HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks

Patrick Brendan Timmons & Chandralal M. Hewage

The growing prevalence of resistance to antibiotics motivates the search for new antibacterial agents. Antimicrobial peptides are a diverse class of well-studied membrane-active peptides which function as part of the innate host defence system, and form a promising avenue in antibiotic drug research. Some antimicrobial peptides exhibit toxicity against eukaryotic membranes, typically characterised by hemolytic activity assays, but currently, the understanding of what differentiates hemolytic and non-hemolytic peptides is limited. This study leverages advances in machine learning research to produce a novel artificial neural network classifier for the prediction of hemolytic activity from a peptide’s primary sequence. The classifier achieves best-in-class performance, with cross-validated accuracy of 85.7% and Matthews correlation coefficient of 0.71. This innovative classifier is available as a web server at https://research.timmons.eu/happen, allowing the research community to utilise it for in silico screening of peptide drug candidates for high therapeutic efficacies.
HAPPENN-RR90, and the negative examples are experimentally validated non-hemolytic peptides exhibiting the greatest compositional similarity to the positive peptides. A negative sequence was deemed to be the most similar to a positive sequence if it possessed the minimum Euclidean distance to the positive sequence. As a number of databases exist that detail the biological activities of peptides, here we present a novel method for the prediction of the hemolytic activity of antimicrobial peptides. A deep neural network that classifies peptides as hemolytic or non-hemolytic based on their primary sequence.

Methods

Datasets. The HAPPENN dataset consists of 3,738 peptide sequences between 7–35 amino acids in length and their corresponding hemolytic activities. All sequences are composed of exclusively natural amino acids, and the only modifications included are N-terminal acetylation and C-terminal amidation. Secondary structure properties are not considered in the creation of the dataset. The dataset is available as supplementary material.

3,408 of these peptide sequences were extracted from the DBAASP database and 1,174 from the Hemolytik database. 844 peptide sequences were present in both databases. Of the sequences extracted from DBAASP, 861 were ribosomally synthesised, while 2,547 were chemically synthesised.

The dataset consists of 1,543 experimentally validated hemolytic peptides and 2,195 experimentally validated non-hemolytic peptides, as determined by criteria detailed in Table 1.

Redundancy reduced dataset. The HAPPENN dataset was internally redundancy reduced using CD-HIT, removing sequences so that no two sequences were ≥ 90% similar to each other, which yielded the HAPPENN-RR90 dataset, which consists of 823 experimentally validated hemolytic peptides, and 1,100 experimentally validated non-hemolytic peptides.

Dataset for additional benchmarking. Discriminating compositionally similar peptides with different biological activities is one of the greatest challenges in developing prediction methods. An additional dataset was created, HAPPENN-hard, wherein positive examples are the experimentally validated hemolytic peptides of HAPPENN-RR90, and the negative examples are experimentally validated non-hemolytic peptides exhibiting the greatest compositional similarity to the positive peptides. A negative sequence was deemed to be the most similar to a positive sequence if it possessed the minimum Euclidean distance to the positive sequence.
Model validation. It is critical that any classifier model created by machine learning is thoroughly validated. For that reason, tenfold cross-validations and validation by an external test set were employed to evaluate the performance of all models presented herein. The HAPPENN dataset was split into twelve parts, ten of which were used for cross-validation, whereby one of the subsets was selected for use in validation while the other nine were employed for training. The resultant models were ensembled and evaluated with an independent test set, which consists of the remaining two of the twelve parts. To avoid possible bias arising from the choice of a randomly selected test set, the procedure is repeated six times in total, allowing for a rigorous assessment of the model’s overall performance.

Validation comparison with HemoPI and HemoPred. The different dataset construction and validation procedure employed by HAPPENN compared to other available tools, namely HemoPI and HemoPred, prevents a direct comparison of their respective validation statistics. To facilitate a more direct comparison with the HemoPI and HemoPred classifiers, a model was trained and tested under equivalent conditions. An altered dataset, HAPPENN-HemoPI3-equiv was created, wherein all the peptide sequences present in the HAPPENN dataset that form part of the HemoPI-3 test dataset were set aside as the test dataset, and the remaining non-test set sequences were used for training and validation as part of a fivefold cross-validation.

Amino acid composition analysis. An analysis of the amino acid composition of the hemolytic and non-hemolytic peptides was carried out, completed by an analysis of peptides randomly extracted from proteins in Swiss-Prot. The analysis comprises the peptides’ full sequences, the 10 N-terminal residues, and the C-terminal 10 residues.

Residue position preference analysis. Enrichment depletion logos (EDLogo) were created to identify preferences for certain amino acid residues at certain positions in the hemolytic peptides’ sequences. The logo plots were constructed using the experimentally validated non-hemolytic peptide sequences as the baseline.

Motif analysis. Motif analysis was carried out on the HAPPENN dataset to identify motifs occurring exclusively in hemolytic and non-hemolytic peptides. Motifs with a length between 2–5 amino acids which occurred in at least ten peptides were considered.

Features extraction. A large selection of features was extracted from the peptides’ primary sequences, which can be divided into two subcategories, amino acid composition based features and physicochemical descriptors.

Physicochemical descriptors. The modlAMP, ChemoPy and RDKit packages were used for the calculation of global physicochemical descriptors, as well as amino acid scale-based descriptors.

Global physicochemical descriptors include sequence length, molecular formula, molecular weight, sequence charge, charge density, isolectric point, instability index, aromaticity index, Boman index and the hydrophobic ratio.

Meanwhile, amino acid scale-based descriptors include AASI, ABHPRK, hydrophobicity, side-chain bulkiness, amino acid charges, COUGAR, Ez, side-chain flexibility, polarity, ISAECI, α-helix propensity, MSS, MSW, pepArc, PPCALI, refractivity, i_scale, transmembrane propensity, z, z5 and z55.

Additionally, physicochemical descriptors were calculated from the amino acid properties in the AAindex. Secondary structure related descriptors were calculated based on the turn, helical, coil and amphiphilic propensities. The sequence hydrophobicity was quantified using the amino acids’ hydrophobies.
Composition descriptors. Amino acid, dipeptide, and tripeptide compositions were calculated for the conventional 20-amino acid alphabet, as well as the reduced alphabets of Veltri et al. 106, Thomas and Dill107, and the conjoint alphabet108. To account for the three-dimensional structure of the peptides, g-gap dipeptide and tripeptide compositions were calculated109. Finally, pseudo amino acid composition110, conjoint triad, composition, transition and distribution111 descriptors were also calculated.

Machine learning approaches. Support vector machine (SVM)112, random forest (RF)113, principal component analysis (PCA)114, t-distributed Stochastic Neighbour Embedding (t-SNE)115 and dense fully connected neural networks116 are employed in this study.

Both a linear and non-linear (RBF) kernel were employed with SVMs. SVM and RF hyperparameters were tuned using a grid search in conjunction with the previously described cross-validation.

Feature selection. Only features which were non-zero for at least 100 samples were retained. Furthermore, features were selected for retention by SVM and random forest.

Features importances were calculated individually for each of the splits during tenfold cross-validation using both support vector machines and random forests. Features which had SVM absolute weights near-zero (< 0.05) were excluded, as practised by Brank et al.117. Features which an ensemble of random forests decided were important (importance > 0.0005) were included.

Neural network architecture. All input features are scaled to have minimum and maximum values of 0 and 1, respectively.

Both a randomized grid search and genetic algorithm were employed to identify the optimal neural network architecture and hyperparameters. The optimized neural network applies a Gaussian noise layer with a standard deviation of 0.03 to the input layers, which mitigates overfitting. The first hidden layer has 1024 nodes and the second hidden layer has 64 nodes. Batch normalization118 is applied before the ReLU activation function. Each hidden layer is followed by a Dropout layer, with a rate of 0.93, which aids in the prevention of overfitting119.

The final output layer consisted of a single node with a sigmoid activation function. A summary of the overall architecture described is shown in Fig. 2.

Implementation. The neural network was implemented with Keras, a popular deep learning framework, using a Tensorflow120 back-end. The binary cross-entropy loss function was employed, which is defined as:

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

(1)

whereby \(y_i\) is the true value of the \(i^{th}\) sample, and \(\hat{y}_i\) is the predicted value of the \(i^{th}\) sample.

This loss function is commonly used in binary classification problems. As the predicted labels of all training data approach their respective true values, the value of the function approaches zero.

The optimizer employed is Adaptive Momentum (Adam), which updates the neural network weights according to the following formula121:

\[\Theta_{t+1} = \Theta_t - \frac{\eta \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}\]  

(2)

whereby 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 network was trained for 600 epochs, without stopping criteria. The model with the highest validation accuracy encountered during training was saved for each of the cross-validation splits.

During training, the loss function was weighted to adjust for the slightly unequal number of positive and negative samples.

Performance evaluation. The robustness of the predictor is evaluated by a number of standard parameters, namely accuracy (Acc), sensitivity (Sn), specificity (Sp), the Matthews correlation coefficient (MCC), and by the receiver operating characteristic (ROC) curve.

The first four of these are defined by the following equations:
Figure 2. Summary of the model development architecture. Peptide sequences and their corresponding activities were extracted from databases, peptides outside the experiment's scope were removed, and descriptors were calculated. The peptides' descriptors are used as training input to a neural network with two hidden layers, which then predicts whether or not the peptide possesses hemolytic activity.

\[ Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \] (3)

\[ Sn = \frac{TP}{TP + FN} \times 100 \] (4)

\[ Sp = \frac{TN}{TN + FP} \times 100 \] (5)

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

whereby
Results

The HAPPENN dataset was constructed from peptide sequences whose hemolytic activity, or lack thereof, has previously been evaluated. Peptides were separated into a positive (hemolytic) class and a negative (non-hemolytic) class based on criteria outlined in Table 1. The peptide sequences were subjected to an amino acid composition analysis, residue position preference analysis and motif analysis. Peptides were then represented by feature vectors composed of physicochemical and compositional descriptors of the peptides. The feature vectors are visualised using principal component analysis (PCA) and t-stochastic neighbour embedding (t-SNE) plots, both of which show an incomplete separation of the positive and negative classes. Finally, the feature vectors are used to train support vector machine, random forest and neural network hemolytic activity classifiers, the prediction results of which are evaluated.

Amino acid composition analysis. To determine whether a preference exists for certain residues in hemolytic peptides compared to non-hemolytic peptides, an amino acid residue composition analysis was performed, the results of which are shown in Fig. 3. It is apparent that hemolytic peptides are most enriched in the hydrophobic leucine and isoleucine residues, and to a lesser extent phenylalanine, tryptophan and glycine. Meanwhile, non-hemolytic peptides are enriched in the positively charged lysine and arginine residues. Interestingly, both hemolytic and non-hemolytic peptides are depleted in the negatively charged aspartic acid and glutamic acid residues compared to the sequences randomly extracted from Swiss-Prot, with the hemolytic peptides exhibiting greater depletion.

Residue position preference analysis. To ascertain whether or not there exists a preference for certain residues at certain positions in the peptide sequence, an enrichment-depletion logo plot was created (Fig. 4) for the hemolytic peptide class, with the non-hemolytic peptide class serving as the baseline for the plot. Enriched residues, therefore, are those which are more common at that position in hemolytic peptides, relative to the non-hemolytic class, and depleted residues are those which are less common.

The first inspection of the EDlogo plot reveals information that is consistent with the amino acid composition analysis: hemolytic peptides are enriched in hydrophobic residues, and predominantly depleted in negatively charged residues. On further inspection, position-specific enrichments become apparent. Hemolytic peptides are enriched in the negatively charged aspartic acid residue at position 4, and at the last, third- and fourth- and eleventh-last position, despite being depleted in this residue for the remainder of the sequence. Hemolytic peptides are depleted in the positively charged arginine residue throughout the sequence, but enriched in lysine at positions 7, 8, 11, 12 and 15. A preference exists at the positions 2 and 3 for tryptophan, position 3 for proline, and positions 3 and 4 for serine. A notable preference exists at position 14 for proline, which is indeed a common feature in the brevinin-1 family of peptides, which do possess hemolytic activity.12 Hemolytic peptides are also enriched in glutamine exclusively at their C-terminus, while being depleted in glutamine throughout the remainder of the sequence.

Motif analysis. A motif analysis was undertaken on the HAPPENN dataset to identify any motifs present exclusively in hemolytic or non-hemolytic peptides. The top twenty motifs occurring exclusively in hemolytic peptides are ‘LKHI’, ‘KIKVK’, ‘TLKIK’, ‘VNWK’, ‘GAIA’, ‘VNWKK’, ‘KLKLG’, ‘VLKAA’, ‘LWKKT’, ‘ALWKKT’, ‘MAL’, ‘KITK’, ‘PKIF’, ‘GKEV’, ‘KIAS’, ‘CKITK’, ‘KHILK’, ‘IKV’, ‘IKVA’, ‘IAS’. The top twenty motifs occurring exclusively in non-hemolytic peptides are ‘PRP’, ‘RPRP’, ‘AAAA’, ‘PRPR’, ‘PRPRP’, ‘AAAAA’, ‘AAFA’, ‘AAFA’, ‘LKYG’, ‘WKI’, ‘KYGK’, ‘ILKYG’, ‘LKYK’, ‘PKIF’, ‘RRKK’, ‘AAFAA’, ‘KPS’, ‘RPG’.

Data visualisation. Principal component analysis (PCA). Principal component analysis (PCA) was undertaken for the full computed dataset, the dataset with only the physicochemical features and the dataset with only composition descriptors (Fig. 5). Inspection of the results of all three indicates that while a separation exists between the hemolytic and non-hemolytic classes, the separation is not clear-cut and a significant overlap exists between the classes. The overlap between classes is most significant for the set consisting of only the physicochemical descriptors, while a greater separation between classes exists in the composition descriptor plot.

$T$-distributed stochastic neighbour embedding (t-SNE). Similarly to the aforementioned PCA analysis, a $t$-distributed Stochastic Neighbour Embedding (t-SNE) analysis was undertaken for the full computed dataset, the dataset with only the physicochemical features and the dataset with only composition descriptors (Fig. 6). As is the case with the PCA results, there exists an incomplete separation between the hemolytic and non-hemolytic classes in all three datasets. In many cases, positive and negatives peptides are near-coincident in the plots, and appear, therefore, to be physicochemically and compositionally similar.

Hemolytic activity prediction. This novel study employed a number of popular machine learning classifiers for predicting peptides’ hemolytic activity on the basis of features calculated from their primary sequence. The predictive power was evaluated using tenfold cross-validation, and the final ensemble of ten neural networks was evaluated by means of external validation. Accuracy, sensitivity, specificity, Matthews correlation coeffi-
Sufficient statistical parameters are reported with their confidence intervals. A receiver operating characteristic curve (ROC) with a calculated area under the curve (AUC) is also reported.

To the authors’ knowledge, three machine-learning based classifiers for the prediction of hemolytic activity peptides are described in the literature, namely HemoPI\textsuperscript{123}, HemoPred\textsuperscript{124}, and HemoPImod\textsuperscript{125}. The former two predict the hemolytic activity of natural amino-acid-based peptides, while the latter specializes in predicting the hemolytic potency of chemically modified peptides. The results of the present study are compared to those of the former two classifiers, HemoPI and HemoPred. As HemoPImod specifically addresses chemically modified peptides, and therefore differs in its aims to HAPPENN, it is excluded from the comparisons.

![Figure 3](https://www.nature.com/scientificreports/) Percentage average amino acid residue composition of the (A) full sequences, (B) N-terminal 10 residues, and (C) C-terminal 10 residues of experimentally validated hemolytic peptides (orange), experimentally validated non-hemolytic peptides (blue) and peptide sequences randomly extracted from Swiss-Prot proteins (green).
Comparison of classifier methods. The prediction statistics achieved by support vector machine (SVM), random forest (RF) and neural network (NN) models are presented in Table 2.

The SVM hyperparameters were optimised using a grid search. The linear kernel SVM achieved its highest performance with the regularization parameter $C = 0.1$. The non-linear RBF kernel SVM achieved its highest performance with the regularization parameter $C = 10$ and the kernel coefficient $\gamma = 2 \times 10^{-4}$. Both the RBF and linear kernel SVM approaches achieve the worst level of performance of the three methods studied, with a validation accuracies of 78% and 81%, and MCCs of 0.54 and 0.61, respectively.

The RF hyperparameters were also optimised using a grid search. The highest performance, with an accuracy and MCC of 83% and 0.65 was achieved with the number of estimators set to be 1024, with unrestricted tree depth. The optimal value for max_features was found to be 70.

The neural network approach, meanwhile, achieves the highest accuracy and MCC score, with scores of 86% and 0.71, respectively, marking it as the most capable predictor. Furthermore, the neural network approach
achieves the best balance between sensitivity and specificity. As it was the most capable, the neural network approach was selected as the classifier of choice for the prediction of hemolytic activity. The predictive power of HAPPENN was further evaluated by means of the receiver operating characteristic (ROC) curve, and its associated area under the curve (AUC) (Fig. 7), which is equivalent to the probability that the predictor will rank a randomly selected positive instance higher than a negative one. We note that the performance is nearly excellent on both the validation and test sets, with both yielding an AUC of 0.90.

Comparison to HemoPI and HemoPred. HemoPI and HemoPred are in silico peptide hemolytic activity prediction models previously reported in the literature, to which the HAPPENN model is compared. The former approach employs a support vector machine (SVM) trained on a combination of single residue-, dipeptide- and property-based features, while the latter employs a random forest (RF) trained on a combination of amino acid and dipeptide composition features. Both models achieve similar cross-validated accuracies and MCC scores not exceeding 78% and 0.56 when trained on the HemoPI-2 and HemoPI-3 datasets.
The HAPPENN model achieves good validation statistics, with a tenfold cross-validated accuracy of 85.66% and MCC value of 0.71. The HAPPENN approach achieves prediction performance that significantly exceeds that of both HemoPI and HemoPred (Table 3), although the cross-validation scheme and test set used differ. In order to facilitate a more direct comparison, an altered dataset was created, termed HAPPENN-HemoPI3-equiv, wherein the test dataset consists exclusively of HAPPENN dataset peptides which also form part of the HemoPI-3 test set. The remaining non-test set peptides were used for training and validation. Using this altered dataset, a neural network sharing the architecture and hyperparameters of the main HAPPENN neural network was trained under fivefold cross-validation, achieving a test set accuracy of 84.96% and an MCC of 0.70. While these results again exceed those of the available classifiers, it should be noted that the test set in this case is not truly independent as its constituent peptides had been previously used in optimising the hyperparameters of the main HAPPENN neural network.

Figure 7. Receiver operating characteristic plot of HAPPENN performance on (A) the tenfold cross-validation sets and (B) the external validation.

Table 3. Validation and test results for HemoPI, HemoPred and HAPPENN. HemoPI, HemoPred and HAPPENN-HemoPI3-Equiv datasets are subjected to fivefold cross-validation, while HAPPENN employs tenfold cross-validation for the HAPPENN, HAPPENN-RR90 and HAPPENN-hard datasets.

| Dataset            | Classifier | Acc (%) | Sn (%) | Sp (%) | MCC  |
|--------------------|------------|---------|--------|--------|------|
| Cross-validation    |            |         |        |        |      |
| HemoPI-2           | HemoPI     | 78.0    | 78.3   | 77.6   | 0.56 |
|                    | HemoPred   | 76.18 ± 0.40 | 76.57 ± 0.34 | 75.66 ± 0.53 | 0.52 ± 0.01 |
| HemoPI-3           | HemoPI     | 77.98   | 79.24  | 76.48  | 0.56 |
|                    | HemoPred   | 77.60 ± 0.70 | 77.91 ± 0.94 | 77.18 ± 0.58 | 0.55 ± 0.01 |
| HAPPENN            | HAPPENN    | 85.66 ± 1.93 | 84.96 ± 3.37 | 86.09 ± 3.43 | 0.71 ± 0.04 |
|                    | HAPPENN-RR90 | 82.73 ± 2.73 | 83.38 ± 4.82 | 82.19 ± 4.22 | 0.65 ± 0.05 |
|                    | HAPPENN-hard | 77.54 ± 3.31 | 82.12 ± 6.41 | 72.45 ± 7.74 | 0.55 ± 0.06 |
|                    | HAPPENN-HemoPI3-equiv | 85.44 ± 1.23 | 83.45 ± 3.16 | 86.79 ± 1.94 | 0.70 ± 0.02 |
| External validation|            |         |        |        |      |
| HemoPI-2           | HemoPI     | 75.7    | 78.2   | 78.3   | 0.51 |
|                    | HemoPred   | 76.82 ± 3.40 | 78.91 ± 3.82 | 74.29 ± 6.62 | 0.53 ± 0.07 |
| HemoPI-3           | HemoPI     | 77.16   | 81.92  | 71.43  | 0.54 |
|                    | HemoPred   | 79.91 ± 0.68 | 85.20 ± 2.09 | 73.33 ± 1.76 | 0.59 ± 0.01 |
| HAPPENN            | HAPPENN    | 84.00 ± 1.67 | 82.85 ± 2.31 | 84.86 ± 2.23 | 0.67 ± 0.03 |
|                    | HAPPENN-RR90 | 80.65 ± 2.41 | 81.75 ± 4.30 | 79.84 ± 1.96 | 0.61 ± 0.05 |
|                    | HAPPENN-hard | 73.94 ± 2.74 | 78.26 ± 3.56 | 69.49 ± 2.46 | 0.48 ± 0.06 |
|                    | HAPPENN-HemoPI3-Equiv | 84.96 ± 0.53 | 84.67 ± 1.19 | 85.27 ± 0.63 | 0.70 ± 0.01 |
Relationship between prediction and hemolytic activity values. Peptides were classified as hemolytic or non-hemolytic based on criteria given in Table 1, which relates hemolytic activity (H) to concentration (c). A peptide's concentration can be expressed as a multiple (x) of the threshold concentration c. For instance, a peptide which exhibits 65% hemolytic activity at 195 μM can be said to have x = 0.5, as 0.5 × 390 μM = 195 μM. As x is less than 1, the concentration is lower than the threshold concentration (390 μM for 65% hemolytic activity), and the peptide is considered hemolytic.

The neural network's output is obtained from the sigmoid activation function of its final layer. As the sigmoid function produces values ranging between 0 and 1, these output values can be interpreted as the probability of a peptide being non-hemolytic (0) and hemolytic (1).

The relationship between the neural network's output values and x, the peptides' multiple of the threshold concentration, are shown in Fig. 8. The upper left quadrant shows the true positives, the lower right quadrant shows the true negatives, and the upper right and lower left quadrant show the false negatives and false positives, respectively.

It can be seen from Fig. 8, that not many peptides are found at the y = 0.5 hemolytic-non-hemolytic prediction boundary. While there are many peptides at the x = 1.0 activity boundary, most peptides are seen to be correctly predicted.

Descriptor-set specific results. Several approaches were trialled for constructing the input feature space (Table 4), namely dipeptide and tripeptide composition, the corresponding g-gap compositions, N- and C-terminus composition and physicochemical features.

Dipeptide and tripeptide composition. Dipeptide composition is defined as the proportion of a given dipeptide in the sequence, while similarly the tripeptide composition is defined as the proportion of a given tripeptide in the sequence. These composition features capture both the chemical nature of the peptide composition while retaining information about the local sequence order. The models achieved respectable accuracies of 82.56% and 82.62%, respectively, and MCC values of 0.65 and 0.64, respectively.

g-gap composition. g-gap dipeptide composition is described as the proportion of a given pair of amino acids separated by 1, 2 or 3 residues, which corresponds to residues which are adjacent in three-dimensional space, especially in regular secondary structures where such non-adjointing residues may be connected by hydrogen bonds. Interestingly, models trained on these features perform better than those trained on the more conventional dipeptide and tripeptide compositions. This can be attributed to these features better capturing the chemical environment that the peptide exposes to the membrane upon contact. For instance, the g-gap feature can represent the spatial adjacency separated by one turn of the α-helix, which has a turn of 3.6 residues.

Termini. A feature set was created to capture the information on the residues specific location in the sequence. This feature set consists of a binary profile for the first and last 15 residues of each peptide. Each peptide is there-
Table 4. Validation statistics achieved by neural networks trained with just single sets of descriptors.

| Dataset        | Features                          | Acc (%) ± | Sn (%) ± | Sp (%) ± | MCC ± |
|----------------|-----------------------------------|-----------|----------|----------|-------|
| **Cross-validation** |                                   |           |          |          |       |
| HAPPENN        | Composition, dipeptide            | 82.56 ± 2.42 | 82.59 ± 3.52 | 82.50 ± 3.12 | 0.65 ± 0.05 |
|                | Composition, tripeptide           | 82.62 ± 1.74 | 80.46 ± 4.45 | 84.06 ± 2.27 | 0.64 ± 0.04 |
|                | g-gap composition, dipeptide      | 84.02 ± 2.38 | 83.87 ± 3.14 | 84.13 ± 3.20 | 0.68 ± 0.05 |
|                | g-gap composition, tripeptide     | 83.55 ± 1.85 | 82.07 ± 3.06 | 84.54 ± 2.86 | 0.66 ± 0.04 |
|                | Termini                           | 82.11 ± 2.00 | 80.92 ± 3.75 | 82.95 ± 2.58 | 0.63 ± 0.04 |
|                | Physicochemical                   | 80.63 ± 2.60 | 82.02 ± 5.13 | 79.64 ± 4.27 | 0.61 ± 0.05 |
| HAPPENN-RR90   | Composition, dipeptide            | 78.05 ± 2.57 | 81.25 ± 4.90 | 75.67 ± 3.86 | 0.56 ± 0.05 |
|                | Composition, tripeptide           | 77.47 ± 2.62 | 72.79 ± 8.53 | 80.92 ± 5.63 | 0.54 ± 0.06 |
|                | g-gap composition, dipeptide      | 79.52 ± 2.79 | 81.50 ± 6.21 | 78.04 ± 4.86 | 0.59 ± 0.06 |
|                | g-gap composition, tripeptide     | 78.61 ± 2.75 | 78.49 ± 6.16 | 78.68 ± 4.90 | 0.57 ± 0.06 |
|                | Termini                           | 79.38 ± 3.01 | 80.39 ± 6.52 | 78.63 ± 3.02 | 0.59 ± 0.06 |
|                | Physicochemical                   | 79.02 ± 4.42 | 81.70 ± 6.77 | 77.01 ± 5.75 | 0.58 ± 0.09 |
| **External validation** |                                   |           |          |          |       |
| HAPPENN        | Composition, dipeptide            | 81.42 ± 1.61 | 81.25 ± 1.51 | 81.58 ± 2.49 | 0.62 ± 0.03 |
|                | Composition, tripeptide           | 81.42 ± 1.01 | 79.91 ± 1.73 | 82.49 ± 1.11 | 0.62 ± 0.02 |
|                | g-gap composition, dipeptide      | 82.98 ± 1.76 | 82.76 ± 2.47 | 83.17 ± 1.64 | 0.65 ± 0.04 |
|                | g-gap composition, tripeptide     | 81.80 ± 1.17 | 80.53 ± 1.59 | 82.72 ± 1.19 | 0.63 ± 0.02 |
|                | Termini                           | 81.34 ± 1.15 | 80.02 ± 3.18 | 82.33 ± 1.26 | 0.62 ± 0.02 |
|                | Physicochemical                   | 79.30 ± 1.82 | 80.93 ± 3.89 | 78.21 ± 2.27 | 0.58 ± 0.04 |
| HAPPENN-RR90   | Composition, dipeptide            | 76.92 ± 1.43 | 80.03 ± 2.43 | 74.62 ± 1.69 | 0.54 ± 0.03 |
|                | Composition, tripeptide           | 76.44 ± 0.96 | 74.63 ± 4.15 | 77.79 ± 2.57 | 0.52 ± 0.02 |
|                | g-gap composition, dipeptide      | 78.14 ± 1.82 | 79.00 ± 2.75 | 77.47 ± 3.37 | 0.56 ± 0.03 |
|                | g-gap composition, tripeptide     | 77.16 ± 1.80 | 77.06 ± 3.52 | 77.28 ± 1.17 | 0.54 ± 0.04 |
|                | Termini                           | 78.07 ± 2.54 | 78.30 ± 3.88 | 77.93 ± 2.11 | 0.56 ± 0.05 |
|                | Physicochemical                   | 77.11 ± 3.58 | 79.33 ± 5.38 | 75.47 ± 3.46 | 0.54 ± 0.07 |

fore represented by a number of vectors, the first of which represents the conventional 20 amino acid alphabet is of length 30 \(\times\) 20, and the remaining vectors represent the conjoint alphabet and the reduced alphabets of Veltri, Thomas and Dill. The model trained on this feature set achieved an accuracy of 82.11% and an MCC of 0.64.

**Physicochemical.** Interestingly, the worst-performing model is the model trained on the physicochemical features, achieving an accuracy of only 80.63% and an MCC of 0.61 on the non-redundancy reduced dataset.

Of the single feature-set approaches trialled, none outperform the model trained on the full feature set. The compositional descriptor-based models are seen to benefit from sequence similarity to an extent, and exhibit somewhat reduced performance on the redundancy-reduced dataset. The physicochemical descriptor-based model, conversely, maintains a comparable performance even on the redundancy-reduced dataset.

**Feature importance analysis.** **Random forest feature importance.** Random forests have the advantage of being easily interpretable and provide an easy method of ranking the importance of input features. The most useful features as determined by cross-validated random forests are the Eisenberg direction of the hydrophobic moment (EISD860103)\(^7\), the Eisenberg normalized hydrophobicity scale (EISD840101)\(^7\), hydrophathy (NADH010102, NADH010103, NADH010104)\(^8\), the hydrophobic parameter pi (FAUJ830101)\(^9\), the Boman index\(^10\), apparent partition energies (GUYH850105)\(^8\), membrane-propensity (PUNT030102)\(^9\), transmembrane propensity\(^9\), side-chain hydrophobicity values (BLAS910101)\(^8\) and Hopp–Woods hydrophobicity (HOPT810101)\(^9\).

Effectively all of these features directly or indirectly quantify hydrophobicity, which points to it being important for hemolytic activity.

**Analysis of neural network weights using Garson’s method.** In order to understand the basis for the neural network’s predictive power, we analysed the importance assigned to various input features by Garson’s method\(^12\), iteratively reducing the feature input space by approximately halving the number of input features, until 300 composition features were retained in each split. 24 features were identified as having large weights in each of the 10 splits.

The occurrences of the FS, LH, KIK, VAK, VLK dipeptides and tripeptides in the peptide sequence were found to be important. Additionally, the LLL Veltri reduced alphabet tripeptide and the RGV and VCR Thomas and Dill (length 3) reduced alphabet tripeptides were found to contribute strongly to the final classification.
Additionally, the occurrence of \( g\)-gap \( i, i+3 \) residue pairs FG, FL, LK, WV, the occurrence of \( g\)-gap \( i, i+4 \) residue pair FK and the occurrence of \( g\)-gap \( i, i+2 \) residue pairs AR, FL, FS, LF were found to be meaningful. Furthermore, the occurrence of the \( g\)-gap \( i, i+3 \) Thomas and Dill (10) reduced alphabet residue pairs PS and VW and \( g\)-gap \( i, i+3 \) Veltri reduced alphabet residue pair QQ were also important.

Finally, the network weights associated with the EstateVSA3 and EstateVSA4 (MOE-type descriptors using Estate indices and surface area contributions), Geary autocorrelation-lag8 weight by atomic polarizabilities, and the acetylation of the N-terminus inputs were also large.

Interestingly, the predictive power of the reduced feature space neural network is nearly as strong as the main network.

**Discussion**

A decreasing number of drug approvals and a rising research and development cost base has contributed to a resurgence of interest in peptide therapeutics. An ideal peptide drug should possess a high therapeutic index, specifically high activity against the biological target and limited toxicity. The therapeutic potential of peptides, however, is highly dependent on it possessing little to no hemolytic activity. Minimizing hemolytic activity is important for improving the therapeutic index of a peptide.

Many research groups have studied the structures of natural peptides as well as engineered peptide analogues in order to characterise how their structure determines their biological activities. A comprehensive understanding of the relationship between structure and function, however, remains elusive. A computational method that can provide information about a peptide’s biological activity from its primary structure prior to chemical synthesis, however, would allow for rapid and efficient exploration of the chemical space and present a significant cost and time saving.

To accelerate the lead molecule design and optimization pipeline, this study aimed to create an in silico method for classifying therapeutic peptides as hemolytic or non-hemolytic based on their primary sequence. The prediction task is challenging, however, as it requires distinguishing between desirable activity at the peptide’s target, the prokaryotic plasma membrane in the case of most antimicrobial peptides, and activity at the membrane of eukaryotic erythrocytes. The task is further complicated by the varying extent to which many peptides display hemolytic activity, which makes arbitrarily classifying them as hemolytic or non-hemolytic challenging. Indeed, there is limited consensus on the most appropriate metric to quantify hemolysis, with many articles reporting only one metric recorded at a single concentration, which consequently precludes a regression approach instead of a classification approach. The topic is complicated further by a lack of consensus on the definition of a key metric, MHC. Most studies define it as the minimum hemolytic concentration, but differ on the specific criteria, with different studies defining it as the concentration at which 5%, 10%, 50% or even 100% hemolysis occurs. Some even define it as the maximum concentration that does not cause any hemolysis. Ideally, studies would present the analysis of hemolytic activity as a series of measurements undertaken at several concentrations, which would allow for a fuller understanding of the toxicity-concentration profile. Until such a time, however, using the MHC values for training classifiers requires investigating its actual meaning on a case by case basis.

In the course of verifying the activity of the sequences in our dataset, and comparing our dataset to the HemoPI dataset, we identified a number of instances of misclassified sequences, sequences whose hemolytic activities were not clear, and sequences whose presence in the literature we were unable to independently verify. These sequences were not included in the HAPENN dataset.

The success and validity of a machine learning classifier are predicated on the correct definition of the problem at hand, which in this case encompasses the definition of positive and negative datasets. Both the positive and negative datasets consist of experimentally validated peptide sequences which exhibit antimicrobial or other biological activities. Unlike HemoPI and HemoPred, we chose not to conduct a machine learning experiment where the negative dataset consists of peptides randomly extracted from proteins in Swiss-Prot, as a machine learning classifier is most dependable when only one property of interest is varied. Using randomly extracted sequences as the negative dataset, and hemolytic antimicrobial peptides as the positive set, likely results in the classifier learning to predict general membrane activity, rather than specifically activity against eukaryotic erythrocytes. The authors believe that the HAPENN dataset represents a major improvement on the HemoPI datasets, both in terms of size and reliability, as it contains 3738 peptides with confirmed biological activities, compared to the 904 and 1623 sequences present in the HemoPI-2 and HemoPI-3 datasets, and therefore has been made available for download both as supplementary information and on the server’s website.

Once a reliable dataset was constructed, the peptide sequences were translated into vectors of physicochemical and composition features, and a number of different machine learning approaches, namely support vector machines, random forests and neural networks, were trialled for relating the peptides’ features to their hemolytic activities. The neural network approach proved most promising, and was therefore retained, further optimised, and had its predictive power thoroughly evaluated by means of tenfold cross-validation and external validation on an independent test set.

The final neural network model achieved a tenfold cross-validated accuracy, sensitivity and specificity of 85.66%, an MCC of 0.71, and an AUC of 0.90. The validation statistics demonstrated that the model is capable of discriminating between hemolytic and non-hemolytic peptides, and that it exhibits minimal bias towards one class or another. The model performs very well compared to the existing methods, with a 35.3% decrease in cross-validated error relative to HemoPI and HemoPred. The model’s residual prediction error rate can likely be attributed to a limited sample size for neural networks to accurately learn from, as well as the fine boundary between the definition of hemolytic and non-hemolytic peptides combined with the margin of error associated with the experimental determination of hemolytic activity. Further improvements to the predictive power of
work model they wish to use for prediction, and the server will return the probability of the peptide being hemolytic propensity, with FS, LH, KIK, VAK and VLK being ranked as important.

Interestingly, the classifier predicts maximin 3 and all of its alanine scan mutants to be non-hemolytic, with the single exception of [E20A]maximin 3, which is consistent with the literature, which acknowledges the relationship between a peptide’s net charge and its hemolytic activity.134

To ascertain the source of misclassification of the wrongly predicted peptides, the main model's false positives and false negatives were highlighted on the PCA and t-SNE plots. It is apparent that many of the misclassifications occur due to the peptides' compositional and/or physicochemical similarity to peptides with differing hemolytic activity. To gain further insight into the source of misclassification, the peptide with the most similar percentage amino acid composition but opposite hemolytic activity was identified for each wrongly predicted peptide. For 16% of misclassified peptides, a compositionally identical peptide with opposite hemolytic activity was identified, compared with just 5% for correctly classified peptides. Overall, misclassified peptides had a smaller Euclidean distance to their most compositionally similar opposite-activity peptide than correctly classified peptides did.

To gain insight into which features were most important for hemolytic activity, the importance assigned to features by random forests was investigated. Hydrophobicity, as quantified by a selection of different metrics, appears to be critical for hemolytic activity, with more hydrophobic sequences generally being found to be more hemolytic than less hydrophobic sequences. These findings are not surprising, and are consistent with the available literature.131,132. A number of compositional descriptors were also found to be indicative of hemolytic propensity, with FS, LH, KIK, VAK and VLK being ranked as important.

HAPPENN's power is demonstrated by an alanine scan applied to maximin 3, a non-hemolytic peptide.127,133 Interestingly, the classifier predicts maximin 3 and all of its alanine scan mutants to be non-hemolytic, with the single exception of [E20A]maximin 3, which is consistent with the literature, which acknowledges the relationship between a peptide's net charge and its hemolytic activity.134

This study presents a significant improvement in the area of in silico hemolytic activity classification, with its results forming the new state-of-the-art. The novel application of a neural network combined with the HAPPENN dataset's superior data quality and quantity has facilitated a 35% decrease in classification error, compared to the results achieved by the best currently available tools.

To conclude, accurate prediction of hemolytic activity of antimicrobial peptides can facilitate in silico design of novel peptide-based therapeutics, thereby accelerating the design phase and reducing its cost. HAPPENN distinguishes itself from existing methods through its focus on antimicrobial peptides, more accurate prediction and incorporation of novel features.

Although HAPPENN displays advantages compared to competing methods, it is limited by the lower interpretability of the neural network's hidden layers. Prediction of hemolytic activity from primary sequence remains a challenging problem, as it is characterised by a complex interplay between numerous features, which also contribute to the desirable antimicrobial activities. Nonetheless, HAPPENN possesses an error rate 35% lower than the most accurate existing classifiers, and we believe that this work will aid future studies focused on the identification and design of novel peptide therapeutics.

Web server implementation

To best serve the scientific community, we have made the classifier algorithm available online at https://research.timmons.eu/happenn in the form of an easy to use web-server, which is available for free use by academic researchers. The web server is capable of predicting the hemolytic activity of peptides' based on their primary sequence, as well as the presence or absence of N-terminal acetylation or C-terminal amidation modifications. Prediction is limited to peptides composed of the 20 natural amino acids; non-natural amino acids are not supported. The web server possesses many features. Neural network models trained on the HAPPENN, HAPPENN-RR90 and HAPPENN-hard datasets are available for prediction.

Hemolytic activity prediction. Hemolytic activity prediction is available for both single and multiple sequences. The user should submit the peptide sequence or sequences in FASTA format, select the neural network model they wish to use for prediction, and the server will return the probability of the peptide being hemo-
lytic, based on the neural network's prediction. This probability is on a scale of 0–1, where 0 is most probably non-hemolytic and 1 is most probably hemolytic.

**Mutation analysis.** Mutation analysis is available for single sequences, provided in FASTA format. After inputting the sequence, the user should select the mutation analysis option, input the residue number that they wish to mutate, and run the prediction. The server will predict the hemolytic activity of each of the peptide’s mutants attained by substituting the residue at the selected position with each of the other natural 20 amino acids.

**Residue scan.** A residue scan, for instance an alanine-scan, is available for single sequences provided in FASTA format. After inputting the sequence, the user should select the residue scan option, choose the residue they wish to scan with and run the prediction. The server will predict the hemolytic activity of each of the peptide’s mutants attained by substituting successive residue positions with the selected residue.

**Data availability**

All data generated or analysed during this study are included in this published article's supplementary data sets.

Received: 3 March 2020; Accepted: 9 June 2020

Published online: 02 July 2020

**References**

1. Hultmark, D. Drosophila immunity: paths and patterns. *Curr. Opin. Immunol.* 15, 12–19 (2003).
2. Yeaman, M. R. & Yount, N. Y. Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev.* 55, 27–55 (2003).
3. Guilhelmelli, F. *et al.* Antibiotic development challenges: the various mechanisms of action of antimicrobial peptides and of bacterial resistance. *Front. Microbiol.* 4, 353 (2013).
4. Vlieghe, P., Lisowski, V., Martinez, J. & Khrestchatisky, M. Synthetic therapeutic peptides: science and market. *Drug Discov. Today* 15, 40–56 (2010).
5. Gordon, Y. J., Romanowski, E. G. & McDermott, A. M. Mini review: A review of antimicrobial peptides and their therapeutic potential as anti-infective drugs. *Curr. Eye Res.* 30, 505–515 (2005).
6. Conlon, J. M., Mechkar, S., Lukic, M. L. & Platt, P. R. Potential therapeutic applications of multifunctional host-defense peptides from frog skin as anti-cancer, anti-viral, immunomodulatory, and anti-diabetic agents. *Peptides* 57, 67–77 (2014).
7. Karapetyan, A. V. *et al.* Bioactive lipids and cationic antimicrobial peptides as new potential regulators for trafficking of bone marrow-derived stem cells in patients with acute myocardial infarction. *Stem Cells Dev.* 22, 1645–1656 (2013).
8. Chow, J. Y. C., Li, Z. J., Kei, W. K. & Cho, C. H. Cathelicidin a potential therapeutic peptide for gastrointestinal inflammation and cancer. *World J. Gastroenterol.* 19, 2731–2735 (2013).
9. Reisner, E. H., Bailey, F. N. & Appelbaum, E. The treatment of pneumonia with bacitracin.
10. Lau, J. L. & Dunn, M. K. Therapeutic peptides: historical perspectives, current development trends, and future directions. *Bioorg. Med. Chem.* 26, 2700–2707 (2018).
11. Sohrabi, C., Foster, A. & Tavazoli, A. Methods for generating and screening libraries of genetically encoded cyclic peptides in drug discovery. *Nat. Rev. Chem.* 4, 90–101 (2020).
12. Bozovičar, K. & Bratkovič, T. Evolving a peptide: library platforms and diversification strategies. *Int. J. Mol. Sci.* 21, 215 (2020).
13. Furka, Ā., Sebestyén, F., Asgedom, M. & Dibó, G. General method for rapid synthesis of multicomponent peptide mixtures. *Int. J. Pept. Protein Res.* 37, 487–493 (1991).
14. Lalehari, J. P. *et al.* A phase II clinical study of the long-term safety and antiviral efficacy of enfuvirtide-based antiretroviral therapy. *AIDS* 17, 691–698 (2003).
15. Heyns, C., Simonin, M.-P., Grosgurin, P., Schall, R. & Porchet, H. Comparative efficacy of triptorelin pamoate and leuprolide acetate in men with advanced prostate cancer. *BJU Int.* 92, 226–231 (2003).
16. Reisner, E. H., Bailey, F. N. & Appelbaum, E. The treatment of pneumonia with bacitracin. *Ann. Intern. Med.* 34, 1232–1242 (1951).
17. Ascione, A. Boceprevir in chronic hepatitis C infection: a perspective review. *Ther. Adv. Chron. Dis.* 3(3), 113–121 (2012).
18. Bruno, B. J., Miller, G. D. & Lim, C. S. Basics and recent advances in peptide and protein drug delivery. *Ther. Deliv.* 4(11), 1443–1467 (2013).
19. Hamamoto, K., Kida, Y., Zhang, Y., Shimizu, T. & Kii, K. Antimicrobial activity and stability to proteolysis of small linear cationic peptides with D-amino acid substitutions. *Microbiol. Immunol.* 46, 741–749 (2002).
20. Wimley, W. C. Describing the mechanism of antimicrobial peptide action with the interfacial activity model. *ACS Chem. Biol.* 5, 905–917 (2010).
21. Hu, Y., Sinha, S. K. & Patel, S. Investigating hydrophilic pores in model lipid bilayers using molecular simulations: correlating bilayer properties with pore-formation thermodynamics. *Langmuir* 31, 6615–6631 (2015).
22. Lai, R., Liu, H., Hui Lee, W. & Zhang, Y. An anionic antimicrobial peptide from toad Bombina maxima. *Biochem. Biophys. Res. Commun.* 295, 796–799 (2002).
23. Matsuzaki, K., Sugishita, K., Itoh, N. & Miyajima, K. Molecular basis for membrane selectivity of an antimicrobial peptide, Magainin 2. *Biochemistry* 34, 3423–3429 (1995).
24. Gomes, B. *et al.* Designing improved active peptides for therapeutic approaches against infectious diseases. *Biotechnol. Adv.* 36, 415–429 (2018).
25. Zeng, M. *et al.* Protein-protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* 36, 1114–1120 (2020).
26. Oti, M., Ballouz, S. & Wouters, M. a. In silico tools for gene discovery. *Methods Mol. Biol.* 760, 175–187 (2011).
27. Holton, T. A., Pollastr, G., Shields, D. C. & Mooney, C. CPPpred: prediction of cell penetrating peptides. *Bioinformatics* 29, 3094–3096 (2013).
28. Pirtskhalava, M. *et al.* Erratum: DDBAASP vol 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.* 44, 6503 (2016).
29. Waghni, B. H. *et al.* CAMP: collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Res.* 42, D1154–8 (2014).
30. Gautam, A. et al. Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Res.* **42**, D444–D449 (2014).
31. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
32. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
33. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
34. Loose, C., Jensen, K., Rigoutsos, I. & Stephanopoulos, G. A linguistic model for the rational design of antimicrobial peptides. *Nature* **443**, 867–869 (2006).
35. Porto, W. F., Pires, A. S. & Franco, O. L. Computational tools for exploring sequence databases as a resource for antimicrobial peptides. *Biotechnol. Adv.* **35**, 337–349 (2017).
36. Kumar, M., Thakur, V. & Raghava, G. P. COPid: composition based protein identification. In *Silico Biol.* **8**, 121–128 (2008).
37. Agrawal, P. et al. In silico approach for prediction of antifungal peptides. *Front. Microbiol.* **9**, 323 (2018).
38. Agrawal, P., Kumar, S., Singh, A., Raghava, G. P. & Singh, I. K. NeuroPlpred: a tool to predict, design and scan insect neuropeptides. *Sci. Rep.* **9**, 5129 (2019).
39. Apweiler, R. et al. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.* **41**, D43–D7 (2013).
40. Dey, K. K., Xie, D. & Stephens, M. A new sequence logo plot to highlight enrichment and depletion. *BMC Bioinf.* **19**, 473 (2018).
41. Müller, A. T., Gabernet, G., Hiss, J. A. & Schneider, G. modlAMP: Python for antimicrobial peptides. *Bioinformatics* (Oxford, England) **33**, 2735–2755 (2017).
42. Cao, D. S., Xu, Q. S., Hu, Q. N. & Liang, Y. V. ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics* **29**, 1092–1094 (2013).
43. Lobry, J. R. & Gautier, C. Hydrophobicity, expressivity and aromaticity to be the major trends of amino acid usage in *999 Escherichia coli* chromosome-encoded genes. *Nucleic Acids Res.* **22**, 3174–3180 (1994).
44. Ikai, A. Thermostability and aliphatic index of globular proteins. *J. Biochem.* **88**, 1895–8 (1980).
45. Boman, H. G., Wade, D., Boman, I. A., Wåhlin, B. & Merrifield, R. B. Antibacterial and antimalarial properties of peptides that are cecropin–melittin hybrids. *FEBS Lett.* **259**, 103–106 (1989).
46. Juretić, D., Vukičević, D., Ilić, N., Antcheva, N. & Tossi, A. Computational design of highly selective antimicrobial peptides. *J. Chem. Inf. Model.* **49**, 2873–2882 (2009).
47. Argos, P., Rao, J. K. & Hargrave, P. A. Structural prediction of membrane-bound proteins. *Eur. J. Biochem.* **128**, 565–575 (1982).
48. Eisenberg, D., Weiss, R. M., Terwilliger, T. C. & Wilcox, W. Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc.* **17**, 109–120 (1982).
49. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
50. Hopp, T. P. & Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* **78**, 3824–3828 (1981).
51. Cornette, J. L. et al. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **195**, 659–685 (1987).
52. Zimmermann, J. M., Eliezer, N. & Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **21**, 170–201 (1968).
53. Senes, A. 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.* **366**, 436–448 (2007).
54. Bhaskaran, R. & Ponnuswamy, P. K. Positional flexibilities of amino acid residues in globular proteins. *Int. J. Pept. Protein Res.* **32**, 241–255 (1988).
55. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864 (1974).
56. Collantes, E. R. & Dunn, W. J. Amino acid side chain descriptors for quantitative structural–activity relationship studies of peptide analogues. *J. Med. Chem.* **38**, 2705–2713 (1995).
57. Levitt, M. & Levitt, M. Conformational preferences of amino acids in globular proteins. *Biochemistry* **17**, 4277–4285 (1988).
58. Raychaudhury, C., Banerjee, A., Bag, P. & Roy, S. Topological shape and size of peptides: identification of potential allele specific helper T cell antigenic sites. *J. Chem. Inf. Comput. Sci.* **39**, 248–254 (1999).
59. 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.* **39**, 525–533 (1999).
60. Koch, C. P. et al. Scrutinizing MHC-I binding peptides and their limits of variation. *PLoS Comput. Biol.* **9**, e1003088 (2013).
61. McMeekin, T. L., Wilensky, M. & Groves, M. L. Refractive indices of proteins in relation to amino acid composition and specific volume. *Biochem. Biophys. Res. Commun.* **7**, 151–156 (1962).
62. Cocchi, M. & Johansson, E. Amino acids characterization by GRID and multivariate data analysis. *Quant. Struct. Act. Relat.* **12**, 1–8 (1993).
63. 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.* **15**, 1987–2001 (2006).
64. Hellberg, S., Sjöström, M., Skagerberg, B. & Wold, S. Peptide quantitative structure–activity relationships, a multivariate approach. *J. Med. Chem.* **30**, 1126–1133 (1987).
65. Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M. & Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **41**, 2481–2491 (1998).
66. Kawashima, S. et al. Amino acid index database, progress report 2008. *Nucleic Acids Res.* **36**, D202–5 (2008).
67. Monné, M., Hermansson, M. & Von Heijne, G. A turn propensity scale for transmembrane helices. *J. Mol. Biol.* **288**, 141–145 (1999).
68. Aurora, R. & Rose, G. D. Helix capping. *Protein Sci.* **7**, 21–38 (1998).
69. Qian, N. & Sejnowski, T. J. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865–884 (1988).
70. Mitaku, S., Hirokawa, T. & Tsuji, T. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics* **18**, 608–616 (2002).
71. Naderi-Manesh, H., Sadeghi, M., Arash, S. & Moosavi Movahedi, A. A. Prediction of protein surface accessibility with information theory. *Proteins: Struct. Funct. Genet.* **42**, 452–459 (2001).
72. 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.* **18**, 369–375 (1983).
73. Eisenberg, D. Three-dimensional structure of membrane and surface proteins. *Annu. Rev. Biochem.* **53**, 595–623 (1984).
74. Ponnuswamy, P. K., Prabhakaran, M. & Manavalan, P. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *BBA Protein Struct.* **623**, 301–316 (1980).
75. Wilce, M. C., Aguilar, M. I. & Hearn, M. T. 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.* **67**, 1210–1219 (1995).
76. Black, S. D. & Mould, D. R. Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. Anal. Biochem. 193, 72–82 (1991).
77. Eisenberg, D. & McLachlan, A. D. Solvation energy in protein folding and binding. Nature 319, 199–203 (1986).
78. Pliska, V., Schmidt, M. & Fauchère, J. L. Partition coefficients of amino acids and hydrophobic parameters of their side-chains as measured by thin-layer chromatography. J. Chromatogr. A 216, 79–92 (1981).
79. Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18, 534–552 (1985).
80. Guy, H. R. Amino acid side-chain partition energies and distribution of residues in soluble proteins. Biophys. J. 47, 61–70 (1985).
81. Meek, J. L. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. Proc. Natl. Acad. Sci. USA 77, 1632–1636 (1980).
82. Parker, J. M., Guo, D. & Hodges, R. S. 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 25, 5425–5432 (1986).
83. Klein, P., Kanehisa, M. & DeLisi, C. Prediction of protein function from sequence properties. Discriminant analysis of a data base. Biochim. Biophys. Acta (BBA) Protein Struct. Mol. 787, 221–226 (1984).
84. Woese, C. R. Evolution of the genetic code. Die Nat. 60, 447–459 (1973).
85. Radlicka, A. & Wollendt, R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry 27, 1664–1670 (1988).
86. Charton, M. & Charton, B. I. The structural dependence of amino acid hydrophobicity parameters. J. Theor. Biol. 99, 629–644 (1982).
87. Eisenberg, D. & Mclachlan, A. D. Solvation energy in protein folding and binding.
Acknowledgements
The authors would like to thank Prof. Denis Shields, Dr. Gianluca Pollastri and Dr. Manaz Kaleel for enlightening conversations. The authors would also like to thank University College Dublin for the Research Scholarship granted to P.B.T. Finally, this work is dedicated to the memory of Janina Jarońska, who inspired a love of learning and a thirst for knowledge.

Author contributions
P.B.T. assembled the dataset, built the feature description and machine learning programmes, trained the machine learning algorithms, created all figures and tables and created the web interface. P.B.T. and C.M.H. determined the scope of the experiments and wrote the manuscript. Both authors contributed to and have approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-67701-3.

Correspondence and requests for materials should be addressed to C.M.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020