for treating human diseases. *Front Bioeng Biotechnol* **9**, 770248 (2021).
32. Yan, J., Zhang, B., Zhou, M., Campbell-Valois, F.-X. & Siu, S. W. I. A deep learning method for predicting the minimum inhibitory concentration of antimicrobial peptides against using multi-branch-CNN and attention. *mSystems* **8**, e0034523 (2023).
33. Kang, X. et al. DRAMP 2.0, an updated data repository of antimicrobial peptides. *Sci Data* **6**, 148 (2019).
34. Patiyal, S., Dhall, A., Kumar, N. & Raghava, G. P. S. HLA-DR4Pred2: An improved method for predicting HLA-DRB1*04:01 binders. *Methods* **232**, 18–28 (2024).
35. Dhall, A., Patiyal, S., Sharma, N., Devi, N. L. & Raghava, G. P. S. Computer-aided prediction of inhibitors against STAT3 for managing COVID-19 associated cytokine storm. *Comput Biol Med* **137**, 104780 (2021).
36. Kertész-Farkas, A. et al. Benchmarking protein classification algorithms via supervised cross-validation. *J Biochem Biophys Methods* **70**, 1215–1223 (2008).
37. Li, Y. et al. Robust and accurate prediction of protein-protein interactions by exploiting evolutionary information. *Sci Rep* **11**, 16910 (2021).
38. Heinson, A. I., Ewing, R. M., Holloway, J. W., Woelk, C. H. & Niranjan, M. An evaluation of different classification algorithms for protein sequence-based reverse vaccinology prediction. *PLoS One* **14**, e0226256 (2019).
39. Ko, C. W., Huh, J. & Park, J.-W. Deep learning program to predict protein functions based on sequence information. *MethodsX* **9**, 101622 (2022).
40. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
41. Pande, A. et al. Pfeature: A tool for computing wide range of protein features and building prediction models. *J Comput Biol* **30**, 204–222 (2023).
42. Kumar, V., Patiyal, S., Dhall, A., Sharma, N. & Raghava, G. P. S. B3Pred: A random-forest-based method for predicting and designing blood–brain barrier penetrating peptides. *Pharmaceutics* **13**, (2021).
43. Agrawal, P., Mishra, G. & Raghava, G. P. S. SAMbinder: A web server for predicting S-Adenosyl-L-methionine binding residues of a protein from its amino acid sequence. *Front Pharmacol* **10**, 1690 (2019).
44. Pokharel, S., Pratyush, P., Ismail, H. D., Ma, J. & Kc, D. B. Integrating embeddings from multiple protein language models to improve protein-GlcNAc site prediction. *Int. J. Mol. Sci.* **24** (2023).
45. Elnaggar, A. et al. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* **44**, 7112–7127 (2022).
46. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding (2018).

47. Marquet, C. et al. Embeddings from protein language models predict conservation and variant effects. *Hum Genet* **141**, 1629–1647 (2022).
48. Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T. & Rost, B. Embeddings from deep learning transfer GO annotations beyond homology. *Sci Rep* **11**, 1160 (2021).
49. Lee, J. et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
50. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
51. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* **27**, 1226–1238 (2005).
52. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python (2012).
53. Peng, C.-Y.J., Lee, K. L. & Ingersoll, G. M. An introduction to logistic regression analysis and reporting. *J. Educ. Res.* https://doi.org/10.1080/00220670209598786 (2002).
54. Awad, M. & Khanna, R. Support vector regression. in *Efficient Learning Machines* 67–80 (Apress, Berkeley, CA, 2015).
55. Cule, E. & De Iorio, M. Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genet Epidemiol* **37**, 704–714 (2013).
56. Muthukrishnan, R. & Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)* (IEEE, 2016). https://doi.org/10.1109/icaca.2016.7887916.
57. Natekin, A. & Knoll, A. Gradient boosting machines, a tutorial. *Front Neurorobot* **7**, 21 (2013).
58. Armano, G. & Manconi, A. Devising novel performance measures for assessing the behavior of multilayer perceptrons trained on regression tasks. *PLoS One* **18**, e0285471 (2023).
59. Solomatine, D. P. & Shrestha, D. L. AdaBoost.RT: a boosting algorithm for regression problems. in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)* (IEEE, 2005). https://doi.org/10.1109/ijcnn.2004.1380102.
60. Zou, H. & Hastie, T. Addendum: Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67**, 768–768 (2005).
61. Vovk, V. Kernel Ridge Regression. in *Empirical Inference* 105–116 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).
62. Imane, M., Aoula, E.-S. & Achouyab, E. H. Using Bayesian ridge regression to predict the overall equipment effectiveness performance. in *2022 2nd International conference on innovative research in applied science, engineering and technology (IRASET)* 1–4 (IEEE, 2022).
63. Fijan, S., Kocbek, P., Steyer, A., Vodičar, P. M. & Strauss, M. The antimicrobial effect of various single-strain and multi-strain probiotics, dietary supplements or other beneficial microbes against common clinical wound pathogens. *Microorganisms* **10**, (2022).
64. Jin, C. et al. Predicting antimicrobial resistance in with discriminative position fused deep learning classifier. *Comput Struct Biotechnol J* **23**, 559–565 (2024).

Acknowledgements

Authors are thankful to the Council of Scientific and Industrial Research (CSIR), University Grants Commission (UGC) and Department of Bio-Technology (DBT) for fellowships and financial support, and the Department of Computational Biology, IIITD New Delhi for infrastructure and facilities. We would like to acknowledge that the Figures were created using BioRender.

Author contributions

NB collected the dataset. NB and GPSR processed the datasets. NB implemented the algorithms and developed the prediction models. NB and GPSR analyzed the results. NB created the front-end user interface, and NK created the back-end of the web server and standalone package. NB and GPSR performed the writing, reviewing and draft preparation of the manuscript. GPSR conceived and coordinated the project and gave overall supervision to the project. All authors have read and approved the final manuscript.

Funding

This work has been supported by the grant (BT/PR40158/BTIS/137/24/2021) received from the Department of Biotechnology (DBT), Govt of India, India.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-86638-z.

Correspondence and requests for materials should be addressed to G.P.S.R.