NAR Genomics and Bioinformatics, 2025, **7**, lqae186
https://doi.org/10.1093/nargab/lqae186
Advance access publication date: 7 January 2025
Standard Article

SProtFP: a machine learning-based method for functional classification of small ORFs in prokaryotes

Akshay Khanduja and Debasisa Mohanty *

National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi 110067, India

*To whom correspondence should be addressed. Tel: +91 11 26703749; Fax: +91 11 26742125; Email: deb@nii.ac.in

Abstract

Small proteins (≤100 amino acids) play important roles across all life forms, ranging from unicellular bacteria to higher organisms. In this study, we have developed SProtFP which is a machine learning-based method for functional annotation of prokaryotic small proteins into selected functional categories. SProtFP uses independent artificial neural networks (ANNs) trained using a combination of physicochemical descriptors for classifying small proteins into antitoxin type 2, bacteriocin, DNA-binding, metal-binding, ribosomal protein, RNA-binding, type 1 toxin and type 2 toxin proteins. We have also trained a model for identification of small open reading frame (smORF)-encoded antimicrobial peptides (AMPs). Comprehensive benchmarking of SProtFP revealed an average area under the receiver operator curve (ROC-AUC) of 0.92 during 10-fold cross-validation and an ROC-AUC of 0.94 and 0.93 on held-out balanced and imbalanced test sets. Utilizing our method to annotate bacterial isolates from the human gut microbiome, we could identify thousands of remote homologs of known small protein families and assign putative functions to uncharacterized proteins. This highlights the utility of SProtFP for large-scale functional annotation of microbiome datasets, especially in cases where sequence homology is low. SProtFP is freely available at http://www.nii.ac.in/sprotfp.html and can be combined with genome annotation tools such as ProsmORF-pred to uncover the functional repertoire of novel small proteins in bacteria.

Introduction

Challenges in genome annotation have hindered the identification and functional characterization of small open reading frames (smORFs) (≤100 amino acids). However, these are being increasingly discovered in eukaryotes (1,2), prokaryotes (3–5) and viruses (6). Recently, 114 new smORFs were identified in *Streptococcus pneumoniae*, including smORFs implicated in virulence and quorum sensing (5). Similarly, >100 unannotated smORFs have been discovered in different strains of *Salmonella enterica* (7,8).

Increasing numbers of smORFs are being functionally characterized experimentally. For example, a conserved smORF family regulates the Ca^{2+} uptake pump SERCA (9) in vertebrates. Amongst prokaryotes, a previously known small RNA encodes a type 1 toxin of 38 amino acids in *S. enterica* (10). YnfU (56 amino acids), discovered in a ribo-seq-based study, had zinc knuckle motifs indicating a possible metal-binding function (4). Another well-known regulator of mRNA translation CsrA (61 amino acids) has an RNA-binding role (11).

Experimental methods for functional characterization of small proteins are resource- and time-intensive. While most studies characterize individual small proteins through knockout or overexpression, high-throughput interaction screens and CRISPR-based screens have been used to assign functions based on interacting partners (12) and the effect on growth/fitness (13), respectively. Though such studies might implicate a small protein in a particular pathway, they may not always indicate its molecular function.

Computational methods are convenient alternatives to experimental methods while functionally annotating a large number of proteins. These methods can be sequence based (14), structure based (15) or expression data based (16). Sequence homology search using BLAST (17) and profile-HMMs (hidden Markov models) and structure-based comparison methods such as DALI (18) and CATHEDRAL (19) have been traditionally used for assigning functions. More recently, sequence- and structure-based approaches in combination with machine learning (ML) have been used for general function prediction in the context of enzymatic activity (20–22) and Gene Ontology (23–25). While several methods exist for general protein function prediction, only a few studies have focused on the annotation of small proteins which often play specialized roles (26). Two such methods based on co-expression data are available, namely FSPP (27) and smORFunction (28). However, these tools are limited by the availability of expression data and focus on human smORFs. In the context of prokaryotic smORFs, a metagenomic data-based protein assembly (metaBP) pipeline and ML annotation tool (metaBP-ML) were developed to profile the small proteome of bacteria (29). MetaBP-ML annotates proteins based on a K-nearest neighbor algorithm by choosing the most frequent annotation among its neighbors in embedding space. However, the study focused on predicting the Enzyme Commission (EC) numbers and functions at the level of individual small proteins instead of modeling broad functional categories having different protein types (e.g. small DNA-binding proteins, bacteriocins, type 1 toxins, small metal-binding proteins, etc.)

Received: August 16, 2024. **Revised:** November 7, 2024. **Editorial Decision:** December 16, 2024. **Accepted:** December 17, 2024

© The Author(s) 2025. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

(29). Hence, lack of automated methods for functional assignment of smORFs in bacteria necessitates the development of novel computational methods to investigate their functional repertoire.

We had previously developed ProsmORF-pred (30) which predicts smORFs in prokaryotic genomes. In this study, we focused on functional characterization and, to automate the task of functional assignment, we have developed an ML-based method called SProtFP (Small Protein Function Predictor) which assigns small proteins from bacteria to eight functional categories, namely antitoxin type 2, bacteriocin, DNA-binding, metal-binding, ribosomal protein, RNA-binding, type 1 toxin and type 2 toxin using sequence-based physicochemical properties. It was encouraging to note that SProtFP achieved an average 10-fold cross-validation ROC-AUC (area under the receiver operator curve) of 0.92 and average test ROC-AUC values of 0.94 and 0.93 on balanced and imbalanced test sets, respectively. To highlight the utility of our approach in annotating sequence data from large-scale microbiome projects, all the ML models in SProtFP were tested on 4 359 796 predicted smORFs from 6503 bacterial isolates (885 species) from the Unified Human Gastrointestinal Genome (UHGG) catalog (31). Modeling the small protein structures of the predictions and comparing them with known small proteins revealed that SProtFP can assign putative functions to small proteins even when sequence homology is low, and hence can be used to discover remote homologs of known proteins as well as novel proteins belonging to different functional categories. Additionally, to predict antimicrobial peptides (AMPs) besides known bacteriocins, we also developed a general AMP predictor and used it to mine the gut microbiome for small ORF-encoded putative AMPs.

Materials and methods

Datasets

For developing the ML classifiers, data were collected from SwissProt (32), BAGEL4 (33) and T1TAdb (34). Small 'Reviewed' bacterial proteins (10–100 amino acids) from UniProt (32) were considered, and keywords associated with them were extracted. Based on the frequency of the keywords, type of function and manual inspection, eight functional categories were chosen, namely antitoxin type 2, bacteriocin, DNA-binding, metal-binding, ribosomal protein, RNA-binding, type 1 toxin and type 2 toxin (Figure 1; Supplementary Table S1) (for details, see the Materials and methods in the Supplementary Information). Only class 1 and class 2 bacteriocins in BAGEL4 and only the type 1 toxins belonging to known categories from the literature in T1TAdb were considered and merged with SwissProt data (Figure 2). It was observed that the small proteins could belong to more than one category so we decided to make independent classifiers. The positive dataset (Supplementary Table S2) was derived from the clustered representatives [after clustering using CD-HIT (35) v4.8.1] of small proteins belonging to a category (Supplementary File S1) while the negative dataset was randomly drawn from clustered representatives of small proteins that did not belong to that category (for details, see the Materials and methods in the Supplementary Information). The train/test split was 80%/20% for the positive datasets of all categories except for RNA-binding (90% training) and type 2 toxins (100% training) (Supplementary Table S2; Supplementary Files S2 and S3). We also explored the effect of undersampling on the model performance by drawing negative training datasets of different sizes (equal, 4× and 8×).

Selection of protein sequence-based descriptors

The Python-based iLearn (36) toolkit was used to compare different feature sets (Figure 2) that could best represent the training datasets using random forest in Weka (37) v3.8 as the base classifier. A variety of different feature sets were used to train the models using balanced training datasets and evaluated using 10-fold cross-validation (for details, see the Materials and methods in the Supplementary Information).

Selection of appropriate machine learning algorithms

Weka v3.8 (37) was used to choose the appropriate ML and deep learning architectures by training different ML classifiers using balanced datasets (Figure 2). The models were evaluated using 10-fold cross-validation (for details, see the Materials and methods in the Supplementary Information).

Evaluation of undersampling for model improvement

Once the feature sets and ML architectures/algorithms were evaluated, the best feature sets in combination with the best ML architecture were used to train three different models (1:1, 1:4 and 1:8) for each functional category by using the same positive training dataset with a different number of negative instances (Figure 2) (for details, see the Materials and methods in the Supplementary Information).

Validation and selection of the final machine learning models

Finally, the best models trained on the optimal ratio of positive and negative data based on 10-fold cross-validation were selected as the final ML models for each category. The ROC-AUC and precision–recall AUC (PR-AUC) were computed using PRROC (38). Two different sets of cut-offs (relaxed and restrictive cut-offs) were selected for the final selected ML models corresponding to the 'relaxed' and 'restrictive' implementation of SProtFP. The code for SProtFP is available at http://www.nii.ac.in/sprotfp.html and in Supplementary File S4.

Testing ML models on held-out test sets

The final selected ML models for each functional category (except for type 2 toxin proteins) were tested on balanced (1:1) and imbalanced (1:10) test sets not used during training and evaluated using ROC curves (for details, see the Materials and methods in the Supplementary Information).

Predicting functional classes of smORFs in human gut microbiome

To highlight the utility of our ML models with respect to large-scale functional characterization of the uncharacterized proteins in microbiome samples, we chose 6503 non-redundant [Mash (39) distance threshold of 0.001, corresponding to a nucleotide identity of 99.9%] bacterial isolates (Supplementary File S5) belonging to 885 different species from the UHGG resource (31). In order to exhaustively predict smORFs, we took the union of the predictions given by ProsmORF-pred (30) and METAPRODIGAL (40) for these

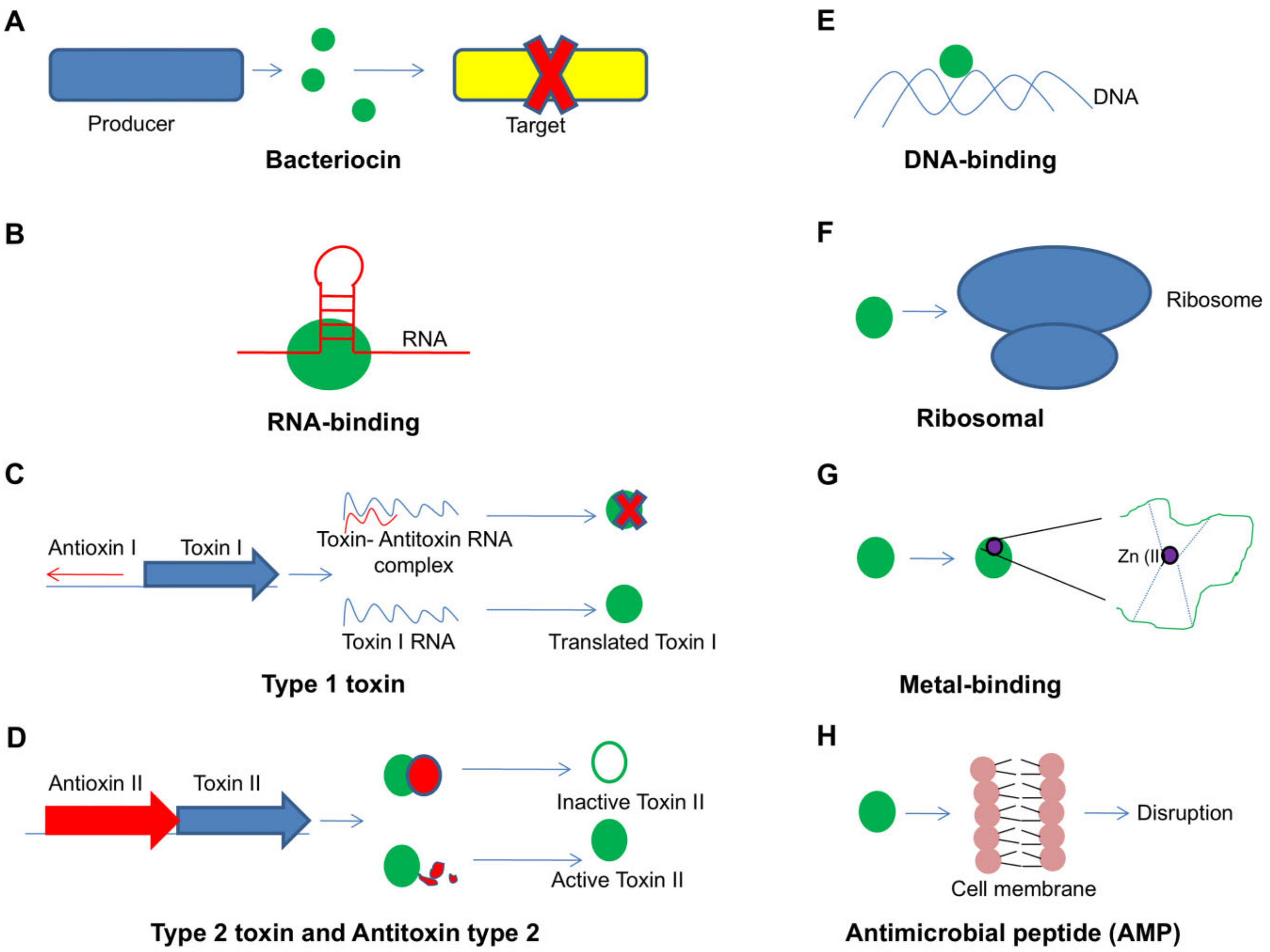


Figure 1. Functional categories of small proteins modeled in the current study.

genomes. To restrict the number of false positives while predicting the functions of the huge number of microbiome-predicted smORFs, restrictive cut-offs were utilized for the final selected ML models for the different categories (for details, see the Materials and methods in the Supplementary Information).

Structural comparison of predicted and known small proteins

To search for structural homologs of known small proteins from different functional categories present amongst our predictions, small proteins in the positive datasets and predictions by SProtFP for all functional categories for the UHGG (31) isolates were modeled using ESMFold (41) and compared using Foldseek (42) (for details, see the Materials and methods in the Supplementary Information).

Training and benchmarking the antimicrobial peptide predictor

AMPs were downloaded from APD3 (43) 2020 release and used for training the AMP predictor employing the selected best feature set and ML architecture (Figure 1). The positive and negative datasets were then split into 80:20 ratios for training and testing (Supplementary File S6) (for details, see the Materials and methods in the Supplementary Information).

Results

Representation of the selected protein categories among SwissProt small proteins

As indicated in the Materials and methods, data were collected from SwissProt (32), BAGEL4 (33) and T1TAdb (34) to build ML classifiers for eight functional categories of smORFs, namely antitoxin type 2, bacteriocin, DNA-binding, metal-binding, ribosomal protein, RNA-binding, type 1 toxin and type 2 toxin (Figures 1 and 2). After filtering the data from SwissProt for sequence issues and fragments, we found 29 051 bacterial small proteins in the length range 10–100 amino acids. The break down of the total number of small proteins belonging to each selected functional category in SwissProt is provided in Supplementary Table S2. It was found that the proteins belonging to the selected categories collectively constitute ∼57% (16 431/29 051) of the total bacterial small proteins in SwissProt. The remaining 43% did not fall under the selected categories.

Sequence-based descriptors capture physicochemical properties of small protein categories

Selection of appropriate features for representing the data is an important task in ML. Structure determination of small proteins may require specialized methods (44) and thus, unlike

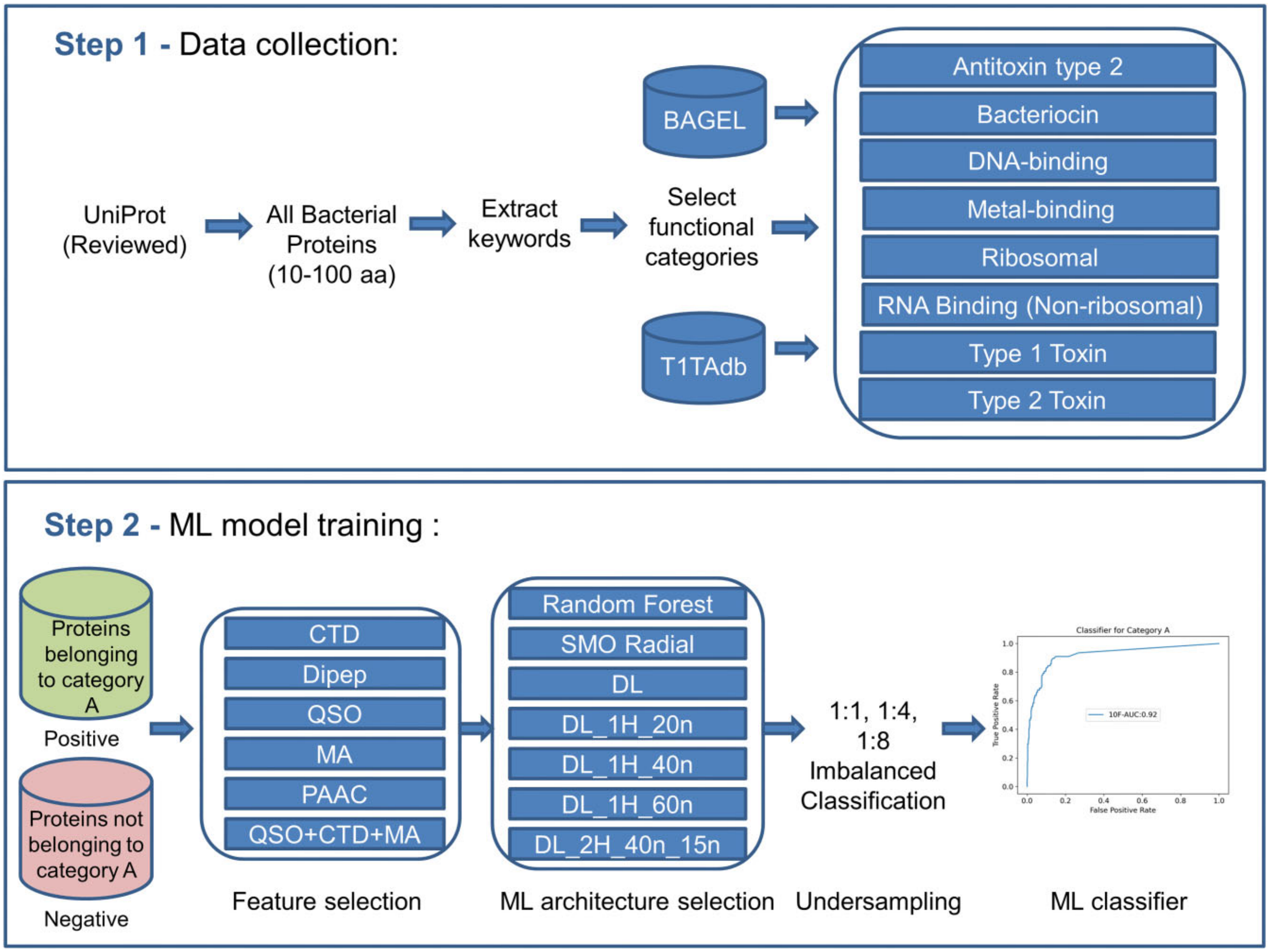


Figure 2. Methodology used for developing SProtFP.

sequence information, structures of these proteins may not be readily available. Hence, we utilized different sequence-based descriptors to train and evaluate ML models using equal-sized positive and negative training datasets and random forest as the base classifier. We used features that capture the composition and physicochemical properties of the protein sequences, i.e. dipeptide frequency (Dipep) (45), Moran autocorrelation (MA) (46), composition, transition and distribution (CTD) (46) features, quasi-sequence-order (QSO) (47) and pseudo amino acid composition (PAAC) (48) (Figure 2). It is seen through 10-fold cross-validation that most feature sets efficiently discriminated the negative and positive instances, except MA (Supplementary Figure S1; Supplementary Table S3; Supplementary File S7). Features such as CTD, Dipep, QSO and PAAC were able to achieve high ROC-AUC values for all the categories. The combination of different types of features such as CTD, QSO and MA (QSO + CTD + MA) gave the best performance overall in terms of the ROC-AUC values, achieving an average 10-fold ROC-AUC of 0.90 and a PR-AUC of 0.90. We chose ROC-AUC as the metric to select the feature set as it depicts the performance of the models across the range of different cut-off scores. All the other measures such as true positive rate (TPR), false positive rate (FPR), precision and MCC (Matthews correlation coefficient) indicated high accuracy of the models (Supplementary Table S4). Though MA did not perform well alone, we added it to increase the descriptive range of the feature sets. After selecting QSO + CTD + MA as our features, we explored several ML and deep learning architectures.

Evaluation of different machine learning algorithms

The choice of ML algorithms plays an important role in the performance of the models. Popular ML algorithms include random forest and support vector machines (SVM) (49–51). Alternatively, deep learning techniques including artificial neural networks (ANNs) are known to provide higher accuracies as compared with traditional ML methods (52). Therefore, we used SMO with radial kernel (SVM), random forest and variations of the Dl4jMlpClassifier available with Weka (37) 3.8v to train ML models on the positive and negative training datasets used earlier (equal set sizes) (Figure 2) utilizing QSO + CTD + MA as the feature set. As seen in 10-fold cross-validation ROC curves in Supplementary Figure S2, Supplementary Table S5 and Supplementary File S8, all the ML algorithms performed well. The ANN with one hidden layer and 40 nodes (DL_1H_40n) had the highest average 10-fold ROC-AUC of 0.91. Interestingly, the predictive power of the models increases with the addition of hidden layers as compared with the base Dl4jMlpClassifier (DL). This is expected as the model without hidden layers is likely to behave as a linear model. We observed that increasing the number of nodes beyond 40 (DL_1H_60n) and adding a second hidden layer (DL_2H_40n_15n) did not lead to any significant

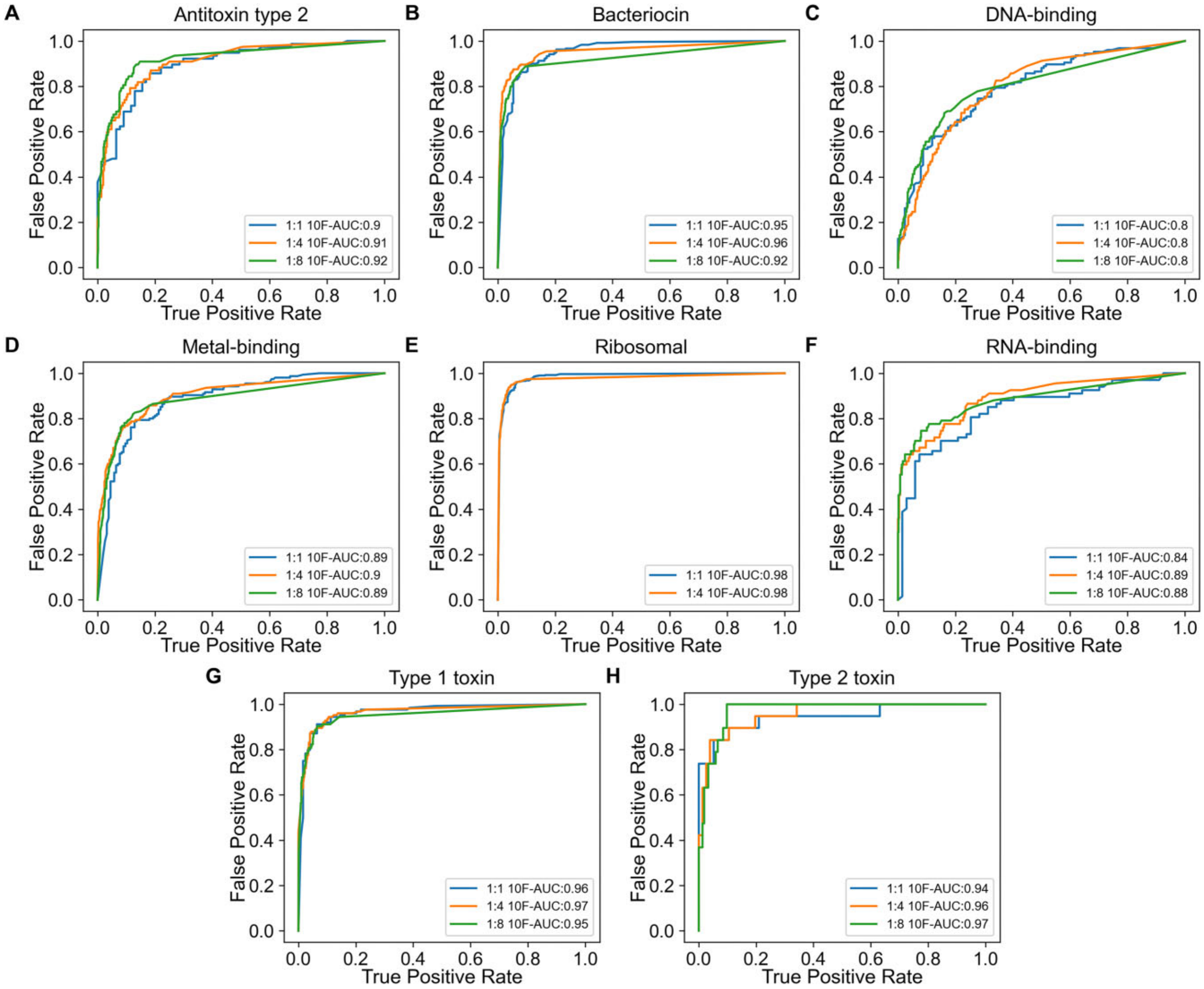


Figure 3. Effect of undersampling on model performance assessed using 10-fold cross-validation ROC curves. 1:1 refers to equal sized positive and negative training datasets, while 1:4 and 1:8 refer to imbalanced training datasets having four and eight times more negative training instances (**A–H**).

improvement in model performance. The DL_1H_40n architecture achieved an average 10-fold ROC-AUC of 0.91 and PR-AUC of 0.90 (Supplementary Table S6). Therefore, we selected this architecture as our final ML training architecture.

Undersampling of the negative dataset improves model performance

Most ML classification tasks involve positive datasets of much lower sizes as compared with the negative datasets. Examples of such tasks include healthy/diseased classification (53), drug toxicity prediction (53), splice site recognition (54), etc. After selecting the appropriate feature set and ML architecture, we wanted to explore if increasing the training dataset size could improve model performance. Undersampling of the negative datasets and the Synthetic Minority Oversampling Technique (SMOTE) (55) are two major ways of dealing with imbalanced classification problems. Since SMOTE augments the data by creating artificial points and its accuracy is known to be limited (55), we decided to undersample the negative datasets to create training datasets in the ratios 1:1, 1:4 and 1:8. Figure 3 shows the performance of these models trained on balanced and imbalanced datasets in the form of 10-fold cross-validation ROC curves. We selected the best performing models for each functional category as our final models and selected relaxed and restrictive cut-offs for the ML models based on 10-fold cross-validation (for details, see the Materials and methods). The cut-offs correspond to the statistical metrics shown in Supplementary Tables S7 and S8. It is seen that the final selected models have an average 10-fold ROC-AUC (Figure 3; Supplementary Tables S7 and S8) of 0.92 and a PR-AUC (Supplementary Figure S3; Supplementary Tables S7 and S8) of 0.78. Comparing the selected models with models trained on balanced datasets, we found that the selected models achieved a higher TPR (Figure 3) at the same FPR. The selected models achieved an average TPR (10-fold cross-validation) of 0.87 at an average FPR of 0.11 at relaxed cut-offs as compared with an average TPR of 0.80 at the same FPR in the case of models trained on balanced datasets (Figure 3; Supplementary Table S7). Similarly, using restrictive cut-offs, the selected models achieved an average TPR (10-fold cross-validation) of 0.72 at an average FPR of 0.05 as compared with an average TPR of 0.66 at the same average FPR in the case of models trained on balanced datasets (Figure 3; Supplementary Table S8). The possible reason for the increase in sensitivity could be that by providing more negative data as

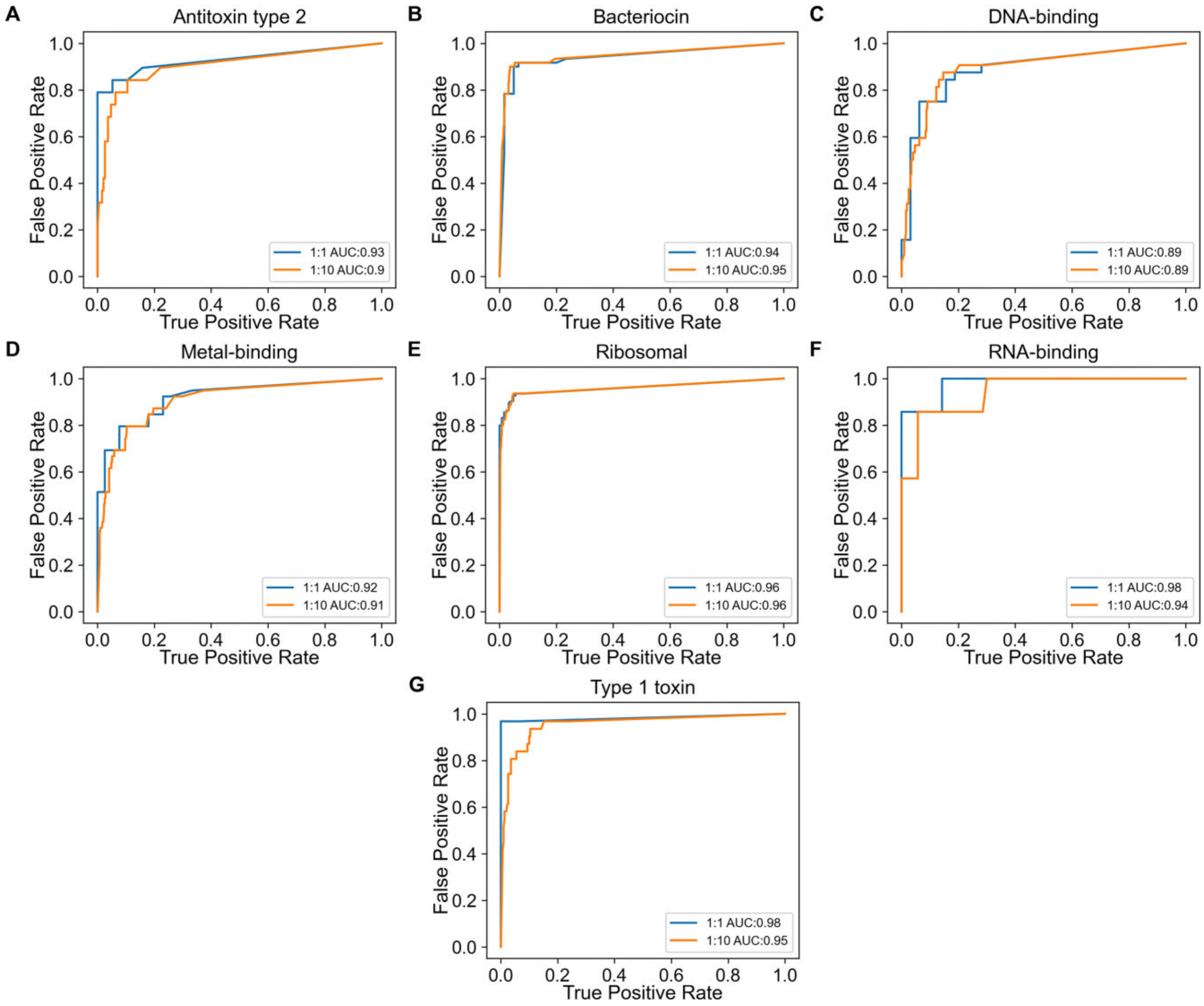


Figure 4. Testing results on held-out balanced (1:1) and imbalanced (1:10) testing datasets. 1:1 refers to the test set having an equal number of positive and negative testing instances, while 1:10 refers to the test set having 10 times more negative test instances as compared with the positive set (**A–G**). Type 2 toxin proteins could not be tested independently due to small dataset size.

compared with positive data, we are able to increase the training dataset size which enhances the models' ability to distinguish positive and negative instances. After selecting the best models, we wanted to assess the models' performance on the held-out test datasets to obtain additional validation on external data not used during training.

Testing the models on held-out test sets

Testing on held-out datasets is essential to ascertain that the models are not overfitting to the training data. Though our cross-validation results indicate high predictive power, we used additional data to test the ML models. As discussed in the Materials and methods, for all the functional categories except RNA-binding proteins and type 2 toxins, we used 80% of the positive dataset for training and 20% for testing (Supplementary Table S2). In the case of RNA-binding proteins, 10% of the positive dataset was selected due to limited data, and in the case of type 2 toxins, we could not test independently due to the low number of positive instances. Nevertheless, it can be concluded that both these prediction models have good predictive power as these models achieve high 10-fold and 2-fold cross-validation ROC-AUC values, displaying 2-fold cross-validation ROC-AUC values of ∼0.9 (Supplementary Tables S7 and S8). We used both equal (1:1) and imbalanced (1:10) negative test sets for benchmarking model performance (for details, see the Materials and methods). Imbalanced datasets were used as, in practical use cases, there will always be a biased distribution of positive and negative instances. Figure 4 shows the test ROC curves of the classifiers for different categories. It was seen that on the balanced (1:1) test sets, the models had an average ROC-AUC of 0.94 and a PR-AUC of 0.95, while the models achieved an average ROC-AUC of 0.93 and PR-AUC of 0.7 on the imbalanced (1:10) test sets (Tables 1 and 2). The models when tested using relaxed and restrictive cut-offs (for details, see the Materials and methods) on the imbalanced test datasets (1:10) had an average sensitivity of 87% at a specificity of 89% (Table 1) and an average sensitivity of 76% at a specificity of 94% (Table 2), respectively. It was encouraging to note that the model trained to predict antitoxin type 2 proteins achieved a sensitivity of 0.84 (84%) at an FPR of 0.11 (11%) on the imbalanced test dataset using relaxed cut-offs, and a sensitivity of 79% at 7% FPR using restrictive cut-offs. Similarly, the bacteriocin

Table 1. Performance of the final selected ML models on the held-out test datasets (balanced and imbalanced) using relaxed cut-offs

Category	Test set type	Positive	Negative	FPR	TPR	ROC-AUC	PR-AUC	Precision	MCC
Antitoxin type 2	1:1	19	19	0.05	0.84	0.93	0.95	0.94	0.79
	1:10	19	190	0.11	0.84	0.9	0.66	0.43	0.55
Bacteriocin	1:1	60	60	0.1	0.92	0.94	0.94	0.9	0.82
	1:10	60	600	0.06	0.92	0.95	0.78	0.6	0.71
DNA-binding	1:1	32	32	0.28	0.91	0.89	0.88	0.76	0.64
	1:10	32	320	0.2	0.91	0.89	0.51	0.31	0.46
Metal-binding	1:1	39	39	0.18	0.79	0.92	0.93	0.82	0.62
	1:10	39	390	0.17	0.79	0.91	0.57	0.32	0.43
Ribosomal	1:1	124	124	0.03	0.9	0.96	0.97	0.97	0.87
	1:10	124	1240	0.04	0.9	0.96	0.88	0.69	0.76
RNA-binding	1:1	7	7	0	0.86	0.98	0.98	1	0.87
	1:10	7	70	0.11	0.86	0.94	0.76	0.43	0.55
Type 1 toxin	1:1	31	31	0	0.9	0.98	0.99	1	0.91
	1:10	31	310	0.1	0.9	0.95	0.74	0.47	0.6

Table 2. Performance of the final selected ML models on the held-out test datasets (balanced and imbalanced) using restrictive cut-offs

Category	Test set type	Positive	Negative	FPR	TPR	ROC-AUC	PR-AUC	Precision	MCC
Antitoxin type 2	1:1	19	19	0.05	0.79	0.93	0.95	0.94	0.75
	1:10	19	190	0.07	0.79	0.90	0.66	0.52	0.6
Bacteriocin	1:1	60	60	0.05	0.78	0.94	0.94	0.94	0.74
	1:10	60	600	0.03	0.78	0.95	0.78	0.72	0.73
DNA-binding	1:1	32	32	0.22	0.88	0.89	0.88	0.8	0.66
	1:10	32	320	0.15	0.88	0.89	0.51	0.37	0.51
Metal-binding	1:1	39	39	0.03	0.69	0.92	0.93	0.96	0.69
	1:10	39	390	0.07	0.69	0.91	0.57	0.48	0.53
Ribosomal	1:1	124	124	0.01	0.8	0.96	0.97	0.99	0.81
	1:10	124	1240	0.01	0.8	0.96	0.88	0.87	0.82
RNA-binding	1:1	7	7	0	0.71	0.98	0.98	1	0.75
	1:10	7	70	0.06	0.71	0.94	0.76	0.56	0.59
Type 1 toxin	1:1	31	31	0	0.65	0.98	0.99	1	0.69
	1:10	31	310	0.03	0.65	0.95	0.74	0.71	0.65

prediction model had a TPR of 0.92 at an FPR of 0.06 and a TPR of 0.78 at an FPR of 0.03 on the imbalanced test dataset. Models for other categories also show high predictive power (Tables 1 and 2; Figure 4) while the DNA-binding model had a slightly lower test ROC-AUC of 0.89. Overall, our analysis indicates that neural networks trained on physicochemical properties of small proteins can help in classifying these proteins into functional families. Next, we wanted to explore the utility of our method for large-scale functional annotation of smORFs in the human gut microbiome.

Function inference for smORFs of the human gut microbiome

One of the advantages of ML models over traditional alignment-based methods is that the former are in general faster and can thus be used for functional annotation of large volumes of data. Hence, we used SProtFP to functionally annotate the smORF predictions from the genomes corresponding to 6503 non-redundant isolates of bacteria (885 species) in the UHGG catalog (31). The taxonomy distribution of the genomic isolates is shown in Supplementary Figure S4. It was seen that Proteobacteria had the maximum sequenced isolates (45.7%), followed by Firmicutes (29%), Bacteriodota (10.1%) and others. Taking the union of smORF predictions from ProsmORF-pred (30) and METAPRODIGAL (40) across these isolates (for details, see the Materials and methods) resulted in 4 359 796 smORFs (Supplementary File S9). From this set, we functionally classified the small proteins into different functional classes using SProtFP (for details, see the Materials and methods). Independently, the ML models classified ∼6% of the total predictions as antitoxin type 2, 3% as bacteriocins, 15% as DNA-binding, 11% as metal-binding, 4% as ribosomal, 6% as RNA-binding, 7% as type 1 toxin and 6% as type 2 toxin proteins (Table 3; Supplementary File S10). Collectively, the models provided functional annotation for 1 996 704 (45.80%) of the smORFs predicted in the microbiome set. Although tools such as BLAST (17) may fail to detect homologous hits when sequences are smaller or homology is remote, a simple homology-based analysis of our predictions against known proteins can give us some idea about the sensitivity of our models with respect to currently available functionally annotated small proteins. We used relatively relaxed parameters (40% identity, 80% query coverage cut-offs and subject length filter of 80–120% of the query) to identify both close and distant homologs. As shown in Table 3, on analyzing the homology of the predictions for each category against known small proteins (used for training and testing—positive datasets) (Supplementary Table S2) using BLASTP (17), we found that the hits of the predicted proteins for most categories represented a significant proportion (>70%) of known small proteins in the positive datasets. Since the sequences in the positive datasets represent different families or clusters according to CD-HIT (35) as they have been clustered, this indicates that our models efficiently capture the protein feature space of known data. As protein structures are known to be more conserved as compared with their sequences, structural modeling allows the detection of

Table 3. Category wise predictions by SProtFP for the 4 359 796 predicted smORFs from the UHGG dataset

Category	smORFs predicted for category	smORFs predicted for category (as % of total 4 359 796 smORFs)	No. of predictions homologous to positive dataset (known proteins) using BLASTP (40% identity, 80% coverage)	% represented by hits of the predictions against positive dataset (known proteins) (BLASTP)	No. of predictions homologous to positive dataset using Foldseek	% represented by hits of the predictions against positive dataset (known proteins) (Foldseek)
Antitoxin type 2	257 306	5.9%	12 835	71.88% (69/96)	25 378	80.21% (77/96)
Bacteriocin	133 982	3%	963	37% (111/300)	39	3.3% (10/300)
DNA-binding	640 978	14.7%	87 443	81.65% (129/158)	139 377	94.94% (150/158)
Metal- binding	492 132	11.28%	42 568	58.25% (113/194)	55 320	64.95% (126/194)
Ribosomal	166 703	3.82%	105 044	94.2% (585/621)	94 865	90.5% (562/621)
RNA-binding	250 061	5.74%	37 266	97.3% (72/74)	36 740	94.59% (70/74)
Type 1 toxin	289 598	6.64%	28 456	78.1% (121/155)	12 609	14.19% (22/155)
Type 2 toxin	251 828	5.77%	8 254	73.68% (14/19)	13 251	89.47% (17/19)

remote homologs not detected by conventional sequence-based methods (56). Therefore, we wanted to look for structural homologs of known proteins among our predictions. Comparing the sequences of the representatives of known small proteins (positive datasets) against the PDB (57) database (January 8, 2024), we observed that <60% of known small proteins from most categories showed at least a hit using the same set of BLASTP filters as used earlier (Supplementary Table S9). This indicates that these functional categories are not adequately represented in the PDB. As a result, we decided to structurally model the three-dimensional structures of known proteins and compare them with the predictions (for details, see the Materials and methods). Since tools such as ESMFold (41) provide opportunities for accurate and large-scale structural modeling of proteins, this tool was chosen for structural modeling. As is seen in Figure 5C and Supplementary Table S10, all the classes of known proteins except bacteriocins were modeled with a high pLDDT (41) score which is an indicator of the quality of the model predictions. Foldseek (42) was used for structural alignment of the modeled three-dimensional structures of the predictions against the modeled structures of known small proteins (Figure 5). As seen in Table 3, compared with the traditional sequence-based homology search, Foldseek detected either a higher or a comparable number of homologs for the majority of the functional categories. Foldseek identified ∼98, 59, 30 and 61% more homologs amongst the predictions belonging to antitoxin type 2, DNA-binding, metal-binding and type 2 toxins, respectively. Moreover, the hits found via Foldseek against the predictions for these categories also represented a higher proportion of known proteins as compared with the BLASTP search (Table 3). As additional homologs were identified by Foldseek despite using relaxed BLASTP identity and query coverage cutoffs, these additional homologs are likely to have remote homology to known proteins. Both BLASTP and Foldseek identified a similar number of homologs and captured an equivalent proportion of known protein clusters in the case of RNA-binding and ribosomal proteins, while Foldseek could not identify more structural homologs for bacteriocins and type 1 toxin proteins due to the challenges in modeling bacteriocin structures (Figure 5C; Supplementary Table S10) and aligning small length type 1 toxin (Figure 5D) structures. Overall, the maximum (BLASTP/Foldseek) proportion/diversity of known proteins represented by the hits of the predictions for different categories was 80.21% (antitoxin type 2), 37% (bacteriocin), 94.94% (DNA-binding), 64.95% (metal-binding), 94.2% (ribosomal), 97.3% (RNA-binding), 78.1% (type 1 toxin) and 89.47% (type 2 toxin). This is encouraging considering the fact that the known proteins in the positive datasets of all small protein categories have been clustered at 50% identity and 80% length difference parameters using CD-HIT (35). Moreover, our models are able to classify a smaller subset of small proteins in the microbiome as putative members of the selected categories. In the process, these independently discard ∼94.1% (antitoxin type 2 model), 97% (bacteriocin model), 85.3% (DNA-binding model), 88.72% (metal-binding model), 96.18% (ribosomal model), 94.26% (RNA-binding model), 93.36% (type 1 toxin model) and 94.23% (type 2 toxin model) of the total small proteins on the initial list (Table 3). Therefore, our models achieve considerable enrichment with respect to the known functional proteins while eliminating a huge number of small proteins. The lowest proportion of known proteins represented among the predictions was ∼37% for the bacteriocin category. A possible reason for this could be the lower length range of these proteins (Figure 5D) which makes homology detection difficult using BLASTP-like sequence-based approaches or their taxonomic restriction to specific species or strains (58,59). Analyzing the taxonomic distribution of the homologs (Supplementary Figure S5) revealed that homologs of known toxin–antitoxin proteins are predominantly found in Proteobacteria while bacteriocins are mostly derived from Firmicutes. Other categories including DNA-binding, metal-binding, RNA-binding and ribosomal small protein homologs are found across the bacterial phyla in the UHGG dataset (Supplementary Figures S4 and S5) which indicates that these may not be particularly taxa specific, and hence are more generally distributed.

Interestingly, our predictions include homologs of known small proteins relevant to the gut microbiome. GUT_GENOME103691_38#16 886_17 113_- is a homolog of Circularin A, a known broad antibacterial range bacteriocin found in conjunction with ABC transporters (60). Interestingly, the producer of this bacteriocin is *Bifidobacterium infantis* which is a known probiotic isolate (61). Hence, bacteriocins such as Circularin may contribute to its probiotic potential. GUT_GENOME096459_183#19 506_19 628_+ is a bL36 ribosomal protein homolog of only 40 amino acids which plays a housekeeping role across members of the microbiome as part of the translational machinery. GUT_GENOME095941_11#55 204_55 437_+ which is a YacG protein homolog, is a metal-binding protein that protects DNA gyrase from the antimicrobial agents targeting it and plays a role in resistance to novobiocin (62). GUT_GENOME239652_1#407 085_407 309_- is a homolog

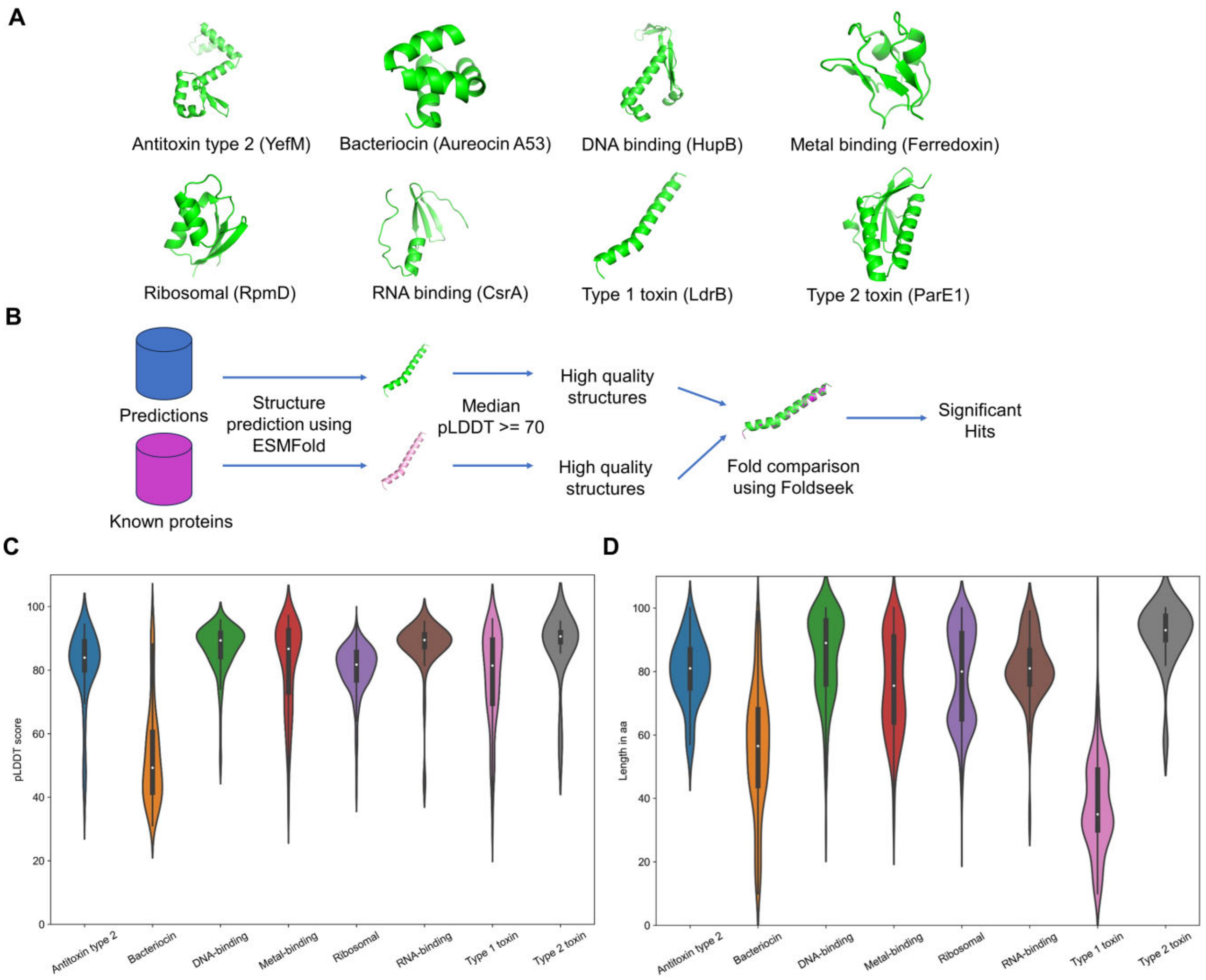
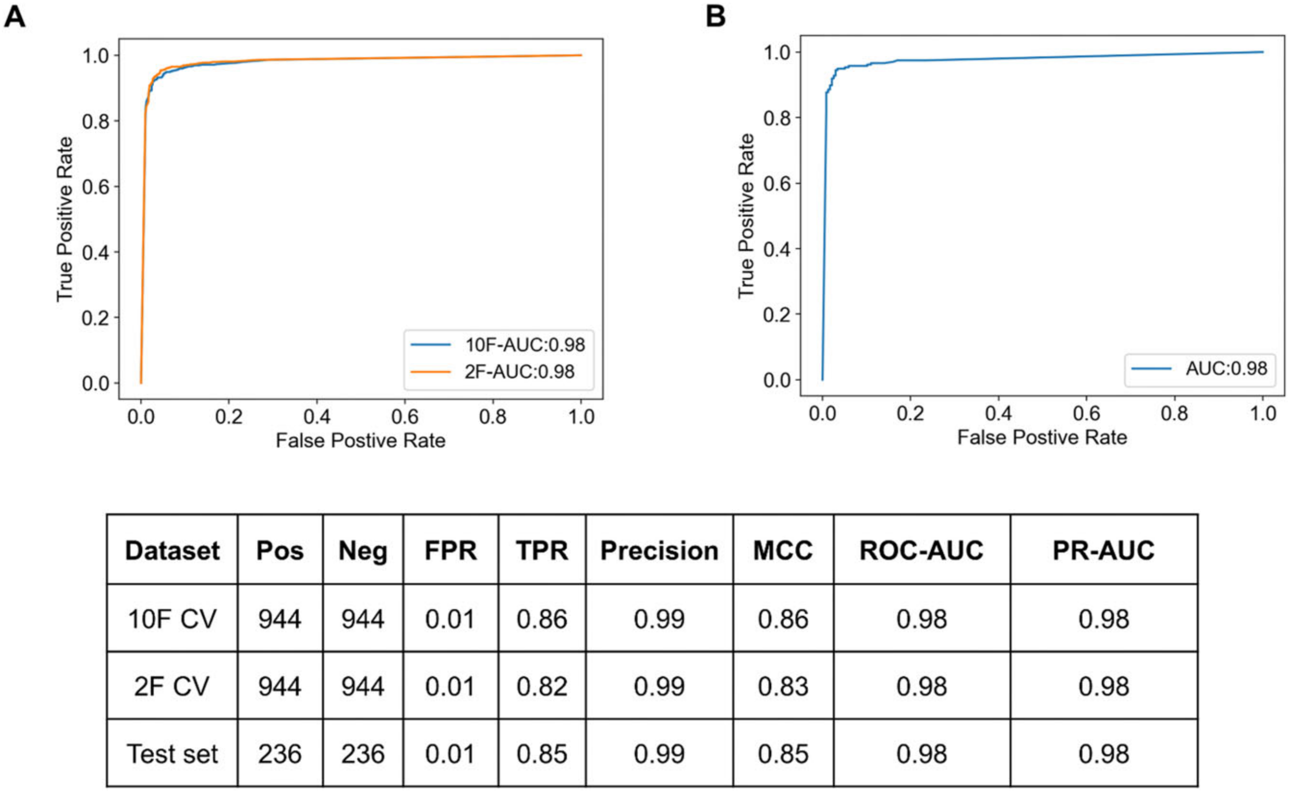


Figure 5. Workflow for structural comparison of the predictions against known proteins. (**A**) ESMFold modeled structures of known proteins from different small protein categories. (**B**) Workflow for structural comparison. (**C**) pLDDT score distribution of known small proteins (Supplementary Table S2) in the positive datasets of the respective categories. (**D**) Length distribution of known small proteins in the positive datasets of the respective categories.

of small acid-soluble spore proteins encoded by some bacteria belonging to the abundant Firmicutes phylum in the gut microbiota (63). Their main role is to protect the endospores against DNA damage by binding to DNA (64). The 66 amino acid GUT_GENOME144635_2#177 604_177 804_- is a homolog of CspD which is a stress-responsive member of the cold shock protein family capable of binding to RNA and single-stranded DNA (65). GUT_GENOME145561_30#24 127_24 267_- is a 46 amino acid homolog of the Hok family of type 1 toxins whose members are known to participate in cell lysis and bacterial persistence (66). GUT_GENOME000052_50#35 051_35 329_+ and GUT_GENOME000052_50#35 326_35 607_+ correspond to the DinJ–YafQ type 2 toxin–antitoxin system proteins where the YafQ toxin has nuclease activity and is capable of cleaving mRNA and inhibiting translation while DinJ counteracts its activity (67). These examples highlight that our predictions harbor homologs of known small proteins belonging to the modeled categories which can play a role in the overall bacterial survival and competition among the gut microbiome members.

Identifying smORF-encoded putative AMPs from the human gut microbiome

More recently, studies focused on mining AMPs at the microbiome scale (68–70) suggest that more AMPs remain to be discovered in bacteria besides the known bacteriocins. To test the validity of our bacteriocin model for predicting AMPs, we tested the bacteriocin model on experimentally validated AMPs from two recent microbiome-based studies (69,70). The model predicted 46/128 (35.94%) and 21/128 (16.41%) of these as positive using the relaxed and restrictive cut-off, respectively, despite showing high accuracy on the bacteriocin set (Figures 3 and 4). This suggests that these AMPs have different sequences or properties as compared with the known conventional bacteriocins. While class 1 and class 2 bacteriocins are encoded by bacteria and form a subset of the known AMPs (71,72), AMPs also include other peptides such as defensins, cathelicidin, etc. from other organisms. In line with this, our BLASTP (17) search of the experimentally validated AMPs against the dataset used for bacteriocin model training did not find any hits (40% identity and 80% query coverage). Hence, we hypothesized that the AMPs in these studies might



Dataset	Pos	Neg	FPR	TPR	Precision	MCC	ROC-AUC	PR-AUC
10F CV	944	944	0.01	0.86	0.99	0.86	0.98	0.98
2F CV	944	944	0.01	0.82	0.99	0.83	0.98	0.98
Test set	236	236	0.01	0.85	0.99	0.85	0.98	0.98

Figure 6. Performance of the machine learning model trained for predicting antimicrobial peptides (AMPs). (**A**) 10F (10-fold) and 2F (2-fold) cross-validation ROC curves. (**B**) ROC curve depicting the performance of the AMP model on the held-out test set. Additionally, the table at the bottom shows statistical metrics for the ML models such as FPR (false positive rate), TPR (true positive rate), MCC (Matthews correlation coefficient), precision, etc. for both the cross-validation and the test set at the optimum cut-off.

have properties and sequences different from those of conventional known bacteriocins and might have properties similar to those of other AMPs present in databases such as APD (43) which cover a broad range of AMPs. Therefore, in order to predict smORF-encoded AMPs, we trained a separate AMP classifier using the AMPs in the APD3 (43) database (for details, see the Materials and methods) and included it as an additional utility in SProtFP. Cross-validation and held-out testing indicated high accuracy of the AMP model which achieved ROC-AUC values of 0.98 (Figure 6). The AMP model also classified 74.22% of the experimentally validated AMP set (128 AMPs) as positive. Interestingly, only 3/128 AMPs from this validated set were homologous to the known AMPs in the positive dataset used for training and testing using BLASTP (40% identity and 80% query coverage cut-offs). This suggests that AMPs in the experimentally validated datasets share physicochemical properties with other known AMPs even in the absence of sequence homology. Finally, 21 800 smORF-encoded putative AMPs were predicted from the UHGG (31) dataset using the AMP classifier (Supplementary File S10). A total of 1132 of these matched (BLASTP search parameters: 50% identity cut-off, 80% query coverage cut-off and e-value 0.01) previously known AMPs from data sources that included the positive dataset from APD, the experimentally validated AMP dataset and the AMPSphere (70) dataset. The remaining 20 668 putative AMPs were novel and, on clustering using CD-HIT (35) at 50% identity with the same parameters as described before, resulted in 5122 putative AMP families (Supplementary File S11). A high-confidence subset (Supplementary File S11) consisting of 1435 putative AMP families was defined by filtering the cluster representatives based on a probability score of 1 and length range between 20 to 50 amino acids, chosen based on the length range of experimentally validated AMPs (19–49 amino acids) (69,70). Interestingly, the phylum Firmicutes represented homologs belonging to the maximum number of high-confidence families (745 = 51.92%), with *Blautia* and *Clostridium* representing the greatest number of families per genus (Figure 7). Other significant phyla included Proteobacteria and Bacteroidota. Despite not being homologous to the known AMPs, the members of these novel putative AMP families might show antimicrobial activity.

Discussion

The presence of small proteins in various organisms, ranging from microscopic viruses and bacteria to macroscopic animals, suggests their important roles in cellular physiology. In this study, based on analysis of keywords associated with bacterial small proteins in SwissProt (32), we have developed and benchmarked SProtFP which consists of independent ML models that can be used for functional annotation of smORFs in bacteria. Moreover, to predict unknown classes of AMPs besides the well-known bacteriocins, we have also developed an smORF-encoded AMP predictor. To the best of our knowledge, SProtFP is the first ML-based tool which assigns prokaryotic small proteins to systematically defined functional categories.

After experimenting with different physicochemical feature sets, different ML architectures and different ratios of posi-

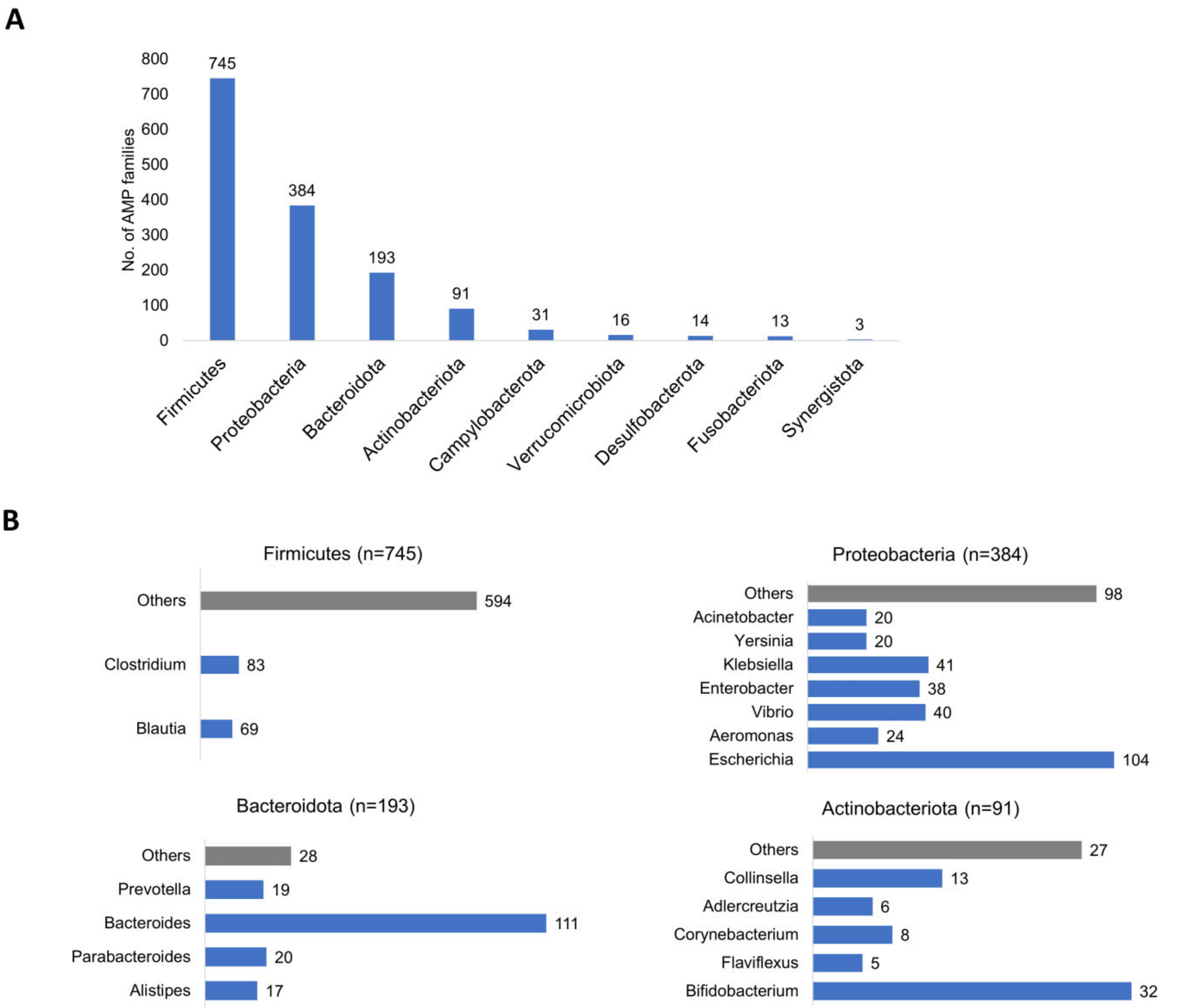


Figure 7. Taxonomic distribution of the 1435 high-confidence AMP families/clusters. (**A**) Phylum-wise distribution of the high-confidence AMP families. (**B**) Genus-wise distribution of the high-confidence AMP families in each phylum. The break down is shown only for the top four phyla having the highest number of AMP families. Only those genera have been plotted that comprise at least 5% of the total AMP families seen in the phylum to which they belong. Other genera have been clubbed together in the 'Others' category. It is important to note that a given AMP family/cluster might have members/homologs across different phyla or genera

tive and negative training datasets, we found that neural networks (DL_1H_40n) trained using a combination of feature sets (QSO + CTD + MA) on imbalanced training datasets gave the best performance. All our models showed high predictive power, as indicated by cross-validation and blind testing on held-out data. One important application of SProtFP is in the functional annotation of data from large-scale microbiome projects to discover novel proteins which may regulate the microbial community composition and dynamics. Using SProtFP, we were able to sample a considerable diversity/proportion of known small protein families across bacterial phyla in addition to assigning putative functions to thousands of uncharacterized proteins (Table 3). Though the analysis of the predictions with respect to their homology to the known set validates the sensitivity of the ML models with respect to currently known data, there could still be additional functional novel proteins which have been assigned functions using the ML models but do not show homology to known small proteins in bacteria (Table 3). A possible reason for this could be that the sensitivity of sequence-based homology detection methods such as BLAST (17) is limited with respect to smORF-encoded proteins (73), while structural homology-based methods such as Foldseek (42) are limited by the quality and confidence of the structural models generated (Supplementary Table S10). Hence, there might be novel uncharacterized small proteins among our predictions which are smaller in length as compared with the known small proteins and represent novel small protein families belonging to these categories. For instance, in a recent study, a novel small protein of 56 amino acids called YnfU was discovered using ribo-seq in *Escherichia coli* which had motifs for zinc binding and thus can participate in metal binding despite being smaller than the majority of known metal-binding proteins (Figure 5D) (4). Additionally, our AMP classifier mined the human gut microbiome isolates for thousands of putative AMP candidates, the majority of which were novel. On the other hand, our predictions are also likely to contain false positives in addition to novel functionally classified proteins. Therefore, at the microbiome scale, while handling large volumes of data, our models can be used to efficiently filter and enrich functional proteins belonging to different functional categories. These predictions can be further refined by intersecting with expression datasets and taxonomic analysis (69) and by using *in silico* approaches such as fold comparison. For instance, in a recent study, Ma *et al*. predicted >20 million smORF-encoded putative antimicrobial peptides using ML from thousands of metagenome-assembled genomes. Further, they utilized metaproteomics along with a network-based analysis of candidate AMPs and bacterial taxa to finally validate a small subset of the candidate AMPs (69). Hence, *in silico* predictions provided by our models can also be combined with experimental data and other types of computational analysis to narrow down the list of predictions.

Although our method has high predictive power, it might have some limitations that are generally associated with ML models. In practical applications, the number of negative instances often outnumber the positive instances. Therefore, depending upon the relative numbers of positive and negative instances, the predictions might also contain false positives in addition to correctly classified proteins. Also, since the models are independent and deal with a multilabel classification problem, they may provide multiple labels to a given small protein, all of which might or might not be true, depending upon the accuracy of each independent ML model. Also, selecting subsets of feature sets amongst the originally chosen feature set (QSO + CTD + MA) might improve model accuracy (74). However, the accuracy of these subsets was either less or comparable (Supplementary Table S11; Supplementary File S7). This observation combined with the cross-validation (Supplementary Tables S7 and S8) and held-out testing (Tables 1 and 2) results of our final models indicate that the original feature set generalizes well. SProtFP is also flexible in the sense that depending upon the requirement, relaxed or restrictive cut-offs (recommended) can be used for the ML models. Therefore, our models can be used to predict putative functions of small proteins at the genomic level and to enrich the small protein sets for proteins belonging to the selected categories at the metagenomic or microbiome level where the dataset size is huge. Our method represents the first step in the systematic categorization and functional annotation of prokaryotic smORFs in an automated fashion using ML. SProtFP can be combined with our previously developed method called ProsmORF-pred (30) to discover novel proteins and analyze them in a functional context. We believe that SProtFP will be a valuable resource for functional characterization of smORFs in prokaryotes.

Data availability

The datasets used for analyses performed in the study including the training datasets, testing datasets as well as the predictions provided by the ML models for the UHGG isolates are provided as Supplementary Files and are also available under the ‘Download’ tab at http://www.nii.ac.in/sprotfp.html. Additionally, the code for the standalone software developed in the current study (SProtFP) is available in Supplementary File S4 and under the ‘Download’ tab at http://www.nii.ac.in/sprotfp.html.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We would like to thank Dr. Luis Pedro Coelho for providing us the list of experimentally validated AMPs in the AMPSphere dataset.

Funding

The Department of Biotechnology, Government of India [grant to the National Institute of Immunology, New Delhi]; the Department of Biotechnology (DBT), India [R&D project grants BICB: BT/PR40325/BTIS/137/1/2020, InTGS: BT/PR50983/MED/29/1650/2023, DIAU: BT/BI/TCB/007/2021, NNP: BT/PR40267/BTIS/137/67/2023 and BT/PR40160/BTIS/137/64/2023 to D.M.]; the National Supercomputing Mission, Ministry of Electronics and Information Technology, India [NSM-NPGDD project MeitY/R&D/HPC/2(1)/2014/ CORP:DG:3191 to DM]; and the CSIR, India [Senior Research fellowship to A.K.].

Conflict of interest statement

None declared.

References

1. Duffy,E.E., Finander,B., Choi,G., Carter,A.C., Pritisanac,I., Alam,A., Luria,V., Karger,A., Phu,W., Sherman,M.A., *et al.* (2022) Developmental dynamics of RNA translation in the human brain. *Nat. Neurosci.*, **25**, 1353–1365.
2. Orr,M.W., Mao,Y., Storz,G. and Qian,S.B. (2020) Alternative ORFs and small ORFs: shedding light on the dark proteome. *Nucleic Acids Res.*, **48**, 1029–1042.
3. Meydan,S., Marks,J., Klepacki,D., Sharma,V., Baranov,P.V., Firth,A.E., Margus,T., Kefi,A., Vazquez-Laslop,N. and Mankin,A.S. (2019) Retapamulin-assisted ribosome profiling reveals the alternative bacterial proteome. *Mol. Cell*, **74**, 481–493.
4. Weaver,J., Mohammad,F., Buskirk,A.R. and Storz,G. (2019) Identifying small proteins by ribosome profiling with stalled initiation complexes. *mBio*, **10**, e02819-18.
5. Laczkovich,I., Mangano,K., Shao,X., Hockenberry,A.J., Gao,Y., Mankin,A., Vazquez-Laslop,N. and Federle,M.J. (2022) Discovery of unannotated small open reading frames in *Streptococcus pneumoniae* D39 involved in quorum sensing and virulence using ribosome profiling. *mBio*, **13**, e0124722.
6. Fremin,B.J., Bhatt,A.S., Kyrpides,N.C. and Global Phage Small Open Reading Frame (GP-SmORF) Consortium.Global Phage Small Open Reading Frame (GP-SmORF) Consortium. (2022) Thousands of small, novel genes predicted in global phage genomes. *Cell Rep.*, **39**, 110984.
7. Baek,J., Lee,J., Yoon,K. and Lee,H. (2017) Identification of unannotated small genes in Salmonella. *G3*, **7**, 983–989.
8. Venturini,E., Svensson,S.L., Maass,S., Gelhausen,R., Eggenhofer,F., Li,L., Cain,A.K., Parkhill,J., Becher,D., Backofen,R., *et al.* (2020) A global data-driven census of Salmonella small proteins and their potential functions in bacterial virulence. *Microlife*, **1**, uqaa002.
9. Anderson,D.M., Makarewich,C.A., Anderson,K.M., Shelton,J.M., Bezprozvannaya,S., Bassel-Duby,R. and Olson,E.N. (2016) Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Sci. Signal*, **9**, ra119.
10. Andresen,L., Martinez-Burgo,Y., Nilsson Zangelin,J., Rizvanovic,A. and Holmqvist,E. (2020) The small toxic

Salmonella protein TimP targets the cytoplasmic membrane and is repressed by the small RNA TimR. *mBio*, **11**, e01659-20.

11. Potts,A.H., Vakulskas,C.A., Pannuri,A., Yakhnin,H., Babitzke,P. and Romeo,T. (2017) Global role of the bacterial post-transcriptional regulator CsrA revealed by integrated transcriptomics. *Nat. Commun.*, **8**, 1596.
12. Sandmann,C.L., Schulz,J.F., Ruiz-Orera,J., Kirchner,M., Ziehm,M., Adami,E., Marczenke,M., Christ,A., Liebe,N., Greiner,J., *et al.* (2023) Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Mol. Cell*, **83**, 994–1011.
13. Chen,J., Brunner,A.D., Cogan,J.Z., Nunez,J.K., Fields,A.P., Adamson,B., Itzhak,D.N., Li,J.Y., Mann,M., Leonetti,M.D., *et al.* (2020) Pervasive functional translation of noncanonical human open reading frames. *Science*, **367**, 1140–1146.
14. Biegert,A. and Soding,J. (2009) Sequence context-specific profiles for homology searching. *Proc. Natl Acad. Sci. USA*, **106**, 3770–3775.
15. Soding,J., Biegert,A. and Lupas,A.N. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**, W244–W248.
16. van Dam,S., Vosa,U., van der Graaf,A., Franke,L. and de Magalhaes,J.P. (2018) Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinf.*, **19**, 575–592.
17. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
18. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
19. Redfern,O.C., Harrison,A., Dallman,T., Pearl,F.M. and Orengo,C.A. (2007) CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. *PLoS Comput. Biol.*, **3**, e232.
20. Dalkiran,A., Rifaioglu,A.S., Martin,M.J., Cetin-Atalay,R., Atalay,V. and Dogan,T. (2018) ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics*, **19**, 334.
21. Kumar,N. and Skolnick,J. (2012) EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. *Bioinformatics*, **28**, 2687–2688.
22. Li,Y., Wang,S., Umarov,R., Xie,B., Fan,M., Li,L. and Gao,X. (2018) DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, **34**, 760–769.
23. Sarac,O.S., Atalay,V. and Cetin-Atalay,R. (2010) GOPred: GO molecular function prediction by combined classifiers. *PLoS One*, **5**, e12382.
24. Zhang,C., Freddolino,P.L. and Zhang,Y. (2017) COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res.*, **45**, W291–W299.
25. Kulmanov,M. and Hoehndorf,R. (2020) DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics*, **36**, 422–429.
26. Duval,M. and Cossart,P. (2017) Small bacterial and phagic proteins: an updated view on a rapidly moving field. *Curr. Opin. Microbiol.*, **39**, 81–88.
27. Li,H., Xiao,L., Zhang,L., Wu,J., Wei,B., Sun,N. and Zhao,Y. (2018) FSPP: a tool for genome-wide prediction of smORF-encoded peptides and their functions. *Front. Genet.*, **9**, 96.
28. Ji,X., Cui,C. and Cui,Q. (2020) smORFunction: a tool for predicting functions of small open reading frames and microproteins. *BMC Bioinformatics*, **21**, 455.
29. Vajjala,M., Johnson,B., Kasparek,L., Leuze,M. and Yao,Q. (2022) Profiling a community-specific function landscape for bacterial peptides through protein-level meta-assembly and machine learning. *Front. Genet.*, **13**, 935351.
30. Khanduja,A., Kumar,M. and Mohanty,D. (2023) ProsmORF-pred: a machine learning-based method for the identification of small ORFs in prokaryotic genomes. *Brief. Bioinf.*, **24**, bbad101.
31. Almeida,A., Nayfach,S., Boland,M., Strozzi,F., Beracochea,M., Shi,Z.J., Pollard,K.S., Sakharova,E., Parks,D.H., Hugenholtz,P., *et al.* (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.*, **39**, 105–114.
32. UniProt,Consortium. (2023) UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
33. van Heel,A.J., de Jong,A., Song,C., Viel,J.H., Kok,J. and Kuipers,O.P. (2018) BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res.*, **46**, W278–W281.
34. Tourasse,N.J. and Darfeuille,F. (2021) T1TAdb: the database of type I toxin–antitoxin systems. *RNA*, **27**, 1471–1481.
35. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
36. Chen,Z., Zhao,P., Li,F., Marquez-Lago,T.T., Leier,A., Revote,J., Zhu,Y., Powell,D.R., Akutsu,T., Webb,G.I., *et al.* (2020) iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinf.*, **21**, 1047–1057.
37. Frank,E., Hall,M., Trigg,L., Holmes,G. and Witten,I.H. (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.
38. Grau,J., Grosse,I. and Keilwagen,J. (2015) PRROC: computing and visualizing precision–recall and receiver operating characteristic curves in R. *Bioinformatics*, **31**, 2595–2597.
39. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
40. Hyatt,D., LoCascio,P.F., Hauser,L.J. and Uberbacher,E.C. (2012) Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, **28**, 2223–2230.
41. Lin,Z., Akin,H., Rao,R., Hie,B., Zhu,Z., Lu,W., Smetanin,N., Verkuil,R., Kabeli,O., Shmueli,Y., *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, **379**, 1123–1130.
42. van Kempen,M., Kim,S.S., Tumescheit,C., Mirdita,M., Lee,J., Gilchrist,C.L.M., Soding,J. and Steinegger,M. (2024) Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.*, **42**, 243–246.
43. Wang,G., Li,X. and Wang,Z. (2016) APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.*, **44**, D1087–D1093.
44. Kubatova,N., Pyper,D.J., Jonker,H.R.A., Saxena,K., Remmel,L., Richter,C., Brantl,S., Evguenieva-Hackenberg,E., Hess,W.R., Klug,G., *et al.* (2020) Rapid biophysical characterization and NMR spectroscopy structural analysis of small proteins from bacteria and archaea. *Chembiochem*, **21**, 1178–1187.
45. Bhasin,M. and Raghava,G.P. (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.*, **279**, 23262–23266.
46. Govindan,G. and Nair,A.S. (2013) Bagging with CTD—a novel signature for the hierarchical prediction of secreted protein trafficking in eukaryotes. *Genomics Proteomics Bioinformatics*, **11**, 385–390.
47. Cai,Y.D., Liu,X.J., Xu,X.B. and Chou,K.C. (2002) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. *J. Cell. Biochem.*, **84**, 343–348.
48. Limongelli,I., Marini,S. and Bellazzi,R. (2015) PaPI: pseudo amino acid composition to score human protein-coding variants. *BMC Bioinf.*, **16**, 123.
49. Meng,C., Jin,S., Wang,L., Guo,F. and Zou,Q. (2019) AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front. Bioeng. Biotechnol.*, **7**, 224.

50. Barradas-Bautista,D., Cao,Z., Vangone,A., Oliva,R. and Cavallo,L. (2022) A random forest classifier for protein–protein docking models. *Bioinform. Adv.*, **2**, vbab042.
51. Couronne,R., Probst,P. and Boulesteix,A.L. (2018) Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinf.*, **19**, 270.
52. Eetemadi,A. and Tagkopoulos,I. (2019) Genetic neural networks: an artificial neural network architecture for capturing gene expression relationships. *Bioinformatics*, **35**, 2226–2234.
53. Lin,W.J. and Chen,J.J. (2013) Class-imbalanced classifiers for high-dimensional data. *Brief. Bioinf.*, **14**, 13–26.
54. Scalzitti,N., Kress,A., Orhand,R., Weber,T., Moulinier,L., Jeannin-Girardon,A., Collet,P., Poch,O. and Thompson,J.D. (2021) Spliceator: multi-species splice site prediction using convolutional neural networks. *BMC Bioinformatics*, **22**, 561.
55. Blagus,R. and Lusa,L. (2013) SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, **14**, 106.
56. Illergard,K., Ardell,D.H. and Elofsson,A. (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins*, **77**, 499–508.
57. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chao,H., Chen,L., Craig,P.A., Crichlow,G.V., Dalenberg,K., Duarte,J.M., *et al.* (2023) RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.*, **51**, D488–D508.
58. Hrala,M., Bosak,J., Micenkova,L., Krenova,J., Lexa,M., Pirkova,V., Tomastikova,Z., Kolackova,I. and Smajs,D. (2021) *Escherichia coli* strains producing selected bacteriocins inhibit porcine enterotoxigenic *Escherichia coli* (ETEC) under both in vitro and in vivo conditions. *Appl. Environ. Microbiol.*, **87**, e0312120.
59. Ahern,M., Verschueren,S. and van Sinderen,D. (2003) Isolation and characterisation of a novel bacteriocin produced by *Bacillus thuringiensis* strain B439. *FEMS Microbiol. Lett.*, **220**, 127–131.
60. Kemperman,R., Kuipers,A., Karsens,H., Nauta,A., Kuipers,O. and Kok,J. (2003) Identification and characterization of two novel clostridial bacteriocins, circularin A and closticin 574. *Appl. Environ. Microbiol.*, **69**, 1589–1597.
61. Han,T., Hu,X., Li,K., Zhang,D., Zhang,Y. and Li,J. (2021) *Bifidobacterium infantis* maintains genome stability in ulcerative colitis via regulating anaphase-promoting complex subunit 7. *Front. Microbiol.*, **12**, 761113.
62. Biswas,P., Sengupta,S. and Nagaraja,V. (2023) Evolution of YacG to safeguard DNA gyrase from external perturbation. *Res. Microbiol.*, **174**, 104093.
63. Browne,H.P., Forster,S.C., Anonye,B.O., Kumar,N., Neville,B.A., Stares,M.D., Goulding,D. and Lawley,T.D. (2016) Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature*, **533**, 543–546.
64. Wetzel,D. and Fischer,R.J. (2015) Small acid-soluble spore proteins of *Clostridium acetobutylicum* are able to protect DNA in vitro and are specifically cleaved by germination protease GPR and spore protease YyaC. *Microbiology*, **161**, 2098–2109.
65. Yamanaka,K., Zheng,W., Crooke,E., Wang,Y.H. and Inouye,M. (2001) CspD, a novel DNA replication inhibitor induced during the stationary phase in *Escherichia coli*. *Mol. Microbiol.*, **39**, 1572–1584.
66. Wilmaerts,D., De Loose,P.J., Vercauteren,S., De Smedt,S., Verstraeten,N. and Michiels,J. (2021) Functional analysis of cysteine residues of the Hok/Gef type I toxins in *Escherichia coli*. *FEMS Microbiol. Lett.*, **368**, fnab069.
67. Armalyte,J., Jurenaite,M., Beinoraviciute,G., Teiserskas,J. and Suziedeliene,E. (2012) Characterization of *Escherichia coli* dinJ–yafQ toxin–antitoxin system using insights from mutagenesis data. *J. Bacteriol.*, **194**, 1523–1532.
68. Gallardo-Becerra,L., Cervantes-Echeverria,M., Cornejo-Granados,F., Vazquez-Morado,L.E. and Ochoa-Leyva,A. (2023) Perspectives in searching antimicrobial peptides (AMPs) produced by the microbiota. *Microb. Ecol.*, **87**, 8.
69. Ma,Y., Guo,Z., Xia,B., Zhang,Y., Liu,X., Yu,Y., Tang,N., Tong,X., Wang,M., Ye,X., *et al.* (2022) Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat. Biotechnol.*, **40**, 921–931.
70. Santos-Junior,C.D., Torres,M.D.T., Duan,Y., Rodriguez Del Rio,A., Schmidt,T.S.B., Chong,H., Fullam,A., Kuhn,M., Zhu,C., Houseman,A., *et al.* (2024) Discovery of antimicrobial peptides in the global microbiome with machine learning. *Cell*, **187**, 3761–3778.
71. Ongpipattanakul,C., Desormeaux,E.K., DiCaprio,A., van der Donk,W.A., Mitchell,D.A. and Nair,S.K. (2022) Mechanism of action of ribosomally synthesized and post-translationally modified peptides. *Chem. Rev.*, **122**, 14722–14814.
72. Yi,Y., Li,P., Zhao,F., Zhang,T., Shan,Y., Wang,X., Liu,B., Chen,Y., Zhao,X. and Lü,X. (2022) Current status and potentiality of class II bacteriocins from lactic acid bacteria: structure, mode of action and applications in the food industry. *Trends Food Sci. Technol.*, **120**, 387–401.
73. Allen,R.J., Brenner,E.P., VanOrsdel,C.E., Hobson,J.J., Hearn,D.J. and Hemm,M.R. (2014) Conservation analysis of the CydX protein yields insights into small protein identification and evolution. *BMC Genomics*, **15**, 946.
74. Pudjihartono,N., Fadason,T., Kempa-Liehr,A.W. and O'Sullivan,J.M. (2022) A review of feature selection methods for machine learning-based disease risk prediction. *Front. Bioinform.*, **2**, 927312.

Received: August 16, 2024. **Revised:** November 7, 2024. **Editorial Decision:** December 16, 2024. **Accepted:** December 17, 2024

© The Author(s) 2025. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.