communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-07615-w

# Prediction of hemolytic peptides and their hemolytic concentration

Check for updates

Anand Singh Rathore, Nishant Kumar, Shubham Choudhury, Naman Kumar Mehta & Gajendra P. S. Raghava

Peptide-based drugs often fail in clinical trials due to their toxicity or hemolytic activity against red blood cells (RBCs). Existing methods predict hemolytic peptides but not the concentration ($HC_{50}$) required to lyse 50% of RBCs. This study develops classification and regression models to identify and quantify hemolytic activity. These models train on 1926 peptides with experimentally determined $HC_{50}$ against mammalian RBCs. Analysis indicates that hydrophobic and positively charged residues were associated with higher hemolytic activity. Among classification models, including machine learning (ML), quantum ML, and protein language models, a hybrid model combining random forest (RF) and a motif-based approach achieves the highest area under the receiver operating characteristic curve (AUROC) of 0.921. Regression models achieve a Pearson correlation coefficient (R) of 0.739 and a coefficient of determination ($R^2$) of 0.543. These models outperform existing methods and are implemented in HemoPI2, a web-based platform and standalone software for designing peptides with desired $HC_{50}$ values (http://webs.iiitd.edu.in/raghava/hemopi2/).

The process of developing and testing drugs is intricate, expensive, time-intensive, and laden with risks. Drug development encompasses several stages, which can be categorized into various phases, including disease-related genomics, identification and validation of targets, lead discovery and optimization, preclinical testing, and clinical trials[1]. In the past few decades, peptide-based drugs have boomed in drug discovery and development because of their advantages over traditional drugs, including greater efficacy, specificity, high tissue penetration ability, low immunogenicity, ease of modification, reduced risk of drug-drug interactions as their degradation product is amino acids, and low cost[2–5]. This trend is evidenced by the approval of 31 peptides as drugs by the Food and Drug Administration (FDA) between 2016 and 2023, alongside 370 new drugs approved during the same period, accounting for more than 8% of the total[6–8]. Additionally, there are over 200 peptides currently in clinical development and approximately 600 peptides undergoing preclinical studies, further highlighting the growing importance and utilization of peptides in pharmaceutical research and development[9,10]. The primary causes of failure of peptide-based drugs during preclinical trials are due to unacceptable safety and efficacy, which are mainly caused by absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiles[11]. Thus, there is a need to effectively screen and enhance the ADMET properties of drugs at an early stage. Many pharmaceutical companies have adopted a “fail early, fail cheap” strategy to address these challenges[12]. Adopting an in-silico strategy for predicting ADMET properties has gained considerable traction due to its cost-saving advantages and ability to provide high-throughput alternatives to traditional experimental measurement methods[13].

Toxicity, a major obstacle in designing peptide-based therapeutics, encompasses three primary categories: cytotoxicity, hemotoxicity responsible for lysing RBCs, and immunotoxicity allergenicity[14–21]. Hemolytic concentration ($HC_{50}$) serves as a common indicator of peptide toxicity, representing the concentration at which 50% lysis of normal human erythrocytes occurs under physiological conditions[22]. Peptides rich in positively charged amino acids can bind to the erythrocyte’s negatively charged lipid bilayer, leading to membrane disintegration and allowing water and solute molecules to enter the cell. This influx increases the osmotic gradient inside the erythrocyte, resulting in cell swelling and ultimately bursting[23]. Several computational methods have been developed for predicting hemolytic peptides in recent years[24–26]. Most of these methods were trained and tested on datasets derived from a dedicated database of hemolytic peptides referred as Hemolytik[17]. In 2016, Chaudhary et al. attempted to develop a machine learning (ML)-based model and motif-based tool for predicting hemolytic peptides called HemoPI[24]. Sung et al. utilized HemoPI, in conjunction with other computational methods, to effectively filter out hemolytic anti-cancer drugs[27]. Similarly, Catania et al.[28] employed HemoPI to assess the hemotoxicity of anticancer peptides derived from Mediterranean Seagrass. These studies highlight the critical importance of HemoPI, demonstrating its utility from computational predictions to practical laboratory applications. A method called HemoPred[25], based on the RF model, was developed utilizing the identical dataset employed in HemoPI. It incorporates a linear

Department of Computational Biology, Indraprastha Institute of Information Technology, Okhla Phase 3, New Delhi, 110020, India. e-mail: raghava@iiitd.ac.in

combination of amino acid composition (AAC) and dipeptide composition (DPC) as features. HLPpred-Fuse[29] proposed a two-layer prediction framework. They share the same positive dataset developed by Chaudhary et al., while the negative dataset is sourced from the PEPred-SUITE method[30]. By using the HemoPI dataset, Plisson et al.[31] developed another ML-based method that predicts the hemolytic nature of peptide sequences using gradient-boosting classifiers. Their multivariate outlier detection models led to the discovery of high-confidence non-hemolytic natural antimicrobial peptides (AMPs), facilitated the de novo design of non-hemolytic peptides, and provided guidelines for designing non-hemolytic peptides. HemoPImod[18] is a prediction model for chemically modified hemolytic peptides. It utilizes a Random Forest (RF) model that integrates various peptide features, including atom and diatom compositions, 2D and 3D descriptors, as well as fingerprints. HAPPENN[26] and HemoNet[32] are two neural network-based classifier models using the DBAASP and Hemolytik databases for dataset construction. HAPPENN represents a state-of-the-art model for predicting hemolytic activity, leveraging features selected via an ensemble of RF model and support vector machines (SVMs). In the HemoNet tool, SMILES-based fingerprints are used as a feature so that it can capture N/C terminal modification. Plisson et al.[31] explored 14 binary classifiers to predict hemolytic activity across three datasets developed by Chaudhary et al. They utilized 56 sequence-based physicochemical descriptors and employed an ML model called Extreme Gradient Boosting Classifier (XGBC) to discover and design non-hemolytic peptides. In 2021, Capecchi et al.[33] introduced a recurrent neural network classifier to identify membrane-disruptive amphiphilic antimicrobial peptides. Additionally, they developed a model for designing short non-hemolytic antimicrobial peptides, leveraging data sourced from DBAASP. Ansari et al.[34] developed a recurrent neural network (Bi-LSTM) with concatenated amino acid frequencies to predict whether a peptide is hemolytic. They utilized the DBAASPv3 database to create their dataset, comprising 9316 positive sequences (length range: 1–77) and negative sequences (length range: 1–190) in the training data. Salem et al.[35] utilized transfer learning to address the issue of small data and employed a protein language model (PLMs) based on LLM (Large Language Model), employing a tool named AMP-Deep. Perveen et al.[36] introduced an ML approach named Hemolytic-pred, designed for predicting hemolytic proteins. The dataset used in this method was collected from UniProtKB-SwissProt[37], and it employs position and composition-based features. Castillo-Mendieta et al.[38] utilize network science and data mining to analyze hemolytic peptides, creating scaffolds to represent their chemical space and uncovering putative hemolytic motifs. Zhuang et al.[39] developed a tool employing the quantum support vector machine (QSVM) to classify peptides as hemolytic or non-hemolytic, utilizing the dataset from the HemoPI tool. A recently developed PeptideBERT[40] based approach for hemolytic peptide identification utilizes the ProtBERT pre-trained transformer model featuring 12 attention heads and 12 hidden layers. Several other computational tools have been developed for different types of toxicity. In Supplementary Table S1, we present a comprehensive list of hemotoxicity prediction methods. While the methods mentioned above have contributed to advancing the discovery of potential hemolytic peptides, further improvement remains necessary. This is primarily because the datasets utilized in these methods are overly generalized to all vertebrates. Moreover, none of these methods have the capability to predict the $HC_{50}$ value of peptides. It is important for drug development to identify the peptide concentration at which any peptide exhibits significant hemolytic activity.

To address challenges faced by the scientific community, we proposed a method for predicting hemolytic peptides as well as $HC_{50}$ value against mammalian RBCs. The proposed method, HemoPI2, is trained and evaluated on experimentally validated 1926 hemolytic peptides. We have developed various classification models and regression models using ML, as well as PLMs. We have also developed ML-based models using word embeddings extracted from LLMs. These models have undergone rigorous benchmarking against independent dataset. Our HemoPI2 proposed in this study is an improved version of HemoPI, which has been widely utilized by the scientific community. Figure 1 provides a visual representation of the algorithm and processes undertaken in the study.

## Results

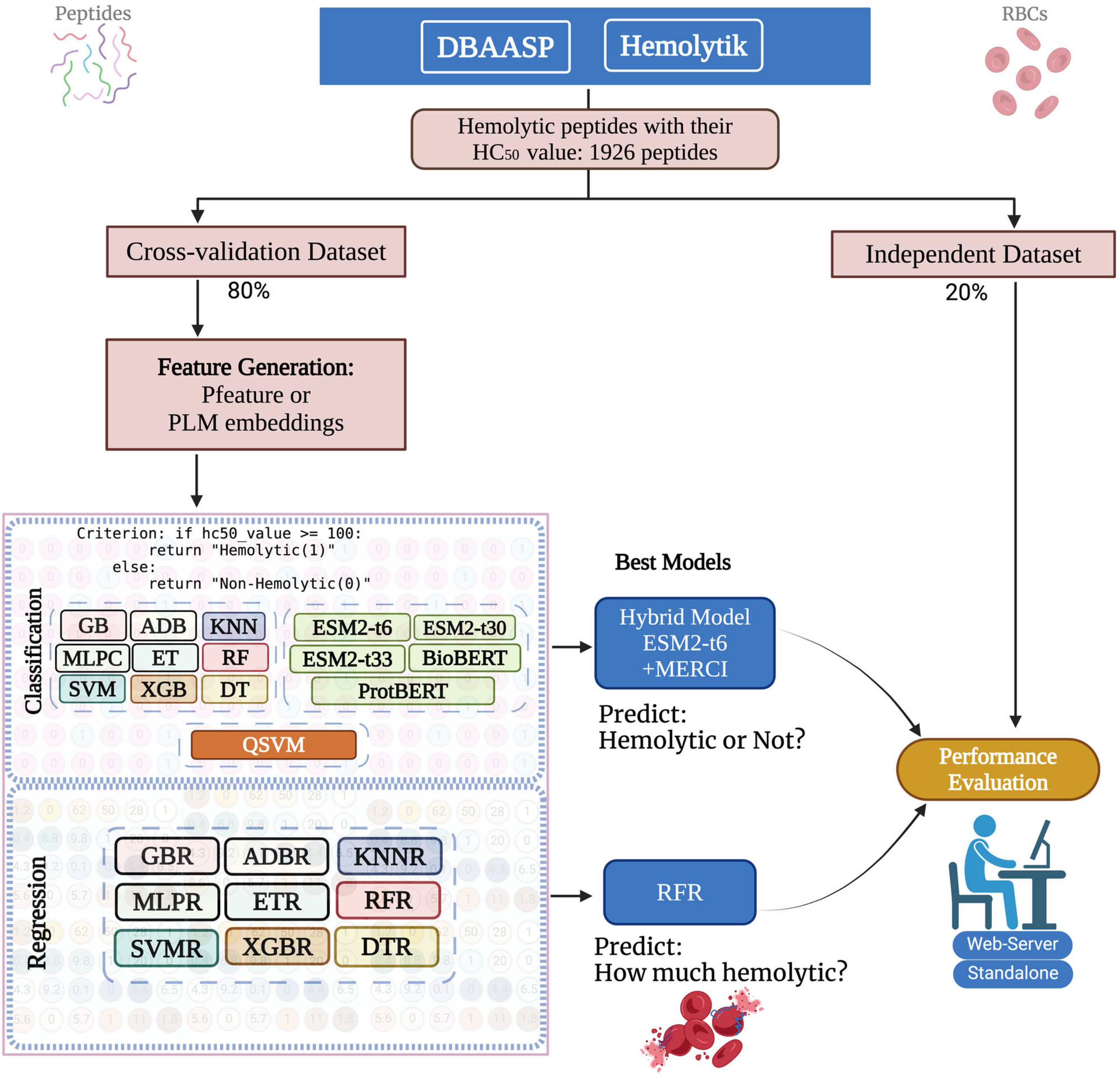
The dataset included peptide sequences with experimentally validated hemolytic activity levels. We performed comprehensive analyses, including compositional, positional, and motif analyses, along with feature extraction using the Pfeature tool. Feature vectors are utilized to study their correlation with $HC_{50}$ values and develop diverse ML models encompassing regression and classification. In addition to traditional ML models, protein language models (PLMs) were also implemented. Embeddings were extracted from large language models (LLMs) and utilized as feature vectors in the ML models. The predictive performance of these models was evaluated using an independent or unseen dataset, providing insights into their efficacy in predicting peptide hemolytic activity.

### Analysis of hemolytic peptides

**Amino acid composition analysis**. The investigation into the amino acid composition of hemolytic and non-hemolytic peptides revealed distinct patterns that underscore their functional differences. In the comparative analysis shown in Fig. 2A, specific amino acid residues Cysteine, Phenylalanine, Glycine, and Serine were identified as significantly more abundant ($p < 0.001$) in hemolytic peptides, highlighting their dominant presence. These residues exhibited significantly higher proportions within hemolytic peptides compared to their non-hemolytic counterparts. Furthermore, the study noted the presence of other residues, such as Lysine, Methionine, Asparagine, and Tryptophan, albeit in lesser quantities, further distinguishing hemolytic peptides. Interestingly, the comparison extended to the termini composition, where the N and C termini exhibited almost identical overall compositions to the entire peptide shown in Fig. 2B. However, disparities surfaced in the distribution of Phenylalanine and Glycine, with elevated concentrations observed exclusively at the N-terminal end. Conversely, the C-terminal region exhibited an increased abundance of Asparagine, Glutamine, Cysteine, and Tyrosine in hemolytic peptides. These findings highlight the amino acid profiles that contribute to the hemolytic properties of peptides, offering insights into their molecular mechanisms and potential applications in various biomedical contexts.

### Positional analysis

During our analysis, we aimed to identify any potential preferences for specific amino acid residues at particular positions within the peptide sequence. We constructed TSL for both hemolytic and non-hemolytic peptides, as illustrated in Fig. 3. The TSL provides valuable insights into the relative abundance of amino acid residues and their significance within the sequence. Enriched residues are more prevalent at a given position in hemolytic peptides compared to the non-hemolytic class, while depleted residues are more prevalent at a given position in non-hemolytic peptides. Upon initial examination of the TSL, we observed findings consistent with the amino acid composition analysis: hemolytic peptides exhibited enrichment in hydrophobic residues and were predominantly depleted in positively charged residues. Further scrutiny revealed position-specific enrichments. In the N-terminal 15 residues (Fig. 3A), hemolytic peptides were enriched in hydrophobic residues like Phenylalanine in position 1; Leucine in positions 2, 6, 8, and 12; Glycine in position 3; Isoleucine in position 5, and Proline in position 15, while in non-hemolytic peptides Lysine is preferred at most positions. Similarly, in the C-terminal 15 residues (Fig. 3B), amino acids such as Alanine at position 1, Proline at position 5, Lysine at positions 6 and 14, and Arginine at positions 12 and 15 were more prevalent in hemolytic peptides, while Lysine is preferred in most positions but at positions 6 and 14 Alanine is preferred in non-hemolytic peptides. These observations highlight the differential distribution of amino acids along the peptide sequence and provide valuable insights into the structural and functional characteristics of hemolytic as well as non-hemolytic peptides.



```
Criterion: if hc50_value >= 100:
        return "Hemolytic(1)"
    else:
        return "Non-Hemolytic(0)"
```

**Fig. 1 |** Illustration of the comprehensive workflow adopted throughout this study.

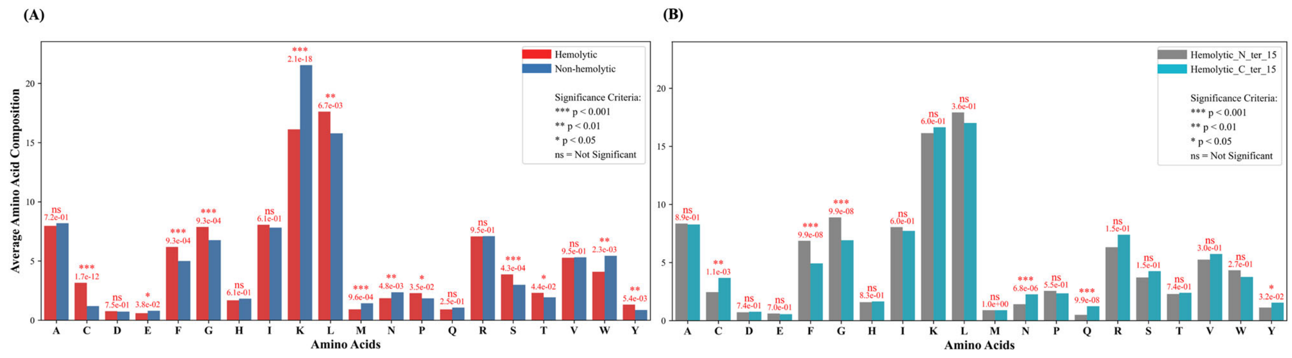


**Fig. 2 | Representation of the average percentage composition of amino acid residues in various parts of experimentally validated peptides. A** Comparison of average amino acid composition between hemolytic and non-hemolytic peptides. **B** Comparison of average amino acid composition between N-terminal 15 residues, and C-terminal 15 residues of hemolytic peptides.

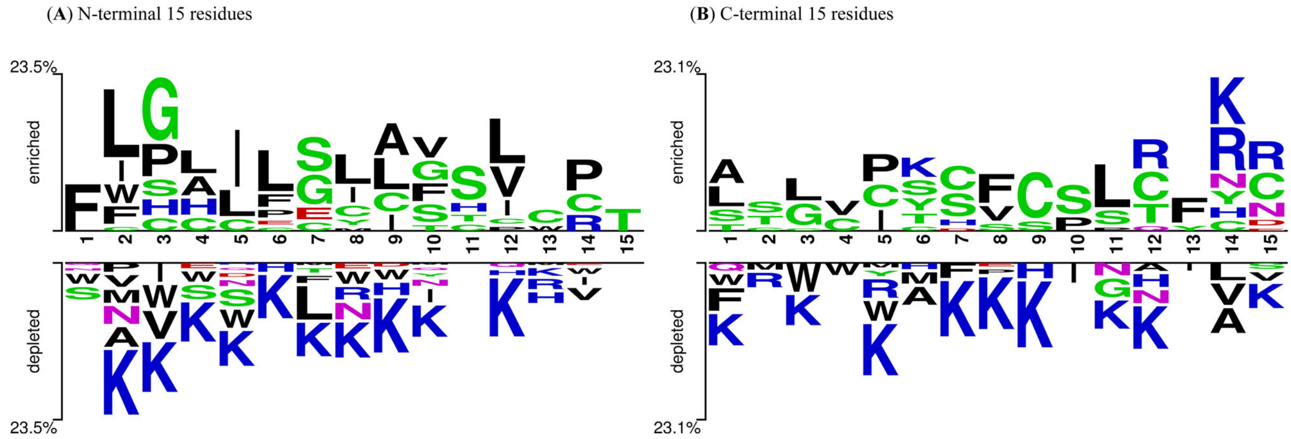


**Fig. 3 | This illustration presents TLS that depicts the residue preferences in hemolytic and non-hemolytic peptides.** Logo **A** represents the first 15 residues of the N-terminal, while Logo **B** represents the last 15 residues of the C-terminal.

**Table 1 | Top single amino acid composition, di-peptide composition, and physico-chemical features correlated to the $HC_{50}$ value of hemolytic peptides**

| Single amino acid composition | Correlation | Di-peptide composition | Correlation | Physico-chemical Features | Correlation |
|---|---|---|---|---|---|
| Lysine (AAC_K) | 0.214 | Lysine-Lysine (DPC1_KK) | 0.207 | Positively charged residues | 0.236 |
| Arginine (AAC_R) | 0.043 | Isoleucine-Lysine DPC1_IK | 0.191 | Basic residues | 0.236 |
| Tryptophan (AAC_W) | 0.018 | Alanine-Lysine DPC1_AK | 0.165 | Hydrophilic residues | 0.220 |
| Alanine (AAC_A) | 0.011 | Leucine-Tyrosine DPC1_LY | 0.143 | Exposed | 0.166 |
| Tyrosine (AAC_Y) | 0.010 | Tyrosine-Lysine DPC1_YR | 0.101 | Helix | 0.157 |
| Proline (AAC_P) | −0.075 | Phenylalanine-Leucine DPC1_FL | −0.088 | Hydrophobic residues | −0.154 |
| Cysteine (AAC_C) | −0.075 | Valine-Leucine DPC1_VL | −0.091 | Buried residue | −0.167 |
| Serine (AAC_S) | −0.092 | Glycine-Glycine DPC1_GG | −0.094 | PCP_Z3 | −0.204 |
| Phenylalanine (AAC_F) | −0.093 | Leucine-Phenylalanine DPC1_LF | −0.102 | Neutral charged residues | −0.233 |
| Glycine (AAC_G) | −0.102 | Isoleucine-Glycine DPC1_IG | −0.132 | Neutral residues based on pH | −0.233 |

## Motif analysis

Motif analysis was conducted to identify motifs present exclusively in either hemolytic or non-hemolytic peptides. This analysis identifies specific segments or patterns within peptides that contribute to their hemolytic activity. Consequently, motif analysis not only enhances prediction accuracy but also aids in identifying the precise motif or region responsible for hemolysis. We selected only exclusive motifs that appeared with a minimum frequency of 12. The top ten motifs unique to hemolytic peptides include "CGET", "CGETC", "HHIIGG", "SAGKA", "RLIR", "GETC", "TLLKKVLKA", "GGLFS", "IGGLF" and "HRLIRR". Conversely, the top ten motifs exclusive to non-hemolytic peptides consist of "AKD", "SKIK", "DLA", "LKHII", "HVQ", "NKL", "HRK", "INKQ", "KDLA" and "KINKQ". Supplementary Table S2 contains the full list of motifs. In addition, we explored classification schemes like *Koolman-Röhm* and *Betts-Russell*, which categorize amino acids based on their physicochemical properties, providing unique frameworks for amino acid classification. The *Koolman-Röhm*[41] scheme groups amino acids into categories such as aliphatic, sulfur-containing, aromatic, neutral, acidic, basic, and proline. In contrast, the Betts-Russell classification, developed by *Betts and Russell*[42], organizes amino acids into three main groups: Polar (charged: positive and negative), Hydrophobic (aromatic and aliphatic), and Small (tiny). By incorporating these MERCI approaches with the classification techniques, we enhanced the predictive capabilities of our model. A detailed discussion of this hybrid classification model is provided under the heading "Hybrid Model for Classification".

## Correlation analysis

Past research shows correlation analysis is crucial in predicting protein functions, understanding disease mechanisms, and discovering drugs[43]. It quantifies the strength and direction of relationships between variables, highlighting key influencers and potential patterns. In Table 1, we have highlighted the top features of AAC, DPC, and physico-chemical properties that exhibit a correlation with the $HC_{50}$ concentration of hemolytic peptides. Features that exhibit a positive correlation with $HC_{50}$ concentration of hemolytic peptides indicate that as specific attributes (e.g., composition of positively charged residues) increase, the $HC_{50}$ value of hemolytic peptides also rises, suggesting a reduction in hemolytic potency. On the other hand, a negative correlation (e.g., the composition of neutral and hydrophobic residues) indicates that an increase in a feature's value leads to a decrease in the $HC_{50}$ value, implying an increase in hemolytic potency. A comprehensive list of correlation analyses of each feature is provided in Supplementary Table S3.

## Classification models

We developed various classifiers using a combination of ML techniques and PLMs. In order to develop a classification model, we employed a dataset comprising 1926 distinct experimentally validated hemolytic peptides. We assigned binary labels to differentiate between strong hemolytic (positive) and weak hemolytic (negative) peptides. Our classification criteria categorized peptides with an $HC_{50}$ of ≤100 µM as hemolytic.

## Machine learning models

We developed various classifiers to classify high hemolytic peptides and weak hemolytic peptides, including ET, SVM, XGBC, RF, MLPC, GB, DT, and LR. Initially, we calculated the features of the hemolytic peptides utilizing the compositional-based module of Pfeature. Additionally, molecular weight and peptide length were included as feature vector components. This procedure yielded a comprehensive set of 1192 feature vectors for each peptide. Once again, the tree-based classifiers outperformed other classifiers, particularly RF. In Table 2, the performance was compared based on the AUC using different features and evaluated on an independent dataset. RF classifier achieved the best performance on the ALLCOMP excluding SOC (1190 descriptors) with an AUC of 0.888, which is very close to the performance on combined features AAC + DPC + PCP (450 descriptors) with an AUC of 0.881, AAC + DPC (420 descriptors) with an AUC of 0.875, and CeTD (189 descriptors) with an AUC score of 0.856. The detailed evaluation metrics are provided in Supplementary Table S4.

## Protein language models

Our study evaluated the performance of various PLMs in classifying hemolytic peptides, measured by the evaluation metric. The results highlight the effectiveness of different PLMs in accurately classifying hemolytic peptides, with ProtBERT exhibiting particularly strong performance with AUC 0.875. The varied AUC values provide insights into the relative strengths of each model in our classification task, allowing for informed model selection and optimization in future studies. Table 3 shows the detailed performance of each model.

## ML models based on word embeddings

Over the past decade, the application of language models in the field of bioinformatics has seen a significant surge. A multitude of protein and nucleotide language models have been developed with the aim of predicting the function of biological macromolecules[44–46]. In this study, we employed PLMs such as ESM-2 (with checkpoints: esm2_t36_3B_UR50D, esm2_t33_650M_UR50D, and esm2_t6_8M_UR50D), ProtBERT, and BioBERT to predict the hemolytic and non-hemolytic peptides. Given that these models are not specifically tuned for any particular property of peptides, we optimized the hyperparameters of these models on our dataset of hemolytic peptides for classification tasks. Subsequently, these tuned models were utilized to extract embeddings from hemolytic peptides. These embeddings were then used as input for ML classification models.

We generated 2560, 1280, 320, 1024, and 767 embeddings from ESM2-t36, ESM2-t33, ESM2-t6, ProtBERT, and BioBERT, respectively, for each peptide. In the case of ProtBERT, we also employed a combination of a fine-tuned ProtBERT model with BiLSTM to extract high-quality embeddings. These embeddings were used as features for developing ML models to predict the hemolytic peptides. The ML classifiers, including ET, SVM, XGB, RF, MLPC, GB, and DT, were trained to predict the hemolytic activity of peptides. MLPC with ProtBERT embedding achieved maximum performance with AUC 0.882 (see Table 4).

**Table 2 | Evaluation performance metrics of RF classifier model on independent dataset using various features derived from Pfeature**

| Features | Sp (%) | Sn (%) | Acc (%) | MCC | AUC |
|---|---|---|---|---|---|
| AAC | 83.9 | 66.3 | 75.9 | 0.512 | 0.831 |
| DPC | 87.7 | 65.1 | 77.5 | 0.547 | 0.837 |
| PCP | 84.8 | 67.4 | 76.9 | 0.534 | 0.854 |
| CeTD | 86.7 | 72.0 | 80.1 | 0.597 | 0.856 |
| AAC + DPC | 74.9 | 85.5 | 80.6 | 0.609 | 0.875 |
| AAC + DPC+CeTD | 76.2 | 78.8 | 77.5 | 0.598 | 0.841 |
| AAC + DPC + PCP | 73.7 | 85.0 | 79.8 | 0.623 | 0.881 |
| ALLCOMP | 74.9 | 83.9 | 80.1 | 0.610 | 0.878 |
| **ALLCOMP-ex SOC** | **76.5** | **86.9** | **82.1** | **0.640** | **0.888** |

*Sn* sensitivity, *Sp* specificity, *Acc* accuracy, *MCC* Matthews correlation coefficient, *AUC* area under receiver operating characteristic, Bold values indicate the best-performing model.

**Table 3 | Comprehensive performance analysis of PLMs on an independent dataset**

| PLM | Sp (%) | Sn (%) | Acc (%) | MCC | AUC |
|---|---|---|---|---|---|
| ESM2-t33 | 86.1 | 70.0 | 79.1 | 0.573 | 0.870 |
| ESM2-t30 | 84.2 | 78.0 | 81.3 | 0.599 | 0.851 |
| ESM2-t12 | 70.0 | 79.0 | 74.2 | 0.490 | 0.831 |
| ESM2-t6 | 76.1 | 84.0 | 79.5 | 0.591 | 0.870 |
| **ProtBERT** | **84.9** | **73.7** | **80.0** | **0.602** | **0.875** |
| BioBERT | 63.7 | 80.5 | 71.5 | 0.474 | 0.800 |

*Sn* sensitivity, *Sp* specificity, *Acc* accuracy, *MCC* Matthews correlation coefficient, *AUC* area under receiver operating characteristic, Bold values indicate the best-performing model.

## Quantum support vector machine

In this study, we utilize the Quantum Support Vector Machine (QSVM) on features extracted using the pfeature method. To facilitate a comparison between the QSVM and classical SVM models, we also developed various classical SVM models incorporating different kernels, including linear, radial basis function, polynomial, and sigmoid. Among the various kernels used in QSVM, the linear kernel showed better performance. Using AAC features, the QSVM achieved a peak AUC of 0.782, outperforming most classical SVM models except for the RBF kernel (AUC 0.797). However, with DPC features, the classical SVM model demonstrated superior performance. Similarly, for AAC + DPC and ALLCOMP features, the QSVM's performance fell below that of the classical SVM models. These results suggest that QSVM performs better than classical SVM when the feature count is low. The detailed performance metrics from both quantum and classical kernels across different feature sets on cross-validation and independent datasets are available in Supplementary Table S5.

## Hybrid model for classification

In order to boost the predictive power of our top-performing models, we developed a hybrid strategy for classifying hemolytic peptides. Table 5 shows the best hybrid model's performance on an independent dataset. When combined with MERCI, traditional ML models outperformed fine-tuned PLMs. Among the ML models, the RF model using ALLCOMP-ex SOC features achieved the highest AUC at 0.921. For embedding-based approaches, the Extra Trees (ET) model performed best with ESM2-t33 embeddings, reaching an AUC of 0.905. In the case of PLMs, the ESM2-t6 model achieved an AUC of 0.919, the highest among PLMs. Supplementary Table S6 provides detailed metrics for these models.

## Threshold optimization for model performance

Adjusting the threshold in a classification model is crucial for achieving the desired balance between sensitivity (true positive rate) and specificity

**Table 4 | Evaluation of ML classifiers' performance utilizing various PLM embedding sources on an independent dataset**

| Embedding source | ML classifier | Sp (%) | Sn (%) | Acc (%) | MCC | AUC |
|---|---|---|---|---|---|---|
| ESM2-t36 | XGBC | 86.0 | 74.9 | 80.8 | 0.614 | 0.869 |
| ESM2-t33 | ET | 85.5 | 73.2 | 79.8 | 0.604 | 0.873 |
| ESM2-t6 | ET | 73.2 | 84.5 | 79.3 | 0.583 | 0.857 |
| **ProtBERT** | **MLPC** | **83.9** | **84.0** | **83.9** | **0.677** | **0.882** |
| BioBERT | MLPC | 81.5 | 68.5 | 75.6 | 0.507 | 0.792 |
| ProtBERT +BiLSTM | ET | 82.8 | 71.4 | 77.4 | 0.547 | 0.859 |

*Sn* sensitivity, *Sp* specificity, *Acc* accuracy, *MCC* Matthews correlation coefficient, *AUC* area under receiver operating characteristic, Bold values indicate the best-performing model.

**Table 5 | Performance of hybrid model on independent dataset developed by combining best-performing classification models with MERCI**

| Model | | Threshold | Sp (%) | Sn (%) | Acc (%) | MCC | AUC |
|---|---|---|---|---|---|---|---|
| ML | **RF<br>(ALLCOMP-ex SOC)** | **0.45** | **83.4** | **83.8** | **83.5** | **0.670** | **0.921** |
| | RF<br>(AAC + DPC + PCP) | 0.49 | 82.6 | 77.6 | 83.9 | 0.621 | 0.899 |
| | XGBC<br>(ESM2t36 embeddings) | 0.47 | 84.0 | 80.8 | 82.9 | 0.658 | 0.900 |
| | ET<br>(ESM2t33 embeddings) | 0.49 | 85.6 | 82.4 | 82.8 | 0.651 | 0.905 |
| | MLPC<br>(ProtBERT embeddings) | 0.48 | 86.4 | 77.4 | 81.1 | 0.636 | 0.891 |
| | ET<br>(ESM2t6 embeddings) | 0.51 | 87.7 | 76.9 | 82.4 | 0.646 | 0.902 |
| PLM | ESM2-t33 | 0.57 | 92.7 | 68.5 | 83.5 | 0.668 | 0.901 |
| | ESM2-t30 | 0.48 | 79.7 | 82.7 | 80.2 | 0.634 | 0.885 |
| | **ESM2-t6** | **0.58** | **83.0** | **82.0** | **82.5** | **0.649** | **0.919** |
| | ProtBERT | 0.44 | 65.9 | 90.3 | 79.0 | 0.585 | 0.871 |

*Sn* sensitivity, *Sp* specificity, *Acc* accuracy, *MCC* Matthews correlation coefficient, *AUC* area under receiver operating characteristic, Bold values indicate the best-performing model in ML and PLM.

**Table 6 | Evaluation of ML regressor models constructed using word embeddings derived from PLMs on an independent dataset**

| Embedding source | ML model | *R* | $R^2$ | MAE | MSE |
|---|---|---|---|---|---|
| ESM2-t36 | ETR | 0.706 | 0.486 | 0.808 | 1.105 |
| **ESM2-t33** | **ETR** | **0.711** | **0.495** | **0.786** | **1.084** |
| ESM2-t6 | ETR | 0.677 | 0.452 | 0.820 | 1.177 |
| ProtBERT | RFR | 0.410 | 0.115 | 1.076 | 1.407 |
| BioBERT | ETR | 0.616 | 0.366 | 0.927 | 1.362 |
| ProtBERT +BiLSTM | RFR | 0.646 | 0.449 | 0.825 | 1.265 |

*R* Pearson correlation coefficient, $R^2$ coefficient of determination, *MAE* mean absolute error, *MSE* mean squared error, Bold values indicate the best-performing model.

(false positive rate avoidance). The threshold represents the cutoff value for the probability at which the model categorizes an instance as positive or negative. As presented in Supplementary Table S7, optimizing the threshold for our best-performing RF model (ALLCOMP-ex SOC), with an AUC of 0.888, enables a tailored balance between high coverage and precise prediction. Lower thresholds (e.g., 0.1) result in a higher rate of negative predictions, meaning the model captures all possible true positives (high sensitivity) but sacrifices precision or positive predictive value. This approach broadens coverage, identifying most cases but also allowing more false positives. On the other hand, setting the threshold high, close to 1, raises the rate of correctly predicted positives (high specificity and positive predictive value), but coverage narrows as fewer instances are classified as positive. This trade-off means that for applications requiring higher sensitivity (e.g., screening tests), lower thresholds are often preferred, while applications needing high specificity might benefit from a higher threshold.

### Regression models

Initial investigations have suggested the feasibility of distinguishing hemolytic peptides from non-hemolytic ones based on factors such as amino acid composition, binary profiles, motifs, and physicochemical properties. Our study utilized several popular ML regressors to predict the hemolytic activity of peptides using features derived from their primary sequences. The predictive models underwent training and testing via fivefold cross-validation on the training dataset (80%), with the final models being evaluated using an independent dataset (20%). We reported key statistical parameters, including *R*, $R^2$, MAE, and MSE.

### ML regressors using Pfeature

In this study, we developed prediction models using a range of regressors, including XGBR, RFR, GBR, ETR, DTR, ADBR, SVR, KNNR, and LR. The various features we used for the regression models are the same as those used for the classification models. This comprehensive approach ensures the robustness and accuracy of our hemolytic activity prediction model. While experimenting with various categories of features and feature combinations, tree-based regressors such as RFR and ETR consistently emerged as the top performers among the various ML algorithms we evaluated. Detailed performance of each model is provided in Supplementary Table S8. Although the performance of RFR slightly outshone ETR, the difference was marginal. The performance of the RFR with different sets of features on an independent dataset. The RFR demonstrated its superior performance when employing the ALLCOMP excluding SPC (1167 descriptors), achieving a *R* of 0.739 and an $R^2$ value of 0.543.

### ML models based on word embeddings

Word embeddings were generated using various checkpoints of ESM2 (ESM2-t36, ESM2-t33, ESM2-t6), ProtBERT, BioBERT, and ProtBERT + BiLSTM, mirroring the approach adopted to develop classification models. The number of embeddings extracted was 2560, 1280, 320, 1024, and 767 for ESM2-t36, ESM2-t33, ESM2-t6, ProtBERT, and BioBERT models, respectively, consistent with that used in classification models. These embeddings were used as features for developing ML regressor models like XGBR, RFR, GBR, ETR, DTR, ADBR, SVR, KNNR, LR, and MLPR to predict the $HC_{50}$ peptides. Table 6 illustrates the performance of the top-performing ML regression model with the corresponding PLMs on an independent dataset, utilizing embeddings derived from various language model architectures. Once more, among the ML regressors, tree-based methods like the EFR and ETR consistently stood out as strong performers. ESM embeddings, particularly those generated by ESM2-t33, demonstrated the best performance, especially with the Extra Trees Regressor achieving an *R* of 0.711 and an $R^2$ of 0.495.

### Model finalization

In our research, we constructed several regression and classification models, utilizing a diverse set of features derived from Pfeature. Additionally, we incorporated word embeddings extracted from PLMs as input features. We experimented with various combinations of these features to optimize our models. All developed models were ultimately evaluated on an independent dataset, and their performances were compared. In our classification models developed using features extracted from Pfeature, the RF classifier achieved an AUC of 0.888, and after combining these features with MERCI, the AUC improved to 0.921. When embeddings were used as features to develop machine learning classifiers, the performance was comparable to the model developed using Pfeature but did not outperform it in either case, before or after combining with MERCI. Among the PLMs, ProtBERT performed the best, with an AUC of 0.875. However, after combining with MERCI, its performance did not improve and was outperformed by the ESM-t6 model, which achieved an AUC of 0.919. Among all the regression models, the RFR model exhibited superior performance with an $R$ 0.739 and an $R^2$ 0.543. A comparative analysis of the performance of the best ML and PLM regression models is presented in the tables.

In our classification model development, we diversified our approach similar to the regression models. Notably, we observed promising performance with RF (ALLCOMP-ex SOC) and ProtBERT models. Moreover, we extended our classification methodologies by integrating hybrid approaches that combine ML and PLM classifiers with MERCI. These hybrid strategies were meticulously crafted to leverage the complementary strengths of both methods. Subsequent evaluation revealed enhancements in model performance following the incorporation of MERCI. A comparative assessment of these hybrid models on independent datasets is detailed in the table. The RF (ALLCOMP-ex SOC) and ESM2-t6 models are particularly noteworthy, which demonstrated optimal performance with AUC scores of 0.921 and 0.919, respectively. These regression and classification models have been seamlessly integrated into our prediction software and web services, streamlining the quantification and classification of hemolytic peptides.

## Benchmarking

Benchmarking of developed method against existing ones is crucial to comprehend its significance and potential improvements. In the context of a regression model, to the best of our knowledge, no tool has been specifically designed to quantify hemolytic peptides. However, several traditional machine learning-based tools have been developed for classification (as detailed in Supplementary Table S1). We conducted an evaluation of these current tools using an independent/unseen dataset from our study. This evaluation provides a comparative analysis with other existing approaches, offering insights into their relative effectiveness. Table 7 compares the performance of our proposed approach, HemoPI2, with other existing methods reported in the literature. On the independent dataset used in HemoPI2, methods like HAPPENN, PeptideBERT, and Plisson et al. achieved AUC values of 0.736, 0.739, and 0.740, respectively. HemoPI2 significantly outperformed these methods, achieving AUC values of 0.888 with the RF model and 0.882 with the ProtBERT model. We further improved performance with hybrid models by incorporating MERCI. Our hybrid models, RF and ESM2-t6, achieved AUCs of 0.921 and 0.919, respectively. The subpar performance of tools like HemoPI and HemoPred could potentially be attributed to the utilization of randomly generated negative datasets during training. Furthermore, we faced challenges in comparing the performance of certain tools such as HLPpred-FUSE, HemoNet, hemolytic-Pred, RNN (Capecchi et al.), and QSVM due to limitations in their services. HLPpred-FUSE and hemolytic-Pred had issues with their web services, while HemoNet and RNN (Capecchi et al.) had non-functional GitHub code due to missing input files. Additionally, QSVM currently offers no model or webserver, limiting our ability to make a full assessment. However, we were able to compare our results with the QSVM model using their GitHub code, which was based on 56 physicochemical descriptors, though their primary model was built using 40 descriptors. On the HemoPI-3 dataset, a QSVM achieved an AUC of 0.705, demonstrating that our classical ML models and PLMs outperform this quantum ML approach. HemoPI2 demonstrates superior performance in predicting hemolytic activity compared to other existing methods, with our Random Forest and ESM-2t6 models leading the way. The ESM-2 model utilizes a transformer architecture to generate rich protein embeddings, capturing more intricate relationships in protein structure compared to the earlier approaches, while the RF model enhances prediction accuracy through ensemble learning, reducing overfitting. Together, these advanced techniques provide robust and reliable predictions, substantiating the effectiveness of our updated method over existing tools. This superior performance underscores the potential of HemoPI2 as a valuable tool in the field of therapeutic peptide development, particularly in the classification and quantification of hemolytic peptides.

**Table 7 | Benchmarking of classification tools on an independent dataset of HemoPI2**

| Tools | Model | Sp (%) | Sn (%) | Acc (%) | MCC | AUC |
|---|---|---|---|---|---|---|
| HemoPI | D1 | 16.7 | 90.8 | 58.4 | 0.099 | 0.538 |
| | D2 | 59.9 | 60.8 | 60.3 | 0.207 | 0.662 |
| | ModelAll | 49.2 | 69.8 | 58.8 | 0.194 | 0.669 |
| HemoPred | Default | 31.5 | 87.0 | 57.2 | 0.221 | – |
| HAPPENN | Main_Dataset | 53.7 | 84.9 | 68.1 | 0.402 | 0.736 |
| AMP_deep | Prot-BERT-BFD | 67.0 | 53.0 | 59.0 | 0.413 | 0.602 |
| Ansari et al. | Embedding + Bi-LSTM | 73.3 | 32.3 | 54.5 | 0.07 | 0.512 |
| PeptideBERT | Default | 79.1 | 62.4 | 71.0 | 0.422 | 0.739 |
| Plisson et al. | XGBC | 69.0 | 61.4 | 70.0 | 0.470 | 0.740 |
| QSVM(56 D Descriptors) | HemoPI-1 | 92.7 | 13.0 | 50.0 | 0.095 | 0.642 |
| | HemoPI-2 | 73.2 | 49.8 | 60.6 | 0.234 | 0.701 |
| | HemoPI-3 | 81.0 | 41.5 | 59.8 | 0.243 | 0.705 |
| HemoPI2 (Without MERCI) | ESM2-t6 | 76.1 | 84.0 | 79.5 | 0.591 | 0.870 |
| | RF | 76.5 | 86.9 | 82.1 | 0.640 | 0.888 |
| **HemoPI2(With MERCI)** | **ESM2-t6** | **80.5** | **88.8** | **84.3** | **0.691** | **0.919** |
| | **RF** | **88.3** | **79.8** | **84.3** | **0.685** | **0.921** |

*Sn* sensitivity, *Sp* specificity, *Acc* accuracy, *MCC* Matthews correlation coefficient, *AUC* area under receiver operating characteristic, Bold values indicate the best-performing model.

### Community contribution through HemoPI2

To contribute to the scientific community, we have made our regression and classification algorithms accessible online via our user-friendly web server, HemoPI2, which is freely available for academic researchers at https://webs.iiitd.edu.in/raghava/hemopi2/. The web server is organized into four main modules: Home, Prediction, Protein scanning, Motif scan, and Design. The prediction modules enable users to predict the hemolytic potency of their peptides by submitting multiple peptide sequences in FASTA format. However, predictions are restricted to peptides composed of the 20 natural amino acids, with non-canonical amino acids not being supported. The protein scanning module provides the functionality to identify or scan hemolytic regions within a protein. The motif scan module enables users to scan or map hemolytic motifs within the query sequence using MERCI. The design module creates non-hemolytic peptides from the primary sequence, generating mutant peptides with a single mutation for hemolytic activity prediction. Additionally, Download and Help modules are available to facilitate data download and provide user assistance. Alongside the web server, we have also developed a standalone package that can be downloaded from our web server. This standalone software and pip package, designed for large-scale hemolytic peptide prediction, offers a versatile solution for comprehensive analysis.

## Discussion

In recent years, the focus on therapeutic peptide development has grown, but many potential peptides face challenges due to hemolytic toxicity, which hinders their approval as a drug molecules. Clinically approved peptides typically have a high therapeutic index, with minimal hemolytic activity. However, many promising peptides still exhibit varying degrees of hemolysis. Thus, developing peptides with a high therapeutic index is crucial. While numerous tools exist to classify hemolytic peptides, this study aims to enhance the design and optimization process by introducing an improved in silico method to classify and quantify peptides as hemolytic or non-hemolytic based on their sequence. This approach aims to advance therapeutic peptide development by improving the understanding of hemolytic toxicity and aiding in the creation of safer, more effective drugs.

We used two types of features for model development: features from Pfeature and embeddings from fine-tuned PLMs. The RF classifier achieved an AUC of 0.888 using ALLCOMP-ex SOC features. After combining with MERCI, the AUC increased to 0.921 on an independent dataset. QSVM performs well with lower-dimensional features, outperforming most classical SVMs with AAC features, but requires further optimization to consistently match or exceed traditional models in complex scenarios. For regression, RFR outperformed other models, achieving an $R$ of 0.739 and an $R^2$ of 0.543 using ALLCOMP-ex SPC features. The logarithmic transformation of $HC_{50}$ values ($pHC_{50}$) improved model performance, though its effectiveness can vary. While beneficial in some cases, as noted by Feng et al.[47], log transformation may sometimes increase variability or skewness[48]. Therefore, it should be applied carefully based on dataset characteristics. The MAE and MSE were 0.734 and 0.981, respectively. Residual errors may result from terminal modifications like amidation and acetylation, variations in hemolytic concentration reporting, or noise in experimental datasets. Some sequences, like "RVKRVWPLVIRTVIAGYNLYRAIKKK," are reported as both hemolytic and non-hemolytic in different studies[34]. To reduce errors, we could normalize hemolytic data, include features for modifications, and incorporate structural data to improve sensitivity.

We found that tree-based methods, especially RF, outperformed other models when applied to features like AAC, DPC, PCP, and CeTD, indicating a strong correlation with hemolytic activity. The combination of these features further improved performance. RF's ensemble learning, which uses multiple decision trees to reduce variance and prevent overfitting, is particularly effective with high-dimensional data[49]. Additionally, RF can capture complex, non-linear relationships common in biological data, making it a robust choice for both regression and classification tasks[50,51]. ESM-2 embeddings showed similar performance across both classifier and regressor categories, regardless of checkpoint size. Typically, larger checkpoints have more layers, contain more parameters, and are trained on larger datasets, which can enhance performance, though this also increases computational demands.

Various analytical methods were employed to analyze experimentally validated hemolytic peptides, focusing on sequence-based features like amino acid composition (AAC), positional distribution, and motif patterns. TSL was used to identify amino acid preferences at different positions. Motif analysis with the MERCI tool revealed recurring patterns associated with hemolytic activity. Correlation analysis between Pfeature-extracted features and $HC_{50}$ concentrations identified key features linked to hemolytic activity. Hydrophobic residues such as Cysteine, Phenylalanine, Glycine, and Leucine were predominant in hemolytic peptides. Positional analysis showed enriched hydrophobic residues in the N-terminal, and motif analysis identified unique motifs in hemolytic peptides (e.g., "CGET", "TLLKKVLKA") and non-hemolytic ones (e.g., "AKD", "SKIK"). These motifs, rich in hydrophobic amino acids, promote membrane disruption through increased interaction with the lipid bilayer, supporting hemolytic activity. Models describing peptide-lipid bilayer interactions, such as the carpet, barrel-stave, and toroidal pore models, explain these mechanisms[26]. Correlation analysis also showed that certain residues, like positively charged ones, negatively correlate with hemolytic activity, while hydrophobic residues positively correlate with it. Our analysis highlights the key role of hydrophobicity in hemolytic activity, with peptides exhibiting higher hydrophobicity showing greater hemolytic potential. Amino acids like Cysteine, Phenylalanine, Glycine, and Leucine, due to their chemical properties, prefer lipid bilayers, leading to membrane disruption and cell lysis[52]. The amphipathic nature of many hemolytic peptides, with both hydrophobic and hydrophilic regions, further enhances membrane disruption[53,54]. Peptides containing cysteine in β-sheets are also linked to increased hemolytic activity[55–57]. These findings emphasize the critical role of hydrophobic residues in membrane interactions and hemolysis.

### Challenges and future directions

This study presents significant advancements in predicting and quantifying the hemolytic potential of peptides. However, several limitations require further attention. One key limitation is the size and recency of the datasets used for training. While the Hemolytik database is valuable, it has not been updated, highlighting the need for a more comprehensive and current dataset. Additionally, HemoPI2 is limited to natural peptides, excluding non-canonical amino acids, modified peptides, and sequences with fewer than six residues. This restricts its applicability to synthetic peptides, which are increasingly relevant in therapeutics. Future work should extend the model to include these peptide types for broader use. Furthermore, HemoPI2, which predicts hemolytic activity based solely on peptide sequences, faces challenges from sequence and structural biases. Integrating structural data, as recent machine learning research suggests, could improve predictive accuracy and applicability across diverse peptide structures[58]. Lastly, the potential taxonomic bias in the training data should be examined to ensure robust predictions across underrepresented taxa[59]. Addressing these challenges will enhance the model's overall utility and accuracy.

## Methods

### Data collection

We acquired experimentally validated hemolytic peptides from DBAASP version 3 and Hemolytik database[17,60]. These databases offer comprehensive details on the hemolytic activity of peptides that have undergone experimental validation. Peptide activity is assessed by extrapolating measurements from dose-response curves to determine the concentration at which 50% of RBCs are lysed, known as the $HC_{50}$ value. We collected 3147 peptides from DBAASP and 560 peptides from the Hemolytik database, whose $HC_{50}$ value is available. We implemented several preprocessing steps that included the removal of peptides containing non-natural amino acids and the removal of peptides containing less than six residues. In cases where a peptide sequence has multiple $HC_{50}$ values or a range of $HC_{50}$ values, we computed the average of these values. This mean activity measure represents the overall hemolytic activity of the peptide under various experimental conditions. By averaging, we ensure that our model captures the general behavior of the peptide's hemolytic activity rather than specific instances, thereby enhancing the robustness of our predictions. The final HemoPI2 dataset comprises 1926 unique experimentally validated hemolytic peptides along with their corresponding hemolytic concentrations measured in µM. For classification purposes, we establish binary labels distinguishing strong hemolytic (positive) from weak hemolytic (negative) peptides; we utilized the criterion that peptides with an $HC_{50}$ of ≤100 µM are classified as hemolytic. Peptides with $HC_{50}$ values ≤ 100 µM were labeled as hemolytic, while those with values > 100 µM were classified as non-hemolytic. Following this criterion, we got 891 hemolytic peptides and 1035 non-hemolytic peptides in the dataset.

We standardized the $HC_{50}$ values by converting them into a uniform measurement unit (µM). Following this, we transformed these $HC_{50}$ values into $pHC_{50}$ values using a specific equation (Eq. (1)). This standardization and transformation process ensures consistency in our data, which is crucial for the accuracy and reliability of our regression model. It allows us to compare and analyze the hemolytic activity of different peptides on a common scale, thereby enhancing the predictive capabilities of our model.

## Predictive target for regression analysis

In this study, we have chosen the negative logarithmic $HC_{50}$ ($pHC_{50}$) as the target for our regression prediction. The $pHC_{50}$ value is calculated using the following equation:

$$pHC_{50} = -\log_{10}(HC_{50}) \tag{1}$$

This equation transforms the $HC_{50}$ values into a logarithmic scale, which can help in handling a wide range of $HC_{50}$ values and can potentially improve the performance of the regression model. This transformation is commonly used in bioinformatics and cheminformatics for handling bioactivity data[61–63]. It allows us to compare and analyze the hemolytic activity of different peptides on a common scale, thereby enhancing the predictive capabilities of our model.

## Cross-validation approach

We have ensured that the study followed established bioinformatics protocols. Initially, our data was randomly divided into training and independent datasets, with the cross-validation set comprising 80% of the data and the independent set containing the remaining 20%. We employed a five-fold cross-validation technique within the cross-validation dataset for training and testing to assess model performance. This process involved randomly dividing sequences into five subsets, using four for training and one for testing iteratively across five repetitions. Notably, the independent dataset was held aside throughout training, testing, and hyperparameter tuning. Only the final models were evaluated using the independent or unseen dataset. Comparison of models on this independent dataset is critical to any predictive methodology.

## Feature extraction

We have used two approaches to extract features from peptide sequences. Initially, we used the Pfeature, which extracts composition-based features from a given peptide sequence, as well as captures the information of the composition along with its distribution in the given sequence. Next, fine-tuned pre-trained language models (PLMs) to generate embeddings, representing the sequences in a dense vector space that encapsulates semantic and syntactic information. This integrated approach, as illustrated in Supplementary Fig. S1, enables a comprehensive representation of the feature extraction approach.

## Feature extraction using Pfeature

Feature extraction is important for predicting peptide properties, as it converts raw sequence data into meaningful representations that machine learning models can accurately analyze and learn. To extract features, we utilized the Pfeature[64] tool, which yielded a diverse range of composition-based descriptors. These include AAC, which computes the frequency of each of the 20 amino acids in a peptide sequence, resulting in a 20-dimensional feature vector; DPC, which considers the coupling of adjacent amino acids and their positional information, yielding 400 features; Atom Type Composition (ATC) and Bond Type Composition (BTC), which calculate the fractions of different atoms (C, H, N, O, S) and bond types (total, aromatic, hydrogen, single, double), generating 5 and 4 features respectively; and Physico-chemical Properties composition, which encompass 30 descriptors that quantify various physical and chemical characteristics of the peptides, such as charge, polarity, aromaticity, etc. Additionally, we included various indices such as the Residue Repeats Index (RRI) (20 descriptors), which counts consecutive runs of each amino acid type; the Property Repeats Index (PRI) (25 descriptors), assessing the repetitiveness of physicochemical properties; and the Distance Distribution of Residues (DDR) (20 descriptors), measuring distances from the N-terminal and C-terminal as well as between identical residues. We also computed Shannon Entropy measures at different levels: overall peptide level individual amino acid level, and specific physicochemical property level (46 descriptors). The Conjoint Triad Descriptors (CTD) analyze the frequency of three consecutive amino acids based on their dipoles and side chain volumes, resulting in 343 descriptors. Pseudo Amino Acid Composition (PAAC) uses 21 descriptors to capture amino acid frequencies and biochemical properties, enhancing sequence analysis. Amphiphilic Pseudo Amino Acid Composition (APAAC) expands this with 23 descriptors to include hydrophilic and hydrophobic traits, useful for membrane proteins. Quasi-sequence Order (QSO) uses 42 descriptors to encode residue correlations, representing both composition and spatial arrangement for complex structural insights. Furthermore, the Sequence Order Coupling Number (SOC) utilizes dissimilarity matrices to compute a 2-dimensional feature vector. Lastly, the Composition-enhanced Transition Distribution (CeTD) describes amino acid distribution patterns along the peptide sequence based on seven physicochemical properties and generates 189 features. We also incorporated molecular weight and peptide length into the feature vector, resulting in a total of 1192 features per peptide, which serve as the basis for our predictive models. All features are further detailed and formulated in Supplementary Information S1.

## Word embeddings from protein language models

Recent strides in natural language processing (NLP) have catalyzed the emergence of PLMs, which harness individual amino acids and their combinations (doublets or triplets) as tokens or words. These models yield fixed-size vectors, referred to as embeddings, to encapsulate specific peptide sequences. These protein embeddings serve as pivotal inputs for a spectrum of tasks, spanning structure prediction, novel sequence generation, and protein classification[65]. In our study, we employed three widely recognized LLMs: ESM-2[66], ProtBERT, and BioBERT, to produce embeddings for peptide sequences. ProtBERT and BioBERT are built upon the BERT model[67] and are pre-trained on extensive datasets of protein sequences in a self-supervised manner. Conversely, ESM-2 (Evolutionary Scale Modeling) is a transformer-based PLMs initially developed for protein structure prediction, trained on sequences sourced from the UniRef protein sequence database[68]. ESM-2, renowned as a state-of-the-art protein model, is trained on a masked language modeling objective. This model proves adept at fine-tuning across an extensive array of tasks that entail protein sequences as inputs. Several ESM-2 checkpoints with varying sizes are available on HuggingFace, where larger sizes generally yield slightly better accuracy but necessitate significantly more memory and training time. We opted to utilize the ESM-2_t36_3B_UR50D, which consists of 36 transformer blocks with 3B parameters trained from UniRef50[69], and ESM-2_t33_650 M_UR50D[70], which consists of 33 transformer blocks with 650M parameters trained from UniRef50[69], checkpoints to generate embeddings. These checkpoints were deemed suitable for our objectives, offering a balance between accuracy and resource requirements. Subsequently, the embeddings derived from these ESM-2 checkpoints were employed as features for an ML regressor and classifier, facilitating the development of a robust model. This approach empowered us to exploit the rich contextual information encoded within the embeddings to enhance the predictive capabilities of the regressor and classifier models.

### Comprehensive analysis

To conduct the preliminary analysis of experimentally validated hemolytic peptides, we employed several analytical approaches. Initially, we conducted sequence-based analysis, which involved scrutinizing amino acid composition, positional distribution, and motif patterns. This exploration includes both the N-terminal and C-terminal regions. Following this, we employed *Two Sample Logo* (TSL)[71] to discern specific preferences for amino acid residues at distinct positions within the peptides. In motif-based analysis, we employed the *Motif-EmeRging and Classes-Identification (MERCI)* tool[72] to uncover recurring patterns contributing to hemolytic activity. We also involved various classification schemes, including *Koolman-Röhm* and *Bets-Russell*, to categorize motifs based on the physicochemical properties of amino acids, such as polarity, charge, hydrophobicity, aromaticity, and aliphatic characteristics. Such motifs serve as pivotal regions within peptides responsible for their hemolytic effects. Identifying these motifs enhances our understanding of the molecular mechanisms underlying toxicity, offering insights for drug development and therapeutic strategies[73–75]. Lastly, we conducted a correlation analysis between features extracted by Pfeature and their corresponding $HC_{50}$ concentrations. This facilitated the identification of features closely associated with peptide hemolytic activity, elucidating meaningful relationships crucial for further research[76].

### Classification models

**Machine learning models**. In our study, we employed a diverse set of ML classifiers to enhance the robustness and prediction accuracy. These include Extra Trees (ET), which are known for their ability to reduce over-fitting and bias; Support Vector Machines (SVM), effective in high-dimensional spaces; Extreme Gradient Boosting (XGB), renowned for its speed and performance; Random Forest (RF), appreciated for its handling of unbalanced datasets; Multi-Layer Perceptron Classifier (MLPC), a type of neural network known for its flexibility; Gradient Boosting (GB), recognized for reducing errors; and Decision Trees (DT) and Logistic Regressor (LR), both fundamental to understanding feature importance and relationships. Each of these classifiers contributes unique strengths to our ML pipeline, resulting in a more robust and accurate predictive strength.

### Protein language models

In our hemolytic peptide classification study, we utilized PLMs, which are computational frameworks that leverage natural language processing techniques to analyze protein structures, functions, and interactions. Specifically, we employed models such as ESM2-t33, ESM2-t30, ESM2-t12, and ESM2-t6 from the Evolutionary Scale Modeling (ESM) series, which are pre-trained on large protein sequence corpora and excel in tasks like structure prediction and variant effect prediction[77]. We also used BioBERT, a domain-specific model pre-trained on large-scale biomedical corpora[78], and ProtBERT, a protein-specific model from the BERT series[79], pre-trained on a vast corpus of protein sequences[80]. These models, each with their unique strengths, significantly enhanced the precision and reliability of our hemolytic peptide classification model.

### Quantum machine learning

Quantum machine learning (QML) is emerging as a transformative tool in healthcare and biomedical research[81]. Recent studies have demonstrated the potential of quantum models to perform sequence classification tasks that are crucial for designing therapeutic proteins[39,82]. After using classical machine learning and LLMs, we employ quantum kernel methods (QKMs)[83] with a focus on the quantum support vector machine (QSVM), to develop classification models. The quantum kernel function used in this study was derived from a feature mapping technique inspired by Zhuang et al.[39] The feature mapping encodes classical descriptors into a high-dimensional quantum Hilbert space[84], where the quantum kernel measures the similarity between pairs of data points based on their mapped representations. The quantum kernels were executed through a noiseless classical simulation with the help of the *PennyLane*[85] Python library, which was specifically employed to compute the quantum kernel matrix. This matrix was subsequently utilized in the scikit-learn function *svm.SVC*, where we experimented with different kernels, including precomputed, linear, radial basis function (RBF), polynomial, and sigmoid. These trials allowed for a comprehensive evaluation of the performance of different kernel types within the context of the quantum-enhanced machine learning framework. Before training, we standardized each descriptor in the training sets using *z*-score normalization, resulting in a mean of 0 and a standard deviation of 1. The independent datasets were then standardized using the same parameters derived from the training sets. Hyperparameter tuning was conducted for all models through stratified 5-fold cross-validation and grid search across various hyperparameter values, allowing us to identify the optimal model for each dataset based on average accuracy across the folds. Different hyperparameters, such as the number of qubits, time (circuit depth), and optimization steps, play a crucial role in fine-tuning the model. The number of qubits determines the dimensionality of the quantum feature space, time influences the complexity of the quantum circuits by controlling the depth of entanglement and gate operations, and optimization steps govern the convergence of the training process. The specific range of hyperparameters used in this study is detailed in Supplementary Table S5. To compare the QSVM with classical SVM, we implemented SVM kernels using scikit-learn's *svm.svc*.

### Hybrid/ensemble model for classification

MERCI identifies motifs composed of specific amino acids and physicochemical properties, which can serve as markers to recognize new sequences involved in biological processes of interest, such as hemolytic activity in this case. In our further investigation, we looked into the approach by combining two approaches to build the final model for enhanced predictive performance. This hybrid strategy merges a motif-based technique utilizing MERCI and models constructed using ML and PLMs. First, we assess the performance of classical motifs and motifs based on physicochemical properties from the *Koolman-Röhm* and *Betts-Russell* classification schemes. Due to the large number of property-based motifs, we limited our selection to motifs with a length of five or more. In the initial phase, each classification approach was tested independently to evaluate its predictive value. Then we experimented with different combinations of these methods, ultimately finding that integrating classical motifs with property-based motifs, especially those derived from the *Koolman-Röhm* classification scheme, significantly improved classifier model performance. We used a weighted scoring method for integrating MERCI with ML models and fine-tuned PLMs. MERCI predictions were weighted as follows: "+0.5" for hemolytic, "−0.5" for non-hemolytic, and "0" for no match, providing confidence scores for each prediction. This weighting, combined with ML and PLMs, significantly boosted area under the curve (AUC) values, an essential metric for model evaluation. This combined approach has been widely applied in numerous scientific investigations, validating its effectiveness in enhancing prediction accuracy.

### Regression models

To predict the $HC_{50}$ of hemolytic peptides, we employed a diverse array of ML models along with PLMs. Our model development process involved leveraging various categories of features extracted by Pfeature. Additionally, we harnessed embeddings derived from fine-tuned PLMs, incorporating them as features for ML model development. This comprehensive approach enabled us to explore the predictive potential of both traditional features and advanced language model embeddings in modeling peptide hemolytic activity.

### Machine learning regressor models

The predictive performance of the computational models developed in this study relies not only on the chosen feature representations but also on the specific regression models employed. To explore the predictive capabilities comprehensively, we utilized a range of popular regressors, including XGBoost Regressor (XGBR), Random Forest Regressor (RFR), Gradient Boosting Regressor (GBR), Extra Trees Regressor (ETR), Decision Tree Regressor (DTR), AdaBoost Regressor (ADBR), Support Vector Regressor (SVR), K-Nearest Neighbors Regressor (KNNR), Linear Regressor (LinR) and Multi-Layer Perceptron Regressor (MLPR).

Each of these regression models operates differently and leverages distinct mathematical algorithms to make predictions. For instance, XGBR is an implementation of gradient-boosted decision trees designed for speed and performance, while RFR utilizes an ensemble of decision trees to improve prediction accuracy and mitigate overfitting. GBR sequentially fits weak learners to the residuals of the previous models, gradually improving prediction accuracy. Similarly, ETR builds an ensemble of randomized decision trees to enhance prediction robustness further. DTR constructs a tree-like model of decisions based on feature inputs, recursively splitting data into subsets to minimize variance. ADBR combines multiple weak learners to create a strong learner iteratively, focusing on instances that previous models misclassified. SVR identifies the optimal hyperplane that maximizes the margin between data points and minimizes prediction error. KNNR predicts the output of a query point based on the majority vote of its k nearest neighbors, while LinR assumes a linear relationship between input features and the target variable. These regression techniques have demonstrated success in predicting various functions and properties of peptides, as well as other biological or chemical entities in previous studies, as evidenced by the cited literature[86–89].

### Performance metrics

The validation of the empirical predictive model is of paramount importance for evaluating its robustness. In the realm of pattern recognition, predicting hemolytic activity is approached as both as a regression and a classification problem. The regression analysis employs four standard statistical parameters. These include the Pearson Correlation Coefficient ($R$), which measures the linear correlation between predicted and actual values, and the Coefficient of Determination ($R^2$), which indicates the fit of the data to the regression model (the closer to 1, the better the fit). Additionally, the Mean Absolute Error (MAE) provides the average of the absolute differences between predicted and actual values, while the Mean Squared Error (MSE) calculates the average of the squared differences, giving more weight to larger errors.

The efficacy of different classification models was gauged using established evaluation metrics, both threshold-dependent and independent. Threshold-dependent metrics, including sensitivity, specificity, accuracy, and the Matthews correlation coefficient (MCC), are influenced by the classification threshold. Conversely, the area under the receiver operating characteristic curve (AUC) is a threshold-independent metric that offers a holistic view of a model's discriminative capacity. These metrics, extensively validated in previous studies, are crucial for reliable performance assessment[21,63].

### Statistics and reproducibility

This study focuses on computational analyses, with no wet lab experiments conducted. We compared amino acid compositions between hemolytic and non-hemolytic peptides and examined N- and C-terminal residues within hemolytic peptides. Statistical analyses were performed using *scipy.stats* for independent t-tests and the *Benjamini-Hochberg* procedure (*statsmodels.stats.multitest*) for false discovery rate correction, ensuring control of type I errors. Replicates correspond to individual peptide observations, enhancing the reproducibility of findings through well-documented statistical methods.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The datasets generated for this study can be accessed on the "HemoPI2" web server at https://webs.iiitd.edu.in/raghava/hemopi2/download.php, publicly available on GitHub https://github.com/raghavagps/HemoPI2, and from Zenodo (https://zenodo.org/records/14676712)[90].

## Code availability

The source code for this study is publicly available on GitHub and can be found at https://github.com/raghavagps/HemoPI2, "HemoPI2" web server at https://webs.iiitd.edu.in/raghava/hemopi2/download.php, and from Zenodo (https://zenodo.org/records/14676712)[90].

Received: 8 August 2024; Accepted: 28 January 2025;
Published online: 04 February 2025

## References

1. Tang, Y., Zhu, W., Chen, K. & Jiang, H. New technologies in computer-aided drug design: toward target identification and new chemical entity discovery. *Drug Discov. Today Technol.* **3**, 307–313 (2006).
2. Wang, L. et al. Therapeutic peptides: current applications and future directions. *Signal Transduct. Target. Ther.* **7**, 1–27 (2022).
3. Rossino, G. et al. Peptides as therapeutic agents: challenges and opportunities in the Green transition era. *Molecules* **28**, 7165 (2023).
4. Muttenthaler, M., King, G. F., Adams, D. J. & Alewood, P. F. Trends in peptide drug discovery. *Nat. Rev. Drug Discov.* **20**, 309–325 (2021).
5. Wang, L. et al. Therapeutic peptides: current applications and future directions. *Signal Transduct. Target. Ther.* **7,1**, 48 (2022).
6. Al Musaimi, O., Al Shaer, D., Albericio, F. & de la Torre, B. G. FDA TIDES (peptides and oligonucleotides) harvest. Pharmaceuticals **16**, 2022 (2023).
7. Mullard, A. 2023 FDA approvals. *Nat. Rev. Drug Discov.* **23**, 88–95 (2024).
8. Al Shaer, D., Al Musaimi, O., Albericio, F. & de la Torre, B. G. 2023 FDA TIDES (peptides and oligonucleotides) harvest. Pharmaceuticals **17**, 243 (2024).
9. Usmani, S. S. et al. THPdb: database of FDA-approved peptide and protein therapeutics. *PLoS ONE* **12**, e0181748 (2017).
10. Jain, S., Gupta, S., Patiyal, S. & Raghava, G. P. S. THPdb2: compilation of FDA approved therapeutic peptides and proteins. *Drug Discov. Today* **29**, 104047 (2024).
11. Caldwell, G. W., Yan, Z., Tang, W., Dasgupta, M. & Hasting, B. ADME optimization and toxicity assessment in early- and late-phase drug discovery. *Curr. Top. Med. Chem.* **9**, 965–980 (2009).
12. Yu, H. & Adedoyin, A. ADME-Tox in drug discovery: integration of experimental and computational technologies. *Drug Discov. Today* **8**, 852–861 (2003).
13. Wu, F. et al. Computational approaches in preclinical studies on drug discovery and development. *Front. Chem.* **8**, 726 (2020).
14. Gupta, S. et al. Peptide toxicity prediction. *Methods Mol. Biol.* **1268**, 143–157 (2015).
15. Gupta, S. et al. In silico approach for predicting toxicity of peptides and proteins. *PLoS ONE* **8**, e73957 (2013).
16. Sharma, N., Naorem, L. D., Jain, S. & Raghava, G. P. S. ToxinPred2: an improved method for predicting toxicity of proteins. *Brief. Bioinform.* **23**, bbac174 (2022).
17. Gautam, A. et al. Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Res.* **42**, D444–D449 (2014).
18. Kumar, V., Kumar, R., Agrawal, P., Patiyal, S. & Raghava, G. P. S. A method for predicting hemolytic potency of chemically modified peptides from its structure. *Front. Pharmacol.* **11**, 54 (2020).

19. Sharma, N. et al. AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Brief. Bioinform.* **22**, bbaa294 (2021).
20. Saha, S. & Raghava, G. P. S. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* **34**, W202–W209 (2006).
21. Rathore, A. S., Choudhury, S., Arora, A., Tijare, P. & Raghava, G. P. S. ToxinPred 3.0: an improved method for predicting the toxicity of peptides. *Comput. Biol. Med.* **179**, 108926 (2024).
22. Ruiz, J., Calderon, J., Rondón-Villarreal, P. & Torres, R. Analysis of structure and hemolytic activity relationships of antimicrobial peptides (AMPs). in *Advances in Intelligent Systems and Computing* 253–258 (Springer International Publishing, 2014).
23. Repsold, L. & Joubert, A. M. Eryptosis: an erythrocyte's suicidal type of cell death. *Biomed. Res. Int.* **2018**, 9405617 (2018).
24. Chaudhary, K. et al. A web server and mobile app for computing hemolytic potency of peptides. *Sci. Rep.* **6**, 22843 (2016).
25. Win, T. S. et al. HemoPred: a web server for predicting the hemolytic activity of peptides. *Future Med. Chem.* **9**, 275–291 (2017).
26. Timmons, P. B. & Hewage, C. M. HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks. *Sci. Rep.* **10**, 10869 (2020).
27. Shin, M. K. et al. De Novo design of AC-P19M, a novel anticancer peptide with apoptotic effects on lung cancer cells and anti-angiogenic activity. *Int. J. Mol. Sci.* **23**, 15594 (2022).
28. Punginelli, D. et al. New bioactive peptides from the Mediterranean seagrass Posidonia oceanica (L.) delile and their impact on antimicrobial activity and apoptosis of human cancer cells. *Int. J. Mol. Sci.* **24**, 5650 (2023).
29. Hasan, M. M. et al. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics* **36**, 3350–3356 (2020).
30. Wei, L., Zhou, C., Su, R. & Zou, Q. PEPred-Suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. *Bioinformatics* **35**, 4272–4280 (2019).
31. Plisson, F., Ramírez-Sánchez, O. & Martínez-Hernández, C. Machine learning-guided discovery and design of non-hemolytic peptides. *Sci. Rep.* **10**, 16581 (2020).
32. Yaseen, A., Gull, S., Akhtar, N., Amin, I. & Minhas, F. HemoNet: predicting hemolytic activity of peptides with integrated feature learning. *J. Bioinform. Comput. Biol.* **19**, 2150021 (2021).
33. Capecchi, A. et al. Machine learning designs non-hemolytic antimicrobial peptides. *Chem. Sci.* **12**, 9221–9232 (2021).
34. Ansari, M. & White, A. D. Serverless prediction of peptide properties with recurrent neural networks. *J. Chem. Inf. Model.* **63**, 2546–2553 (2023).
35. Salem, M., Keshavarzi Arshadi, A. & Yuan, J. S. AMPDeep: hemolytic activity prediction of antimicrobial peptides using transfer learning. *BMC Bioinformatics* **23**, 389 (2022).
36. Perveen, G., Alturise, F., Alkhalifah, T. & Daanial Khan, Y. Hemolytic-Pred: a machine learning-based predictor for hemolytic proteins using position and composition-based features. *Digit. Health* **9**, 20552076231180740 (2023).
37. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. *Methods Mol. Biol.* **406**, 89–112 (2007).
38. Castillo-Mendieta, K. et al. Peptide hemolytic activity analysis using visual data mining of similarity-based complex networks. *NPJ Syst. Biol. Appl.* **10**, 115 (2024).
39. Zhuang, S. et al. Non-hemolytic peptide classification using a quantum support vector machine. https://doi.org/10.48550/ARXIV.2402.03847 (2024).
40. Guntuboina, C., Das, A., Mollaei, P., Kim, S. & Barati Farimani, A. PeptideBERT: a language model based on transformers for peptide property prediction. *J. Phys. Chem. Lett.* **14**, 10427–10434 (2023).
41. Koolman, J. & Röhm, K. Color Atlas of Biochemistry. In *Georg Thieme Verlag eBooks*. https://doi.org/10.1055/b-005-148878 (2004).
42. Betts, M. J. & Russell, R. B. Amino acid properties and consequences of substitutions. in *Bioinformatics for Geneticists* 289–316 (John Wiley & Sons, Ltd, 2003).
43. Rodríguez-Pérez, R. & Bajorath, J. Feature importance correlation from machine learning indicates functional relationships between proteins and similar compound binding characteristics. *Sci. Rep.* **11**, 14245 (2021).
44. Karollus, A. et al. Species-aware DNA language models capture regulatory elements and their evolution. *Genome Biol.* **25**, 83 (2024).
45. Li, H.-L., Pang, Y.-H. & Liu, B. BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models. *Nucleic Acids Res.* **49**, e129 (2021).
46. Liu, J. et al. Large language models in bioinformatics: applications and perspectives. https://doi.org/10.48550/ARXIV.2401.04155 (2024).
47. Feng, C. et al. Log-transformation and its implications for data analysis. *Shanghai Arch. Psychiatry* **26**, 105–109 (2014).
48. Feng, C., Wang, H., Lu, N. & Tu, X. M. Log transformation: application and interpretation in biomedical research. *Stat. Med.* **32**, 230–239 (2013).
49. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
50. Dhall, A., Patiyal, S., Sharma, N., Usmani, S. S. & Raghava, G. P. S. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Brief. Bioinform.* **22**, 936–945 (2021).
51. Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D. & Levy, S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* **21**, 631–643 (2005).
52. Aftabuddin, M. & Kundu, S. Hydrophobic, hydrophilic, and charged amino acid networks within protein. *Biophys. J.* **93**, 225–231 (2007).
53. Oda, T. et al. Temperature- and pH-dependent cytotoxic effect of the hemolytic lectin CEL-III from the marine invertebrate Cucumaria echinata on various cell lines. *J. Biochem.* **121**, 560–567 (1997).
54. Hatakeyama, T., Kohzaki, H., Nagatomo, H. & Yamasaki, N. Purification and characterization of four Ca(2+)-dependent lectins from the marine invertebrate, Cucumaria echinata. *J. Biochem.* **116**, 209–214 (1994).
55. Ehret-Sabatier, L. et al. Characterization of novel cysteine-rich antimicrobial peptides from scorpion blood. *J. Biol. Chem.* **271**, 29537–29544 (1996).
56. Silva, P. I. Jr, Daffre, S. & Bulet, P. Isolation and characterization of gomesin, an 18-residue cysteine-rich defense peptide from the spider Acanthoscurria gomesiana hemocytes with sequence similarities to horseshoe crab antimicrobial peptides of the tachyplesin family. *J. Biol. Chem.* **275**, 33464–33470 (2000).
57. Ramamoorthy, A. et al. Deletion of all cysteines in tachyplesin I abolishes hemolytic activity and retains antimicrobial activity and lipopolysaccharide selective binding. *Biochemistry* **45**, 6529–6540 (2006).
58. Durairaj, J., de Ridder, D. & van Dijk, A. D. J. Beyond sequence: structure-based machine learning. *Comput. Struct. Biotechnol. J.* **21**, 630–643 (2023).
59. Rádai, Z., Kiss, J. & Nagy, N. A. Taxonomic bias in AMP prediction of invertebrate peptides. *Sci. Rep.* **11**, 17924 (2021).
60. Pirtskhalava, M. et al. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* **49**, D288–D297 (2021).
61. Liang, L. et al. Bioactivity-explorer: a web application for interactive visualization and exploration of bioactivity data. *J. Cheminform.* **11**, 47 (2019).
62. Abdi, G. et al. Unveiling the dynamic role of bioinformatics in automation for efficient and accurate data processing and

interpretation. in *Advances in Bioinformatics* 279–319 (Springer Nature Singapore, 2024).
63. Yan, J., Zhang, B., Zhou, M., Campbell-Valois, F.-X. & Siu, S. W. I. A deep learning method for predicting the minimum inhibitory concentration of antimicrobial peptides against Escherichia coli using Multi-Branch-CNN and Attention. *mSystems* **8**, e0034523 (2023).
64. Pande, A. et al. Pfeature: a tool for computing wide range of protein features and building prediction models. *J. Comput. Biol.* **30**, 204–222 (2023).
65. Tran, C., Khadkikar, S. & Porollo, A. Survey of protein sequence embedding models. *Int. J. Mol. Sci.* **24**, 3775 (2023).
66. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv* https://doi.org/10.1101/622803 (2019).
67. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA***118**, e2016239118 (2021).
68. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).
69. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
70. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
71. Vacic, V., Iakoucheva, L. M. & Radivojac, P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* **22**, 1536–1537 (2006).
72. Vens, C., Rosso, M.-N. & Danchin, E. G. J. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* **27**, 1231–1238 (2011).
73. Mooney, C. et al. In silico protein motif discovery and structural analysis. *Methods Mol. Biol.* **760**, 341–353 (2011).
74. Corbi-Verge, C. & Kim, P. M. Motif mediated protein-protein interactions as drug targets. *Cell Commun. Signal.* **14**, 8 (2016).
75. Corbi-Verge, C., Garton, M., Nim, S. & Kim, P. M. Strategies to develop inhibitors of motif-mediated protein-protein interactions as drug leads. *Annu. Rev. Pharmacol. Toxicol.* **57**, 39–60 (2017).
76. Janse, R. J. et al. Conducting correlation analysis: important limitations and pitfalls. *Clin. Kidney J.* **14**, 2332–2337 (2021).
77. Lin, Z. et al. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv* https://doi.org/10.1101/2022.07.20.500902 (2022).
78. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
79. Elnaggar, A. et al. ProtTrans: towards cracking the language of life's code through self-supervised learning. *bioRxiv* https://doi.org/10.1101/2020.07.12.199554 (2020).
80. Singer, J. M. et al. Large-scale design and refinement of stable proteins using sequence-only models. *bioRxiv* https://doi.org/10.1101/2021.03.12.435185 (2021).
81. Durant, T. J. S. et al. A primer for quantum computing and its applications to healthcare and biomedical research. *J. Am. Med. Inform. Assoc.* **31**, 1774–1784 (2024).
82. London, C. et al. Peptide binding classification on quantum computers. *arXiv [quant-ph]* https://doi.org/10.48550/ARXIV.2311.15696 (2023).
83. Jerbi, S. et al. Quantum machine learning beyond kernel methods. *Nat. Commun.* **14**, 517 (2023).
84. Schuld, M. & Killoran, N. Quantum machine learning in feature Hilbert spaces. *Phys. Rev. Lett.* **122**, 040504 (2019).
85. Bergholm, V. et al. PennyLane: automatic differentiation of hybrid quantum-classical computations. *arXiv [quant-ph]* https://doi.org/10.48550/ARXIV.1811.04968 (2018).
86. Mishra, G., Sehgal, D. & Valadi, J. K. Quantitative structure activity relationship study of the anti-hepatitis peptides employing random forests and extra-trees regressors. *Bioinformation* **13**, 60–62 (2017).
87. Qiu, J., Sheffler, W., Baker, D. & Noble, W. S. Ranking predicted protein structures with support vector regression. *Proteins* **71**, 1175–1182 (2008).
88. Chen, C. et al. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. *Comput. Biol. Med.* **123**, 103899 (2020).
89. Giard, J., Ambroise, J., Gala, J.-L. & Macq, B. Regression applied to protein binding site prediction and comparison with classification. *BMC Bioinforma.* **10**, 276 (2009).
90. Rathore, A. S., Kumar, N., Choudhury, S., Kumar Mehta, N. & Raghava, G. Prediction of hemolytic peptides and their hemolytic concentration. *Zenodo*. https://doi.org/10.5281/ZENODO.14676712 (2025).

## Acknowledgements

The authors express their gratitude to the University Grants Commission (UGC), Council of Scientific and Industrial Research (CSIR), Department of Science and Technology (DST-INSPIRE), and Department of Biotechnology (DBT) for their generous fellowships and financial support. They also thank the Department of Computational Biology, IIITD, New Delhi, for the excellent infrastructure and facilities. The authors would like to acknowledge the Department of Biotechnology (DBT) for the infrastructure grant awarded to the institute. The current work has been supported by the Department of Biotechnology (DBT) grant BT/PR40158/BTIS/137/24/2021. Furthermore, they would like to acknowledge BioRender.com for creating the figures utilized in this work.

## Author contributions

Anand Singh Rathore (A.S.R.) collected and processed the datasets. A.S.R. and Nishant Kumar (N.K.) implemented the algorithms and developed the prediction models. Gajendra P. S. Raghava (G.P.S.R.) analyzed the results. A.S.R. and S.C. (Shubham Choudhury) created the back end of the web server, and S.C. and N.K. created the front-end user interface. Naman Kumar Mehta (N.K.M.) did the benchmarking. A.S.R. and G.P.S.R. penned the manuscript. G.P.S.R. conceived and coordinated the project. All authors have read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Inclusion and ethics

The authors have committed to upholding the principles of research inclusion & ethics as advocated by the Nature Portfolio.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-07615-w.

**Correspondence** and requests for materials should be addressed to Gajendra P. S. Raghava.

**Peer review information** *Communications Biology* thanks Fabien Plisson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Jasmine Pan. A peer review file is available.

**Reprints and permissions information** is available at http://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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025