Received: 4 January 2024 | Revised: 29 April 2024 | Accepted: 13 May 2024

DOI: 10.1002/pmic.202400004

Proteomics
Proteomics and Systems Biology

RESEARCH ARTICLE

Prediction of peptide hormones using an ensemble of machine learning and similarity-based methods

Dashleen Kaur | **Akanksha Arora** | **Palani Vigneshwar** | **Gajendra P. S. Raghava**

Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India

Correspondence
Gajendra P. S. Raghava, Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, Okhla Industrial Estate, Phase III, New Delhi 110020, India.
Email: raghava@iiitd.ac.in

Funding information
Department of Biotechnology (DBT), Grant/Award Number: BT/PR40158/BTIS/137/24/2021

Abstract

Peptide hormones serve as genome-encoded signal transduction molecules that play essential roles in multicellular organisms, and their dysregulation can lead to various health problems. In this study, we propose a method for predicting hormonal peptides with high accuracy. The dataset used for training, testing, and evaluating our models consisted of 1174 hormonal and 1174 non-hormonal peptide sequences. Initially, we developed similarity-based methods utilizing BLAST and MERCI software. Although these similarity-based methods provided a high probability of correct prediction, they had limitations, such as no hits or prediction of limited sequences. To overcome these limitations, we further developed machine and deep learning-based models. Our logistic regression-based model achieved a maximum AUROC of 0.93 with an accuracy of 86% on an independent/validation dataset. To harness the power of similarity-based and machine learning-based models, we developed an ensemble method that achieved an AUROC of 0.96 with an accuracy of 89.79% and a Matthews correlation coefficient (MCC) of 0.8 on the validation set. To facilitate researchers in predicting and designing hormone peptides, we developed a web-based server called HOPPred. This server offers a unique feature that allows the identification of hormone-associated motifs within hormone peptides. The server can be accessed at: https://webs.iiitd.edu.in/raghava/hoppred/.

KEYWORDS
alignment-based method, ensemble method, machine learning techniques, peptide hormones

1 | INTRODUCTION

The peptide hormones are a varied group of genome-encoded regulatory molecules with a specialized and important purpose of transferring specific information between cells and organs. This sort of molecular communication emerged and developed into a sophisticated system for regulating growth, development, and homeostasis in the early stages of evolution [1]. A wide range of reasons, including metabolite deposition in cells, surgical removal of glands, or destruction of glands, can cause inadequate production of hormones. A frequent cause of hormone reduction is autoimmunity, and less frequently, it is also caused by genetic anomalies. Altered levels of hormones can disrupt the delicate equilibrium of the body and lead to endocrine disorders like endocrine neoplasia and diabetes [2]. It is thus indispensable to focus on the therapeutics of endocrine disorders associated with hormonal imbalances.

Abbreviations: AAC, Amino Acid Composition; CD-HIT, Cluster Database at High Identity with Tolerance; CSS, Cascading Style Sheets; CTC, Conjoint Triad Descriptors; CeTD, Composition enhanced Transition and Distribution; DL, Deep Learning; DT, Decision Tree; DDR, Distance Distribution; DPC, Dipeptide Composition; GNB, Gaussian Naïve Bayes; HTML, HyperText Markup Language; KNN, k-Nearest Neighbour; LR, Logistic Regression; ML, Machine LearningMachine Learning; PCP, Physicochemical Properties; PHP, Hypertext Preprocessor; QSO, Quasi-sequence Order; RF, Random Forest; SEP, Shannon Entropy; TabNet, Tabular Neural Network; TPC, Tripeptide Composition; TextCNN, Text Convolutional Neural Network; XGB, eXtreme Gradient Boosting.

Dashleen Kaur and Akanksha Arora contributed equally to this work.

Keypoints

- An ensemble method for predicting hormonal peptides with high precision.
- BLAST-based similarity for searching hormonal peptides.
- MERCI software for searching hormone-associated motifs in a peptide.
- Prediction models using machine and deep learning techniques.
- A standalone software package and a webserver.

Significance Statement

Problem: Peptide hormones have demonstrated promising results in medical therapies since the very beginning. The demand for efficient in-silico techniques arises from the time-consuming and resource-intensive nature of identifying and screening potential peptide hormones in wet lab settings.
What is already known: Several methods have been proposed in the past for predicting and designing therapeutic peptides. As far as we know, no method has been created particularly to predict peptide hormones.
What this paper adds: This study introduces a unique ensemble approach employing machine learning and deep learning methods for predicting hormonal peptides. The development of our method, HOPPred, not only provides a reliable prediction tool but also offers a unique feature for identifying hormone-associated motifs within peptides. We hope this tool will aid researchers in accurately predicting hormones and open avenues for unraveling intricate signaling pathways.

Peptide hormones, like all other peptides, are effective targeted signaling molecules that attach to specific cell surface receptors and cause intracellular effects. They are a great place to start while developing novel treatments because of their appealing pharmacological profile, target specificity, and inherent qualities. It was, in fact, the foundational studies into natural hormones like insulin and gonadotropin-releasing hormone (GnRH), which marked the beginning of research into therapeutic peptides [3]. Their exceptional tolerability and efficacy have been demonstrated to be a direct result of their specificity. This feature may also distinguish between peptides and conventional small molecules. The depiction of the mechanism of peptide hormones is shown in Figure 1.

Many peptide medications act as "replacement therapies," restoring or supplanting peptide hormones when endogenous levels are insufficient or absent. They are intrinsic signaling molecules for a variety of physiological functions, which opens the possibility for therapeutic intervention emulating natural routes [4]. One such example is insulin which has helped thousands of people with diabetes since its introduction as the first peptide medication for commercial use in 1923 [3]. While replicating peptides generated from nature was formerly the discovery and development trend in the therapeutic peptide sector, it has recently changed to the rational design of peptides with desirable physiological and biochemical properties. Peptide-based natural hormone analogs have been developed as drug candidates. Given their enormous therapeutic potential, it can be anticipated that therapeutic hormone peptides will continue to draw research attention and see long-term success.

Several methods have been proposed in the past for predicting and designing therapeutic peptides. TPpred-ATMV improves the prediction of therapeutic peptides, THPep predicts the tumor-homing peptide (THP), AntiCP 2.0 for the prediction of anti-cancer peptides, and PrMFTP has been developed to predict multi-functional therapeutic peptides [5–8]. As far as we know, no method has been created particularly to predict peptide hormones. In this study, we have proposed an approach HOPPred, to classify the peptide hormone and non-hormone peptide sequences.

2 | METHODS

2.1 | Dataset compilation and preprocessing

We took 5729 peptide hormone sequences for both plants and animals from the Hmrbase2 database [9]. The Hmrbase2 database contains comprehensive and updated information for peptide and non-peptide hormones as well as their receptors for about 562 organisms. These are mature hormone sequences devoid of the signal and precursor areas. All duplicate peptides were removed from the data to eliminate redundancy from the dataset. Additionally, it was observed that the peptides ranged from the length of 11 to 41 peptides. Redundant peptides were removed using CD-HIT with a cut-off of 60% to make sure there was no peptide in the dataset that had more than 60% similarity with any other peptide [10]. CD-HIT is a standard practice used to get rid of the similarity within dataset. Hence, we applied CD-HIT at various cut-offs—90%, 80%, 70%, and so on. It was observed that 60% similarity was retaining the most peptides and very few were the outliers that were more than 60% similar. Ultimately, we obtained 1174 peptide hormones, which we refer to in this work as the positive dataset. For the non-hormone peptide dataset, that is, the negative dataset, we used the PeptideAtlas database. PeptideAtlas is a freely available database containing peptide sequences from a number of experiments for a variety of organisms [11]. About 1174 non-hormone peptides were selected randomly from PeptideAtlas, and the amino acid composition (AAC) for these peptides was calculated to assess if the composition matches the standard AAC of proteins/peptides present in nature to avoid the random non-biologically significant sequences.

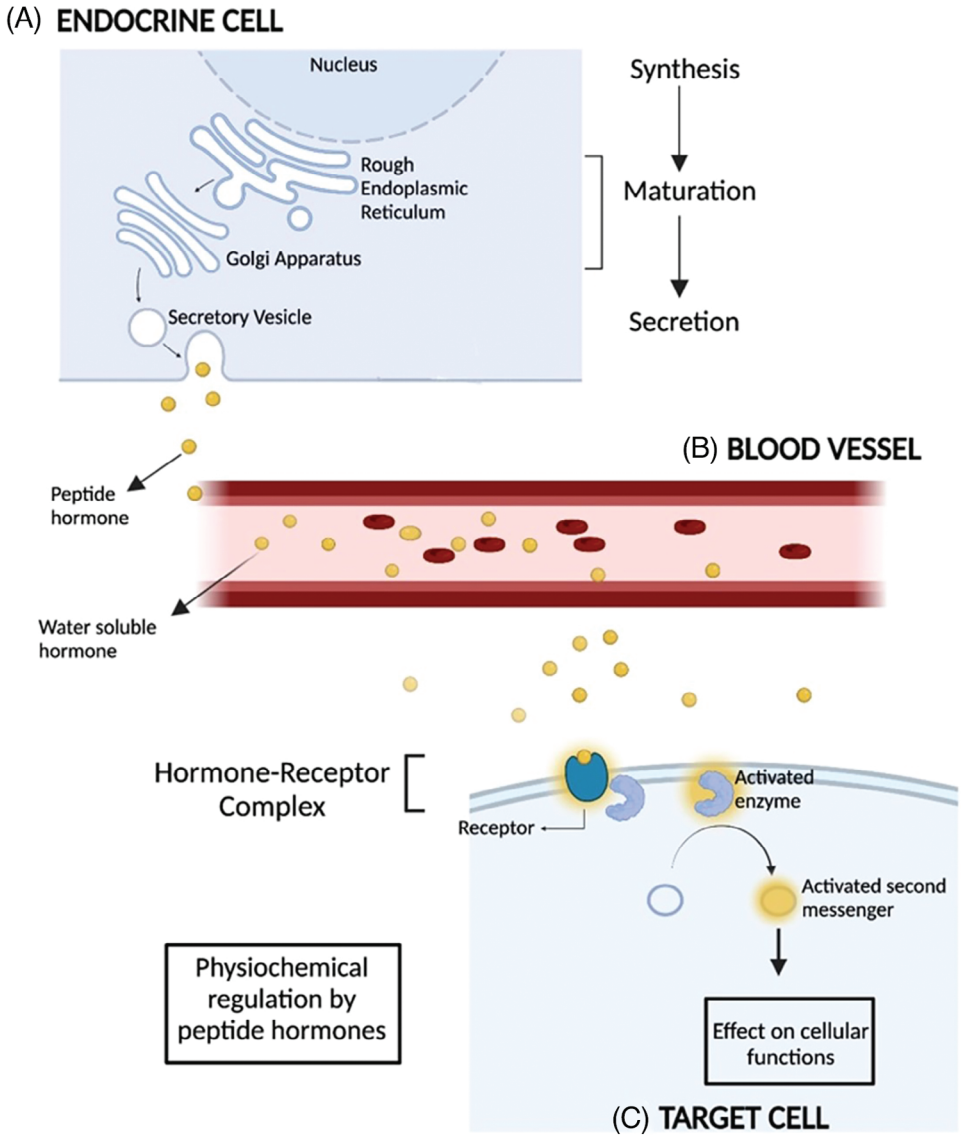

FIGURE 1 Mechanism of action of peptide hormones.

The complete architecture and methodology followed in this study is given in Figure 2.

2.2 | Feature generation and selection

Structural and functional annotations of peptide sequences were performed using Pfeature. It computes the descriptors of a sequence and gives a vector of 9149 features [12]. Using Pfeature, we calculated all 9149 features, which included 15 categories of characteristics and descriptors, including AACs, tripeptide compositions, dipeptide compositions, etc. It is important to determine the few significant features out of the plethora of a vast number of features for several reasons [13]. Feature selection techniques allow us to select the most relevant feature, which allows us to reduce noise in prediction. In addition, feature selection can help reduce the overfitting of the model, enhance interpretability, and improve model performance [14]. In order to perform feature selection, we deployed different feature selection methods like filter methods, wrapper methods, and embedded methods. In this study, Recursive Feature Elimination (RFE)—a wrapper method, performed better than other feature selection techniques [15]. Hence, we utilized the Python programming language's Scikit-learn package to implement RFE, with Logistic Regression serving as the estimator [16]. The features were chosen from the standardized data obtained using the StandardScaler from the Scikit-learn package [16]. Until a certain number of features have been attained, this feature selection approach recursively deletes the weakest features from the set. Unlike the other feature selection methods, which focus on individual properties of features, RFE focuses on features that affect the performance of the model [17]. We performed RFE on different numbers of features—10, 20, 30, 40, and so on. It was observed that the top 50 most relevant features performed best in the machine learning models; detailed description of these features is shown in Table S1 [18].

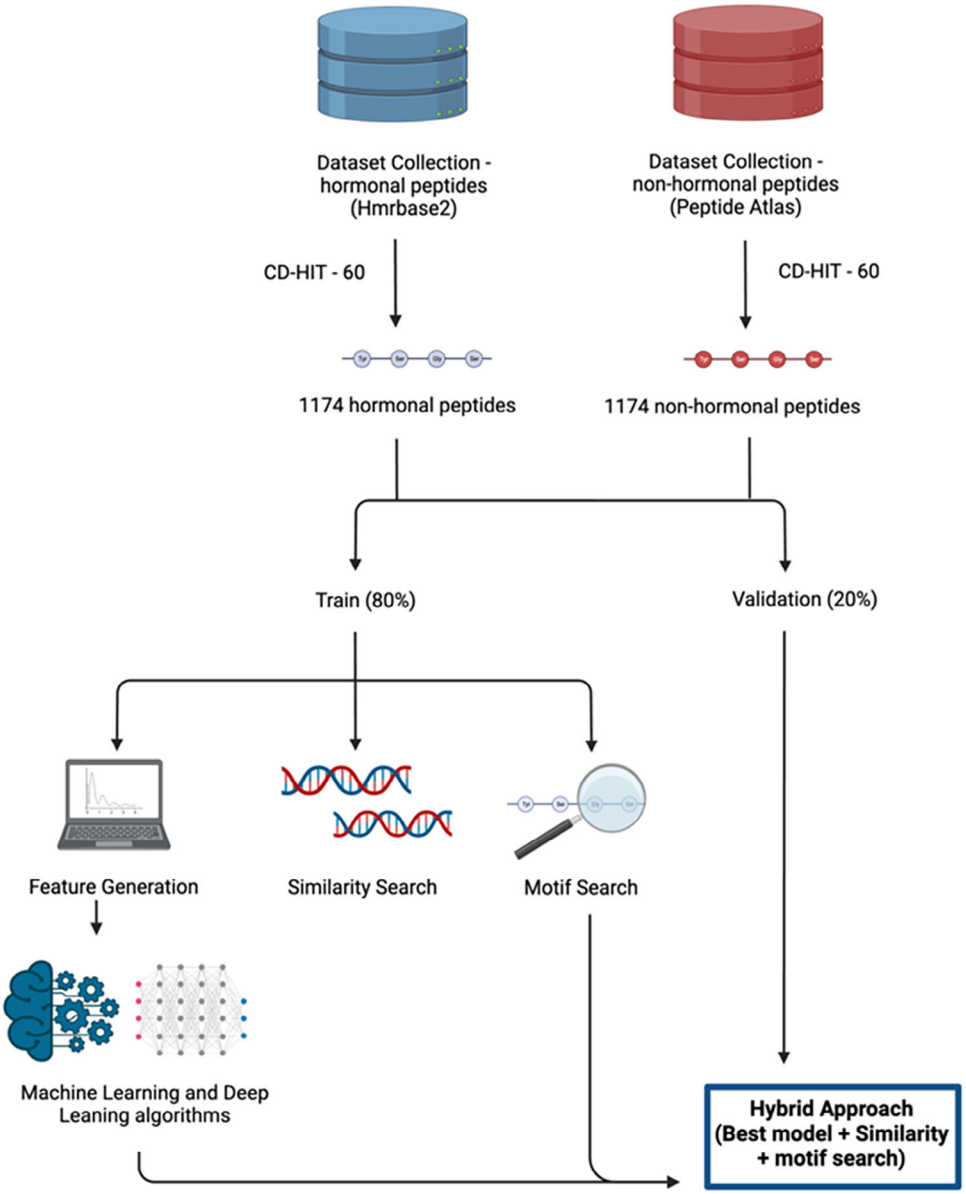

FIGURE 2 Architecture and methodology followed in HOPPred—a hybrid approach to predict and design hormonal peptides.

2.3 | Machine learning techniques

We classified peptide hormones and non-hormonal peptides using a variety of machine-learning algorithms like Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), Logistic Regression (LR), and XGBoost (XGB) classifiers, using a python-based toolkit called Scikit-learn. DT classifier is a non-parametric supervised machine learning model that identifies the output by learning decision rules from input, ensemble-based RF classifier trains several decision trees to prodigy a single tree, LR classifier calculates the likelihood of an event using a logistic function, KNN classifier bases its forecast on the highest number of votes cast in favor of the class that is closest to the nearest neighboring data point, GNB classifier is a probabilistic classifier based on Bayes' theorem, and XGB classifier is a distributed gradient-boosted decision tree machine learning library that avails parallel tree boosting [19–24].

2.4 | Deep learning techniques

2.4.1 | Tabnet

We applied TabNet on a set of 9149 biological features, which were extracted using Pfeature. TabNet is a deep learning model applied on tabular data. At each stage of the model, it chooses which features to use by providing sequential attention. This facilitates interpretability and improved learning as the most useful features are used. Tabnet

utilizes a single deep learning architecture for both feature selection and reasoning, known as soft feature selection [25].

2.4.2 | TextCNN

We also used the TextCNN method, where we considered sequences as text and classified them using convolution neural networks. The padding of sequences was done to make the lengths the same for all sequences. Then, we took the word length of three alphabets and divided the sequences into sets of words. We then applied TextCNN to the matrix of 8000 features (20*20*20 = 8000), where 20 is the possible amino acids that can be present in a single place [26].

2.5 | Five-fold cross-validation

We implemented a five-fold cross-validation procedure to ensure that our derived models are not biased and do not suffer from overfitting. This is a standard technique commonly used to evaluate the performance of newly developed methods. In this technique, the dataset is randomly divided into five sets; four sets are used for training, and the remaining set is for testing. This process is repeated five times in such a way that each set is used once for testing. Finally average performance is computed on all sets. As models are trained and tested on different sets, the chances of overfitting are negligible. It ensures that the model's performance is consistent across different sets of data. Five-fold cross-validation strikes a balance between computational efficiency and reliable estimation of the model's performance. It provides a good compromise between the number of folds and computational resources required for training multiple models [27, 28]. The entire dataset was divided into training data and validation data, following an 80:20 ratio. The 80:20 ratio for training and validation datasets is a common choice in machine learning. It strikes a balance between having enough data for training to learn complex patterns and having enough data for validation to reliably estimate the model's performance [29]. We employed the 5-fold cross-validation technique on the training data, where the training data was further divided into five folds. In each iteration, four folds were used for training the model, while the remaining fold served as the test set for internal validation. This process was repeated five times, allowing each fold to serve as the test set once. The 20% of data that was kept aside during the initial split was used for the external validation of the model. This held-out dataset provided an independent evaluation of the model's performance. This approach is widely accepted and has been utilized in various studies for robust validation of data [29–31].

2.6 | Evaluation of parameters

The evaluation of the prediction model is one of the most important processes. Several parameters, which may be threshold-dependent or threshold-independent, can be used to evaluate the prediction models. Threshold-dependent parameters are sensitivity, which measures how well the model can predict the hormones (Equation (1)); specificity for the correct prediction of non-hormones (Equation (2)), accuracy which shows the proportion of hormones and non-hormones that were successfully predicted (Equation (3)); and Matthew's correlation coefficient (MCC) shows the relationship between predicted and observed values (Equation (4)). Sensitivity measures the percent of correctly predicted hormonal peptides, whereas specificity estimates correctly predicted non-hormonal peptides. Accuracy measures the overall correctness of the model's predictions, calculated as the ratio of correctly predicted peptides to the total number of peptides. MCC is the most robust parameter for evaluating a method; it imposes a penalty on false positive and false negative predictions. The threshold-independent parameter is the area under the receiver operating characteristic curve (AUC) between the sensitivity and 1-specificity. It provides an aggregate measure of the model's performance across all possible classification thresholds. It is insensitive to class imbalance and provides a single scalar value that summarizes the model's discriminatory power. These evaluation metrics have been commonly used in various bioinformatics studies [32, 33].

$$\text{Sensitivity} = \frac{T_P}{T_P + F_N} \times 100 \tag{1}$$

$$\text{Specificity} = \frac{T_N}{T_N + F_P} \times 100 \tag{2}$$

$$\text{Accuracy} = \frac{T_P + T_N}{T_N + T_P + F_N + F_P} \times 100 \tag{3}$$

$$\text{MCC} = \frac{(T_N \times T_P) - (F_N \times F_P)}{\sqrt{(F_P + T_P)\,(F_N + T_P)\,(F_P + T_N)\,(F_N + T_N)}} \tag{4}$$

where T_P, T_N, F_P, and F_N stand for true positive, true negative, false positive, and false negative, respectively.

2.7 | Similarity search

The Basic Local Alignment Search Tool (BLAST) is a commonly used tool for annotating protein and nucleotide sequences [34]. In this study, we utilized BLAST shortp, which is specifically designed for peptide sequences, to identify hormonal sequences based on sequence similarity. To ensure the reliability and accuracy of our methodology, we followed established practices used by other researchers [31, 35]. Specifically, we created the BLAST database using all the sequences present in the training dataset. During the evaluation of the training set sequences, we considered the first hit after excluding self-hits. For the validation set sequences, we used the first hit

to determine their classification as hormonal or non-hormonal. By employing BLAST and adopting these annotation procedures, we aimed to leverage sequence similarity to identify and annotate peptide hormones accurately. We performed BLAST for a set of e-values, including 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, and 10^{2}, using the default parameters. The results for BLAST are shown in Table 3.

2.8 | Motif analysis

The Motif-EmeRging and Classes-Identification (MERCI) tool is utilized in this study to identify motifs within the set of peptide sequences [27]. Motif analysis provides valuable information about recurring patterns observed in peptide hormone sequences. The input for the MERCI tool is a FASTA file containing the peptide sequences, and the output consists of a collection of recurring sequence patterns known as motifs, along with their respective positions of occurrence within the sequences. In this particular study, we focused on identifying motifs that exclusively appear in hormonal and non-hormonal sequences. By analyzing these motifs, we can gain insights into unique sequence patterns associated with each class of peptides. By employing the MERCI tool and examining the identified motifs, we aim to uncover important information about the distinctive sequence characteristics of hormonal and non-hormonal peptides.

2.9 | Ensemble approach

We combined three different techniques to calculate the score: (i) BLAST search, (ii) MERCI motifs, and (iii) Machine learning models. First, BLAST was used to categorize the protein sequence with an E-value of 10^{-1}. We gave the positive predictions (peptide hormones) a score of "+0.5," the negative predictions (non-hormone peptides) a score of "–0.5," and no hits a score of "0." Second, MERCI was used to categorize the same peptide sequence. When the hormonal motifs were discovered, we gave them a score of "+0.5," and when they were not, we gave them a score of "0." Third, ML techniques were used to classify hormonal and non-hormonal peptides, where we obtain a set of prediction probabilities for each class that varies between 0 to 1, which are obtained using the predict_proba() function in Python. We added the scores obtained from MERCI and BLAST to the prediction probabilities obtained by ML model prediction. The scoring system is explained in detail in Equations (5) and (6). In the hybrid technique, the total score was computed by combining the results of all three approaches. The peptide sequence is classified as hormonal or non-hormonal on the overall hybrid score (M'') as defined in Equation (6). The overall hybrid score varies from the range of –1 to 2.

$$M' = \begin{cases} M + 0.5 \; If \; BLAST \; hit \; is \; against \; hormonal \; sequence \\ M - 0.5 \; If \; BLAST \; hit \; is \; against \; non - hormonal \; sequence \\ M \; If \; no \; BLAST \; hit \; is \; found \end{cases} \quad (5)$$

Here, M = prediction probability score obtained from ML-based approach and M' = score obtained after adding scores from BLAST-based approach.

$$M'' = \begin{cases} M' + 0.5 \; If \; hormonal \; motif \; present \\ M' - 0.5 \; If \; non - hormonal \; motifpresent \\ M' \; If \; no \; motif \; is \; found \end{cases} \quad (6)$$

Here, M'' = final score obtained from the ML-based approach, BLAST-based approach, and motif-based approach ranging from –1 to 2.

3 | RESULTS

3.1 | Analysis of amino acid composition

In our study, we conducted an AAC analysis to investigate the differences between hormonal and non-hormonal peptides. Figure 3 illustrates the AAC of peptide hormones and non-hormonal peptides, along with the corresponding *p*-values for each amino acid. The compositional differences between the two classes are evident from the figure. To statistically compare the AAC values of the two groups, namely hormonal peptides and non-hormonal peptides, we performed a two-sided Mann–Whitney U test. This test allows us to determine if there are significant differences between the groups. Our analysis revealed that the composition of certain amino acids was significantly higher in hormonal peptide sequences, including phenylalanine and proline (p-values < 0.05). On the other hand, the composition of amino acids such as glutamic acid, isoleucine, lysine, leucine, methionine, glutamine, threonine, and valine was found to be significantly lower in hormonal peptide sequences compared to non-hormonal peptide sequences (p-values < 0.05). These findings indicate that the relative abundance of specific amino acids differs between hormonal and non-hormonal peptides, potentially contributing to their functional distinctions.

3.2 | Machine learning-based models

First, we computed a variety of features using the Pfeature software and removed the unnecessary features using the RFE feature selection methodology, as mentioned in the Materials and Methods section. For comparison purposes, we developed multiple models based on the top 30 and top 50 features in the dataset and assessed them on the training and validation datasets. All the machine learning methods were used on the datasets with AAC, top 30, and top 50 features to create peptide hormone prediction models. Hyperparameter tuning was used to optimize parameters using a grid-search algorithm on the sci-kit Learn [36]. Grid search is used to find the optimal hyperparameters for the machine learning methods from a set of hyperparameter values to explore [37]. Grid search is then performed on the cross-validation fold of the model. Cross-validation involves splitting the training data into multiple subsets (folds), training the

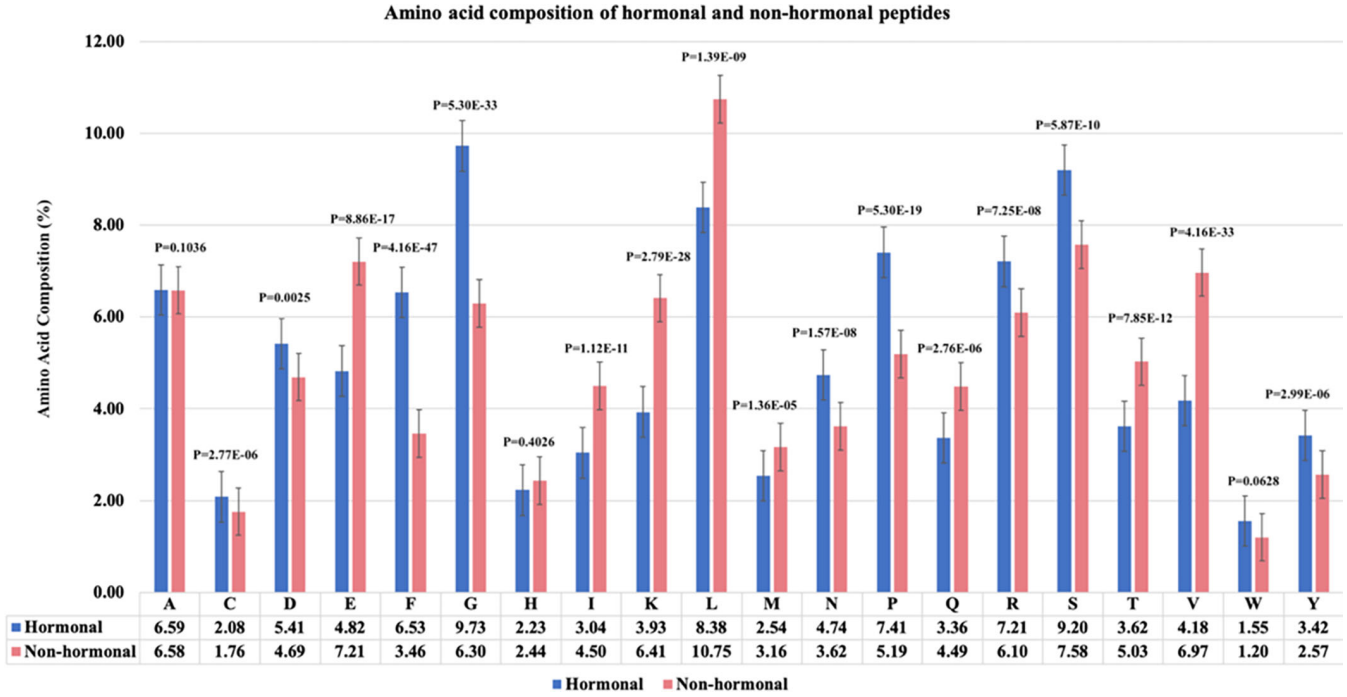

FIGURE 3 Amino acid composition analysis for peptide hormones and non-hormone peptides.

model on some folds, and evaluating it on the remaining fold. This process is repeated multiple times, with different folds used for training and evaluation each time. For each combination of hyperparameters, grid search calculates the performance metric (e.g., accuracy) based on the average performance across all cross-validation folds. Once all combinations have been evaluated, the grid search selects the combination of hyperparameters that results in the best performance metric. This combination is typically chosen based on maximizing accuracy or minimizing error on a validation set. We specified all possible values for each parameter for the grid search algorithm. The best model—Logistic regression (C = 1.623776739188721, class_weight = 'balanced', solver = 'liblinear') was employed in both the web server and standalone software. The hyperparameters for other ML models are given in Supplementary Table S3. This model was chosen due to its high performance and robustness. It gave a standard deviation of 0.01494, a standard error of 0.00668, and an average AUC of 0.94 for five folds in the training dataset. It also performed well on the test (independent validation) dataset with an AUC of 0.93.

The set of top 50 features that performed best was chosen as the final features in our prediction model, as evidenced by the results—LR (training: 0.94 AUROC and validation: 0.93 AUROC) is the best model, followed by RF (training: 0.91 AUROC and validation: 0.90 AUROC). The performance of various models employed on AAC, top 30, and top 50 features are demonstrated in Figure 4, and detailed information is given in Table 1. The parameters for best performing (50 features) models are given in Table S2.

3.3 | Performance of deep learning techniques

3.3.1 | TextCNN

TextCNN is a deep learning method to perform text classification tasks [26]. We divided the sequences into the words of length = 3 and used them to build our deep learning model. This model was able to achieve an AUROC of 0.98 on the training dataset. However, it was not able to perform well on the testing dataset, giving an AUROC of 0.90. The MCC was also seen to decrease from 0.88 in the training dataset to 0.67 in the testing dataset.

3.3.2 | TabNet

TabNet is a deep learning method that uses tabular data as input and is trained using gradient-descent-based optimization [25]. The 9149 features generated from the Pfeature were passed as features through the TabNet model to get the train AUROC as 0.81 and the validation AUROC as 0.75 with an MCC of 0.61 for the training dataset and 0.57 for the validation dataset. The performances of the deep learning models TextCNN and TabNet are summarized in Table 2.

3.4 | Motif performance

The motifs exclusively seen in peptide hormones have been found using the MERCI software. The list of motifs found in hormone peptides is

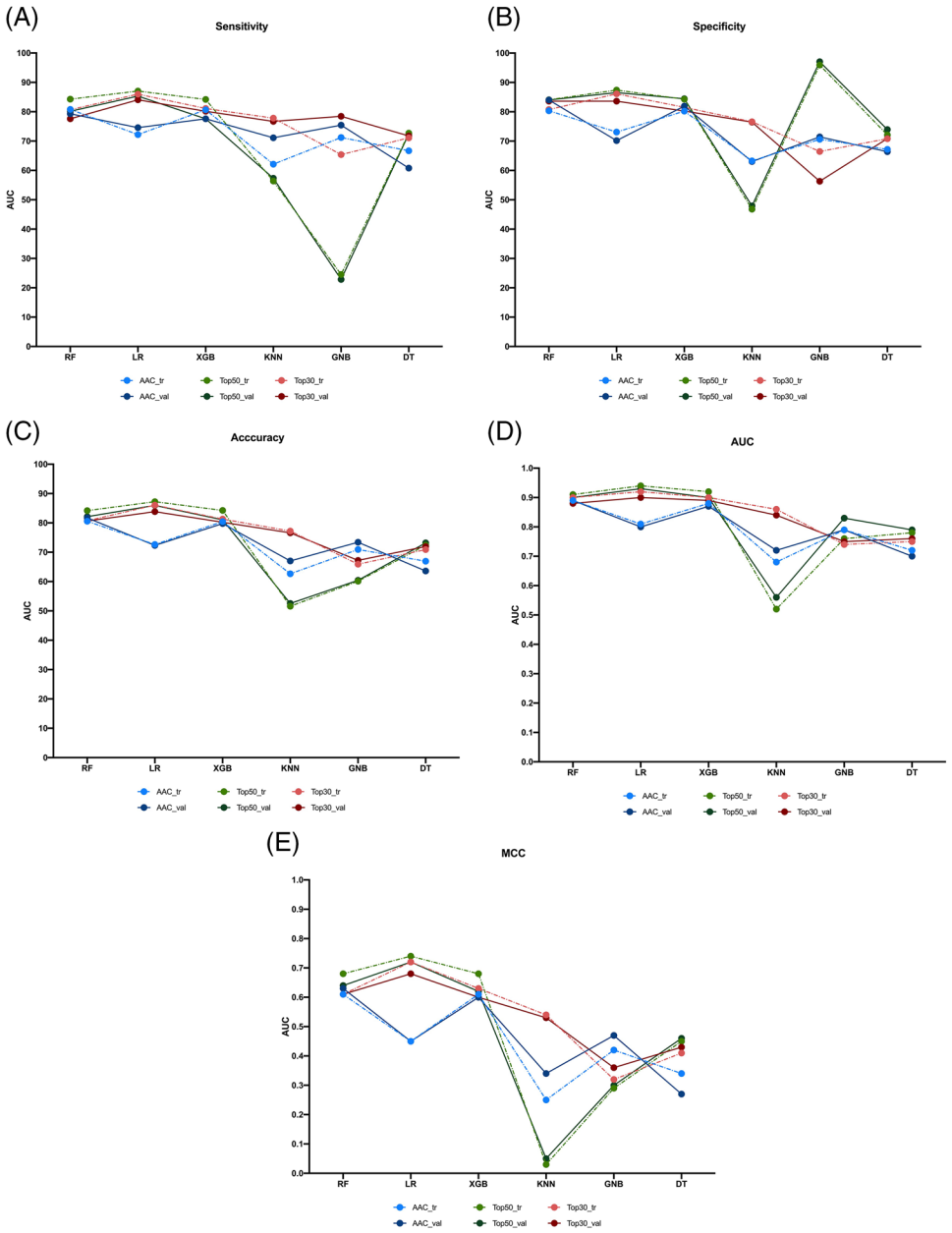

FIGURE 4 Comparison of performance for ML models developed on amino acid composition, top 30 features, and top 50 features on the basis of (A) sensitivity, (B) specificity, (C) accuracy, (D) AUC, and (E) MCC (here tr, training set; val, validation set).

mentioned in Table 3. A total of 12 motifs have been identified in hormones with a coverage of 59 hormone sequences. It was observed that amino acids like phenylalanine (F), glycine (G), leucine (L), proline (P), arginine (R), tryptophan (W), and methionine (M), Cysteine (C), and Serine (S), are the most recurring amino acids in the motifs of hormonal peptide sequences.

3.5 | Performance of BLAST

In this study, we performed a BLAST search to identify hormonal sequences based on the similarity between the two sequences. To perform this algorithm, we used BLAST shortp, which is used for the small protein sequences, that is, peptides. We used the standard top-hit approach on various e-values. The results of BLAST for e-values varying from 10^{-6} to 10^{2} are explained in Table 4.

3.6 | Ensemble or hybrid approach

In the end, various strategies were combined to get around the shortcomings of distinct approaches. These methods were created to more accurately detect the peptide hormones. This combines BLAST and motif-based methods with machine learning models. At first, peptides were categorized using the motif-search method followed by BLAST at various e-values. Not all peptides were categorized

TABLE 1 The performance of machine learning-based models developed for amino acid composition, with the top 30 and top 50 features selected using RFE.

Machine learning methods										
Amino acid composition (AAC)										
	Training					**Validation**				
Model	**Sens (%)**	**Spec (%)**	**Acc (%)**	**AUROC**	**MCC**	**Sens (%)**	**Spec (%)**	**Acc (%)**	**AUROC**	**MCC**
RF	**80.79**	**80.34**	**80.56**	**0.89**	**0.61**	**79.31**	**84.03**	**81.70**	**0.89**	**0.63**
LR	72.19	73.08	72.63	0.81	0.45	74.57	70.17	72.34	0.80	0.45
XGB	80.68	80.23	80.46	0.88	0.61	77.59	81.93	79.79	0.87	0.60
KNN	62.10	63.25	62.67	0.68	0.25	71.12	63.02	67.02	0.72	0.34
GNB	71.231	70.62	70.93	0.79	0.42	75.43	71.43	73.40	0.79	0.47
DT	66.67	67.20	66.93	0.72	0.34	60.78	66.39	63.62	0.70	0.27
SVC	74.10	72.33	73.22	0.81	0.46	76.72	68.49	72.55	0.79	0.45
Top 30 features										
	Training					**Validation**				
Model	**Sens (%)**	**Spec (%)**	**Acc (%)**	**AUC**	**MCC**	**Sens (%)**	**Spec (%)**	**Acc (%)**	**AUC**	**MCC**
RF	80.68	80.77	80.72	0.90	0.61	77.59	83.61	80.64	0.88	0.61
LR	**85.99**	**86.11**	**86.05**	**0.92**	**0.72**	**84.05**	**83.61**	**83.83**	**0.90**	**0.68**
XGB	81.10	81.52	81.31	0.90	0.63	80.17	80.25	80.21	0.89	0.60
KNN	77.81	76.60	77.21	0.86	0.54	76.72	76.47	76.60	0.84	0.53
GNB	65.39	66.45	65.92	0.74	0.32	78.45	56.30	67.23	0.75	0.36
Top 50 features										
	Training					**Validation**				
Model	**Sens (%)**	**Spec (%)**	**Acc (%)**	**AUC**	**MCC**	**Sens (%)**	**Spec (%)**	**Acc (%)**	**AUC**	**MCC**
RF	84.29	84.08	84.18	0.91	0.68	80.17	84.03	82.13	0.90	0.64
LR	**87.05**	**87.40**	**87.22**	**0.94**	**0.74**	**85.34**	**86.55**	**85.96**	**0.93**	**0.72**
XGB	84.18	84.29	84.24	0.92	0.68	77.59	84.45	81.06	0.90	0.62
KNN	56.37	46.79	51.60	0.52	0.03	57.33	47.90	52.55	0.56	0.05
GNB	24.52	95.94	60.12	0.76	0.29	22.84	97.06	60.43	0.83	0.30
DT	72.72	72.11	72.42	0.78	0.45	72.41	73.95	73.19	0.79	0.46

TABLE 2 The performance of deep learning (DL) models (TextCNN and TabNet).

Deep Learning Methods										
	Training					**Validation**				
Model	**Sens (%)**	**Spec (%)**	**Acc (%)**	**AUC**	**MCC**	**Sens (%)**	**Spec (%)**	**Acc (%)**	**AUC**	**MCC**
Text CNN	96.00	93.00	94.00	0.98	0.88	87.00	79.00	83.00	0.90	0.67
TabNet	79.00	80.00	8.00	0.81	0.61	73.00	75.00	74.00	0.75	0.57

using these methods. Hence ML approach was also used in combination. The coverage and precision were much improved by the hybrid approach, which was not possible when using each of these approaches separately.

Firstly, we combined the motif-search approach with ML models to boost the performance of the classification model. The best-performing algorithm—Logistic Regression (LR) achieves an AUROC of 0.94, which was increased from an AUC of 0.93 on the validation set when only ML was applied. In addition, after adding motif search to ML, the MCC was increased from 0.72 to 0.74 for the validation set.

Secondly, we combined both BLAST for e-value = 10^{-1} and motif-search with the ML model to boost the performance of our classification model. The ML model, when combined with BLAST at e-value = 10^{-1}, gave the best results for all ML algorithms compared to other e-values. The performance of the LR-based model increased from 0.94 to 0.96 on the validation dataset for the performance when

TABLE 3 List of motifs and their frequency hormone peptides.

Motifs in hormone sequences	No of hormone sequences	Motifs in hormone sequences	No of hormone sequences
FGPR	32	LCGS	27
WFGP	32	MWF	25
WFGPR	32	MWFG	25
FGPRL	30	MWFGP	25
GPRL	30	MWFGPR	25
WFGPRL	30	MWFGPRL	25

we combined BLAST with the previous model (ML + Motif-search +BLAST). The combined performance for both hybrid approaches (ML + Motif-search) and (ML + Motif-search + BLAST) is displayed in Table 5, and the AUROC plots comparing only the ML model and our best performing hybrid model (ML + Motif-search + BLAST) is given in Figure 5.

3.7 | Webserver and standalone software

The web server that we created for HOPPred (https://webs.iiitd.edu.in/raghava/hoppred/) can distinguish between peptide hormones and non-hormone peptides. The front end and back end of the web server were created using HTML5, Java, CSS3, and PHP scripts. All of the most recent gadgets, including smartphones, tablets, iMacs, and desktop PCs, are all compatible with this web server. Predict, design, and peptide design modules are the primary modules on HOPPred.

4 | DISCUSSIONS

Over the past few decades, peptide therapeutics have attracted much attention. The rise in publications and the creation of in silico tools and databases are evidence of this. Several peptide-based medications have previously received FDA approval due to their benefits over conventional small molecule-based medications. Peptide hormones have demonstrated promising results in replacement therapies since the very beginning [3]. There is a need for in-silico techniques that can predict peptide hormones with high confidence because the identification and screening of putative peptide hormones as drugs in the wet lab is a time-consuming, expensive, and labor-intensive operation.

We created a novel model for the current study using 1174 peptide hormones and 1174 non-hormone peptides for the prediction of peptide hormones. We have used multiple approaches to classify hormonal and non-hormonal peptides—a) motif-search, b) BLAST search, c) Machine Learning models, d) Deep Learning models, and e) hybrid models. In motif search, we found motifs that were exclusively present in hormonal sequences. However, this approach only covered 59 hormonal sequences in total. It was found that phenylalanine (F), glycine (G), leucine (L), proline (P), arginine (R), tryptophan (W), Cysteine (C), Serine (S), and methionine (M) amino acids recurred in the most of the motifs found in hormonal sequences. We identified motifs like—LCGS, and MWFGPRL with various variations, for example,—FGPR, WFGP, WFGPR, etc., present in about 59 hormonal peptide sequences. LCGS is also one of the known motif found in chain B of Insulin, a well-known peptide hormone. In BLAST-search, we used the top-hit method on various e-values ranging from 10^{-6} to 10^{2}. Similar to the motif approach, BLAST was not able to cover all sequences.

Despite being powerful techniques, MERCI and BLAST have some limitations. Merci might struggle with detecting complex motifs that vary significantly from the canonical forms or are hidden within highly variable regions of the sequences. Similarly, high background noise in the sequence data can obscure motif signals [38]. High sequence similarity does not always correlate with similar functions, especially for sequences that have undergone neofunctionalization. BLAST may identify sequences as similar without acknowledging functional divergence. BLAST's effectiveness is contingent upon the sequences available in the database. Novel or poorly represented sequences may not be identified, limiting coverage. BLAST might struggle with extremely short sequences or those that have diverged significantly from any database sequence, resulting in incomplete or inaccurate alignments.

Therefore, in addition to BLAST and MERCI, we extracted 9149 compositional features using Pfeature software and selected the top-performing features to develop our model. We used an array of ML and DL models to classify hormonal and non-hormonal peptides, namely—RF, DT, KNN, GNB, LR, XGB, TextCNN, and TabNet. Each of these brings a unique set of advantages and capabilities to the analysis of peptide sequences. RF is an ensemble method that is great for handling the complexity that might arise from the diverse and high-dimensional space of peptide sequences. DTs offer interpretability, which is valuable in a biological context where understanding the decision process can provide insights into which features (amino acids or motifs) are most important for classification. Although they are prone to overfitting, they serve as a good baseline and can be used as a part of RFs. KNN is a non-parametric method that makes no assumptions about the underlying data distribution [39]. It is intuitive and can be particularly useful when the classification decision is based on the similarity of sequence patterns, which is often the case with peptides. GNB works on the assumption of feature independence and can be quite effective when the number of features (amino acids in the peptide sequence) is large, which is often the case in bioinformatics [23]. LR is a simple and fast method providing probabilistic results, which is useful for binary classification tasks like hormonal versus non-hormonal peptide classification [21]. XGB is a gradient-boosting framework known for its performance and speed [25, 40]. It can handle a variety of data structures and distributions, which makes it suitable for the complex patterns in peptide sequences. Its regularization component helps to prevent overfitting, making it robust for biological data. TabNet, a deep learning-based model designed for tabular data, leverages the sequen-

TABLE 4 BLAST-based search results for training and validation dataset (here, Whits, wrong hits; Chits, correct hits; Nhits, no hits).

	Training						Validation					
	Hormonal			Non-hormonal			Hormonal			Non-hormonal		
e-values	Chits	Whits	Nhits	Chits	Whits	Nhits	Chits	Whits	Nhits	Chits	Whits	Nhits
10^{-6}	269 (14.32%)	0 (0.00%)	673 (35.84%)	0 (0.00%)	0 (0.00%)	936 (49.84%)	65 (13.83%)	0 (0.00%)	167 (35.53%)	0 (0.00%)	0 (0.00%)	238 (50.64%)
10^{-5}	302 (16.08%)	0 (0.00%)	640 (34.08%)	4 (0.21%)	0 (0.00%)	932 (49.63%)	78 (16.60%)	0 (0.00%)	154 (32.77%)	0 (0.00%)	0 (0.00%)	238 (50.64%)
10^{-4}	350 (18.64%)	0 (0.00%)	592 (31.52%)	4 (0.21%)	0 (0.00%)	932 (49.63%)	90 (19.15%)	0 (0.00%)	142 (30.21%)	0 (0.00%)	0 (0.00%)	238 (50.64%)
10^{-3}	427 (22.74%)	0 (0.00%)	515 (27.42%)	8 (0.43%)	0 (0.00%)	928 (49.41%)	107 (22.77%)	0 (0.00%)	125 (26.60%)	0 (0.00%)	0 (0.00%)	238 (50.64%)
10^{-2}	527 (28.06%)	0 (0.00%)	415 (22.1%)	8 (0.43%)	0 (0.00%)	928 (49.41%)	126 (26.81%)	0 (0.00%)	106 (22.55%)	2 (0.43%)	0 (0.00%)	236 (50.21%)
10^{-1}	628 (33.44%)	1 (0.05%)	313 (16.67%)	8 (0.43%)	1 (0.05%)	927 (49.36%)	144 (30.64%)	0 (0.00%)	88 (18.72%)	2 (0.43%)	0 (0.00%)	236 (50.21%)
1	705 (37.54%)	3 (0.16%)	234 (12.46%)	21 (1.12%)	10 (0.53%)	905 (48.19%)	163 (34.68%)	4 (0.85%)	65 (13.83%)	10 (2.13%)	2 (0.43%)	226 (48.09%)
10^{1}	766 (40.79%)	40 (2.13%)	136 (7.24%)	165 (8.79%)	114 (6.07%)	657 (34.98%)	180 (38.3%)	19 (4.04%)	33 (7.02%)	51 (10.85%)	34 (7.23%)	153 (32.55%)
10^{2}	754 (40.05%)	179 (9.53%)	9 (0.48%)	520 (27.69%)	374 (19.91%)	42 (2.24%)	176 (37.45%)	55 (11.70%)	1 (0.21%)	140 (29.79%)	89 (18.94%)	9 (1.91%)

TABLE 5 The performance of ensemble method that combine ML-based models, Motif-search and BLAST (e-value = 10^{-1}) based similarity.

ML + Motif-search										
	Training					Validation				
Model	Sens (%)	Spec (%)	Acc (%)	AUC	MCC	Sens (%)	Spec (%)	Acc (%)	AUC	MCC
RF	84.5	83.55	84.03	0.92	0.68	81.03	86.13	83.62	0.92	0.67
LR	**87.90**	**87.93**	**87.91**	**0.95**	**0.76**	**86.64**	**87.39**	**87.02**	**0.94**	**0.74**
XGB	85.03	84.51	84.77	0.92	0.70	80.17	85.29	82.77	0.91	0.66
KNN	58.39	52.46	55.43	0.59	0.11	58.19	57.56	57.87	0.67	0.16
GNB	26.86	96.05	61.34	0.78	0.32	25.43	97.9	62.13	0.85	0.34
DT	73.78	73.82	73.8	0.80	0.48	64.22	75.21	69.79	0.76	0.40
ML + Motif-search + BLAST										
	Training					Validation				
Model	Sens (%)	Spec (%)	Acc (%)	AUC	MCC	Sens (%)	Spec (%)	Acc (%)	AUC	MCC
RF	88.96	85.90	87.43	0.95	0.75	86.21	88.66	87.45	0.95	0.75
LR	**91.19**	**90.38**	**90.79**	**0.97**	**0.82**	**90.09**	**89.5**	**89.79**	**0.96**	**0.80**
XGB	88.96	87.71	88.71	0.96	0.77	85.34	87.82	86.6	0.95	0.73
KNN	77.18	72.54	73.87	0.85	0.50	74.14	75.63	74.89	0.86	0.50
GNB	69.43	96.26	82.80	0.90	0.68	64.66	97.9	81.49	0.91	0.67
DT	81.95	79.70	80.83	0.89	0.62	77.16	76.89	77.02	0.86	0.54

tial attention mechanism to choose which features to reason from at each decision step [25]. It was particularly useful for our big set of 9149 features. TextCNN applies convolutional neural networks to text classification, treating peptide sequences as a kind of language [26]. This model can capture local patterns (motifs) within the peptide sequences, and due to the weight sharing of convolutional layers, it can efficiently process variable-length sequences.

The ML model alone was able to achieve an AUROC of 0.89, 0.90, and 0.93 for independent validation set on AAC, top 30 features and top 50 features, respectively. For DL models—TextCNN and TabNet,

the performance of AUROC were 0.90 and 0.75, respectively, for the independent validation set, as it was observed that the top 50 features were giving the best performance on logistic regression (ML model). The top 50 compositional features included features like AAC, dipeptide composition (DPC), tripeptide composition (TPC), physicochemical properties (PCP), distance distribution (DDR), Shannon entropy (SEP), conjoint triad descriptors (CTC), composition enhanced transition and distribution (CeTD), and quasi-sequence order (QSO). These top 50 features also align with the motifs identified in this study. The motifs identified include WFGPRL, WFGP, FGPRL, LCGS, etc., and some of the top features include—DPC1_CF (Dipeptide composition of Cysteine-Phenylalanine), TPC_FRP (Tripeptide Composition of Phenylalanine, Arginine, Proline), TPC_GNF (Tripeptide Composition of Glycine, Asparagine, Phenylalanine), TPC_LMG (Tripeptide Composition of Leucine, Methionine, Glycine), and TPC_RGL (Tripeptide Composition of Arginine, Glycine, Leucine). The Motif 'WFGP' shares 'F' and 'P' with TPC_FRP, indicating the potential importance of these residues.'G' and 'P' are also present in the motifs 'FGPRL', 'WFGPR', and 'GPR', which suggests that these amino acids are key components in the hormone function, supported by their presence in the tripeptide composition features like TPC_GNF. TPC_RGL has 'R', 'G', and 'L', which are present in the 'FGPRL' motif. This can indicate the significance of Arginine (R) and Glycine (G) in the structure or function of these hormones, as these residues also appear in the motifs with high frequency. Phenylalanine (F) and Proline (P) are found in multiple motifs, and their involvement in dipeptide and tripeptide features suggests they might play a critical role in hormone activity or stability. Arginine (R), seen in TPC_FRP and TPC_RGL and also part of the 'WFGPR' motif, could indicate a role in binding or molecular interaction given its positive charge. There's a recurring pattern of 'W', 'F', 'G', 'P', and 'R' in both motifs and top features. This suggests these motifs could be involved in critical interactions or structural configurations necessary for the peptides' hormonal activity. DDR_C (Distance distribution of Cysteine) might be relevant for motifs with 'C', as the spatial distribution of Cysteine can be crucial for disulfide bond formation, influencing the three-dimensional structure of peptide hormones. Features like dipeptide and tripeptide compositions reflect the frequency of specific amino acid pairings and triplets, which have been determined to be significant through machine learning techniques. Shannon entropy and distance distribution features provide insight into the variability and spatial arrangement of these amino acids within the hormones.

We decided to combine the best-performing ML model with other approaches used in the study and called it a hybrid model. Firstly, we combined the ML model with motif search, which achieved an AUROC of 0.94 on the validation set. Secondly, we merged our best ML model with both motif-search and BLAST-search (e-value = 10^{-1}), which achieved the highest AUROC of 0.96 with an accuracy of 89.70%, specificity of 90.09% and sensitivity of 89.50%. In addition, we performed AAC analysis on the sequences to identify which amino acids are more or less likely to be found in hormones. It was observed that cysteine, aspartic acid, phenylalanine, glycine, arginine, serine, asparagine, proline, and tyrosine were significantly increased in hormonal peptide sequences, whereas the compositions of amino acids like glutamic acid, isoleucine, leucine, methionine, glutamine, lysine, threonine, and valine were significantly decreased in hormonal peptide sequences than non-hormonal peptide sequences. In the comparative analysis of AAC between hormonal and non-hormonal peptides, significant differences might suggest specialized roles for these residues in hormone function. Statistically heightened levels of hydrophobic amino acids, such as leucine and valine, in hormonal peptides could imply an evolutionary adaptation for interactions with lipid membranes and hydrophobic receptor domains, essential for hormone signaling pathways [41]. Elevated proportions of charged residues like lysine and arginine could indicate a propensity for binding and active sites recognition, reflecting the necessity for precise cellular interactions [42, 43]. The presence of cysteine, potentially more abundant in hormonal peptides, can point to the importance of disulfide bridges in maintaining structural integrity for hormonal activity [44]. Aromatic amino acids, if increased, might play a role in receptor binding through complex interaction mechanisms like π-π stacking, which could enhance the specificity and strength of hormone-receptor interactions [45, 46]. The presence of amino acids like serine (S), threonine (T), and tyrosine (Y), which can be phosphorylated, may suggest that hormonal peptides could be regulated through post-translational modifications [47]. These compositional characteristics underscore the biological implications: hormones are fine-tuned for high-affinity receptor interactions, possess structural features conducive to stability and storage, and are primed for regulation via post-translational modifications. This detailed understanding paves the way for more focused hypotheses on hormone function and interaction dynamics.

The study is one of its kind as it tries to classify peptides into hormonal and non-hormonal, which has not been attempted before. Moreover, it also identifies unique motifs found in hormonal peptides. We have employed our best-performing hybrid model, which achieves an AUROC of 0.96 and is chosen to be put on the web server of HOPPred as it is able to differentiate between hormonal and non-hormonal peptides accurately and does not take much more time than only the ML model being applied. We have developed a comprehensive platform that enables users to categorize both hormone and non-hormone peptide sequences. We have provided the users with a website and a standalone version for a better user experience. It provides users with a choice of using the ML model or an ensemble model according to the user's needs with no significant difference in the amount of time taken by the web server in the prediction. The users can either predict or design novel hormonal peptides and also display various physicochemical properties like charge, hydrophobicity, hydrophilicity, molecular weight, steric hindrance, etc. We believe our research will also be helpful for a variety of areas of biological research, like the development of peptide-based treatments, synthetically designing a novel peptide with hormonal activity, development of growth enhancers or inhibitors for agricultural biotechnology, and more. The server can be accessed at https://webs.iiitd.edu.in/raghava/hoppred/, with the codes available on Github: https://github.com/raghavagps/HOPPRED.

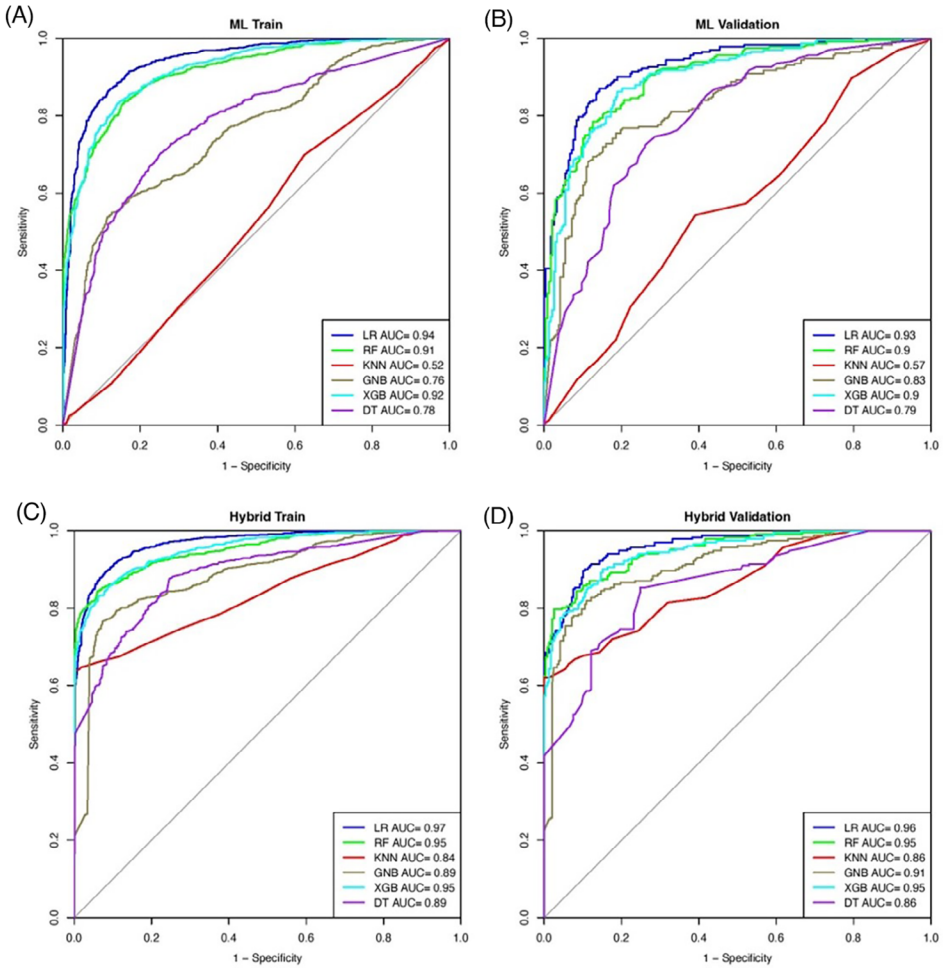

FIGURE 5 AUC plots for (A) training set in the ML model, (B) validation set in the ML model, (C) training set in the hybrid model (ML + Motif-search +BLAST), and (D) validation set in the hybrid model (ML + Motif-search +BLAST).

5 | LIMITATIONS

We acknowledge potential biases arising from random negative data selection and the inherent assumptions of our computational model. In addition, when we develop models on AAC, the model assumes a direct correlation between amino acid frequency and peptide function, which simplifies the complex nature of peptide bioactivity. Furthermore, post-translational modifications, crucial in modulating peptide functions, were not accounted for in this study. Hence, experimental validation can add value to the computational predictions made by our model.

AUTHOR CONTRIBUTIONS

Dashleen Kaur and Akanksha Arora collected and processed the data. Dashleen Kaur, Palani Vigneshwar, and Akanksha Arora implemented the algorithms. Dashleen Kaur, Palani Vigneshwar, and Akanksha Arora developed the prediction models. Dashleen Kaur and Akanksha Arora developed the front end and back end of the web server. Dashleen Kaur, Akanksha Arora, and Gajendra P.S. Raghava prepared the manuscript. Gajendra P.S. Raghava conceived and coordinated the project. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors are thankful to the Council of Scientific and Industrial Research (CSIR) and the All India Council for Technical Education for providing fellowships and financial support. The authors are also thankful to the Department of Computational Biology, IIITD, New Delhi for its infrastructure and facilities. We thank the Department of Biotechnology (DBT) for providing an infrastructure grant to the institute (Grant BT/PR40158/BTIS/137/24/2021). We would like to acknowledge that figures were created using BioRender, and English was corrected using Grammarly.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

All the datasets generated in this study are available at https://webs.iiitd.edu.in/raghava/hoppred/dataset.php. The codes are available at—https://github.com/raghavagps/HOPPRED.

BIORXIV LINK

https://www.biorxiv.org/content/10.1101/2023.05.15.540764v1

ORCID

Gajendra P. S. Raghava https://orcid.org/0000-0002-8902-2876

REFERENCES

1. Kołodziejski, P. A., Pruszyńska-Oszmałek, E., Wojciechowicz, T., Sassek, M., Leciejewska, N., Jasaszwili, M., Billert, M., Małek, E., Szczepankiewicz, D., Misiewicz-Mielnik, M., Hertig, I., Nogowski, L., Nowak, K. W., Strowski, M. Z., & Skrzypski, M. (2021). The Role of peptide hormones discovered in the 21st century in the regulation of adipose tissue functions. *Genes (Basel)*, *12*, 756.
2. Thakur, S. S. (2021). Proteomics and its application in endocrine disorders. *Biochimica et Biophysica Acta-Proteins and Proteomics*, *1869*, 140701.
3. Wang, L., Wang, N., Zhang, W., Cheng, X., Yan, Z., Shao, G., Wang, X., Wang, R., & Fu, C. (2022). Therapeutic peptides: Current applications and future directions. *Signal Transduction and Targeted Therapy*, *7*, 48.
4. Lau, J. L., & Dunn, M. K. (2018). Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorganic & Medicinal Chemistry*, *26*, 2700–2707.
5. Yan, K., Lv, H., Guo, Y., Chen, Y., Wu, H., & Liu, B. (2022). TPpred-ATMV: Therapeutic peptide prediction by adaptive multi-view tensor learning model. *Bioinformatics*, *38*, 2712–2718.
6. Shoombuatong, W., Schaduangrat, N., Pratiwi, R., & Nantasenamat, C. (2019). THPep: A machine learning-based approach for predicting tumor homing peptides. *Computational Biology and Chemistry*, *80*, 441–451.
7. Agrawal, P., Bhagat, D., Mahalwal, M., Sharma, N., & Raghava, G. P. S. (2021). AntiCP 2.0: An updated model for predicting anticancer peptides. *Briefings in Bioinformatics*, *22*(3), bbaa153.
8. Yan, W., Tang, W., Wang, L., Bin, Y., & Xia, J. (2022). PrMFTP: Multi-functional therapeutic peptides prediction based on multi-head self-attention mechanism and class weight optimization. *Plos Computational Biology*, *18*, e1010511.
9. Kaur, D., Arora, A., Patiyal, S., & Raghava, G. P. S. (2023). Hmrbase2: a comprehensive database of hormones and their receptors. *Hormones (Athens)*, *22*(3), 359–366.
10. Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*, 3150–3152.
11. Deutsch, E. W., Lam, H., & Aebersold, R. (2008). PeptideAtlas: A resource for target selection for emerging targeted proteomics workflows. *European Molecular Biology Organization Reports*, *9*, 429–434.
12. Pande, A., Patiyal, S., Lathwal, A., Arora, C., Kaur, D., Dhall, A., Mishra, G., Kaur, H., Sharma, N., Jain, S., Usmani, S. S., Agrawal, P., Kumar, R., Kumar, V., & Raghava, G. P. S. (2023). Pfeature: A tool for computing wide range of protein features and building prediction models. *Journal of Computational Biology*, *30*, 204–222.
13. Ren, K., Zeng, Y., Cao, Z., & Zhang, Y. (2022). ID-RDRL: A deep reinforcement learning-based feature selection intrusion detection model. *Scientific Reports*, *12*, 15370.
14. Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, *2*, 927312.
15. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer New York.
16. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Front Neuroinform*, *8*, 14.
17. Di Noto, T., von Spiczak, J., Mannil, M., Gantert, E., Soda, P., Manka, R., & Alkadhi, H. (2019). Radiomics for distinguishing myocardial infarction from myocarditis at late gadolinium enhancement at MRI: Comparison with subjective visual analysis. *Radiol Cardiothorac Imaging*, *1*, e180026.
18. Arora, A., Patiyal, S., Sharma, N., Devi, N. L., Kaur, D., & Raghava, G. P. S. (2024). A random forest model for predicting exosomal proteins using evolutionary information and motifs. *Proteomics*, *24*, e2300231.
19. Flayer, C. H., Perner, C., & Sokol, C. L. (2021). A decision tree model for neuroimmune guidance of allergic immunity. *Immunology and Cell Biology*, *99*, 936–948.
20. Yi, Y., Sun, D., Li, P., Kim, T.-K., Xu, T., & Pei, Y. (2022). Unsupervised random forest for affinity estimation. *Computational Visual Media (Beijing)*, *8*, 257–272.
21. Stoltzfus, J. C. (2011). Logistic regression: A brief primer. *Academic Emergency Medicine*, *18*, 1099–1104.
22. Miao, Y., Hunter, A., & Georgilas, I. (2021). An occupancy mapping method based on K-nearest neighbours. *Sensors (Basel)*, *22*(1), 139.
23. Joshi, D., Mishra, A., & Anand, S. (2012). A naïve Gaussian Bayes classifier for detection of mental activity in gait signature. *Computer Methods in Biomechanics and Biomedical Engineering*, *15*, 411–416.
24. Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z., & Wang, K. (2020). Predicting 30-days mortality for MIMIC-III patients with sepsis-3: A machine learning approach using XGboost. *Journal of Translational Medicine*, *18*, 462.
25. Arık, S. Ö., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*, 6679–6687.
26. Kalchbrenner, N., Grefenstette, E., & Blunsom, P., in: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Stroudsburg, PA, USA*. (2014), pp. 655–665.
27. Rathore, A. S., Arora, A., Choudhury, S., Tijare, P., & Raghava, G. P. S. (2023). ToxinPred 3.0: An improved method for predicting the toxicity of peptides. *bioRxiv* 2023.08.11.552911.
28. Wang, Z., Wu, M., Liu, Q., Wang, X., Yan, C., & Song, T. (2024). Multi-classification of hepatic cystic echinococcosis by using multiple kernel learning framework and ultrasound images. *Ultrasound in Medicine & Biology*. https://doi.org/10.1016/j.ultrasmedbio.2024.03.018
29. Sharma, N., Naorem, L. D., Jain, S., & Raghava, G. P. S. (2022). ToxinPred2: An improved method for predicting toxicity of proteins. *Briefings in Bioinformatics*, *23*, bbac174.
30. Sharma, N., Patiyal, S., Dhall, A., Pande, A., Arora, C., & Raghava, G. P. S. (2021). AlgPred 2.0: An improved method for predicting allergenic proteins and mapping of IgE epitopes. *Briefings in Bioinformatics*, *22*(4), bbaa294.
31. Aggarwal, S., Dhall, A., Patiyal, S., Choudhury, S., Arora, A., & Raghava, G. P. S. (2023). An ensemble method for prediction of phage-based therapy against bacterial infections. *Frontiers in Microbiology*, *14*, 1148579.
32. Le, N. Q. K., Li, W., & Cao, Y. (2023). Sequence-based prediction model of protein crystallization propensity using machine learning and two-level feature selection. *Briefings in Bioinformatics*, *24*(5), bbad319.
33. Kha, Q.-H., Ho, Q.-T., & Le, N. Q. K. (2022). Identifying SNARE proteins using an alignment-free method based on multiscan convolutional neural network and PSSM profiles. *Journal of Chemical Information and Modeling*, *62*, 4820–4826.
34. Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., & Zaretskaya, I. (2013). BLAST: A more efficient report with usability improvements. *Nucleic Acids Research*, *41*, W29–W33.
35. Arora, A., Patiyal, S., Sharma, N., Devi, N. L., Kaur, D., & Raghava, G. P. S. (2024). A random forest model for predicting exosomal proteins using evolutionary information and motifs. *Proteomics*, *24*(6), e2300231.
36. Agrawal, T. (2021). *Hyperparameter optimization in machine learning* (pp. 31–51) Apress.
37. Belete, D. M., & Huchaiah, M. D. (2022). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS

test results. *International Journal of Computers and Applications*, *44*, 875–886.
38. Vens, C., Rosso, M.-N., & Danchin, E. G. J. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, *27*, 1231–1238.
39. Wu, Y., Ianakiev, K., & Govindaraju, V. (2002). Improved k-nearest neighbor classification. *Pattern Recognition*, 2311–2318.
40. Chen, T., & Guestrin, C. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2016).
41. Al Mughram, M. H., Catalano, C., Herrington, N. B., Safo, M. K., & Kellogg, G. E. (2023). 3D interaction homology: The hydrophobic residues alanine, isoleucine, leucine, proline and valine play different structural roles in soluble and membrane proteins. *Frontiers in Molecular Biosciences*, *10*, 1116868.
42. Vidya, J., Ushasree, M. V., & Pandey, A. (2014). Effect of surface charge alteration on stability of L-asparaginase II from Escherichia sp. *Enzyme and Microbial Technology*, *56*, 15–19.
43. Strickler, S. S., Gribenko, A. V., Gribenko, A. V., Keiffer, T. R., Tomlinson, J., Reihle, T., Loladze, V. V., & Makhatadze, G. I. (2006). Protein stability and surface electrostatics: A charged relationship. *Biochemistry*, *45*, 2761–2766.
44. Wiedemann, C., Kumar, A., Lang, A., & Ohlenschläger, O. (2020). Cysteines and disulfide bonds as structure-forming units: Insights from different domains of life and the potential for characterization by NMR. *Frontiers in Chemistry*, *8*, 280.
45. Meyer, E. A., Castellano, R. K., & Diederich, F. (2003). Interactions with aromatic rings in chemical and biological recognition. *Angewandte Chemie (International ed in English)*, *42*, 1210–1250.
46. Shao, J., Kuiper, B. P., Thunnissen, A.-M. W. H., Cool, R. H., Zhou, L., Huang, C., Dijkstra, B. W., & Broos, J. (2022). The role of tryptophan in π interactions in proteins: An experimental approach. *Journal of the American Chemical Society*, *144*, 13815–13822.
47. Santos, A. L., & Lindner, A. B. (2017). Protein posttranslational modifications: Roles in aging and age-related disease. *Oxidative Medicine and Cellular Longevity*, *2017*, 5716409.

SUPPORTING INFORMATION

Additional supporting information may be found online https://doi.org/10.1002/pmic.202400004 in the Supporting Information section at the end of the article.

How to cite this article: Kaur, D., Arora, A., Vigneshwar, P., & Raghava, G. P. S. (2024). Prediction of peptide hormones using an ensemble of machine learning and similarity-based methods. *Proteomics*, *24*, e2400004. https://doi.org/10.1002/pmic.202400004