Briefings in Bioinformatics, 22(6), 2021, 1–17

https://doi.org/10.1093/bib/bbab258
Problem Solving Protocol

ENNAVIA is a novel method which employs neural networks for antiviral and anti-coronavirus activity prediction for therapeutic peptides

Patrick Brendan Timmons and Chandralal M. Hewage

Corresponding author: Chandralal M. Hewage, UCD School of Biomolecular and Biomedical Science, UCD Centre for Synthesis and Chemical Biology, UCD Conway Institute, University College Dublin, Dublin 4, Ireland. Tel: +353 1 716 6870; E-mail: chandralal.hewage@ucd.ie

Abstract

Viruses represent one of the greatest threats to human health, necessitating the development of new antiviral drug candidates. Antiviral peptides often possess excellent biological activity and a favourable toxicity profile, and therefore represent a promising field of novel antiviral drugs. As the quantity of sequencing data grows annually, the development of an accurate *in silico* method for the prediction of peptide antiviral activities is important. This study leverages advances in deep learning and cheminformatics to produce a novel sequence-based deep neural network classifier for the prediction of antiviral peptide activity. The method outperforms the existent best-in-class, with an external test accuracy of 93.9%, Matthews correlation coefficient of 0.87 and an Area Under the Curve of 0.93 on the dataset of experimentally validated peptide activities. This cutting-edge classifier is available as an online web server at https://research.timmons.eu/ennavia, facilitating *in silico* screening and design of peptide antiviral drugs by the wider research community.

Key words: Neural network; machine learning; antiviral drugs; peptides; in silico screening

Introduction

Viruses are an ancient infection agent that replicate inside the cells of living organisms. They are ubiquitous, affecting all species, from bacteria to plants and animals [1], and are incredibly successful due to their genetic diversity, non-uniformity of mode of transmission, efficient replication and capacity for persistence in their hosts [2–4]. Viral diseases are difficult to control due to their potential for high pathogenicity, increased resistance to antiviral drugs, continuous evolution of existing viruses and the emergence of novel viruses [5]. Viruses are responsible for many human diseases and are the cause of many death annually. Cold sores, influenza, AIDS and the current coronavirus disease 2019 (COVID-19) pandemic are all caused by viral infection. Zoonotic viruses, such as the Ebola, Zika, West Nile, HIV, SARS-CoV and SARS-CoV-2 viruses, are especially dangerous, as they are not well adapted to the human hosts' immune systems, and consequently, cause life-threatening diseases. The World Health Organisation estimated in 2017 that influenza alone is responsible for up to 645 832 death annually, and at the time of publication, the COVID-19 pandemic has been responsible for 3,995,000 deaths. Therefore, the development of novel antiviral drugs, including anti-coronavirus drugs, is important to control emerging viral pathogens.

Host defence peptides (HDPs) are ubiquitous elements of the immune system, having been identified in all living species [6]. Indeed, the induction pathways of HDPs are highly conserved among the genomes of animal and plant genomes [6, 7]. Many HDPs have been found to possess antiviral activity. These antiviral peptides (AVPs) are short, typically 8–40 amino acids, cationic and α-helical, although AVPs with an overall

Patrick Brendan Timmons is a researcher at the Conway Institute, University College Dublin. His research focuses on the application of deep learning to the study of therapeutic peptide structure and function.
Chandralal M. Hewage is a principal investigator at the Conway Institute, University College Dublin. His research focuses on the use of NMR spectroscopy to study the structures of bioactive peptides.
Submitted: 2 April 2021; **Received (in revised form):** 9 June 2021

negative charge and other secondary structures have also been identified [8]. Most importantly, AVPs are a promising resource for the development of novel antiviral drugs for the prevention or treatment of viral diseases [8], including those caused by coronaviruses. For example, a subsequence derived from β-defensin, P9, possesses potent inhibitory activity against the SARS-CoV and MERS-CoV viruses [9]. Other anti-coronavirus peptides include Mucroporin-M1 and HR2P, which inhibit the SARS-CoV and MERS-CoV viruses, respectively [10, 11].

This class of potential antiviral agents possesses a number of advantages over conventional non-peptide drugs, as they are highly specific, cost-effective to produce while remaining easy to modify and synthesize, and possess a limited susceptibility to drug resistance [12]. Although initially AVPs were isolated from plant and animal secretions where they formed part of the host defence mechanism [13], AVPs have also been derived from chemical [14], genetic [15] and recombinant [16] libraries, as well as from rational design [17]. AVPs can be divided into two classes based on their mechanism of action: virus-targeting and host-targeting [18]. AVPs belonging to the former class focus on the inhibition of viral enzymes involved in transcription and replication [19, 20], or the inactivation of viral structural proteins [21]. AVPs of the latter class act as immunomodulators, like interferons [22, 23], or target cyclophilins, which are important cellular factors that are hijacked by viruses during replication cycle [18, 24]. The currently identified AVPs, however, represent only a small subset of a largely unexplored chemical space, with only a few of those being peptide-based antiviral drugs available on the market. Those drugs include Enfuvirtide, the first peptide inhibitor of HIV-1, Boceprevir and Telaprevir, which both act against hepatitis-C [25]. A number of databases exist which detail the antiviral activities of AVPs, such as AVPdb [26], DBAASP [27, 28], CAMP [29] and APD3 [30].

In silico methods offer a fast, efficient way of exploring the large chemical space that AVPs inhabit, by minimizing the quantity of peptides that need to be synthesized and experimentally assayed for antiviral activity. A few methods for the prediction of peptide antiviral activity exist, namely AVPpred [31], AntiVPP 1.0 [32], Meta-iAVP [33], Firm-AVP [34] and the method of Chang et al. [35]. Antiviral peptide prediction methods have been comprehensively reviewed by Charoenkwan et al. [36]. Furthermore, Pang et al. recently developed a novel method for the prediction of peptides with specifically anti-coronavirus activity [37]. The most popular machine learning methods employed are support vector machines (SVMs) or random forests, although a number of others have also been trialled. Many areas of bioinformatics have benefited from the predictive power of deep learning; neural network-based methods exist for many tasks, such as DeepPPISP for the prediction of protein–protein interaction sites [38], SCLpred and SCLpred-EMS for protein subcellular localization prediction [39, 40], CPPpred for the prediction of cell-penetrating peptides [41], HAPPENN for the prediction of peptide hemolytic activity [42], ENNAACT for the prediction of peptide anticancer activity [43] and APPTEST for the prediction of peptide tertiary structure [44]. As the quantity of antiviral peptide sequence data continuously increases, we have exploited the available data to create a deep neural network method for the identification of AVPs from the primary sequence. Herein, we describe ENNAVIA, a novel neural network peptide antiviral and anti-coronavirus activity predictor. ENNAVIA (**E**mploying **N**eural **N**etworks for **A**nti**vi**ral **A**ctivity Prediction for Therapeutic Peptides) is available as a free-to-use online webserver for the benefit of the academic community at https://research.timmons.eu/ennavia.

Methods

Datasets

To facilitate easy comparison with existing peptide antiviral activity predictors, the two AVPpred datasets of Thakur et al. were used in this work [31]. The first dataset consists of 604 peptides with experimentally validated antiviral activities, and 452 peptides that were experimentally found to have poor or no antiviral activity. This dataset is divided into training and external validation subsets, termed $T^{544p+407n}$ and $V^{60p+45n}$, respectively, where p and n denote the number of positive and negative samples. For brevity, these are collectively referred to as ENNAVIA-A. The second dataset consists of 604 peptides with experimentally validated antiviral activities, and 604 negative peptides from the AntiBP2 negative dataset, which were randomly extracted from non-secretory proteins [45]. This second dataset is similarly divided into training and external validation subsets, termed $T^{544p+544n}$ and $V^{60p+60n}$, respectively, where p and n again denote the number of positive and negative samples. These are collectively referred to as ENNAVIA-B. Peptide sequences in the datasets consist only of natural amino acids; peptides that contain residues not included in the canonical 20 amino acids are excluded, as are peptides with a sequence length shorter than 7 or longer than 40. Information about the peptides' secondary structure is not included in the dataset. The datasets are available for download from the webserver website and as supplementary material to this article.

In order to develop a classifier specific to the prediction of peptides with anti-coronavirus activity, two additional datasets were created, ENNAVIA-C and ENNAVIA-D. The positive samples of both datasets are peptide sequences with anti-coronavirus activity, taken from the dataset created by Pang et al. [37]. The original dataset included 139 peptide sequences with anti-coronavirus activity. Once peptide sequences with a sequence length shorter than 7 or longer than 40 were excluded, 109 peptide sequences remained. The negative samples of ENNAVIA-C and ENNAVIA-D are the same as the negative samples of ENNAVIA-A and ENNAVIA-B, respectively.

Model validation

It is imperative to thoroughly validate classifier models created by machine learning. Tenfold cross-validations and validation by an external test set were employed for the performance evaluation of all models presented herein. The models trained under cross-validation were ensembled and evaluated with the external test sets. For ENNAVIA-A and ENNAVIA-B, the peptides used in the external test sets are those from the $V^{60p+45n}$ and $V^{60p+60n}$ datasets of Thakur et al. [31], in order to facilitate a direct comparison with existing methods.

Peptides with anti-coronavirus activity which are also present in the ENNAVIA-A and ENNAVIA-B datasets are assigned to the same fold as in ENNAVIA-A and ENNAVIA-B. In order to prevent overfitting, the CD-HIT-2D program [46, 47] was used to identify anti-coronavirus peptides that can be matched to anti-virus peptides using a sequence identity cut-off value of 0.9. Anti-coronavirus peptides which had high sequence identity to anti-virus peptides in the ENNAVIA-A and ENNAVIA-B datasets were assigned to the same fold as those peptides. The negative peptides of the ENNAVIA-C and ENNAVIA-D datasets maintained the same fold-assignment as in the ENNAVIA-A and ENNAVIA-B datasets.

Amino acid composition analysis

The amino acid composition of the experimentally verified AVPs was analysed and compared to that of the experimentally verified non-antiviral peptide sequences and the random non-secretory peptide sequences extracted from UniProt. The composition analysis includes the peptides' full sequences, the 10 N-terminal residues and the C-terminal 10 residues.

Residue position preference analysis

Enrichment depletion logos (EDLogo) [48] were created for the AVPs' sequences to identify any position-specific amino acid preferences that may exist. The experimentally validated non-antiviral peptide sequences were used as the baseline in the construction of the logo plots.

Features extraction

A variety of features was extracted from the peptides' primary sequences. These features can be divided into two subcategories, amino acid-based descriptors and physicochemical descriptors. Only features that were non-zero for at least 20 samples were retained in the final feature vector, which has a dimensionality of 6397.

Composition descriptors

The peptides' compositional descriptors were calculated based on the peptides' amino acid, dipeptide and tripeptide compositions for the conventional 20-amino acid alphabet. Additionally, descriptors were also calculated based on the reduced amino acid alphabets of Veltri et al. [49], Thomas and Dill [50] and the conjoint alphabet [51]. g-gap dipeptide and tripeptide compositions were calculated to account for the three-dimensional structure of the peptides [52], with the values of the parameter g being 1, 2 and 3 for the dipeptide compositions, and 3 and 4 for the tripeptide compositions. Furthermore, conjoint triad, composition, transition and distribution [53] and pseudo amino acid composition [54] descriptors were also calculated.

Physicochemical descriptors

The modlAMP package was employed for the calculation of global physicochemical descriptors and amino acid scale-based descriptors [55]. Global physicochemical features include molecular formula, sequence length, molecular weight, sequence charge, charge density, isoelectric point, instability index, aliphatic index [56], aromaticity index [57], hydrophobic ratio and the Boman index [58]. Amino acid scale-based descriptors include hydrophobicity [59–63], side-chain bulkiness [64], refractivity [65], side-chain flexibility [66], α-helix propensity [67], transmembrane propensity [68], polarity [64, 69], amino acid charges, AASI [70], ABHPRK [55], COUGAR [55], Ez [71], ISAECI [72], MSS [73], MSW [74], PPCALI [75], t_scale [76], z3 [77], z5 [78] and pepArc [55].

Additional physicochemical features were calculated based on amino acid properties detailed in the AAindex [79]. The peptides' hydrophobicities were quantified using the amino acids' hydrophobicities [80, 81], hydropathies [82], retention coefficients in HPLC [83] and partition energies [84, 85]. Similarly, the peptide sequences' hydrophilicities were characterized using descriptors based on the amino acid hydrophilicity scale [86], the amino acids' net charges [87], polar requirements [88] and fractions of site occupied by water [89]. Descriptors pertaining to sterics were obtained from the residues' steric hindrance [90] and bulkiness [64] properties, while secondary structure features were calculated based on helical [91] propensities. Furthermore, descriptors were also calculated from the side-chain interaction parameters [92] and membrane-buried preference parameters [93].

Machine learning approaches

Unsupervised and supervised machine learning approaches are employed in the current study. The former includes principal component analysis (PCA) [94] and t-distributed Stochastic Neighbour Embedding (t-SNE) [95] for visualizing the data. The latter includes SVM [96], random forest (RF) [97] and dense fully connected neural networks [98] for creating supervised classifiers. The scikit-learn Python module is used for its PCA, t-SNE, SVM and RF implementations [99].

SVMs were trialled using both a linear and non-linear radial base function (RBF) kernel. A grid search was employed for the tuning of the RF number of estimators, the maximum number of features and the maximum depth hyperparameters, and the SVM regularization parameter C and kernel width parameter γ.

Neural network architecture and implementation

The Keras deep learning framework with a Tensorflow backend was used to build and train the deep-fully connected neural networks [100].

The neural network's input features are scaled to have minimum and maximum values of 0 and 1, respectively.

The optimal combination of neural network architecture and hyperparameters was selected using a randomized grid search strategy.

The first hidden layer has 1024 nodes, and is followed by two layers of 256 nodes each. Batch normalization [101] is applied before the ReLU activation function for each hidden layer. To prevent overfitting to the training data, each hidden layer is followed by a Dropout regularization layer, with a rate of 0.30 [102]. The output layer is a single node activated by the sigmoid function. As is common in binary classification neural networks, the binary cross-entropy loss function is employed.

It is defined as:

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{1}$$

where y_i is the true value of the i^{th} sample, and $\hat{y}_i$ is the predicted value of the i^{th} sample. As the predicted labels of all training data approach their respective true values, the value of the function approaches zero.

The optimal optimizer was found to be Adaptive Momentum (Adam), with an optimal initial learning rate of 0.05 and a decay of 0.0001. Adam utilizes the following formula to update the neural network weights [103]:

$$\Theta_{t+1} = \Theta_t - \frac{\eta \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{2}$$

where the $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected estimates of the mean and the variance of the gradients, respectively.

The neural networks were trained for 600 epochs, without stopping criteria. The model with the highest validation accuracy encountered during training was retained for each of the cross-validation splits.

Transfer learning

As the dataset of peptides with anti-coronavirus is small, numbering only 109 peptides, transfer learning was used to train the models for the ENNAVIA-C and ENNAVIA-D datasets. Models originally trained for each cross-validation fold for ENNAVIA-A and ENNAVIA-B, respectively, were used to initialize the weights for the neural network models of the corresponding cross-validation folds for ENNAVIA-C and ENNAVIA-D, respectively. The neural network models were then trained for 600 epochs, without stopping criteria. The model with the highest validation accuracy encountered during training was retained for each of the cross-validation splits.

Performance evaluation

A number of standard metrics are employed for the evaluation of the presented models' performance, specifically accuracy (Acc), sensitivity (Sn), specificity (Sp), the Matthews correlation coefficient (MCC) and the receiver operating characteristic (ROC) curve. Confidence intervals are provided at the 95% level of significance.

The first four metrics are defined by the following equations:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{3}$$

$$Sn = \frac{TP}{TP + FN} \times 100\% \tag{4}$$

$$Sp = \frac{TN}{TN + FP} \times 100\% \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

where

- TP = True positives: the number of correctly predicted positive (antiviral) peptides.
- FP = False positives: the number of non-antiviral peptides incorrectly predicted as being antiviral.
- TN = True negatives: the number of correctly predicted negative (non-antiviral) peptides.
- FN = False negatives: the number of anticancer peptides incorrectly predicted as being non-antiviral.

Results

The dataset of peptide sequences was subjected to an amino acid composition analysis and residue position preference analysis. Feature vectors comprising the peptides' physicochemical descriptors, compositional descriptors and all descriptors were constructed and visualized in two-dimensional space using PCA and t-SNE plots. Plots created using both methods show an incomplete separation of the positive and negative classes. Finally, three machine learning classifiers, namely SVMs, random forests and neural networks, are trained on the dataset's feature vectors, and the antiviral activity prediction results are evaluated.

Amino acid composition analysis

To identify if particular amino acid residues are more prevalent in antiviral and anti-coronavirus peptides, an amino acid residue composition analysis was performed. The amino acid compositions of anti-coronavirus peptides, AVPs, experimentally validated non-antiviral peptides and random non-antiviral peptide sequences are illustrated in Figure 1. Statistical analysis was carried out using a Chi-squared test; all results are significant at the $P < 0.01$ significance level.

Interestingly, antiviral and anti-coronavirus peptides are enriched in the cysteine and the hydrophobic isoleucine residue, and depleted in proline and histidine. While AVPs in general exhibit enrichment in lysine and tryptophan, this is not observed for the specifically anti-coronavirus peptides. Similarly, AVPs are depleted in glycine and valine, while anti-coronavirus peptides are enriched in these residues. While the amino acid composition for anti-coronavirus peptides is based on a limited sample size, it does suggest that the composition requirements for peptides to possess activity against coronaviruses differ from the composition requirements for activity against viruses in general.

Furthermore, an amino acid composition analysis was carried out for AVPs on the basis of their mode of action (Figure 2). Interestingly, while AVPs are generally not enriched in aspartic acid or tryptophan, AVPs that act at the viral membrane are rich in aspartic acid.

Residue position preference analysis

To assess the possibility of a preference existing for certain amino acid residues at certain positions in the peptides' primary sequence, an enrichment-depletion logo plot was produced (Figure 3) for the experimentally validated AVPs. The experimentally validated non-antiviral peptides were used to establish a baseline for the plot.

The first inspection of the logo plot suggests that AVPs are enriched in tryptophan at most positions. This is consistent with the aforementioned amino acid composition analysis. More specifically, however, AVPs appear to be enriched in glycine at position 1, and have a preference for a positively charged residue at position 4. Conversely, they are enriched in aspartic acid at position 5 and 8, and the third-last residue. Enrichment is also observed in phenylalanine at the three C-terminal positions. Again, in agreement with the amino acid composition analysis, AVPs are depleted in proline and tryptophan at all positions.

Data Visualization

Principal Component Analysis

PCA was carried out on the ENNAVIA-A dataset for all computed descriptors, only the physicochemical descriptors and only the compositional descriptors subsets (Figure 4). While a separation does exist between the experimentally verified antiviral and non-antiviral peptides, it is incomplete, and the two classes are significantly overlapped.

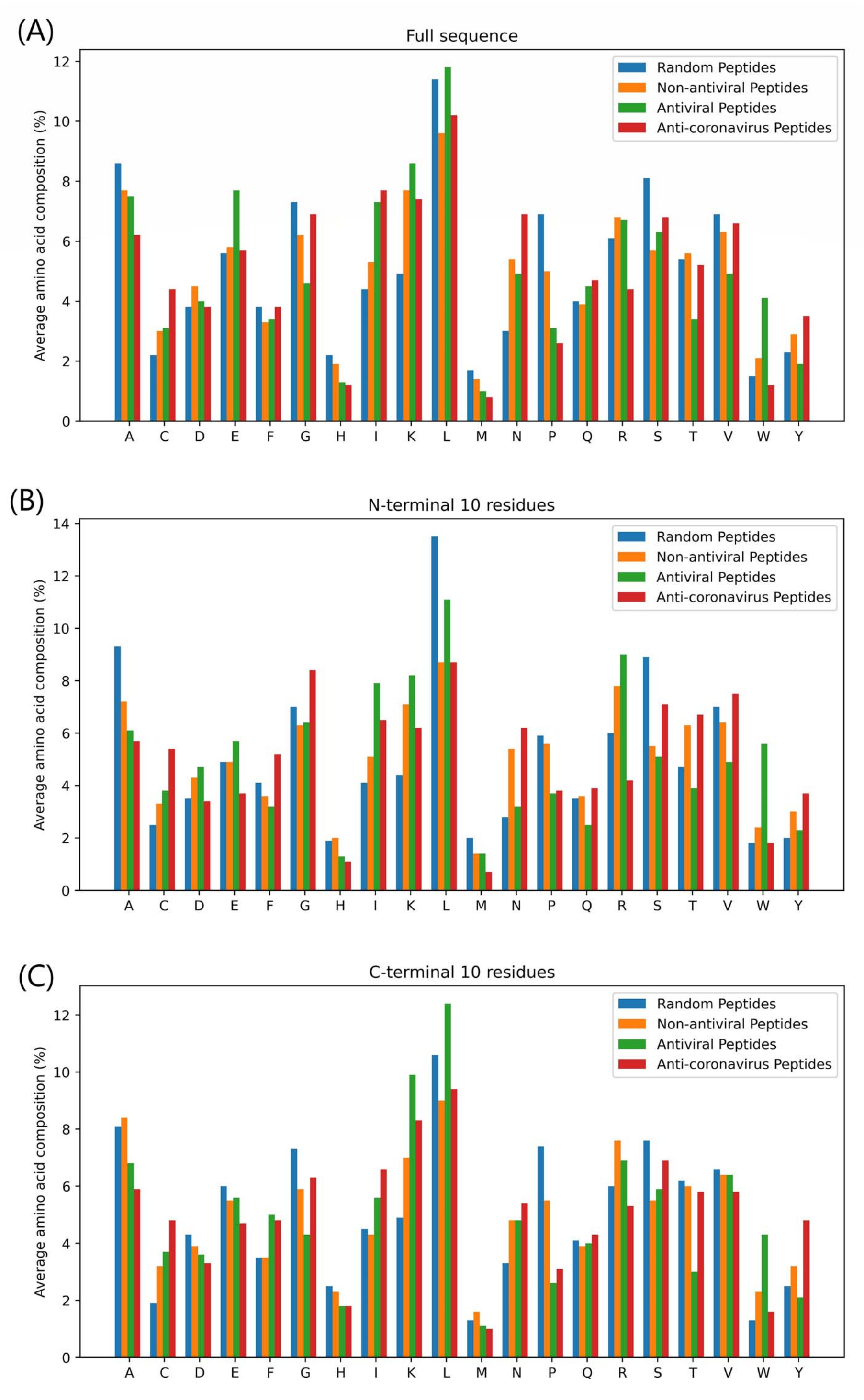

Figure 1. Percentage average amino acid residue composition of the (A) full sequences, (B) N-terminal 10 residues and (C) C-terminal 10 residues of anti-coronavirus peptides (red), experimentally validated antiviral peptides (green), experimentally validated non-antiviral peptide (orange) and non-antiviral peptides randomly extracted from UniProt proteins (blue). One-letter amino acid codes are given for the residues on the x-axis.

T-Distributed Stochastic Neighbour Embedding

To complement the PCA analysis, a t-SNE analysis was conducted for the experimentally verified antiviral and non-antiviral peptides, again for all computed descriptors, only the physicochemical descriptors and only the compositional descriptors subsets (Figure 5). As with the results of the PCA analysis, the inter-class separation is incomplete, although it is clearly greater.

Antiviral activity prediction

The principal aim of this study was to train and evaluate a selection of machine learning classifiers for the prediction of peptide antiviral activity. Tenfold cross-validation was employed for the evaluation of the classifiers' robustness and predictive power. Additionally, the 10 models trained for each classifier under 10-fold cross-validation were ensembled and further evaluated through the use of the external, independent test set. The

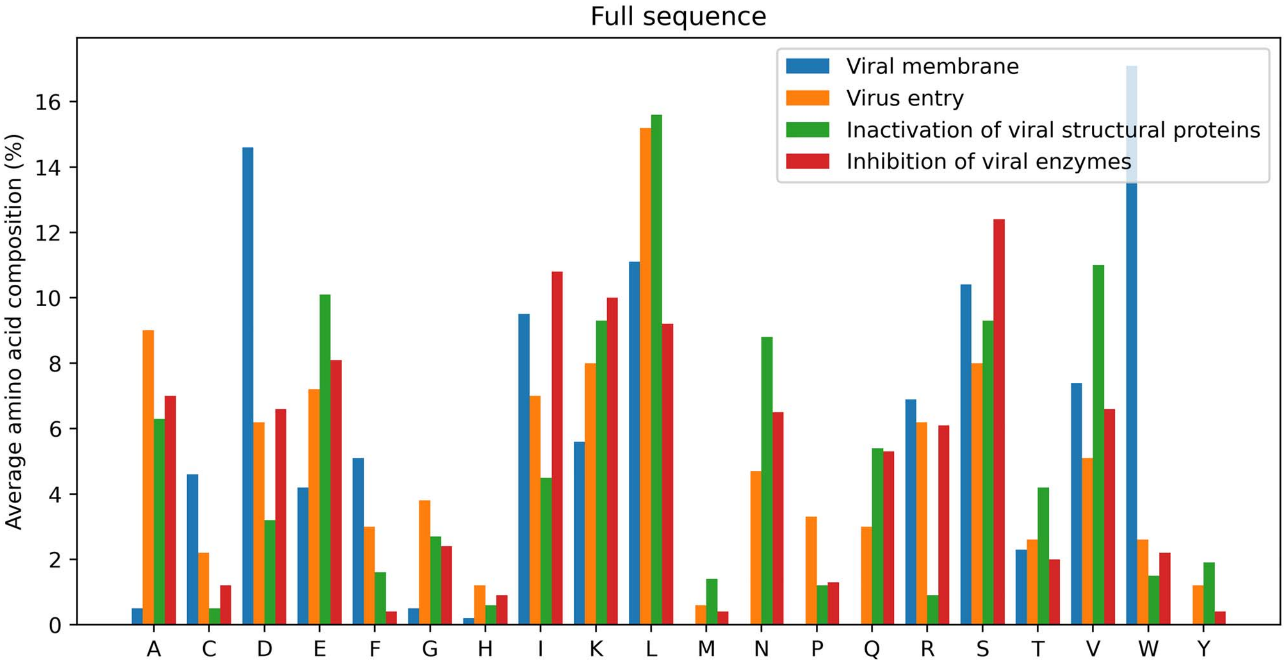

Figure 2. Percentage average amino acid residue composition of the full peptide sequences that exert their antiviral activity by disruption of the viral membrane (blue), prevention of virus entry (orange), inactivation of viral structural proteins (green) and inhibition of viral enzymes (red). One-letter amino acid codes are given for the residues on the x-axis.

accuracy, MCC, sensitivity and specificity parameters, together with their respective confidence intervals, are reported for each model. ROC curves with calculated area under the curve (AUC) values are also given for both final neural network models. The SVM, random forest (RF) and neural network (NN) performance metrics are tabulated in Table 1.

A grid search strategy was employed for the optimization of SVM and RF hyperparameters.

The SVM classifier achieved its best performance with the regularization parameters $C = 1$ and $C = 10$ for the linear and RBF kernels, respectively, and the kernel coefficient $\gamma = 1.5 \times 10^{-4}$ for the non-linear kernel. The SVM classifiers, both with a linear and non-linear kernel, perform worse than the RF and NN approaches, with cross-validation accuracies of 84.2% and 82.8%, and MCCs of 0.68 and 0.65, respectively on the ENNAVIA-A dataset.

The optimal RF hyperparameters differed depending on the dataset used. For the ENNAVIA-A dataset, optimal performance was observed with 124 estimators, a maximum tree depth of 10 and a maximum of 80 features, achieving a cross-validation accuracy and MCC of 84.9% and 0.69, respectively. For the ENNAVIA-B dataset, meanwhile, optimal performance was observed with 512 estimators, unrestricted tree depth and a maximum of 13 features.

The neural network approach, however, achieves the best predictive performance of all machine learning approaches trialled, with an accuracy and MCC scores of 93.88% and 0.87 on the ENNAVIA-A external test set, and 95.65% and 0.91 on the ENNAVIA-B external test set. Furthermore, the neural network achieves a very good balance between sensitivity and specificity, 94.74% and 92.68% for ENNAVIA-A. ROC (Figure 6) curves were produced to further evaluate the neural networks' robustness, as were the corresponding AUC values, which were calculated as 0.93 and 0.98 for the ENNAVIA-A and ENNAVIA-B models, respectively.

As the neural networks' performance was superior to the SVM and RF approaches, it was deemed as the best model for the prediction of peptide antiviral activity and further studied.

Comparison with existing peptide antiviral activity prediction methods

To establish the utility of ENNAVIA in the context of prediction methods already described in the literature, ENNAVIA was benchmarked against three existing antiviral peptide prediction methods, specifically AVPpred [31], the method of Chang et al. [35], AntiVPP [32], Meta-iAVP [104] and FIRM-AVP [34]. Detailed results are given in Table 2.

The results presented in Table 2 are reproduced from the respective articles describing the methods. It must be noted, however, that the results for Meta-iAVP and AntiVPP 1.0 could not be reproduced. Independent evaluation of the Meta-iAVP via its webserver on the $V^{60p+45n}$ dataset resulted in Acc, Sn and Sp values of 81.0%, 83.3% and 77.8%, respectively. Similarly, evaluation of the AntiVPP 1.0 software on the $V^{V60p+60n}$ dataset resulted in Acc, Sn and Sp values of 81.6%, 76.6% and 86.6%, respectively. Contact with the corresponding authors of these articles was attempted prior to publication, however, we have not received a response to our queries prior to publication.

Anti-coronavirus activity prediction

A recent study by Pang et al. described a machine learning method for the identification of anti-coronavirus peptides through imbalanced learning strategies [37]. This study utilizes the datasets created by Pang et al. and employs transfer learning to adapt the ENNAVIA-A and ENNAVIA-B models to the task of anti-coronavirus peptide prediction. For both ENNAVIA-A and ENNAVIA-B, the neural network weights of each of

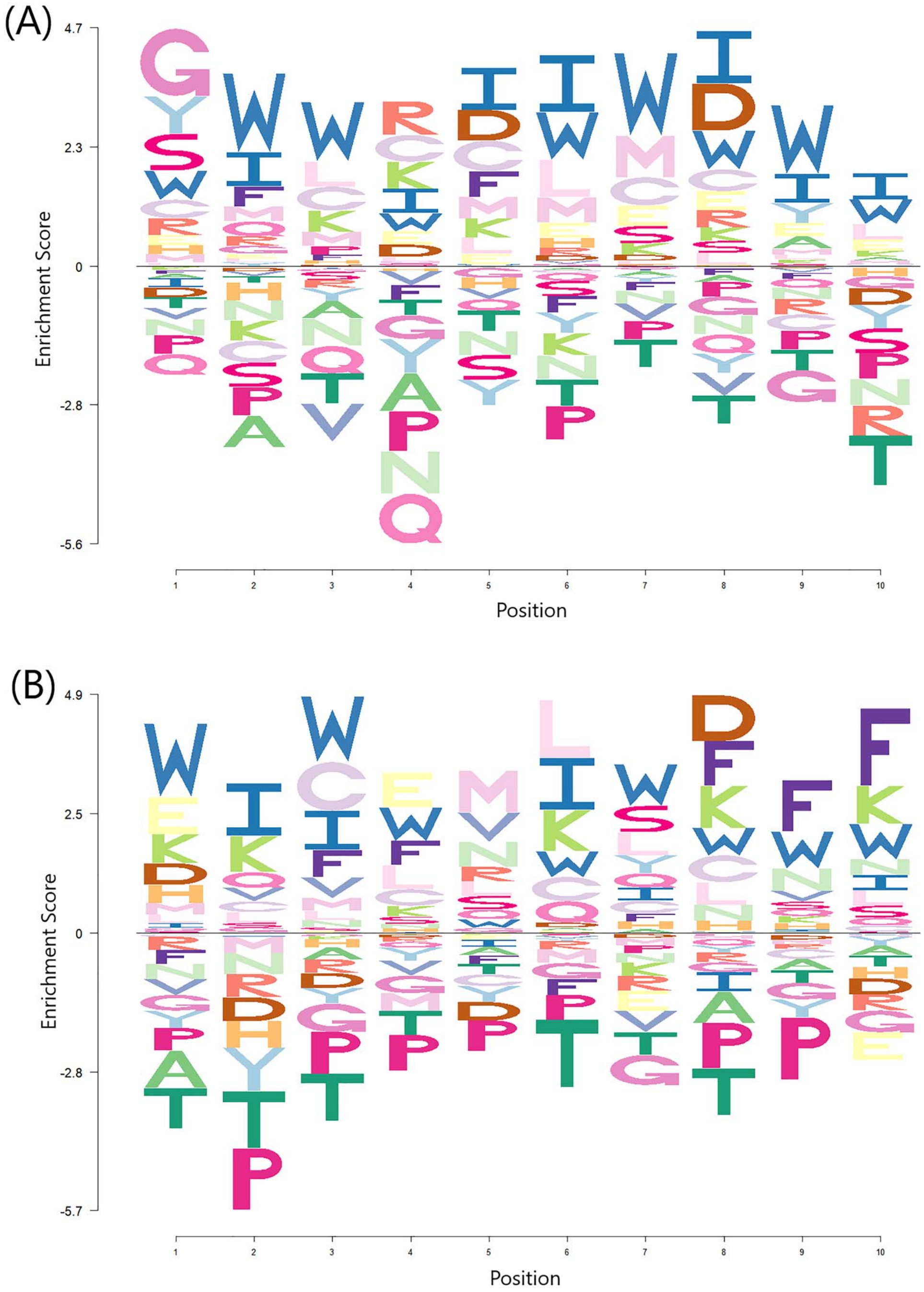

Figure 3. Enrichment-depletion logo plot of (A) N-terminal 10 residues and (B) C-terminal 10 residues of experimentally validated antiviral peptides of the ENNAVIA-A dataset. Data are scaled to account for the background probability of each amino acid, based on the experimentally validated non-antiviral peptides dataset.

the 10 models trained under cross-validation are transferred to their corresponding models for anti-coronavirus peptide prediction, which are then trained on their respective datasets. The accuracy, MCC, sensitivity and specificity parameters, together with their respective confidence intervals, are reported for each model. ROC curves with the calculated AUC values are also given for both final neural network models (Figure 6). The anti-coronavirus peptide prediction performance obtained by each model is compared with the results obtained by Pang et al. As the size of the anti-coronavirus peptide dataset is extremely limited, and neural network performance typically increases with the amount of data available, validation is limited to 10-fold cross-validation. Detailed results are given in Table 3.

Descriptor-set specific results

To ascertain the extent to which a given set of features can contribute to the correct prediction of peptide antiviral activity, neural networks were trained on subsets of the feature space. The validation results obtained by these neural networks trained on the peptides' physicochemical features, dipeptide composition, dipeptide *g*-gap composition and tripeptide composition are detailed in Table 4.

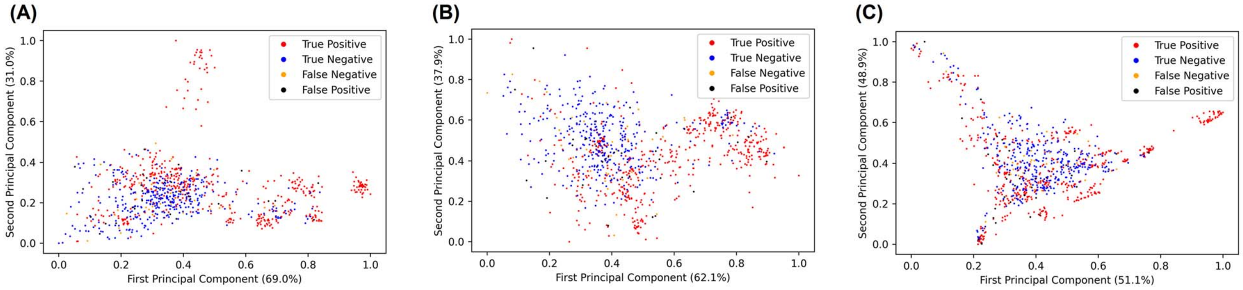

Figure 4. Principal component analysis of (A) all descriptors, (B) only the physicochemical descriptors and (C) only the compositional descriptors. Experimentally validated antiviral peptides (positives) are coloured red, experimentally validated non-antiviral peptides (negatives) are coloured yellow, false-positives are coloured black and false-negatives are coloured blue.

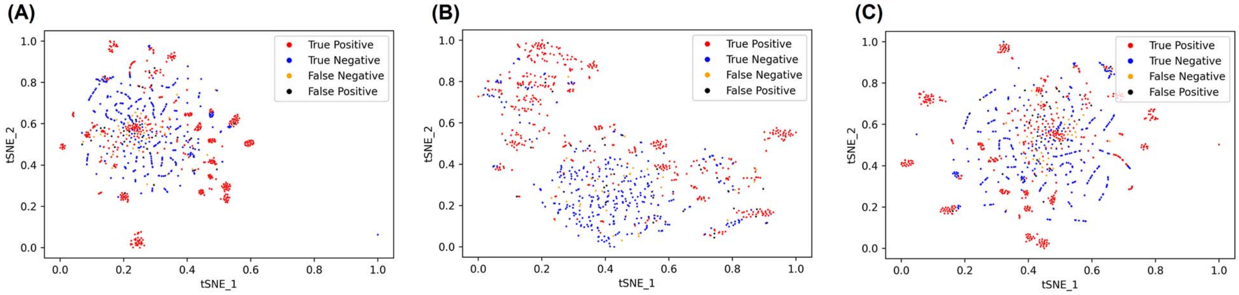

Figure 5. t-SNE visualization of (A) all descriptors, (B) only the physicochemical descriptors and (C) only the compositional descriptors. Experimentally validated antiviral peptides (positives) are coloured red, experimentally validated non-antiviral peptides (negatives) are coloured yellow, false-positives are coloured black and false-negatives are coloured blue.

Table 1. Cross-validation and external validation statistical parameters for SVM with a linear kernel, SVM with a RBF kernel, RF and NN models trained on the ENNAVIA-A and ENNAVIA-B datasets

		Cross-validation			
Model Dataset	Method	Acc	Sn	Sp	MCC
ENNAVIA-A	SVM (linear)	84.20 ± 2.38	86.83 ± 2.91	81.03 ± 3.95	0.68 ± 0.05
	SVM (RBF)	82.80 ± 2.47	85.92 ± 2.99	78.85 ± 4.11	0.65 ± 0.05
	RF	84.87 ± 2.34	89.77 ± 2.60	78.44 ± 4.14	0.69 ± 0.05
	NN	91.25 ± 1.85	90.56 ± 2.51	91.88 ± 2.75	0.82 ± 0.04
ENNAVIA-B	SVM (linear)	92.72 ± 1.56	91.96 ± 2.34	93.47 ± 2.09	0.85 ± 0.03
	SVM (RBF)	90.14 ± 1.80	84.02 ± 3.15	96.20 ± 1.61	0.81 ± 0.04
	RF	92.92 ± 1.54	91.36 ± 2.41	94.42 ± 1.94	0.86 ± 0.03
	NN	95.90 ± 1.19	93.44 ± 2.13	98.35 ± 1.07	0.92 ± 0.02
		External validation			
Model Dataset	Method	Acc	Sn	Sp	MCC
ENNAVIA-A	SVM (linear)	83.77 ± 7.30	91.71 ± 7.16	72.72 ± 13.63	0.67 ± 0.15
	SVM (RBF)	80.24 ± 7.88	91.23 ± 7.34	64.97 ± 14.60	0.59 ± 0.16
	RF	84.23 ± 7.22	93.30 ± 6.49	71.62 ± 13.80	0.68 ± 0.15
	NN	93.88 ± 4.75	94.74 ± 5.80	92.68 ± 7.97	0.87 ± 0.10
ENNAVIA-B	SVM (linear)	87.35 ± 6.07	91.07 ± 7.40	83.70 ± 9.51	0.75 ± 0.12
	SVM (RBF)	89.17 ± 5.68	87.56 ± 8.57	90.75 ± 7.46	0.78 ± 0.11
	RF	89.88 ± 5.51	91.39 ± 7.28	88.40 ± 8.24	0.80 ± 0.11
	NN	95.65 ± 3.73	92.98 ± 6.63	98.28 ± 3.35	0.91 ± 0.07

None of the reduced subset models trained achieve performance better than the hybrid model trained on both compositional and physicochemical descriptors, validating the choice of the hybrid model as the principal approach.

Dipeptide and tripeptide composition

Information about local sequence order can be relayed to a machine learning method through the use of dipeptide and tripeptide composition descriptors. A peptide's dipeptide

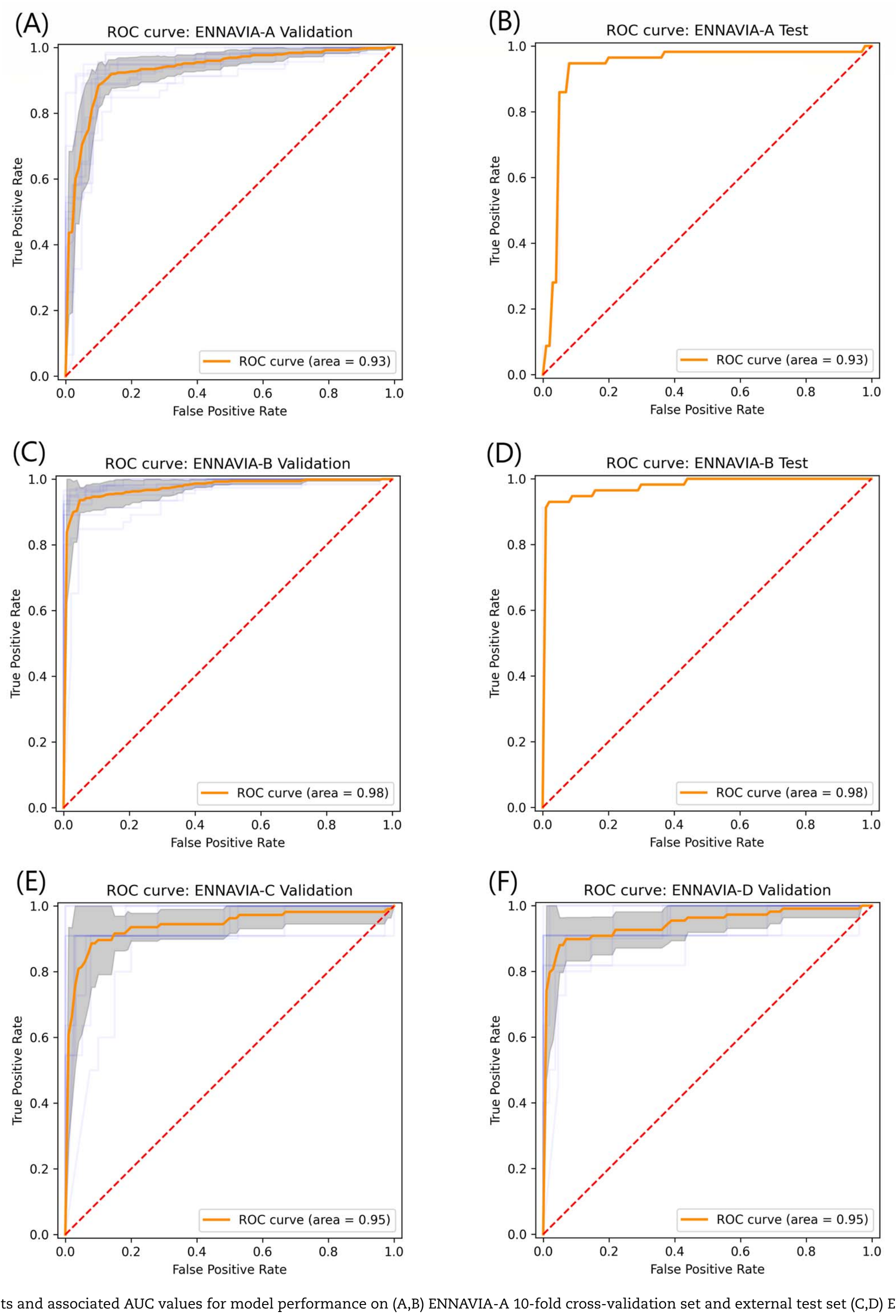

Figure 6. ROC plots and associated AUC values for model performance on (A,B) ENNAVIA-A 10-fold cross-validation set and external test set (C,D) ENNAVIA-B 10-fold cross-validation set and external test set and (E,F) ENNAVIA-C and ENNAVIA-D 10-fold cross-validation sets. The ENNAVIA-A model achieves an AUC value of 0.93, ENNAVIA-B achieves and AUC value of 0.98 and ENNAVIA-C and ENNAVIA-D both achieve values of 0.95.

and tripeptide composition can be defined as the percentage of a given dipeptide or tripeptide in the sequence. These features also have the added benefit of capturing the peptide’s chemical nature. Both the dipeptide-based model and the tripeptide-based model achieve good results, with accuracies of 90.1% and 89.8%, respectively, and MCC values of 0.80.

g-gap composition

g-gap compositions, defined as the proportion of a pair of amino acids separated by 1, 2 or 3 residues, are a useful descriptor as they correspond to residues that may be proximate to one another in three-dimensional space. As peptides often possess

Table 2. Cross-validation and external validation performance comparison between ENNAVIA and existing methods for the prediction of peptide antiviral activity.

Dataset	Method	Cross-validation Classifier[1]	Features[2]	Acc	Sn	Sp	MCC
$T^{544p+407n}$	AVPpred	SVM	AAindex	85.0	82.2	88.2	0.70
	Chang et al’s method	RF	AAC, Aggr.	85.1	86.6	83.0	0.70
	AntiVPP 1.0	RF	PC	–	–	–	–
	Meta-iAVP	M-P	AAC,APseAAC	88.20	89.20	86.90	0.76
	Firm-AVP	SVM	AAC,DPC,PseAAC,APseAAC, PC,SS	–	–	–	–
	ENNAVIA	NN	AAC,DPC,AAindex,PC	91.25	90.56	91.88	0.82
$T^{544p+544n}$	AVPpred	SVM	AAindex	90.0	89.7	90.3	0.80
	Chang et al’s method	RF	AAC, Aggr.	91.5	89.0	94.1	0.83
	AntiVPP 1.0	RF	PC	–	–	–	–
	Meta-iAVP	M-P	AAC,Am-PseAAC	93.20	89.00	97.40	0.87
	Firm-AVP	SVM	AAC,DPC,PseAAC,APseAAC, PC,SS	–	–	–	–
	ENNAVIA	NN	AAC,DPC,AAindex,PC	95.90	93.44	98.35	0.92
		External validation					
$V^{60p+45n}$	AVPpred	SVM	AAindex	85.7	88.3	82.2	0.71
	Chang et al’s method	RF	AAC, Aggr.	89.5	91.7	86.7	0.79
	AntiVPP 1.0	RF	PC	–	–	–	–
	Meta-iAVP	M-P	AAC,Am-PseAAC	95.20	96.70	93.20	0.90
	Firm-AVP	SVM	AAC,DPC,PseAAC,APseAAC, PC,SS	92.4	93.3	91.1	0.84
	ENNAVIA	NN	AAC,DPC,AAindex,PC	93.88	94.74	92.68	0.87
$V^{60p+60n}$	AVPpred	SVM	AAindex	92.5	93.3	91.7	0.85
	Chang et al’s method	RF	AAC, Aggr.	93.0	91.7	95.0	0.87
	AntiVPP 1.0	RF	PC	93	87	97	0.87
	Meta-iAVP	M-P	AAC,Am-PseAAC	94.90	91.70	98.30	0.90
	Firm-AVP	SVM	AAC,DPC,PseAAC,APseAAC, PC,SS	–	–	–	–
	ENNAVIA	NN	AAC,DPC,AAindex,PC	95.65	92.98	98.28	0.91

[1] SVM: Support vector machine, RF: Random forest, M-P: Meta-predictor, NN: Neural network;
[2] AAindex: Amino acid index database, Aggr: Aggregation propensity, PC: Physicochemical properties, AAC: Amino acid composition, PseAAC: Pseudo amino acid composition, APseAAC: Amphiphilic pseudo amino acid composition, DPC: Dipeptide composition, SS: Predicted secondary structure information.

Table 3. Performance evaluation of ENNAVIA in prediction of anti-coronavirus peptides, and comparison with the methods of Pang et al.

Dataset	Method	Classifier[1]	Acc	Sn	Sp	MCC
Anti-CoV vs. Non-AVP	Pang et al’s method	NM,BRF	85.32	85.71	85.31	0.305
ENNAVIA-C	ENNAVIA	NN	94.95 ± 1.99	91.64 ± 5.20	95.96 ± 2.04	0.87 ± 0.05
Anti-CoV vs. Random	Pang et al’s method	NM,BRF	97.72	100	97.66	0.730
ENNAVIA-D	ENNAVIA	NN	97.29 ± 1.25	89.82 ± 5.68	98.77 ± 0.93	0.91 ± 0.03

[1] NM: Near-miss undersampling, BRF: Balanced random-forest, NN: Neural network with transfer learning;
[2] Pang et al.’s method utilizes amino acid composition, pseudo amino acid composition, dipeptide composition and physicochemical features. ENNAVIA utilizes amino acid composition, dipeptide composition, physicochemical features and features calculated from the amino acid index database.

secondary structure upon interaction with their targets, this information allows for the capturing of the chemical environment that the peptide presents to its target. Models trained on *g*-gap dipeptide composition do not perform better than those trained on conventional dipeptide composition, achieving an accuracy and MCC of 90.0% and 0.80.

Physicochemical

Models trained on physicochemical features, such as charge, amphiphilicity and charge, achieve an accuracy and MCC of 88.3% and 0.76, respectively. Although this performance is poorer than that achieved by the models trained on compositional features, it is only marginally so, and still demonstrates predictive capability.

Prediction based on selected features

Feature selection was performed for each validation split using SVMs and random forests; the 500 features with the largest absolute SVM weights, and the 500 features with the largest RF feature importance were selected. Neural network models were

Table 4. Validation statistics achieved by neural network models trained on subsets of the feature space. The g-gap parameter g=1,2,3.

		Cross-validation			
Dataset	Features	Acc (%)	Sn (%)	Sp (%)	MCC
ENNAVIA-A	Composition, dipeptide	90.05 ± 1.96	90.20 ± 2.56	89.96 ± 3.03	0.80 ± 0.04
	Composition, tripeptide	89.82 ± 1.98	88.73 ± 2.72	91.42 ± 2.82	0.80 ± 0.04
	g-gap composition, dipeptide	90.01 ± 1.96	90.85 ± 2.48	88.82 ± 3.17	0.80 ± 0.04
	Physicochemical	88.30 ± 2.10	89.95 ± 2.58	85.69 ± 3.53	0.76 ± 0.04
ENNAVIA-B	Composition, dipeptide	94.84 ± 1.33	92.50 ± 2.26	97.14 ± 1.41	0.90 ± 0.03
	Composition, tripeptide	95.17 ± 1.29	92.30 ± 2.29	97.93 ± 1.20	0.90 ± 0.03
	g-gap composition, dipeptide	94.93 ± 1.32	92.64 ± 2.24	97.21 ± 1.39	0.90 ± 0.03
	Physicochemical	94.00 ± 1.43	92.18 ± 2.31	95.87 ± 1.68	0.88 ± 0.03
		External validation			
Dataset	Features	Acc (%)	Sn (%)	Sp (%)	MCC
ENNAVIA-A	Composition, dipeptide	90.14 ± 5.90	96.49 ± 4.78	81.30 ± 11.94	0.80 ± 0.12
	Composition, tripeptide	90.14 ± 5.90	92.40 ± 6.88	86.99 ± 10.30	0.80 ± 0.12
	g-gap composition, dipeptide	87.76 ± 6.49	91.81 ± 7.12	82.11 ± 11.73	0.75 ± 0.13
	Physicochemical	85.71 ± 6.93	94.15 ± 6.09	73.98 ± 13.43	0.71 ± 0.14
ENNAVIA-B	Composition, dipeptide	87.83 ± 5.98	90.06 ± 7.77	85.63 ± 9.03	0.76 ± 0.12
	Composition, tripeptide	89.28 ± 5.66	88.89 ± 8.16	89.66 ± 7.84	0.79 ± 0.11
	g-gap composition, dipeptide	92.17 ± 4.91	89.47 ± 7.97	94.83 ± 5.70	0.84 ± 0.10
	Physicochemical	87.54 ± 6.04	89.47 ± 7.97	85.63 ± 9.03	0.75 ± 0.12

Table 5. Cross-validation and external validation results achieved with neural network models trained on feature sets reduced by feature selection.

		Cross-validation		
Dataset	Acc	Sn	Sp	MCC
ENNAVIA-A	89.65 ± 1.99	91.00 ± 2.46	87.60 ± 3.32	0.79 ± 0.04
ENNAVIA-B	94.94 ± 1.32	92.83 ± 2.22	96.79 ± 1.49	0.90 ± 0.03
ENNAVIA-C	93.94 ± 2.50	83.64 ± 6.95	96.91 ± 1.80	0.83 ± 0.06
ENNAVIA-D	68.22 ± 4.84	80.61 ± 7.42	64.15 ± 4.98	0.39 ± 0.10
		External validation		
Dataset	Acc	Sn	Sp	MCC
ENNAVIA-A	89.80 ± 5.99	94.74 ± 5.80	82.93 ± 11.52	0.79 ± 0.12
ENNAVIA-B	87.83 ± 5.98	91.23 ± 7.34	84.48 ± 9.32	0.76 ± 0.12

constructed and trained on the sets of selected features, the results are presented in Table 5.

The prediction results obtained in all cases are inferior to those obtained by the models trained on the full feature sets, most notably in the case of the models trained on the ENNAVIA-D dataset. This is not unexpected, considering that the feature selection is performed on the ENNAVIA-B dataset prior to transfer learning, which appears to result in the exclusion of features important for anti-coronavirus activity prediction.

Discussion

The need for novel anti-viral drugs, especially in the context of the COVID-19 pandemic, is great. Interest in the development of novel peptide-based therapeutics has increased in recent years, even as the number of new drugs approved each year declines and the cost of drug research and development grows. More specifically, AVPs represent a promising class of novel drug candidates. Despite extensive research having been conducted on the relationship between the conformations of various bioactive peptides and their biological activities [105–107], understanding of this relationship remains insufficient for the accurate de-novo design of novel peptide drugs, especially antiviral peptide drugs, which compared to antimicrobial peptides are less numerous in the literature, and consequently less studied. Molecular dynamics simulations can reveal insights into activity, but are time-consuming and largely unsuitable for bulk-screening of peptide sequences.

An accurate computational method for the prediction of peptide antiviral activity from the primary sequence alone would facilitate a more rapid exploration of the peptide chemical space, and lower the cost of research and development by reducing the need for chemical synthesis and laboratory evaluation of peptide antiviral activity. With a view to accelerating the screening and design of new antiviral peptide drugs, the present study focuses on the combination of compositional and physicochemical descriptors with a deep neural network architecture to create an *in silico* method for a more accurate classification of peptides as either antiviral or non-antiviral, and additionally the prediction of peptide anti-coronavirus activity specifically, solely on the basis of their primary sequence.

To facilitate as direct a comparison as possible with existing antiviral peptide prediction methods, the dataset of Thakur et al. [31] was adapted for use in this study. Peptide sequences comprising non-natural amino acids or with a length outside the 7–40 amino acid range were excluded. A total of 577 of the original 604 AVPs remain in the ENNAVIA datasets. Two negative datasets are used in this study: the ENNAVIA-A dataset includes 420 experimentally evaluated non-antiviral peptides, while the ENNAVIA-B dataset includes 597 random peptide sequences as the negative samples.

Compositional and physicochemical descriptors were employed for the construction of feature vectors from the peptides' primary sequences, and a selection of machine learning methods were evaluated for the peptide antiviral activity prediction task through both 10-fold cross-validation and validation on an external test set. Deep neural networks proved most promising, and their architecture was, therefore, further optimized and evaluated.

The neural network model with five hidden layers was found to achieve optimal performance. On the ENNAVIA-A dataset, a 10-fold cross-validated accuracy, sensitivity and specificity of 91.3%, 90.6% and 91.9% was achieved, clearly demonstrating that the neural network model is capable of accurately identifying AVPs among non-antiviral peptides. ENNAVIA's predictive performance was compared to existing methods, especially the existent state-of-the-art, Meta-iAVP, which exhibited a cross-validated accuracy, sensitivity and specificity of 88.2%, 89.2% and 86.9%, respectively, on the $T^{504p+407n}$ dataset. ENNAVIA's performance surpasses that of Meta-iAVP and other existent models on all metrics, designating it a new state-of-the-art model for antiviral peptide prediction.

Similarly, neural network models were trained and evaluated on the ENNAVIA-B dataset, achieving cross-validated accuracy, sensitivity and specificity of 95.9%, 93.4% and 98.6%, respectively, demonstrating that ENNAVIA can distinguish between AVPs and random peptide sequences. A comparison of performance on this dataset to existing methods again establishes ENNAVIA as the best-in-class method for antiviral peptide prediction, surpassing the previously best accuracy, sensitivity and specificity of 93.2%, 89.0% and 97.4% achieved by meta-iAVP on the $T^{504p+504n}$ dataset.

Recently, Pang et al. published a study that employed random forests with imbalanced learning strategies for the identification of anti-coronavirus peptides. Notably, the anti-coronavirus peptide dataset is small, with only a total of 139 peptide sequences. Despite the small number of positive samples available for training, respectable validation statistics were achieved, with a sensitivity, specificity and MCC of 85.7%, 85.3% and 0.31 with non-antivirus peptides as the negative dataset, and 100%, 97.7% and 0.73 with random peptide sequences as the negative dataset.

To expand the scope of the current study to include the facilitation of rapid screening of peptides for anti-coronavirus activity specifically, two additional datasets which include the anti-coronavirus peptides from the dataset of Pang et al. as the positive samples were constructed: ENNAVIA-C and ENNAVIA-D, which use the negative peptides from the ENNAVIA-A and ENNAVIA-B datasets, respectively. As the number of positive samples is too small to accurately train neural network models, transfer learning was employed, whereby the already-trained weights of the ENNAVIA-A and ENNAVIA-B models were transferred to the ENNAVIA-C and ENNAVIA-D models, respectively, and further fine-tuned to the anti-coronavirus peptide prediction task. The ENNAVIA-C model achieved a sensitivity, specificity and MCC of 91.6%, 96.0% and 0.87, representing a significant improvement on the work of Pang et al. The ENNAVIA-D model, similarly, achieved good performance, with a sensitivity, specificity and MCC of 89.8%, 98.8% and 0.91, respectively, outperforming the method of Pang et al. in specificity, although not sensitivity.

The ENNAVIA model does possess drawbacks, some of which it shares with the other existing algorithms. Since the publication of the dataset of Thakur et al. [31], the literature on AVPs has expanded, and continues to expand as new AVPs continue to be identified. Consequently, the number of peptide sequences available for training increases. As neural networks' predictive power scales with the quantity of data available for training, further improvements in predictive performance for both the antiviral and anti-coronavirus predictive models could be achieved through the development of an updated, expanded dataset. Neural networks are generally known as non-interpretable black box models, which precludes rigorous analysis of the basis for the model's predictions. Furthermore, as mentioned previously, AVPs can exert their biological activity through a variety of host-targeting and virus-targeting mechanisms of action, which can include the prevention of virus cell-entry, blocking cell receptors, viral lysis or enhancement of host immune response. While it stands to reason that the mechanism of action a given peptide utilizes to exert antiviral activity depends on the peptide's amino acid composition and physicochemical properties, unfortunately the number of known antiviral peptide sequences still cannot be considered plentiful, much less the number of AVPs that utilize a given mechanism of action. For instance, while inhibition of virus entry is the most prevalent mechanism by which AVPs exert their action, accounting for 30% of entries in the AVPdb, only seven peptides are listed in the AVPdb as exerting their antiviral activity through immunostimulation [26]. Consequently, it is not always feasible to analyse the relationships between peptides' properties and their mode of action, nor is it currently feasible to construct machine learning models that are specific to a mode of action. Instead, antiviral activity predictors remain limited to the prediction of the presence or absence of antiviral activity.

To conclude, the limited quantity of available experimentally validated data and the incomplete understanding of the mechanism of peptide antiviral activity continue to pose challenges for the research community. In an effort to overcome these challenges, this study described ENNAVIA, a collection of novel *in silico* peptide antiviral and anti-coronavirus activity classifiers. The classifiers, which employ a deep neural network architecture and benefit from a rich feature-space, achieve predictive power that surpasses the state-of-the-art. This work complements a suite of existing *in silico* classifiers developed by the authors, which includes methods for the prediction of peptide anticancer and hemolytic activity, and peptide tertiary structure. The authors believe that the results of this work, in combination with the aforementioned methods, will enable better *in-silico* design of novel peptide-based antiviral and anti-coronavirus therapeutics, thereby reducing the cost and time required for the design phase, helping to drive medicinal chemistry into an unprecedented revolution.

Web server implementation

For the benefit of the scientific community, the ENNAVIA classifier is available as a user-friendly, publicly accessible web server online at https://research.timmons.eu/ennavia. The web server is capable of predicting peptides' antiviral activity based on the primary sequence. Input peptide sequences are restricted to only the 20 natural amino acids; non-natural amino acids are not supported. The web server includes many features, and models trained on the ENNAVIA-A ($T^{504p+406n}$) , ENNAVIA-B

($T^{504p+504n}$), ENNAVIA-C and ENNAVIA-D datasets are available for prediction.

Peptide antiviral activity prediction

Peptide antiviral activities can be predicted for both a single sequence and a batch of sequences. Peptide sequences should be provided in the standard FASTA format. The maximum batch size is variable depending on the length of the sequences; longer sequences necessitate smaller batch sizes. The prediction will be carried out by the ensemble of trained neural networks, and the average score will be returned, which corresponds to the probability of the peptide sequence possessing antiviral activity. Probabilities are given on a scale of 0–1, whereby 0 and 1 are most probably non-antiviral, and most probably antiviral, respectively.

Mutation analysis

Mutation analysis may be carried out on single peptide sequences, by selecting the mutation analysis option and inputting the residue number to be mutated. Mutant sequences will be created by substituting the residue at the specified position with each of the other 20 natural amino acids. The probability of each of the mutant sequences possessing antiviral activity will be returned by the chosen neural network model.

Residue scan

Residue scans, such as, for instance, an alanine scan, are available for single peptide sequences, by choosing the residue scan option and selecting the amino acid residue to be scanned with. Mutant sequences are attained by substituting successive residues with the selected amino acid residue. The probability of the native and mutant sequences possessing antiviral activity will be returned by the selected neural network model.

Key Points

- An artificial neural network model ENNAVIA was constructed for the prediction of antiviral and anti-coronavirus peptides
- Feature extraction was used to obtain compositional and physicochemical descriptors from the peptide sequences
- Transfer learning was employed to adapt neural networks for anti-coronavirus activity prediction
- ENNAVIA was evaluated by 10-fold cross-validation and an external test set
- ENNAVIA outperforms the current best-in-class methods for antiviral peptide prediction

Data Availability

All data generated or analysed during this study are available for download at https://research.timmons.eu/ennavia.

Acknowledgements

The authors would also like to thank University College Dublin for the Research Scholarship granted to P.B.T.

References

1. Koonin EV, Senkevich TG, Dolja VV. The ancient Virus World and evolution of cells. *Biol Direct* 2006;**1**:29. http://www.biology-direct.com/content/1/1/29.
2. Nichol ST, Arikawa J, Kawaoka Y. Emerging viral diseases. *Proc Natl Acad Sci U S A* 2000;**97**:12411–2. https://pubmed.ncbi.nlm.nih.gov/11035785/.
3. Domingo E. Mechanisms of viral emergence. *Vet Res* 2010;**41**:38. https://pubmed.ncbi.nlm.nih.gov/20167200/.
4. Phan T. Genetic diversity and evolution of SARS-CoV-2. *Infect Genet Evol* 2020;**81**:104260. https://pubmed.ncbi.nlm.nih.gov/32092483/.
5. Goldenthal, K. L., Midthun, K. & Zoon, K. C. *Control of Viral Infections and Diseases* (University of Texas Medical Branch at Galveston, Galveston, TX 1996). URL http://www.ncbi.nlm.nih.gov/pubmed/21413344.
6. Mahlapuu, M., Håkansson, J., Ringstad, L. & Björn, C. Antimicrobial peptides: An emerging category of therapeutic agents. *Front Cell Infect Microbiol* 2016; **6**:194. URL www.frontiersin.org http://www.ncbi.nlm.nih.gov/pubmed/28083516 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5186781.
7. Hancock RE, Diamond G. The role of cationic antimicrobial peptides in innate host defences. *Trends Microbiol* 2000;**8**:402–10. https://www.sciencedirect.com/science/article/pii/S0966842X00018230.
8. Mahendran ASK, Lim YS, Fang CM, *et al.* The Potential of Antiviral Peptides as COVID-19 Therapeutics. *Front Pharmacol* 2020;**11**:575444. https://www.frontiersin.org/article/10.3389/fphar.2020.575444/full.
9. Zhao H, *et al.* A novel peptide with potent and broad-spectrum antiviral activities against multiple respiratory viruses. *Sci Rep* 2016;**6**:1–13. www.nature.com/scientificreports.
10. Li Q, *et al.* Virucidal activity of a scorpion venom peptide variant mucroporin-M1 against measles, SARS-CoV and influenza H5N1 viruses. *Peptides* 2011;**32**:1518–25.
11. Lu L, *et al.* Structure-based discovery of Middle East respiratory syndrome coronavirus fusion inhibitor. *Nat Commun* 2014;**5**:3067. https://pubmed.ncbi.nlm.nih.gov/24473083/.
12. Otvos L. Peptide-based drug design: Here and now. *Methods Mol Biol.* 2008;**494**:1–8. https://pubmed.ncbi.nlm.nih.gov/21413344/.
13. Lau JL, Dunn MK. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic and Medicinal Chemistry* 2018;**26**:2700–7.
14. Furka Á, Sebestyén F, Asgedom M, *et al.* General method for rapid synthesis of multicomponent peptide mixtures. *Int J Pept Protein Res* 1991;**37**:487–93. http://doi.wiley.com/10.1111/j.1399-3011.1991.tb00765.x.
15. Sohrabi C, Foster A, Tavassoli A. Methods for generating and screening libraries of genetically encoded cyclic peptides in drug discovery. *Nature Reviews Chemistry* 2020;**4**:90–101.
16. Bozovičar K, Bratkovič T. Evolving a peptide: Library platforms and diversification strategies. 2020;**21**:215.
17. Larue RC, *et al.* Rationally Designed ACE2-Derived Peptides Inhibit SARS-CoV-2. *Bioconjug Chem* 2021;**32**:215–23. https://dx.doi.org/10.1021/acs.bioconjchem.0c00664.
18. Lou, Z., Sun, Y. & Rao, Z. Current progress in antiviral strategies. *Trends in Pharmacological Sciences* (2014);**35**:86–102. URL https://pubmed.ncbi.nlm.nih.gov/24439476/.

19. McDonald CK. Human Immunodeficiency Virus Type 1 Protease Inhibitors. *Arch Intern Med* 1997;**157**:951. https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/623267.
20. Kiser, J. J. & Flexner, C. Direct-acting antiviral agents for hepatitis c virus infection. *Annual Review of Pharmacology and Toxicology* 2013;**53**:427–449. URL https://pubmed.ncbi.nlm.nih.gov/23140245/.
21. Yu F, Lu L, du L, *et al*. Approaches for Identification of HIV-1 Entry Inhibitors Targeting gp41 Pocket. *Viruses* 2013;**5**:127–49. http://www.mdpi.com/1999-4915/5/1/127.
22. el Raziky M, Fathalah WF, el-akel WA, *et al*. The effect of peginterferon alpha-2a vs. peginterferon alpha-2b in treatment of naive chronic HCV genotype-4 patients: A single centre egyptian study. *Hepatitis Monthly* 2013;**13**:10069. https://sites.kowsarpub.com/hepatmon/articles/70462.html, https://sites.kowsarpub.com/hepatmon/articles/70462.html#abstract.
23. Lin F-c, Young HA. Interferons: Success in anti-viral immunotherapy. *Cytokine & Growth Factor Reviews* 2014;**25**:369–376. https://linkinghub.elsevier.com/retrieve/pii/S135961011400077.
24. Vilas Boas LCP, Campos, ML, Berlanda RLA, *et al*. Antiviral peptides as promising therapeutic drugs. *Cellular and Molecular Life Sciences* 2019;**76**:3525–3542. https://doi.org/10.1007/s00018-019-03138-w.
25. Agarwal, G. & Gabrani, R. Antiviral Peptides: Identification and Validation. *International Journal of Peptide Research and Therapeutics* 2021;**27**:149–168. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7233194/.
26. Qureshi A, Thakur N, Tandon H, *et al*. AVPdb: A database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res* 2014;**42**:D1147–D1153. https://pubmed.ncbi.nlm.nih.gov/24285301/.
27. Pirtskhalava M, Gabrielian A, Cruz P, *et al*. Erratum: DBAASP v.2: An enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides (Nucleic Acids Research 44 (D1104-D1112) DOI: 10.1093/nar/gkv1174). *Nucleic Acids Res* 2016;**44**:6503. http://www.ncbi.nlm.nih.gov/pubmed/27060142 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4994862, https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw243.
28. Pirtskhalava M, Amstrong AA, Grigolava 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* 2021;**49**:D288–97. http://dbaasp.org.
29. Waghu FH, Gopi L, Barai RS, *et al*. CAMP: Collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Res* 2014;**42**:D1154–8. http://www.ncbi.nlm.nih.gov/pubmed/24265220, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3964954.
30. Wang G, Li X, Wang Z. APD3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res* 2016;**44**:D1087–93.
31. Thakur, N., Qureshi, A. & Kumar, M. AVPpred: Collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res* **40**, W199–W204 (2012). URL /pmc/articles/PMC3394244/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3394244/.
32. Beltrán Lissabet JF, Belén LH, Farias JG. AntiVPP 1.0: A portable tool for prediction of antiviral peptides. *Comput Biol Med* 2019;**107**:127–30.
33. Schaduangrat N, Nantasenamat C, Prachayasittikul V, *et al*. Meta-iavp: A sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int J Mol Sci* 2019;**20**:5743. http://codes.bio/meta-iavp/.
34. Chowdhury AS, Reehl SM, Kehn-Hall K, *et al*. Better understanding and prediction of antiviral peptides through primary and secondary structure feature importance. *Sci Rep* 2020;**10**:1–8. https://doi.org/10.1038/s41598-020-76161-8.
35. Chang KY, Yang J-R. Analysis and Prediction of Highly Effective Antiviral Peptides Based on Random Forests. *PLoS ONE* 2013;**8**:e70166. https://dx.plos.org/10.1371/journal.pone.0070166.
36. Charoenkwan P, Anuwongcharoen N, Nantasenamat C, *et al*. In silico approaches for the prediction and analysis of antiviral peptides: a review. *Curr Pharm Des* 2020;**27**:2180–2188. https://www.eurekaselect.com/187420/article.
37. Pang Y, Wang Z, Jhong J-H, *et al*. Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies. *Brief Bioinform* 2021;**22**:1085–1095. https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa423/6120286.
38. Zeng M, Zhang F, Wu FX, *et al*. Protein-protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* 2019;**36**:1114–20. https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz699/5564115.
39. Mooney C, Wang YH, Pollastri G. SCLpred: Protein subcellular localization prediction by N-to-1 neural networks. *Bioinformatics* 2011;**27**:2812–9. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr494.
40. Kaleel M, Zheng Y, Chen J, *et al*. SCLpred-EMS: Subcellular localization prediction of endomembrane system and secretory pathway proteins by Deep N-to-1 Convolutional Neural Networks. *Bioinformatics* 2020;**36**:3343–9. https://academic.oup.com/bioinformatics/article/36/11/3343/5788524.
41. Holton TA, Pollastri G, Shields DC, *et al*. CPPpred: Prediction of cell penetrating peptides. *Bioinformatics* 2013;**29**:3094–6. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt518.
42. Timmons PB, Hewage CM. HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks. *Sci Rep* 2020;**10**:10869. http://www.nature.com/articles/s41598-020-67701-3.
43. Timmons PB, Hewage CM. ENNAACT is a novel tool which employs neural networks for anticancer activity classification for therapeutic peptides. *Biomed Pharmacother* 2021;**133**:111051.
44. Timmons, P. B. & Hewage, C. M. APPTEST is an innovative new method for the automatic prediction of peptide tertiary structures. *bioRxiv* 2021.03.09.434600 (2021). URL https://doi.org/10.1101/2021.03.09.434600.
45. Lata, S., Mishra, N. K. & Raghava, G. P. AntiBP2: Improved version of antibacterial peptide prediction. *BMC Bioinformatics* **11**, S19 (2010). URL /pmc/articles/PMC3009489/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3009489/.
46. Huang Y, Niu B, Gao Y, *et al*. CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**:680–2. http://www.ncbi.nlm.nih.gov/pubmed/20053844, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2828112.

47. Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.
48. Dey KK, Xie D, Stephens M. A new sequence logo plot to highlight enrichment and depletion. *BMC Bioinformatics* 2018;**19**:473. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2489-3.
49. Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* 2018;**34**: 2740–7. http://www.ncbi.nlm.nih.gov/pubmed/29590297, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6084614.
50. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci U S A* 1996;**93**:11628–33. http://www.ncbi.nlm.nih.gov/pubmed/8876187, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC38109, http://www.pnas.org/cgi/doi/10.1073/pnas.93.21.11628.
51. Shen J, Zhang J, Luo X, *et al*. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A* 2007;**104**:4337–41. http://www.ncbi.nlm.nih.gov/pubmed/17360525, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1838603.
52. Ding H, Feng PM, Chen W, *et al*. Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. *Mol Biosyst* 2014;**10**:2229–35. http://lin.uestc.edu.cn/server/PVPred.
53. Dong J, Yao ZJ, Zhang L, *et al*. PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. *J Chem* 2018;**10**:16. http://www.ncbi.nlm.nih.gov/pubmed/29556758, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5861255, https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0270-2.
54. Cao DS, Liang YZ, Yan J, *et al*. PyDPI: Freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies. *J Chem Inf Model* 2013;**53**: 3086–96. http://www.ncbi.nlm.nih.gov/pubmed/24047419, https://pubs.acs.org/doi/10.1021/ci400127q.
55. Müller AT, Gabernet G, Hiss JA, *et al*. modlAMP: Python for antimicrobial peptides. *Bioinformatics (Oxford, England)* 2017;**33**:2753–5. https://academic.oup.com/bioinformatics/article/33/17/2753/3796392.
56. Ikai A. Thermostability and Aliphatic Index of Globular Proteins. *The Journal of Biochemistry* 1980;**88**:1895–8. http://www.ncbi.nlm.nih.gov/pubmed/7462208.
57. Lobry JR, Gautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 escherichia coli chromosome-encoded genes. *Nucleic Acids Res* 1994;**22**:3174–80. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/22.15.3174.
58. Boman HG, Wade D, Boman IA, *et al*. Antibacterial and antimalarial properties of peptides that are cecropin-melittin hybrids. *FEBS Lett* 1989;**259**: 03–106. http://doi.wiley.com/10.1016/0014-5793%2889%2981505-4.
59. Argos P, Rao JK, Hargrave PA. Structural Prediction of Membrane-Bound Proteins. *Eur J Biochem* 1982;**128**:565–75. http://www.ncbi.nlm.nih.gov/pubmed/7151796, http://doi.wiley.com/10.1111/j.1432-1033.1982.tb07002.x.
60. Eisenberg D, Weiss RM, Terwilliger TC, *et al*. Hydrophobic moments and protein structure. *Faraday Symposia of the Chemical Society* 1982;**17**:109–20. http://xlink.rsc.org/?DOI=fs9821700109.
61. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;**157**:105–32. http://www.ncbi.nlm.nih.gov/pubmed/7108955, https://linkinghub.elsevier.com/retrieve/pii/0022283682905150.
62. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci U S A* 1981;**78**:3824–8. http://www.ncbi.nlm.nih.gov/pubmed/6167991, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC319665, http://www.pnas.org/cgi/doi/10.1073/pnas.78.6.3824.
63. Cornette JL, Cease KB, Margalit H, *et al*. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol* 1987;**195**:659–85. http://www.ncbi.nlm.nih.gov/pubmed/3656427, https://linkinghub.elsevier.com/retrieve/pii/0022283687901896.
64. Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol* 1968;**21**:170–201. http://www.ncbi.nlm.nih.gov/pubmed/5700434, https://linkinghub.elsevier.com/retrieve/pii/0022519368900696.
65. McMeekin TL, Wilensky M, Groves ML. Refractive indices of proteins in relation to amino acid composition and specific volume. *Biochem Biophys Res Commun* 1962;**7**:151–6. https://www.sciencedirect.com/science/article/pii/0006291X62901651.
66. Bhaskaran R, Ponnuswamy PK. Positional flexibilities of amino acid residues in globular proteins. *Int J Pept Protein Res* 1988;**32**:241–55. http://doi.wiley.com/10.1111/j.1399-3011.1988.tb01258.x.
67. Levitt M, Levitt M. Conformational Preferences of Amino Acids in Globular Proteins. *Biochemistry* 1978;**17**:4277–85. https://pubs.acs.org/doi/abs/10.1021/bi00613a026.
68. Zhao G, London E. An amino acid "transmembrane tendency" scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: Relationship to biological hydrophobicity. *Protein Sci* 2006;**15**:1987–2001. http://www.ncbi.nlm.nih.gov/pubmed/16877712, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2242586, http://doi.wiley.com/10.1110/ps.062286306.
69. Grantham R. Amino acid difference formula to help explain protein evolution. *Science* 1974;**185**:862–4. http://www.ncbi.nlm.nih.gov/pubmed/4843792.
70. Juretić D, Vukičević D, Ilić N, *et al*. Computational design of highly selective antimicrobial peptides. *J Chem Inf Model* 2009;**49**:2873–82. https://pubs.acs.org/doi/10.1021/ci900327a.
71. Senes A, Chadi DC, Law PB, *et al*. Ez, a Depth-dependent Potential for Assessing the Energies of Insertion of Amino Acid Side-chains into Membranes: Derivation and Applications to Determining the Orientation of Transmembrane and Interfacial Helices. *J Mol Biol* 2007;**366**:436–48. http://www.ncbi.nlm.nih.gov/pubmed/17174324, https://linkinghub.elsevier.com/retrieve/pii/S0022283606012095.
72. Collantes ER, Dunn WJ. Amino Acid Side Chain Descriptors for Quantitative Structure-Activity Relationship Studies of Peptide Analogues. *J Med Chem* 1995;**38**:2705–13. http://www.ncbi.nlm.nih.gov/pubmed/7629809, https://pubs.acs.org/doi/abs/10.1021/jm00014a022.
73. Raychaudhury C, Banerjee A, Bag P, *et al*. Topological shape and size of peptides: Identification of potential allele specific helper T cell antigenic sites. *J Chem Inf Comput Sci* 1999;**39**:248–54. https://pubs.acs.org/doi/abs/10.1021/ci980052w.

74. Zaliani A, Gancia E. MS-WHIM scores for amino acids: A new 3D-description for peptide QSAR and QSPR studies. *J Chem Inf Comput Sci* 1999;**39**:525–33. https://pubs.acs.org/doi/abs/10.1021/ci980211b.
75. Koch CP, Perna AM, Pillong M, *et al.* Scrutinizing MHC-I Binding Peptides and Their Limits of Variation. *PLoS Comput Biol* 2013;**9**:e1003088. http://www.ncbi.nlm.nih.gov/pubmed/23754940, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3674988, https://dx.plos.org/10.1371/journal.pcbi.1003088.
76. Cocchi M, Johansson E. Amino Acids Characterization by GRID and Multivariate Data Analysis. *Quantitative Structure-Activity Relationships* 1993;**12**:1–8. http://doi.wiley.com/10.1002/qsar.19930120102.
77. Hellberg S, Sjöström M, Skagerberg B, *et al.* Peptide Quantitative Structure-Activity Relationships, a Multivariate Approach. *J Med Chem* 1987;**30**:1126–35. http://www.ncbi.nlm.nih.gov/pubmed/3599020, https://pubs.acs.org/doi/abs/10.1021/jm00390a003.
78. Sandberg M, Eriksson L, Jonsson J, *et al.* New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J Med Chem* 1998;**41**:2481–91. https://pubs.acs.org/doi/10.1021/jm9700575.
79. Kawashima S, Pokarowski P, Pokarowska M, *et al.* AAindex: Amino acid index database. *progress report 2008 Nucleic Acids Research* 2008;**36**:D202–5. http://www.ncbi.nlm.nih.gov/pubmed/17998252, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2238890.
80. Fauchere J-L, Pliska V. Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur J Med Chem* 1983;**18**: 369–75.
81. Wilce MC, Aguilar MI, Hearn MT. Physicochemical Basis of Amino Acid Hydrophobicity Scales: Evaluation of Four New Scales of Amino Acid Hydrophobicity Coefficients Derived from RP-HPLC of Peptides. *Anal Chem* 1995;**67**:1210–9. https://pubs.acs.org/doi/abs/10.1021/ac00103a012.
82. Naderi-Manesh H, Sadeghi M, Arab S, *et al.* Prediction of protein surface accessibility with information theory. *Proteins: Structure, Function. Genetics* 2001;**42**:452–9. http://www.ncbi.nlm.nih.gov/pubmed/11170200.
83. Parker JM, Guo D, Hodges RS. New Hydrophilicity Scale Derived from High-Performance Liquid Chromatography Peptide Retention Data: Correlation of Predicted Surface Residues with Antigenicity and X-ray-Derived Accessible Sites. *Biochemistry* 1986;**25**:5425–32. http://www.ncbi.nlm.nih.gov/pubmed/2430611, https://pubs.acs.org/doi/abs/10.1021/bi00367a013.
84. Pliška, V., Schmidt, M. & Fauchère, J.-L. Partition coefficients of amino acids and hydrophobic parameters π of their side-chains as measured by thin-layer chromatography. *J Chromatogr A* **216**, 79–92 (1981). URL https://linkinghub.elsevier.com/retrieve/pii/S0021967300823377.
85. Guy HR. Amino acid side-chain partition energies and distribution of residues in soluble proteins. *Biophys J* 1985;**47**:61–70. http://www.ncbi.nlm.nih.gov/pubmed/3978191, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1435068, https://linkinghub.elsevier.com/retrieve/pii/S0006349585838777.
86. Kuhn LA, Swanson CA, Pique ME, *et al.* Atomic and residue hydrophilicity in the context of folded protein structures. *Proteins: Structure, Function. Bioinformatics* 1995;**23**: 536–47.
87. Klein P, Kanehisa M, DeLisi C. Prediction of protein function from sequence properties. Discriminant analysis of a data base. *Biochimica et Biophysica Acta (BBA)/Protein Structure and Molecular* 1984;**787**:221–6. http://www.ncbi.nlm.nih.gov/pubmed/6547351, https://linkinghub.elsevier.com/retrieve/pii/0167483884903121.
88. Woese CR. Evolution of the genetic code. *Naturwissenschaften* 1973;**60**:447–59. http://www.ncbi.nlm.nih.gov/pubmed/4588588.
89. Krigbaum WR, Komoriya A. Local interactions as a structure determinant for protein molecules: II. *BBA - Protein Structure* 1979;**576**:204–28. https://linkinghub.elsevier.com/retrieve/pii/0005279579904987.
90. Charton M. Protein folding and the genetic code: An alternative quantitative model. *J Theor Biol* 1981;**91**:115–23. https://linkinghub.elsevier.com/retrieve/pii/0022519381903775.
91. Aurora R, Rose GD. Helix capping. *Protein Sci* 1998;**7**:21–38. http://www.ncbi.nlm.nih.gov/pubmed/9514257, http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2143812, http://doi.wiley.com/10.1002/pro.5560070103.
92. Oobatake M, Kubota Y, Ooi T. Optimization of Amino Acid Parameters for Correspondence of Sequence to Tertiary Structures of Proteins. *Tech. Rep.* 1985;**63**:82–94. https://repository.kulib.kyoto-u.ac.jp/dspace/bitstream/2433/77104/1/chd063_2_082.pdf.
93. Zhou H, Zhou Y. Quantifying the Effect of Burial of Amino Acid Residues on Protein Stability. *Proteins: Structure, Function and Genetics* 2004;**54**:315–22. http://www.ncbi.nlm.nih.gov/pubmed/14696193, http://doi.wiley.com/10.1002/prot.10584.
94. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 1901;**2**: 559–72.
95. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of Machine Learning Research* 2008;**9**:2579–625.
96. Cortes C. Support-Vector Networks. *Tech Rep* 1995;**20**: 273–297.
97. Ho, T. K. Random decision forests. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, vol. **1** of *ICDAR* '95, 278–82 (IEEE Computer Society, Washington, DC, USA, 1995). URL http://dl.acm.org/citation.cfm?id=844379.844681.
98. White BW, Rosenblatt F. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. *Spartan Books, New York* 1963;**76**:1–616.
99. Pedregosa F, *et al.* Scikit-learn: Machine Learning in Python. *Tech Rep* 2011;**85**:2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html.
100. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems (2016). URL http://arxiv.org/abs/1603.04467. 1603.04467.
101. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015* **1**, 448–56 (2015). URL http://arxiv.org/abs/1502.03167. 1502.03167.
102. Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res* 2014;**15**:1929–58.
103. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, in San Diego. Conference Track Proceedings* (2015). Cornell

University, New York. URL http://arxiv.org/abs/1412.6980. 1412.6980.

104. Schaduangrat N, Nantasenamat C, Prachayasittikul V, *et al*. ACPred: A computational tool for the prediction and analysis of anticancer peptides. *Molecules* 2019;**24**:1973.
105. Benetti S, Timmons PB, Hewage CM. NMR model structure of the antimicrobial peptide maximin 3. *Eur Biophys J* 2019;**48**:203–12. http://link.springer.com/10.1007/s00249-019-01346-7.
106. Timmons PB, O'Flynn D, Conlon JM, *et al*. Structural and positional studies of the antimicrobial peptide brevinin-1BYa in membrane-mimetic environments. *J Pept Sci* 2019;**25**:e3208.
107. Timmons PB, O'Flynn D, Conlon JM, *et al*. Insights into conformation and membrane interactions of the acyclic and dicarba-bridged brevinin-1BYa antimicrobial peptides. *Eur Biophys J* 2019;**48**:701–10. http://link.springer.com/10.1007/s00249-019-01395-y.