Computers in Biology and Medicine 183 (2024) 109297

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

mHPpred: Accurate identification of peptide hormones using multi-view feature learning

Shaherin Basith [a,*], Vinoth Kumar Sangaraju [b], Balachandran Manavalan [b,**], Gwang Lee [a,c,***]

[a] Department of Physiology, Ajou University School of Medicine, Suwon, 16499, Republic of Korea
[b] Department of Integrative Biotechnology, College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon, 16419, Republic of Korea
[c] Department of Molecular Science and Technology, Ajou University, Suwon, 16499, Republic of Korea

ARTICLE INFO

Keywords:
Peptide hormones
Meta-model
Multi-view learning
Hybrid approach
Integrative framework
Machine learning

ABSTRACT

Peptide hormones were first used in medicine in the early 20th century, with the pivotal event being the isolation and purification of insulin in 1921. These hormones are integral to a sophisticated system that emerged early in evolution to regulate growth, development, and homeostasis. They serve as targeted signaling molecules that transfer specific information between cells and organs, ensuring coordinated and precise physiological responses. While experimental methods for identifying peptide hormones present challenges such as low abundance, stability issues, and complexity, computational methods offer promising alternatives. Advances in machine learning and bioinformatics have facilitated the prediction of peptide hormones, further enhancing their therapeutic potential. In this study, we explored three different computational frameworks for peptide hormone identification and determined that the meta-approach was the most suitable. Firstly, we evaluated the discriminative power of 26 feature descriptors using a series of baseline models and identified seven feature descriptors with high predictive potential. Through a systematic approach, we then selected the top 20 performing baseline models and integrated their predicted probabilities to train a meta-model, leveraging the strengths of multiple prediction strategies. Our final light gradient boosting-based meta-model, mHPpred, significantly outperformed the existing method, HOPPred, on both benchmarking and independent datasets. Notably, mHPpred also demonstrated superior performance compared to the hybrid and integrative framework approaches employed in this study. This superiority demonstrates the effectiveness of our multi-view feature learning strategy in capturing discriminative features and providing a more accurate prediction model for peptide hormones. mHPpred is publicly accessible at: https://balalab-skku.org/mHPpred.

1. Introduction

The use of peptide hormones in medicine dates back to the early 20th century, with the isolation and purification of insulin in 1921 marking a significant milestone [1–3]. This breakthrough paved the way for the therapeutic use of peptide hormones and spurred advancements in biology, chemistry, and pharmacology [4–6]. Since the discovery of insulin and its inception into medicine 100 years ago, peptide hormones and their analogs and mimetics have been a key component of modern drug development [7]. Peptide hormones are an important class of chemical signaling molecules composed of short chain amino acids, typically ranging from 10 to 100 in length [8]. These hormones are synthesized as part of larger precursor proteins which are subsequently cleaved by endoproteases to form the smaller active peptide hormone. They function as signaling molecules by binding to specific receptors on target cells, initiating a cascade of cellular responses. These peptide hormones play a crucial role in regulating energy homeostasis and metabolism [9]. They are also involved in controlling appetite, managing the functions of the gastrointestinal and cardiovascular systems, regulating energy expenditure, and reproductive processes [9].

Key studies of natural human hormones including insulin, oxytocin, vasopressin, and gonadotropin-releasing hormone (GnRH) sparked

* Corresponding author. Department of Physiology, Ajou University School of Medicine, 206 World cup-ro, Yeongtonggu, Suwon 16499, Republic of Korea.
** Corresponding author. Department of Integrative Biotechnology, College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon, 16419, Republic of Korea.
*** Corresponding author. Department of Physiology, Ajou University School of Medicine, 206 World cup-ro, Yeongtonggu, Suwon 16499, Republic of Korea.
E-mail addresses: sbasith@ajou.ac.kr (S. Basith), bala2022@skku.edu (B. Manavalan), glee@ajou.ac.kr (G. Lee).

https://doi.org/10.1016/j.compbiomed.2024.109297
Received 20 August 2024; Received in revised form 4 October 2024; Accepted 15 October 2024
Available online 23 October 2024

interest in therapeutic peptides [10]. Over 80 peptide drugs have been approved worldwide since the first therapeutic peptide, insulin, was synthesized in 1921 [11]. Consequently, the development of peptide drugs has become one of the most prominent topics in pharmaceutical research [11,12]. Currently, there are over 40 commercially available peptide-based drugs, including insulin, atrial natriuretic peptide (ANP), glucagon-like peptide-1 (GLP-1) analogs (exendin-4), and thymosin alpha 1 [13]. Additionally, more than 100 novel peptide therapeutics are undergoing evaluation in clinical trials [14]. The molecular weights of these hormones vary significantly, ranging from small molecules like tripeptide thyrotropin-releasing hormone (TRH) analog to longer polypeptides such as human insulin (51 residues) and parathyroid hormone (PTH (1–84)).

Despite recent progress in the therapeutic application of peptide hormones, however identification and characterization of these peptides using experimental techniques are facing several challenges including high cost and time, low abundance, low stability, isolation and specificity issues, and post-translational modifications [9,15]. To overcome these problems, computational approaches could complement and enhance traditional experimental methods. Machine learning (ML), one of the most powerful bioinformatics approaches could be utilized for the prediction of peptide hormones. By leveraging sequence data and advanced algorithms, researchers can uncover new insights into the roles and mechanisms of these crucial biomolecules, paving the way for innovative therapeutic strategies and a deeper understanding of endocrine physiology. Although peptide hormones have been established for over a century, computational identification using ML is still in its infancy.

While several computational methods [16–19] have been developed for identifying therapeutic peptides, such as anti-cancer peptides [20–22], anti-viral peptides, anti-inflammatory peptides [23], anti-hypertensive peptides [24], and anti-microbial peptides [25–28], to date, only one ML method has been developed specifically for identifying peptide hormones. Recently, Kaur et al. [15] developed an ensemble-based ML method, HOPPred for predicting peptide hormones. While this method demonstrated excellent performance on both the training and independent dataset, there is still room for improvement in terms of its robustness. In this study, we investigated three distinct computational frameworks for predicting peptide hormones: meta-modeling, feature fusion, and an integrative approach. Fig. 1 illustrates the workflow for the development of peptide hormone prediction tool, which is structured according to Chou's five-step rule [29–31]. Initially, a non-redundant dataset was created using a balanced dataset of positive and negative peptide sequences. Subsequently, we

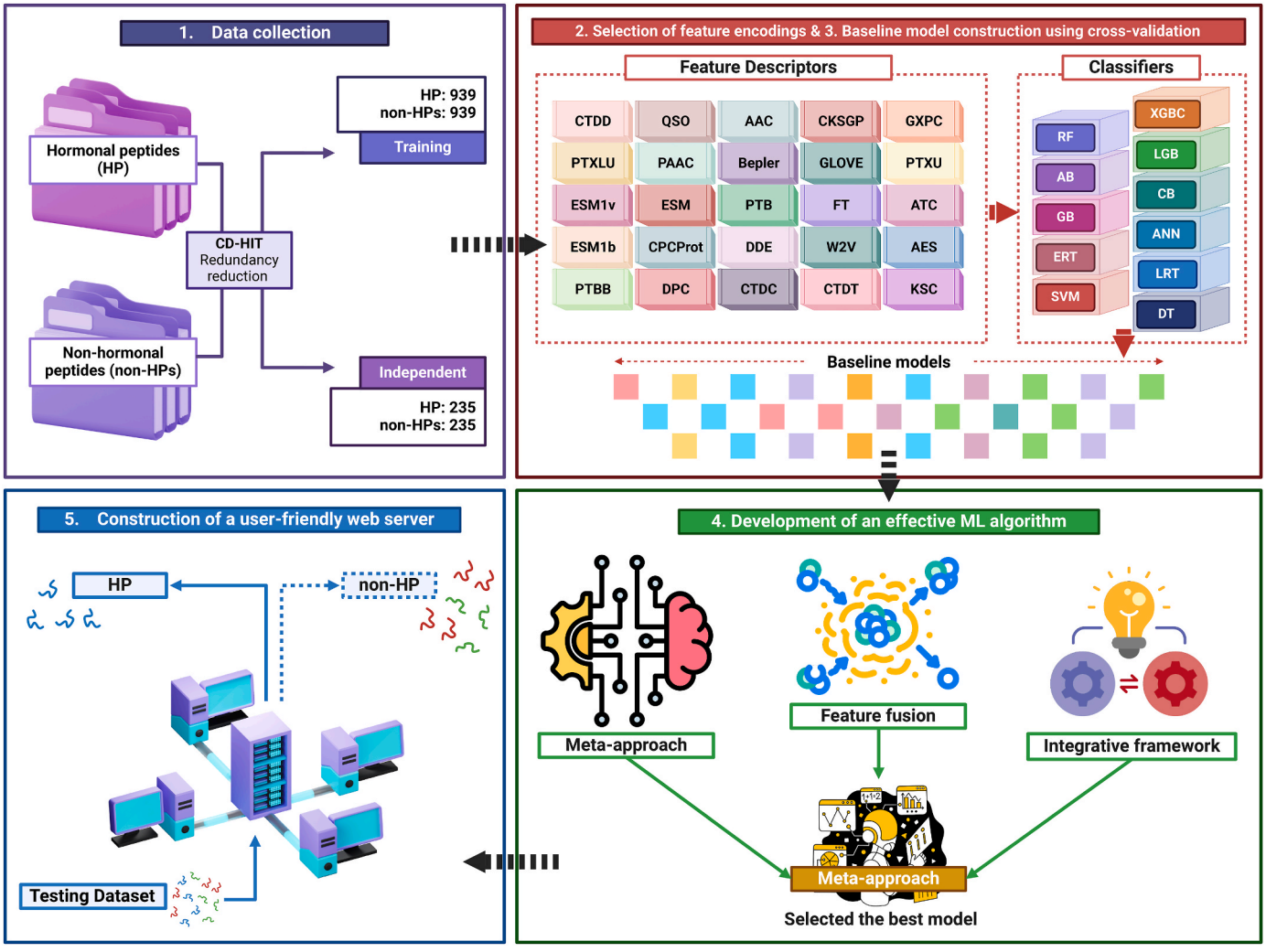

Fig. 1. Overall workflow methodology of mHPpred. The schematic workflow shows the incorporation of Chou's five-step rule involved in the development of mHPpred: (1) Construction of non-redundant benchmark and independent datasets; (2 & 3) Construction of 286 baseline models using 26 feature descriptors and 11 machine learning (ML) classifiers; (4) Evaluation of the top-performing baseline models and features using three different multi-view learning: meta-learning, feature fusion learning, and integrative framework. Following extensive validations, the optimal model, mHPpred, was created utilizing a meta-learning approach; and (5) Webserver development.

evaluated the discriminative power of 26 feature descriptors using 11 different ML classifiers and developed single baseline models. From this analysis, we identified one conventional feature descriptor and six protein language model (PLM)-based embeddings that demonstrated superior performance. These features and their corresponding models were then applied to the three different frameworks, with the meta-modeling approach emerging as marginally better than the integrative framework and significantly outperforming the hybrid approach. The resulting meta-model, named mHPpred, demonstrated superior performance compared to existing predictor in both cross validation and independent tests. The final model, mHPpred, is then deployed as a user-friendly tool to facilitate peptide hormone prediction. By leveraging the strengths of multiple feature descriptors and combining them through an advanced meta-modeling framework, mHPpred effectively captured the complex patterns necessary for accurate peptide hormone prediction. This comprehensive evaluation confirms that mHPpred is a highly effective tool, offering significant improvements over other prediction tools.

2. Methods

2.1. Construction of benchmarking and independent datasets

Herein, we address the first step of Chou's five-step rule which involves constructing a high-quality, non-redundant dataset. We employed the same benchmark and independent datasets used in the previous study [15] for both model development and evaluation. Initially, a total of 5729 mature hormone peptide sequences (excluding signal and precursor regions) were retrieved from the Hmrbase2 database [32]. To remove redundancy from the dataset, all duplicate peptides were eliminated. The lengths of these peptides ranged from 11 to 41 amino acids. To identify unique peptides, redundant peptides were filtered using CD-HIT [33] cut-off of 0.6, meaning no peptide in the dataset shared more than 60% similarity to another peptide. This rigorous filtering process resulted in a final set of 1174 unique peptide hormones, which is referred to as the positive dataset in this study. To generate a negative dataset of non-peptide hormones, the authors randomly selected 1174 non-peptide hormones from the PeptideAtlas database [34]. The combined positive and negative datasets were then randomly split into 80% benchmarking (training) dataset for model development and the remaining 20% independent dataset for evaluating the model's performance and generalizability.

2.2. Feature encodings

The second step of Chou's five-step rule involves selecting appropriate feature representations for the dataset. Herein, a diverse set of 13 protein language model (PLM)-based feature encodings [35] and 13 conventional feature encodings were utilized [36,37]. PLM-based feature encodings included evolutionary scale modeling (ESM) and their variants ESM1b, and ESM1v, ProtTransT5XLU50 (PTXLU), ProtTransT5UniRef50 (PXTU), ProtTransBertBFD (PTBB), CPCProt, ProtTransAlbertBFD (PTAB), Bepler, PTB, GLoVE, FastText (FT), and Word2Vec (W2V). Conventional feature encodings included composition-transition-distribution descriptors (CTDC, CTDD, and CTDT), quasi-sequence-order (QSO), pseudo-amino acid composition (PAAC), dipeptide composition (DPC), amino acid composition (AAC), DDE, composition of k-spaced amino acid group pairs (CKSGP), combination of grouped dipeptide and tripeptide compositions (GXPC), ATC, AES, and KSC. The feature encodings and their corresponding dimensions (D) are provided in detail in Table S1. Out of these 26 feature encodings, we identified the top seven feature descriptors (CTDD, PTXLU, ESM1v, PTBB, ESM, PTAB, and ESM1b) that demonstrated outstanding discriminatory power and were instrumental in the development of mHPpred. The biological significance of these top seven descriptors has been briefly described. CTDD captures the distribution of amino acid properties that influence peptide structure, stability, and function [38]. PTXLU combines evolutionary information with local structural dynamics, providing a comprehensive understanding on how specific regions of a peptide contribute to its overall function [39]. ESM, ESM1b, and ESM1v are advanced deep learning models that accurately predict peptide sequences by analyzing evolutionary patterns and structural characteristics [40]. PTAB and PTBB are essential descriptors for peptide prediction, capturing crucial binding-related features and improving the accuracy and interpretability of ML models for peptide interactions [41]. Furthermore, a concise explanation of how these top seven feature encodings work is provided in the supplementary material.

2.3. Construction of baseline models

The third step of Chou's five-step rule involves selecting or developing a powerful algorithm for making predictions. In this study, we employed 26 feature descriptors, each subjected individually to 11 ML classifiers. Chou's fourth step of five-step rule emphasizes the importance of conducting proper cross validation (CV) to objectively assess the prediction model's accuracy. To assess the generalizability of the predicted models when applied to unknown data, we utilized CV. Although various CV techniques are available, we opted for 10-randomized 10-fold CV [42,43] due to its numerous advantages, including lower bias, improved performance estimation, better model selection, and strong resilience to outliers. In the 10-fold CV technique, the benchmarking dataset is randomly divided into 10 parts. Nine parts are used for training the model, while the remaining part is used for testing. This process is repeated until each part has been used as a testing data at least once, ensuring that all parts contribute to the evaluation. The overall performance is then assessed based on the results from all 10 parts. To optimize the hyperparameters of each classifier during training, we employed grid search, a technique recommended by previous studies [20,44–48]. This method systematically explores a range of hyperparameter values to identify the best combination for each classifier. The baseline models were constructed based on the median parameters obtained from the 10-randomized 10-fold CVs. In total, we constructed 286 baseline models (26 feature descriptors x 11 ML classifiers) on the training dataset and evaluated its transferability on independent dataset.

2.4. Construction of meta-models

To develop a robust meta-model, we first selected baseline models that achieved an accuracy (ACC) of over 84% on the training data. This selection resulted in 52 models, encompassing seven distinct feature descriptors (CTDD, PTXLU, ESM1v, PTBB, ESM, PTAB, and ESM1b) paired with various ML classifiers (excluding DT). Subsequently, we ranked these 52 models based on their performance. To leverage the strength of multiple models, we extracted the predicted probability scores from each model and constructed five different feature sets: 10D, 20D, 30D, 40D, and 52D. Each feature set represented a different subset of the top-ranked models. For instance, 10D feature set combined the predicted probabilities from the top 10 baseline models. Each feature set was then used to train a new meta-model using all 11 classifiers and 10-randomized 10-fold CV, resulting in 55 meta-models. These meta-models were then rigorously evaluated to identify the best-performing final model.

2.5. Feature fusion

In addition to the meta-model approach, we also employed feature fusion, a technique that linearly integrates two or more features to potentially improve the predictive power. Based on the performance of our baseline models, we identified the top seven features (CTDD, PTXLU, ESM1v, PTBB, ESM, PTAB, and ESM1b), that showed excellent

capability in predicting peptide hormones. To explore the advantages of feature fusion, we systematically integrated various combinations of these top features. We started by combining the first two features, followed by the first three, and continued this process up to all seven features. These hybrid feature sets were named Hyb_2, Hyb_3, Hyb_4, Hyb_5, Hyb_6, and Hyb_7, respectively. For each of these hybrid feature sets, we trained models using all 11 classifiers and optimized their hyperparameters (settings that control the learning process) using a rigorous 10-randomized 10-fold CV procedure.

2.6. Integrative framework

In addition to the meta-model and feature fusion approaches, we also implemented an integrated approach, a widely adopted strategy that combines the strengths of multiple models and features. This approach has often demonstrated superior performance compared to individual meta-models.

As described in the meta-model section, we integrated the predicted probability scores from the 55 meta-models we developed previously. These probability scores were then used as input features to train a new set of models using 11 different classifiers and a rigorous 10-randomized 10-fold CV procedure. From these integrated models, we selected the top five performing models based on their accuracy and other performance metrics.

2.7. Implementation of machine learning algorithms

In this study, we employed 11 ML classifiers, such as random forest (RF), extremely randomized tree or extra trees (ERT), gradient boosting (GB), adaptive boosting or AdaBoost (AB), extreme gradient boosting (XGB), support vector machine (SVM), light gradient boosting (LGB), artificial neural network or neural net (ANN), decision tree (DT), logistic regression (LR), and category boosting or CatBoost (CB) for model development and validation. RF and ERT are both tree-based ensemble learning methods. RF constructs multiple decision trees during training and outputs the mode of the classes or mean prediction of the individual trees [49]. ERT is similar to RF but introduces more randomness in tree splitting to reduce overfitting [50]. GB, AB, CB, XGB, and LGB are all ensemble techniques that combines weak classifiers to create a strong classifier. GB builds models sequentially, where each new model corrects the errors of the previous ones to optimize overall prediction accuracy [51]. AB also combines weak classifiers to create a strong classifier by focusing more on hard-to-classify instances [52]. CB is also a GB algorithm that automatically handles categorical features, making it robust and effective for various types of data [53]. XGB is an efficient implementation of GB known for high performance and speed, often used for large-scale datasets [54]. LGB is another highly efficient GB framework that uses tree-based learning algorithms and is optimized for speed and memory efficiency [55]. SVM is a different type of ML classifier that finds the hyperplane which best separates the data into different classes with maximum margin [56]. ANN is a computational model inspired by the human brain consisting of interconnected nodes (neurons) that can learn from data through training [57]. DT is a simple yet powerful model that splits the data into subsets based on feature values, forming a tree-like structure to make predictions [58]. LR is a statistical model that uses a logistic function to model a binary dependent variable and predict the probability of a certain class [59]. Importantly, these classifiers have proven effective in a wide range of bioinformatics function prediction tasks [60–62].

2.8. Evaluation metrics

The performance of the models were evaluated using various evaluation parameters, including Matthews' correlation coefficient (MCC), ACC, sensitivity (Sn), specificity (Sp), and area under the ROC curve (AUC) [63–66]. The mathematical equation for each metric is given below:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$MCC = \frac{(TP\mathrm{x}TN) - (FP\mathrm{x}FN)}{\sqrt{(TP + FP)\mathrm{x}(TP + FN)\mathrm{x}(TN + FP)\mathrm{x}(TN + FN)}}$$

where TP, TN, FP and FN represent true positives, true negatives, false positives, and false negatives, respectively.

3. Results and discussion

3.1. Performance evaluation of baseline models using different features and ML classifiers

We assessed the intrinsic discriminative patterns of 13 PLM- and 13 conventional feature-based descriptors by inputting them into 11 different ML classifiers. The predictive performance of 11 ML-based classifiers along with 26 feature encodings were examined using 10-randomized 10-fold CV. A total of 286 baseline models were developed using the benchmarking dataset (Table S2). Their overall performance, as measured by AUC, MCC, and ACC, is illustrated in Fig. 2A, B, and C. These models were subsequently evaluated on independent datasets (Table S3), as shown in Fig. 2D, E, and F. Among these models, LGB model utilizing CTDD demonstrated the best performance on the training dataset with the MCC, ACC, Sn, Sp, and AUC of 0.808, 0.904, 0.916, 0.891, and 0.957, respectively. The corresponding metrics for this model on independent dataset were 0.787, 0.894, 0.897, 0.891, and 0.950. Notably, the performance of this LGB-based CTDD model is on par with that of the existing method, HOPPred.

To evaluate the ability of each feature to distinguish between peptide hormones and non-peptide hormones, we averaged 11 classifiers training performance. Fig. S1 shows that seven descriptors (CTDD, PTXLU, ESM1V, PTBB, ESM1b, ESM, and PTAB) exhibit exceptional capability in distinguishing peptide hormones from non-peptide hormones, with average ACC ranging from 83.30% to 87.60%. In contrast, five descriptors (CTDT, GXPC, ATC, AES, and KSC) show limited ability to distinguish between the two classes, resulting in the average ACC below 75%. The remaining descriptors exhibit moderate discrimination, with the average ACC falling between 76.50% and 81.10%. Importantly, the top seven high-performing descriptors consistently maintained their strong performance when tested on an independent dataset, confirming the robustness of their discriminative ability (Fig. S1). Upon further examination of these descriptors, we observed that the conventional descriptors, CTDD, captures the distribution of physicochemical properties within peptides. The remaining six are derived from PLMs. Specifically, ESM and ESM1v encode sequence information, while PTBB, PTXLU, ESM1b, and PTAB represent biophysical and structural features of proteins. These results underscore the importance of physicochemical, sequential, biophysical, and structural properties in accurately differentiating peptide hormones from non-peptide hormones. The diverse nature of the high-performing descriptors suggests that combining these properties contributes to optimal classification. To integrate this information and further enhance model performance, we explored three different computational frameworks: meta-model construction [24,67], feature fusion [68], and an integrative approach [69].

3.2. Construction of mHPpred and comparison with top baseline models

To construct the best model, we employed a meta-predictor

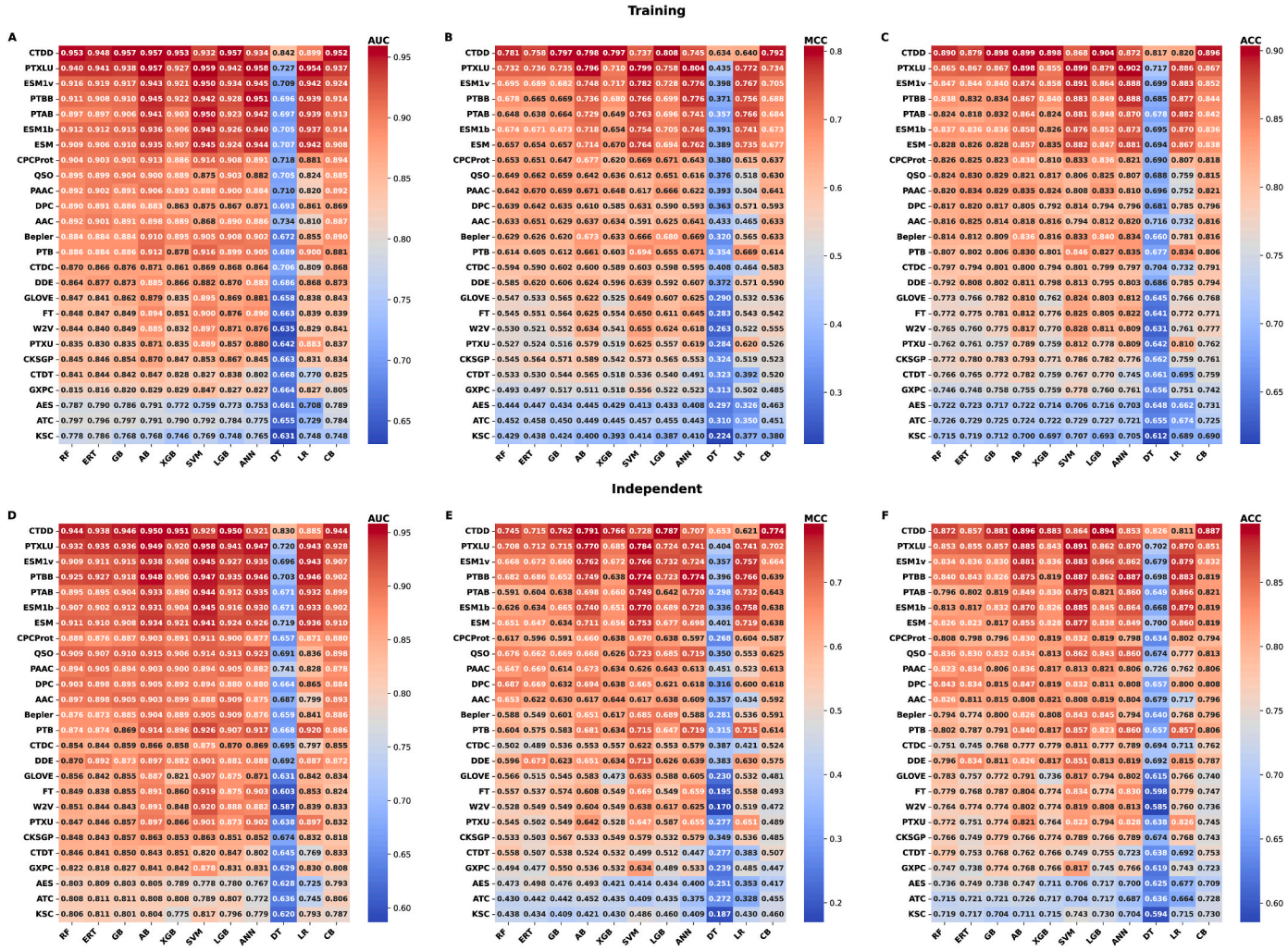

Fig. 2. Performance comparison of baseline models for each feature descriptor. The evaluation performances of baseline models for each descriptor during training are shown as follows: (A) area under the receiver operating characteristic (ROC) curve (AUC), (B) Matthews' correlation coefficient (MCC), and (C) accuracy (ACC) performance metrics of each classifier across all feature descriptors. Similarly, the performance evaluation during the independent testing in terms of (D) AUC, (E) MCC, and (F) ACC is also shown.

approach. For each classifier, we generated 26 models using different feature descriptors and selected only those with an ACC over 84%, resulting in 52 baseline models (Table S4). Specifically, we selected RF, ERT, GB, and XGB classifiers with CTDD, PTXLU, ESM1v encodings, AB classifier with CTDD, PTXLU, ESM1v, PTBB, PTAB, ESM1b, and ESM encodings, SVM classifier with PTXLU, ESM1v, PTBB, ESM, PTAB, ESM1b, CTDD, and PTB encodings, LGB classifier with CTDD, PTXLU, ESM1v, ESM1b, PTBB, PTAB, and ESM encodings, ANN classifier with PTXLU, ESM1v, PTBB, ESM, ESM1b, CTDD, and PTAB encodings, LR classifier with PTXLU, ESM1v, PTAB, PTBB, ESM1b, and ESM encodings, and CB classifier with CTDD, PTXLU, ESM1v, PTBB, and PTAB encodings. To optimize model selection and harness the strength of multiple models, we implemented a hierarchical meta-learning strategy. This involved grouping 52 baseline models into five different sets (10D, 20D, 30D, 40D and 52D) based on their ACC ranking, with each set progressively incorporating models with lower ACC. For instance, the 10D set included only the top 10 performing models, while the 52D set encompassed all 52 models. Each set of predicted scores was then used to train 55 meta-models, employing 11 different classifiers and 10-randomized 10-fold CV. For each set, we selected the best classifier (Fig. S2). The result shows that the overall performance of the various meta-models, ranging from 10D to 52D, was comparable across both training and independent datasets. Notably, LGB using the 20D set and SVM using the 52D set achieved the highest performance, though marginally better than the other meta-models. Given the smaller input feature dimension of the LGB meta-model compared to the SVM, we chose the LGB model and designated it as mHPpred.

We compared our developed model mHPpred with the top five baseline models including LGB_CTDD, ANN_PTXLU, SVM_PTXLU, AB_CTDD, and GB_CTDD (Fig. 3). As shown in Fig. 3A, mHPpred outperformed the top five baseline models, achieving AUC, MCC, ACC, Sn, and Sp of 0.977, 0.864, 0.932, 0.932, and 0.932, respectively on the training dataset. These evaluation metrics represent improvements of 1.7–2.0% in AUC, 5.6–6.7% in MCC, and 2.8–3.4% in ACC over the top five baseline models. To evaluate the robustness of mHPpred, we tested its performance on their respective independent dataset and compared mHPpred performance with the top five baseline models. It can be observed from Fig. 3B that mHPpred model consistently outperformed the top five baseline models on independent dataset. mHPpred achieved the ACC, MCC, Sn, Sp, and AUC values of 0.966, 0.868, 0.934, 0.935, and 0.933 on the independent dataset, respectively. The overall improvements of mHPpred in terms of AUC, MCC, and ACC compared to the top five baseline models on the independent dataset are 0.8–2.0%, 7.7–12.8%, and 3.8–6.4%, respectively. Overall, mHPpred consistently exceeded the performance of the top five baseline models on both training and independent datasets, showcasing its exceptional

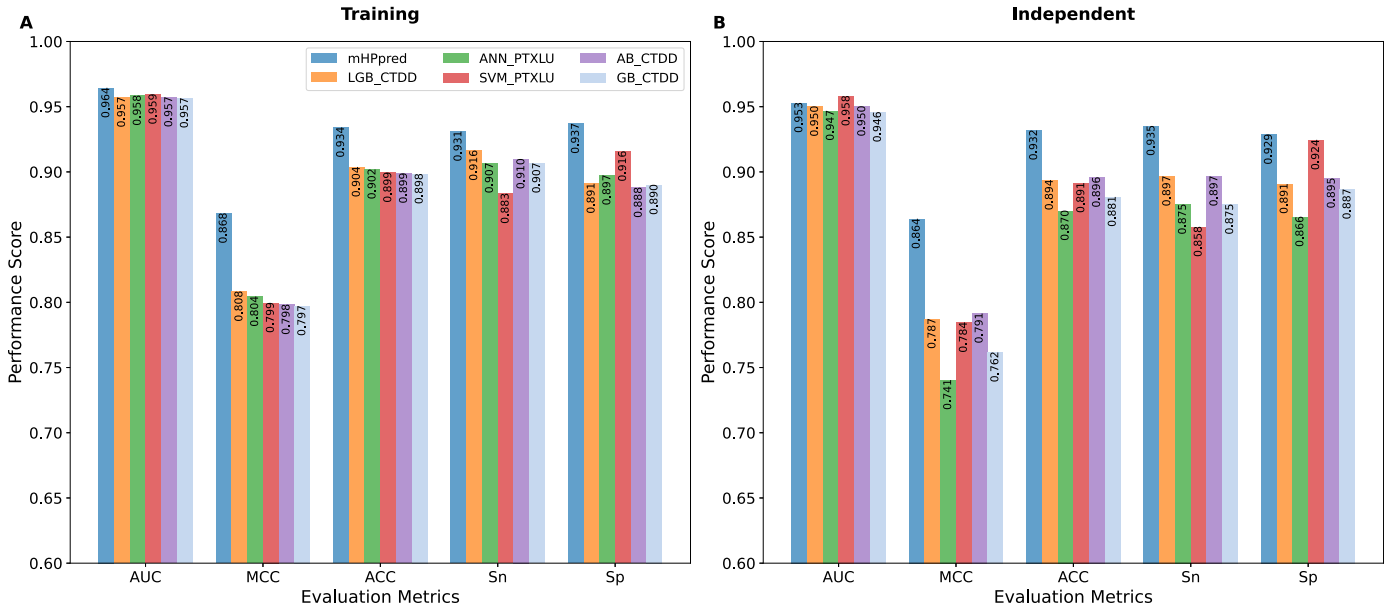

Fig. 3. Performance evaluation of mHPpred with the five top-performing baseline models during (A) training and (B) independent testing. The models were evaluated in terms of evaluation metrics such as AUC, MCC, ACC, Sn, and Sp.

reliability, convergence, and generalization abilities.

3.3. Benchmarking mHPpred against feature fusion and integrative framework approaches

Since the top seven feature encodings such as CTDD, PTXLU, ESM1v, PTBB, ESM, PTAB, and ESM1b exhibit superior discriminative capability in distinguishing between peptide hormones and non-peptide hormones, we investigated their synergistic potential. We trained 11 different ML classifiers on each set of hybrid features ranging from top 2 to top 7 features (Hyb_2 to Hyb_7) and evaluated the model performance on independent dataset. As shown in Table S5, LGB-based models demonstrated superior performance on all hybrid models. We therefore focused on comparing the meta-model, mHPpred, against these top-performing LGB-based hybrid models. mHPpred consistently outperformed the hybrid models on both the training and independent datasets. As evident in Fig. 4A, the overall improvements of mHPpred in terms of ACC and MCC compared to the hybrid models (Hyb_2 to Hyb_7) on the training dataset are 0.4–1.0% and 0.9–1.9%, respectively. We also evaluated the performance of mHPpred and hybrid models on independent dataset to check our model stability. Fig. 4B demonstrated the superior performance of mHPpred, consistently showing the best results on the independent dataset as well. The overall improvements of mHPpred in terms of ACC and MCC compared to the hybrid models on the independent dataset are 1.0–2.8% and 2.1–5.5%, respectively.

As detailed in the methods section about the integrative framework approach, we trained 11 classifiers using the predicted probability scores from the meta-models and subsequently selected the top five

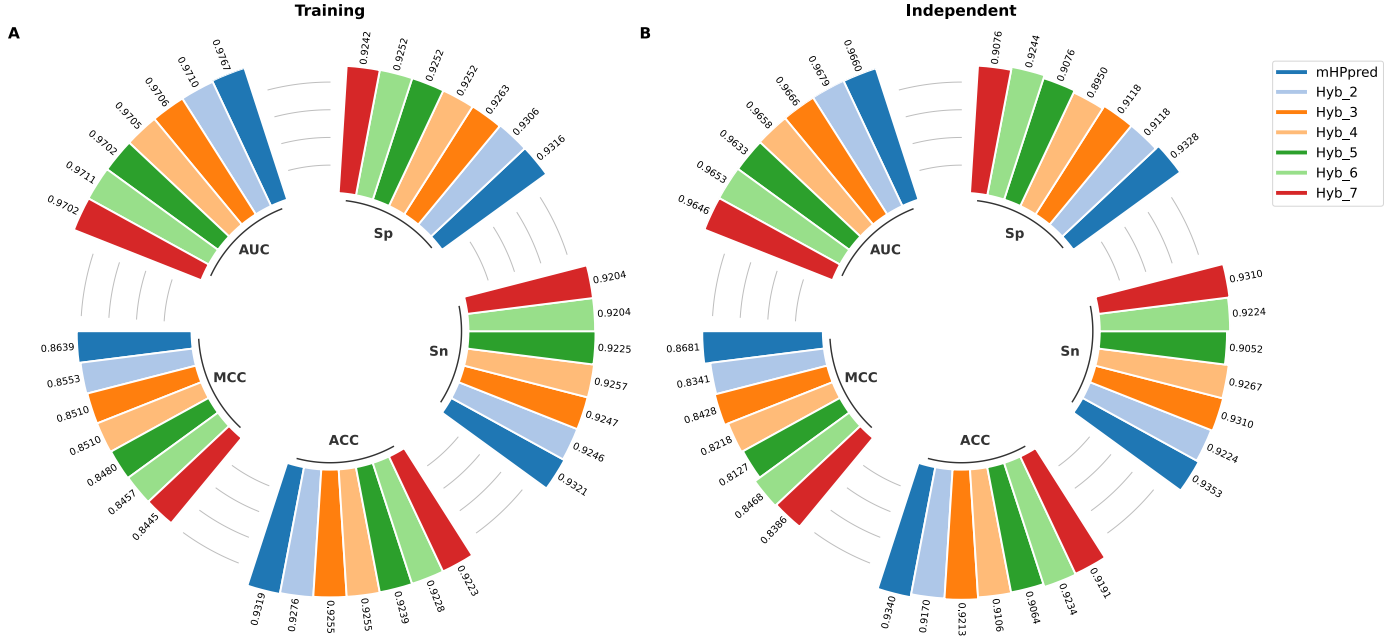

Fig. 4. Performance evaluation of mHPpred with the hybrid models (Hyb_2 to Hyb_7) that were developed using a feature fusion approach during (A) training and (B) independent testing. The models were assessed based on evaluation metrics such as AUC, MCC, ACC, Sn, and Sp.

super meta-models. Subsequently, we compared mHPpred's performance with the top five super meta-models derived from the integrative framework approach. On the training dataset, these five super meta-models exhibited slightly higher performance metrics compared to mHPpred (Fig. 5A). Out of the five models, CB-based super meta-model showed the top performance with ACC, MCC, Sn, Sp, and AUC values of 0.979, 0.880, 0.940, 0.940, and 0.940, respectively. However, its performance slightly declined when tested on an independent dataset, with AUC, MCC, ACC, Sn, and Sp values of 0.960, 0.860, 0.930, 0.935, and 0.924, respectively. A similar pattern was also observed with other top integrative super meta-models when tested on an independent dataset. In terms of consistency between the training and independent datasets, mHPpred outperforms the other integrative models (Fig. 5A and B). The overall improvements of mHPpred in terms of ACC and MCC compared to the top 5 integrative models are 0.4–1.9% and 0.9–3.8%, respectively on the independent dataset. This demonstrates that the meta-approach used in mHPpred can effectively achieve a higher level of discriminative capability. By combining various feature descriptors and leveraging advanced ML techniques, mHPpred is able to capture more nuanced and complex patterns in the data. Although integrative approaches and feature fusion have demonstrated promising results in various applications, our study indicates that these models, while advantageous over existing methods are slightly less consistent than mHPpred. This suggests that mHPpred is more effective at capturing subtle and complex patterns in the data resulting in better generalization and robustness on unseen datasets.

3.4. Comparison of mHPpred with the state-of-the-art predictor

Despite the abundance of peptide hormones data, only one ML-based method, namely HOPPred has been developed so far. Initially, we compared the training performance between HOPPred and mHPpred since both were trained and developed on the same dataset (Table 1). Compared to HOPPred, mHPpred showed improvements of 4.4% and 2.4% in MCC and ACC values, respectively. These findings illustrate that the meta-approach through systematic analysis significantly boosted performance over the existing method on the training dataset.

Notably, mHPpred outperformed existing predictor by 3.6% in ACC and 6.8% in MCC metrics. Moreover, McNemar's chi-square test was applied to evaluate whether the differences between mHPpred and HOPPred were statistically significant. At a p-value threshold of 0.05, the results demonstrated that mHPpred significantly outperformed HOPPred (Table 1). The superlative performance can be attributed to mHPpred's ability to effectively learn and integrate predictions from a diverse set of baseline models, capturing a broader range of patterns within the data. By leveraging highly discriminative feature descriptors, multiple classifiers, and a robust meta-learning framework, mHPpred achieves significantly improved accuracy and reliability compared to existing methods for peptide hormone prediction.

3.5. Feature selection analysis

To evaluate the effectiveness of our features in distinguishing between peptide hormones and non-peptide hormones, we applied t-distributed stochastic neighbor embedding (t-SNE) to visualize the relationship among samples in 20D probabilistic feature (PF) vector and compared it with the top four individual feature descriptors including CTDD, PTXLU, ESM1v, and PTBB. While the individual feature descriptors showed some high degree of overlap between peptide hormones and non-peptide hormones (Fig. 6A–D), the PF vector showed a more pronounced separation (Fig. 6E). This clear separation, with minimal overlap between two classes, highlights the superior performance on the training dataset. Importantly, the pattern of enhanced separation of PF was consistently observed when evaluated on independent dataset (Fig. 6F–J). These findings suggest that the PF vector generated by our multi-view feature learning is more effective at differentiating between peptide hormones and non-peptide hormones samples compared to the individual feature descriptors. The superior performance of our approach on both datasets demonstrates its ability to capture and leverage diverse features, suggesting broader applicability for identifying various biological patterns.

3.6. Development of mHPpred web server

In bioinformatics, providing publicly accessible databases and web servers aids biomedical researchers in conducting experimental analyses. Chou's final step in the five-step rule involves establishing a user-friendly, publicly accessible web server. To assist users in identifying

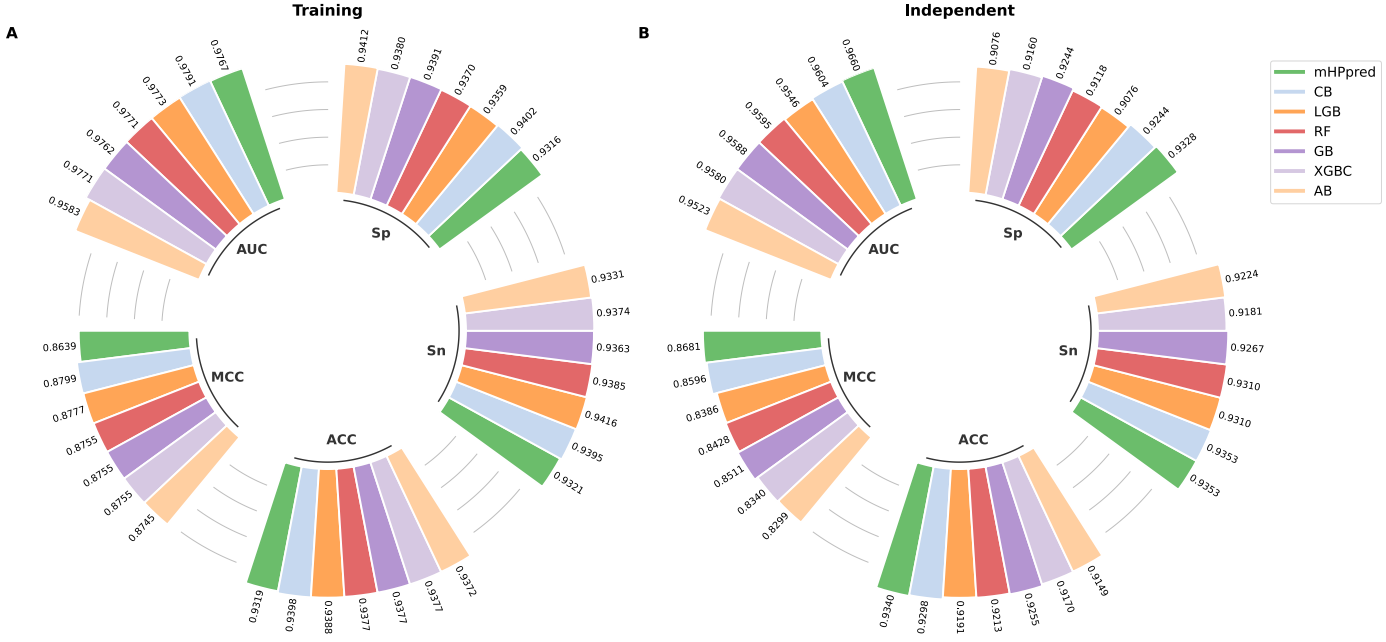

Fig. 5. Performance evaluation of mHPpred with the top-performing five super meta-models that were developed using an integrative framework during (A) training and (B) independent testing. The models were assessed based on evaluation metrics such as AUC, MCC, ACC, Sn, and Sp.

Table 1
Performance comparison of mHPpred with the state-of-the-art predictor.

Dataset	Model	AUC	MCC	ACC	Sn	Sp	p Value
Benchmarking	mHPpred	0.977	0.864	0.932	0.932	0.932	<0.0001[a]
	HOPPred	0.970	0.820	0.908	0.912	0.904	
Independent	mHPpred	0.966	0.868	0.934	0.935	0.933	<0.0001[a]
	HOPPred	0.960	0.800	0.898	0.901	0.895	

[a] Both p-values suggest that the difference is statistically significant at the standard significance level of 0.05.

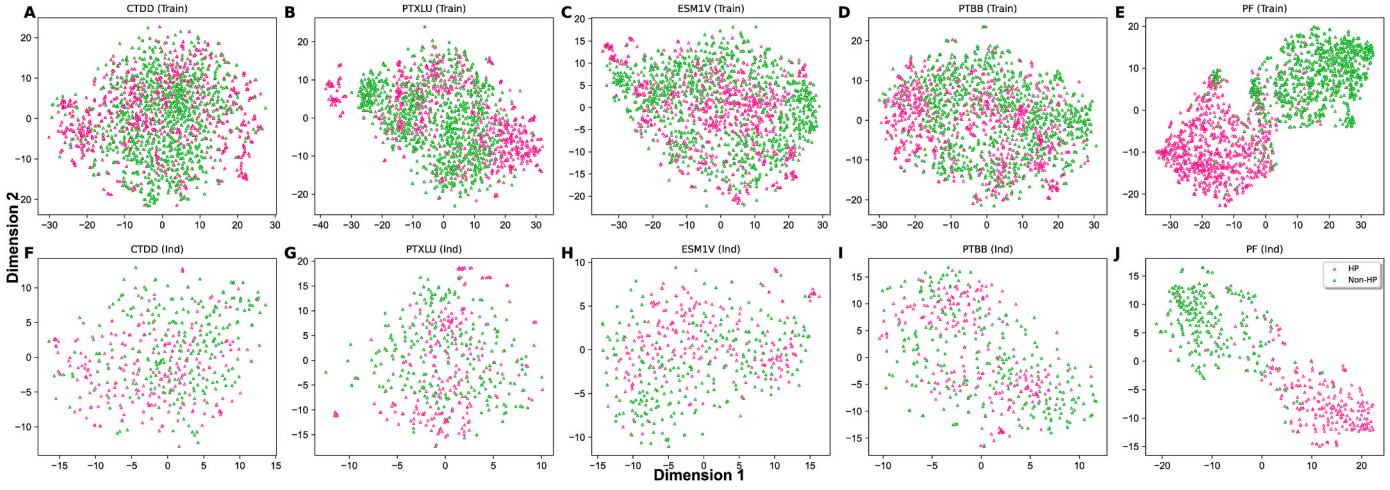

Fig. 6. t-SNE distribution of hormonal peptides and non-hormonal peptides using 20D probabilistic feature (PF) vector and the top four individual feature descriptors. Panels A–E represent the distributions of CTDD, PTXLU, ESM1v, PTBB, and PF, respectively.

peptide hormones, we have developed the mHPpred online web server, available for free at https://balalab-skku.org/mHPpred. Users can either paste their sequences directly into the provided textbox in FASTA format or upload them via the file selection dialog box. mHPpred analyzes the input and provides a prediction in an easy-to-interpretable table, containing four columns: serial number, FASTA ID, predicted class (HP or non-HP), and predicted probability of being a peptide hormone. Generally, the probability score (PS) for peptide hormone ranges between 0 and 1, reflecting the confidence of prediction. A PS closer to 1 suggest a higher probability of being peptide hormone. Regarding the web server performance and response times, users typically receive results within 1–5 min for standard input sequences comprising up to 100 amino acids upon submitting a prediction task. For larger or more complex sequences, particularly those exceeding 200 amino acids, the computational time may extend to 10 min or longer. This variability is primarily attributed to the increased computational resources required to analyze longer sequences. If users encounter any issues with job submission or results retrieval on the mHPpred web server, they are encouraged to reach out to the support team via the contact email listed on the server's webpage for assistance. This ensures smooth operation and addresses any technical difficulties promptly. Detailed instructions for using the mHPpred web server are available on the help page.

4. Conclusions

Due to their appealing pharmacological profile, target specificity, and inherent qualities, peptide hormones are a valuable starting point for developing novel treatments. Peptide hormones have been identified to possess a wide range of biological properties, including anti-angiogenic, anti-bacterial, anti-cancer, and anti-fungal activities, making them highly attractive for therapeutic applications. In this study, we proposed mHPpred, a cutting-edge predictor designed for the accurate identification of peptide hormones. Our approach involved a systematic exploration of 26 feature descriptors and 11 ML classifiers, resulting in the generation of 286 baseline models. Through rigorous analysis, we selected the top 20 performing baseline models and leveraged them to develop a powerful meta-model using the LGB algorithm. This meta-model consistently surpassed the performance of others on both the training and testing datasets. In addition to the meta-approach, we also employed feature fusion and integrative framework, which have proven effective in various prediction tasks. Although these approaches excelled well in predicting peptide hormones compared to the existing predictor, they fell slightly short compared to mHPpred. mHPpred significantly outperformed the state-of-the-art predictor HOPPred, on both training and independent datasets. The significant improvement is attributable to our systematic integration of various feature descriptors, leveraging advanced ML algorithms, and rigorous validation techniques. Feature analysis on both training and testing demonstrates that the top 20 baseline models have an excellent ability to discriminate between peptide hormones and non-peptide hormones, resulting in an improved performance compared to the top five descriptors. The superior performance of mHPpred demonstrates its potential as a valuable tool for identifying peptide hormones, with wide-ranging applications in biological research and drug discovery. Moreover, the success of our multi-view learning suggests its broader applicability to other bioinformatics prediction tasks [35,63,70–72], paving the way for advancements in the field.

CRediT authorship contribution statement

Shaherin Basith: Writing – original draft, Software, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Vinoth Kumar Sangaraju:** Software, Formal analysis. **Balachandran Manavalan:** Writing – original draft, Supervision, Funding acquisition, Conceptualization. **Gwang Lee:** Writing – original draft, Supervision, Funding acquisition, Conceptualization.

Funding

This work was supported by grants from the National Research Foundation (NRF), funded by the Ministry of Science and ICT (MSIT) in Korea (2022R1I1A1A01071228, 2020M3E5D9080661, RS-2024-00344752, and RS-2024-00416536).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the Korea BioData Station (K-BDS) for providing computational resources. We extend our gratitude to Nattanong Bupi for his assistance in the creation of Fig. 1.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.109297.

References

[1] P. Falcetta, M. Aragona, A. Bertolotto, C. Bianchi, F. Campi, M. Garofolo, et al., Insulin discovery: a pivotal point in medical history, Metabolism 127 (2022) 154941.
[2] S. Ghosh, S. Mahalanobish, P.C. Sil, Diabetes: discovery of insulin, genetic, epigenetic and viral infection mediated regulation, Nucleus (Calcutta). 65 (2022) 283–297.
[3] R. Seetharaman, S. Pawar, M. Advani, One hundred years since insulin discovery: an update on current and future perspectives for pharmacotherapy of diabetes mellitus, Br. J. Clin. Pharmacol. 88 (2022) 1598–1612.
[4] C. Huang, A. Palani, Z. Yang, Q. Deng, V. Reddy, R.P. Nargund, et al., Discovery of insulin/GLP-1/glucagon triagonists for the treatment of diabetes and obesity, ACS Med. Chem. Lett. 13 (2022) 1255–1261.
[5] R.P. Mishra, S. Gupta, A.S. Rathore, G. Goel, Multi-level high-throughput screening for discovery of ligands that inhibit insulin aggregation, Mol. Pharm. 19 (2022) 3770–3783.
[6] D.A. Pissarnitski, A. Kekec, L. Yan, Y. Zhu, D.D. Feng, P. Huo, et al., Discovery of insulin receptor partial agonists MK-5160 and MK-1092 as novel basal insulins with potential to improve therapeutic index, J. Med. Chem. 65 (2022) 5593–5605.
[7] O. Racz, [How was it? Contributions to the history of insulin discovery], Orv. Hetil. 163 (2022) 201–205.
[8] O. Mirabeau, E. Perlas, C. Severini, E. Audero, O. Gascuel, R. Possenti, et al., Identification of novel peptide hormones in the human proteome by hidden Markov model screening, Genome Res. 17 (2007) 320–327.
[9] P.A. Kolodziejski, E. Pruszynska-Oszmalek, T. Wojciechowicz, M. Sassek, N. Leciejewska, M. Jasaszwili, et al., The role of peptide hormones discovered in the 21st century in the regulation of adipose tissue functions, Genes (Basel) 12 (2021).
[10] L. Wang, N. Wang, W. Zhang, X. Cheng, Z. Yan, G. Shao, et al., Therapeutic peptides: current applications and future directions, Signal Transduct Target Ther 7 (2022) 48.
[11] D.J. Craik, D.P. Fairlie, S. Liras, D. Price, The future of peptide-based drugs, Chem. Biol. Drug Des. 81 (2013) 136–147.
[12] X. Luo, H. Chen, Y. Song, Z. Qin, L. Xu, N. He, et al., Advancements, challenges and future perspectives on peptide-based drugs: focus on antimicrobial peptides, Eur. J. Pharmaceut. Sci. 181 (2023) 106363.
[13] P. Barman, S. Joshi, S. Sharma, S. Preet, S. Sharma, A. Saini, Strategic approaches to improvise peptide drugs as next generation therapeutics, Int. J. Pept. Res. Therapeut. 29 (2023) 61.
[14] L. Otvos Jr., Wade JD. Big peptide drugs in a small molecule world, Front. Chem. 11 (2023) 1302169.
[15] D. Kaur, A. Arora, P. Vigneshwar, G.P.S. Raghava, Prediction of peptide hormones using an ensemble of machine learning and similarity-based methods, Proteomics (2024) e2400004.
[16] Y. Zhu, C. Jia, F. Li, J. Song, Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling, Anal. Biochem. 593 (2020) 113592.
[17] R. Xie, J. Li, J. Wang, W. Dai, A. Leier, T.T. Marquez-Lago, et al., DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy, Briefings Bioinf. 22 (2021).
[18] B. Manavalan, J. Lee, FRTpred: a novel approach for accurate prediction of protein folding rate and type, Comput. Biol. Med. 149 (2022) 105911.
[19] S. Basith, N.T. Pham, M. Song, G. Lee, B. Manavalan, ADP-Fuse: a novel two-layer machine learning predictor to identify antidiabetic peptides and diabetes types using multiview information, Comput. Biol. Med. 165 (2023) 107386.
[20] L. Thi Phan, H. Woo Park, T. Pitti, T. Madhavan, Y.J. Jeon, B. Manavalan, Mlacp 2.0: an updated machine learning tool for anticancer peptide prediction, Comput. Struct. Biotechnol. J. 20 (2022) 4473–4480.
[21] V. Boopathi, S. Subramaniyam, A. Malik, G. Lee, B. Manavalan, D.C. Yang, mACPpred: a support vector machine-based meta-predictor for identification of anticancer peptides, Int. J. Mol. Sci. 20 (2019).
[22] B. Manavalan, S. Basith, T.H. Shin, S. Choi, M.O. Kim, G. Lee, MLACP: machine-learning-based prediction of anticancer peptides, Oncotarget 8 (2017) 77121–77136.
[23] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest, Front. Pharmacol. 9 (2018) 276.
[24] B. Manavalan, S. Basith, T.H. Shin, L. Wei, G. Lee, mAHTPred: a sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation, Bioinformatics 35 (2019) 2757–2765.
[25] S. Basith, B. Manavalan, T. Hwan Shin, G. Lee, Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening, Med. Res. Rev. 40 (2020) 1276–1314.
[26] G. Zhong, H. Liu, L. Deng, Ensemble machine learning and predicted properties promote antimicrobial peptide identification, Interdiscip. Sci. (2024). https://doi.org/10.1007/s12539-024-00640-z.
[27] J. Xu, F. Li, C. Li, X. Guo, C. Landersdorfer, H.H. Shen, et al., iAMPCN: a deep-learning approach for identifying antimicrobial peptides and their functional activities, Briefings Bioinf. 24 (2023).
[28] J. Xu, F. Li, A. Leier, D. Xiang, H.H. Shen, T.T. Marquez Lago, et al., Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides, Briefings Bioinf. 22 (2021).
[29] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236–247.
[30] W. Lu, Z. Song, Y. Ding, H. Wu, Y. Cao, Y. Zhang, et al., Use Chou's 5-Step Rule to Predict DNA-Binding Proteins with Evolutionary Information, vol. 2020, 2020 6984045.
[31] M.A. Akmal, W. Hussain, N. Rasool, Y.D. Khan, S.A. Khan, K.C. Chou, Using CHOU'S 5-steps rule to predict O-linked serine glycosylation sites by blending position relative features and statistical moment, IEEE ACM Trans. Comput. Biol. Bioinf 18 (2021) 2045–2056.
[32] D. Kaur, A. Arora, S. Patiyal, G.P.S. Raghava, Hmrbase2: a comprehensive database of hormones and their receptors, Hormones (Athens) 22 (2023) 359–366.
[33] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.
[34] E.W. Deutsch, H. Lam, R. Aebersold, PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows, EMBO Rep. 9 (2008) 429–434.
[35] Z. Yan, F. Ge, Y. Liu, Y. Zhang, F. Li, J. Song, et al., TransEFVP: a two-stage approach for the prediction of human pathogenic variants based on protein sequence embedding fusion, J. Chem. Inf. Model. 64 (2024) 1407–1418.
[36] Z. Chen, P. Zhao, F. Li, T.T. Marquez-Lago, A. Leier, J. Revote, et al., iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data, Briefings Bioinf. 21 (2020) 1047–1057.
[37] Z. Chen, X. Liu, P. Zhao, C. Li, Y. Wang, F. Li, et al., iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets, Nucleic Acids Res. 50 (2022) W434–W447.
[38] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, Prediction of protein folding class using global description of amino acid sequence, Proc. Natl. Acad. Sci. U.S.A. 92 (1995) 8700–8704.
[39] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.
[40] W. Wang, Z. Peng, J. Yang, Single-sequence protein structure prediction using supervised transformer protein language models, Nat Comput Sci 2 (2022) 804–814.
[41] Y. Lei, S. Li, Z. Liu, F. Wan, T. Tian, S. Li, et al., A deep-learning framework for multi-level peptide-protein interaction prediction, Nat. Commun. 12 (2021) 5465.
[42] Y. Zhang, R. Xie, J. Wang, A. Leier, T.T. Marquez-Lago, T. Akutsu, et al., Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework, Briefings Bioinf. 20 (2019) 2185–2199.
[43] W. Shoombuatong, S. Basith, T. Pitti, G. Lee, B. Manavalan, THRONE: a new approach for accurate prediction of human rna N7-methylguanosine sites, J. Mol. Biol. 434 (2022) 167549.
[44] S. Basith, B. Manavalan, T.H. Shin, G. Lee, SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome, Mol. Ther. Nucleic Acids 18 (2019) 131–141.
[45] B. Manavalan, S. Basith, T.H. Shin, G. Lee, Computational prediction of species-specific yeast DNA replication origin via iterative feature representation, Briefings Bioinf. 22 (2021).
[46] B. Manavalan, M.C. Patra, Mlcpp 2.0: an updated cell-penetrating peptides and their uptake efficiency predictor, J. Mol. Biol. 434 (2022) 167604.
[47] M.M. Hasan, N. Schaduangrat, S. Basith, G. Lee, W. Shoombuatong, B. Manavalan, HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation, Bioinformatics 36 (2020) 3350–3356.

[48] B. Manavalan, T.H. Shin, M.O. Kim, G. Lee, PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions, Front. Immunol. 9 (2018) 1783.
[49] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[50] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42.
[51] J.H. Friedman, Greedy function approximation: a gradient boosting machine, 29 % J The Annals of Statistics 44 (2001) 1189–1232.
[52] F. Yoav, E.S. Robert, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1997) 119–139.
[53] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: Unbiased Boosting with Categorical Features, vol. 18, Nips, 2018, pp. 6639–6649.
[54] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, vol. 16, Kdd, 2016, pp. 785–794.
[55] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., LightGBM: a highly efficient gradient boosting decision tree, NIPS (News Physiol. Sci.) 17 (2017) 3149–3157.
[56] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[57] W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, Bull. Math. Biophys. 5 (1943) 115–133.
[58] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[59] D.R. Cox, The regression analysis of binary sequences, J. Roy. Stat. Soc. B 20 (2018) 215–232.
[60] Z.F. Gu, Y.D. Hao, T.Y. Wang, P.L. Cai, Y. Zhang, K.J. Deng, et al., Prediction of blood-brain barrier penetrating peptides based on data augmentation with Augur, BMC Biol. 22 (2024) 86.
[61] Z.Y. Zhang, Z. Zhang, X. Ye, T. Sakurai, H. Lin, A BERT-based model for the prediction of lncRNA subcellular localization in Homo sapiens, Int. J. Biol. Macromol. 265 (2024) 130659.
[62] W. Zhu, S.S. Yuan, J. Li, C.B. Huang, H. Lin, B. Liao, A first computational frame for recognizing heparin-binding protein, Diagnostics 13 (2023).
[63] N.T. Pham, R. Rakkiyapan, J. Park, A. Malik, B. Manavalan, H2Opred: a robust and efficient hybrid deep learning model for predicting 2'-O-methylation sites in human RNA, Briefings Bioinf. 25 (2023).
[64] A. Malik, S. Subramaniyam, C.B. Kim, B. Manavalan, SortPred: the first machine learning based predictor to identify bacterial sortases and their classes using sequence-derived information, Comput. Struct. Biotechnol. J. 20 (2022) 165–174.
[65] X. Zou, L. Ren, P. Cai, Y. Zhang, H. Ding, K. Deng, et al., Accurately identifying hemagglutinin using sequence information and machine learning methods, Front Med (Lausanne) 10 (2023) 1281880.
[66] H. Zulfiqar, Z. Guo, R.M. Ahmad, Z. Ahmed, P. Cai, X. Chen, et al., Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings, Front Med (Lausanne) 10 (2023) 1291352.
[67] N.T. Pham, L.T. Phan, J. Seo, Y. Kim, M. Song, S. Lee, et al., Advancing the accuracy of SARS-CoV-2 phosphorylation site detection via meta-learning approach, Briefings Bioinf. 25 (2023).
[68] M.J. Sabir, M.R. Kamli, A. Atef, A.M. Alhibshi, S. Edris, N.H. Hajarah, et al., Computational prediction of phosphorylation sites of SARS-CoV-2 infection using feature fusion and optimization strategies, Methods 229 (2024) 1–8.
[69] S. Basith, M.M. Hasan, G. Lee, L. Wei, B. Manavalan, Integrative machine learning framework for the identification of cell-specific enhancers from the human genome, Briefings Bioinf. 22 (2021).
[70] N. Bupi, V.K. Sangaraju, L.T. Phan, A. Lal, T.T.B. Vo, P.T. Ho, et al., An effective integrated machine learning framework for identifying severity of tomato yellow leaf curl virus and their experimental validation, Research 6 (2023) 16.
[71] N.T. Pham, Y. Zhang, R. Rakkiyappan, B. Manavalan, HOTGpred: enhancing human O-linked threonine glycosylation prediction using integrated pretrained protein language model-based features and multi-stage feature selection approach, Comput. Biol. Med. 179 (2024) 108859.
[72] X. Fu, H. Duan, X. Zang, C. Liu, X. Li, Q. Zhang, et al., Hyb_SEnc: an antituberculosis peptide predictor based on a hybrid feature vector and stacked ensemble learning, IEEE ACM Trans. Comput. Biol. Bioinf (2024), https://doi.org/10.1109/TCBB.2024.3425644.