Methods 234 (2025) 264–274

Contents lists available at ScienceDirect

Methods

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

AntiT2DMP-Pred: Leveraging feature fusion and optimization for superior machine learning prediction of type 2 diabetes mellitus☆

Shaherin Basith [a,*], 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:
Diabetes
Machine learning
Antidiabetic peptides
Type 2 diabetes mellitus
Feature selection
Baseline models

ABSTRACT

Pancreatic α-amylase breaks down starch into isomaltose and maltose, which are further hydrolyzed by α-glucosidase in the intestine into monosaccharides, rapidly raising blood sugar levels and contributing to type 2 diabetes mellitus (T2DM). Synthetic inhibitors of carbohydrate-digesting enzymes are used to manage T2DM but may harm organ function over time. Bioactive peptides offer a safer alternative, avoiding such adverse effects. Computational methods for predicting antidiabetic peptides (ADPs) can significantly reduce the time and cost of experimental testing. While machine learning (ML) has been applied to identify ADPs, advancements in data analysis and algorithms continue to drive progress in the field. To address this, we developed AntiT2DMP-Pred, the first ML-based tool specifically designed for predicting type 2 antidiabetic peptides (T2ADPs). This tool employs a feature fusion strategy, combining ten highly discriminative feature descriptors chosen from a pool of 32 descriptors and eight ML algorithms, tested across a range of baseline models. AntiT2DMP-Pred demonstrated excellent performance, surpassing both baseline and feature-optimized models, with an accuracy (ACC) and Matthews' correlation coefficient (MCC) of 0.976 and 0.953 on the training dataset, and an ACC and MCC of 0.957 and 0.851 on the independent dataset. The web server (https://balalab-skku.org/AntiT2DMP-Pred) is freely accessible, enabling researchers worldwide to utilize it in their experimental workflows and contribute to the discovery and understanding of T2ADPs, ultimately supporting peptide-based therapeutic development for diabetes management.

1. Introduction

Consumption of excessive carbohydrates and fats without proper exercise results in an increase in blood sugar levels and other health complications, which in turn lead to high low-density lipoproteins, blood pressure, and triglycerides levels [1]. It is well known that rapid socioeconomic progress, genetic factors, and lifestyle changes over the last few decades have contributed to obesity and an increase in morbidity linked to type 2 diabetes mellitus (T2DM). T2DM has emerged as a common, chronic non-infectious disease that poses a serious public health concern [2]. This disease is primarily caused by either the inability of insulin-sensitive tissues to respond to insulin, or inadequate insulin secretion by pancreatic cells [1,3]. T2DM encompasses several dysfunctions, primarily marked by hyperglycemia, which arises from insulin resistance, insufficient insulin secretion, and excessive glucagon production. The clinical manifestations of T2DM include blurred vision, polyphagia, polyuria, polydipsia, weight loss, and lower extremity paresthesia [4]. As per the 2021 statistics from the International Diabetes Federation (IDF), around 537 million adults aged 20 to 79 are currently living with this condition. It is projected that by 2030, the total number of adults with diabetes will rise to 643 million, and by 2045, this figure is expected to increase by 46 %, meaning that 1 in 8 adults will have this disease [4].

Diabetes can be controlled by targeting the enzymes involved in carbohydrate hydrolysis, such as α-amylase, α-glucosidase, and dipeptidyl peptidase IV (DPP-IV) [5]. Thus, a number of clinical drugs have been developed to control blood glucose levels in T2DM patients by inhibiting these enzymes, though side effects or toxicity cannot be ruled out [6–8]. In order to minimize drug side effects, researchers have focused on identifying safer natural alternatives with little or no toxicity

☆ This article is part of a special issue entitled: 'Epigenetic Modification YMETH' published in Methods.

* Corresponding authors.
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.ymeth.2025.01.003
Received 31 October 2024; Received in revised form 26 December 2024; Accepted 4 January 2025
Available online 10 January 2025

for treating T2DM. There has been an increased interest in bioactive peptides derived from animal and plant proteins due to their multifaceted health benefits [9]. Bioactive peptides derived from food have been extensively studied as hypoglycemic drugs due to their safety, low side effects, and easy absorption [10]. Similarly, the peptides derived from different organisms also exert inhibitory effects on the key enzymes involved in T2DM regulation. Few of those organism-derived inhibitory peptides include LPIIDI and APGPAGP (isolated from hydrolysates of silver carp) [11], QPGR, SQSPA, QPPT, and NSPR (isolated from silkworm chrysalis) [12], and LAPSLPGKPKPD (isolated from egg york protein) [13]. However, the process of screening peptides with antidiabetic activity from crude protein hydrolysates was complex and time-consuming. Furthermore, the hypoglycemic process of these peptides could not be studied or analyzed further due to certain limitations [14]. To identify new and effective therapeutic peptides for diabetes, computational predictions of antidiabetic peptides (ADPs) can facilitate peptide-based drug discovery. In this context, bioinformatics analysis including machine learning (ML), may be crucial for predicting and identifying ADPs. With the recent advancements in data analysis and algorithm development, researchers are increasingly utilizing ML techniques to uncover novel ADPs.

Although there is a vast amount of diabetes-related data, only two ML-based methods have been created to date. AntiDMPpred is the first random forest (RF)-based predictor developed by He et al. [15] for the identification of ADPs using sequence data. In this method, the authors utilized RF and hybrid features to build the final model using 5-fold cross-validation (CV). The results from the nested 5-fold CV indicated that AntiDMPpred attained an accuracy of 77.12 % and an area under the receiver operating characteristic curve (AUC-ROC) of 0.8193, which was superior to the performance of other classifiers. Drawback of this method is the collection of negative datasets (experimentally unverified) sourced from database like AVPdb [16], which have a lower probability of including non-ADPs. Furthermore, the developed model has not been validated using an external dataset, demonstrating its lack of robustness. Recently, we developed ADP-Fuse, a two-layer ML predictor that identifies and classifies ADPs into type 1 and type 2 ADPs by utilizing multi-view information [2]. Using a multi-view feature-learning approach, ADP-Fuse extracts distinguishing characteristics from the original features, including composition, physicochemical properties, evolutionary data, and position-specific information. Classifiers and single models based on feature descriptors with excellent discriminative capabilities were systematically identified through CV. In layer 1, ADP-Fuse achieved MCC, ACC, and AUC values of 0.841, 0.940, and 0.986, respectively. In layer 2, it reached MCC, ACC, sensitivity (Sn), specificity (Sp), and AUC values of 0.858, 0.966, 0.968, 0.954, and 0.993. ADP-Fuse outperformed both single-feature models and feature fusion approaches in predicting ADPs and their classifications, as demonstrated by CV and independent testing.

Here, we introduce AntiT2DMP-Pred, an innovative prediction tool specifically designed for the accurate identification of type 2 antidiabetic peptides (T2ADPs) using sequence data. We explored various computational frameworks, including single feature (baseline), feature fusion, and feature optimization strategies, to develop the most effective model for T2ADP prediction. The top-performing feature descriptors were selected from a pool of 32 descriptors by constructing a series of baseline models. By integrating a fusion of these top descriptors with advanced ML techniques (feature fusion), AntiT2DMP-Pred achieves high accuracy and robustness in differentiating between T2ADPs and non-T2ADPs. Comprehensive evaluation and comparative analysis with baseline and feature-optimized models have demonstrated the superior generalization capability, stability, and reliability of AntiT2DMP-Pred. While other models performed well on training data, they often exhibited diminished performance on independent datasets, highlighting issues with overfitting and lack of robustness. In contrast, AntiT2DMP-Pred has consistently shown strong performance across diverse datasets, making it an invaluable resource for identifying potential peptide-based therapeutics for diabetes. By providing a user-friendly interface and delivering reliable predictions, AntiT2DMP-Pred significantly accelerates the peptide discovery process. It not only supports researchers in efficiently identifying promising T2ADPs, but also establishes a new benchmark in the computational prediction of ADPs.

2. Materials and methods

2.1. Workflow of AntiT2DMP-Pred

The AntiT2DMP-Pred workflow, illustrated in Fig. 1, includes four main stages: (1) developing a training dataset with 384 peptides (192 T2ADPs and 192 non-T2ADPs) and an independent test dataset of 324 peptides (56 T2ADPs and 268 non-T2ADPs); (2) extracting diverse features from peptide sequences, such as compositional, positional, and natural-language processing (NLP)-based features that were subjected to eight popular ML classifiers; (3) conducting a systematic evaluation across three ML frameworks (single feature, feature fusion, and feature optimization) to identify the best-performing model; and (4) creating an accessible web server for predicting T2ADPs along with their associated prediction probability score.

2.2. Dataset construction

In this study, we collected 248 T2ADPs identified using ADP-Fuse layer2 predictions [2] and 460 non-T2ADPs, which included various peptide types such as antimicrobial peptides, random functional peptides, and type 1 antidiabetic peptides (T1ADPs) [17,18]. For model development, 80 % of the dataset was allocated as the training set, while the remaining 20 % served as an independent test set to assess model performance. The training dataset was balanced, comprising 192 T2ADPs and 192 non-T2ADPs, while the independent dataset included 56 T2ADPs and 268 non-T2ADPs to ensure a comprehensive validation of predictive accuracy. All peptide samples used in this study are publicly accessible for download at https://balalab-skku.org/AntiT2DMP-Pred/download/.

2.3. Feature encodings and ML algorithms

Here, we utilized 32 feature descriptors, split evenly between 16 conventional encodings and 16 NLP-based encodings, using multiple packages [19–22]. The conventional feature descriptors included both composition-specific and position-specific encodings. Composition-based conventional feature descriptors encompassed amino acid composition (AAC), composition (CTDC), distribution (CTDD), transition (CTDT), dipeptide composition (DPC), quasi-sequence-order (QSO), amino acid index (AAI), DDE, k-spaced amino acid group pairs composition (CKSGP), KSC, and atomic composition (ATC). Position-specific based conventional feature descriptors included Bit3, Bit5, Bit7, BLOSUM62 (BLOS), and binary (BINA) encodings. For NLP-based feature descriptors, we incorporated ProtTransBertBFD (PTBB), ProtTransT5UniRef50 (PXTU), ProtTransAlbertBFD (PTAB), evolutionary scale modeling (ESM) and its variants ESM1b and ESM1v, ProtTransT5XLU50 (PTXLU), Bepler, GLOVE, FastText (FT), S2V, PRNN, PTU, PTU5, CPCP, and Word2Vec (W2V) [22]. For model development and validation, we applied eight popular ML classifiers: random forest (RF) [23], adaboost (AB) [24], gradient boost (GB) [25], extremely randomized trees (ERT) [26], extreme gradient boost (XGB) [27], light gradient boost (LGB) [28], support vector machines (SVM) [29], and catboost (CB) [30].

2.4. Baseline model construction

We utilized 32 feature descriptors, each tested individually with eight ML classifiers. We applied cross-validation (CV) to assess the models' ability to generalize to new, unseen data. While there are various CV methods available, we chose 10-randomized 10-fold CV due

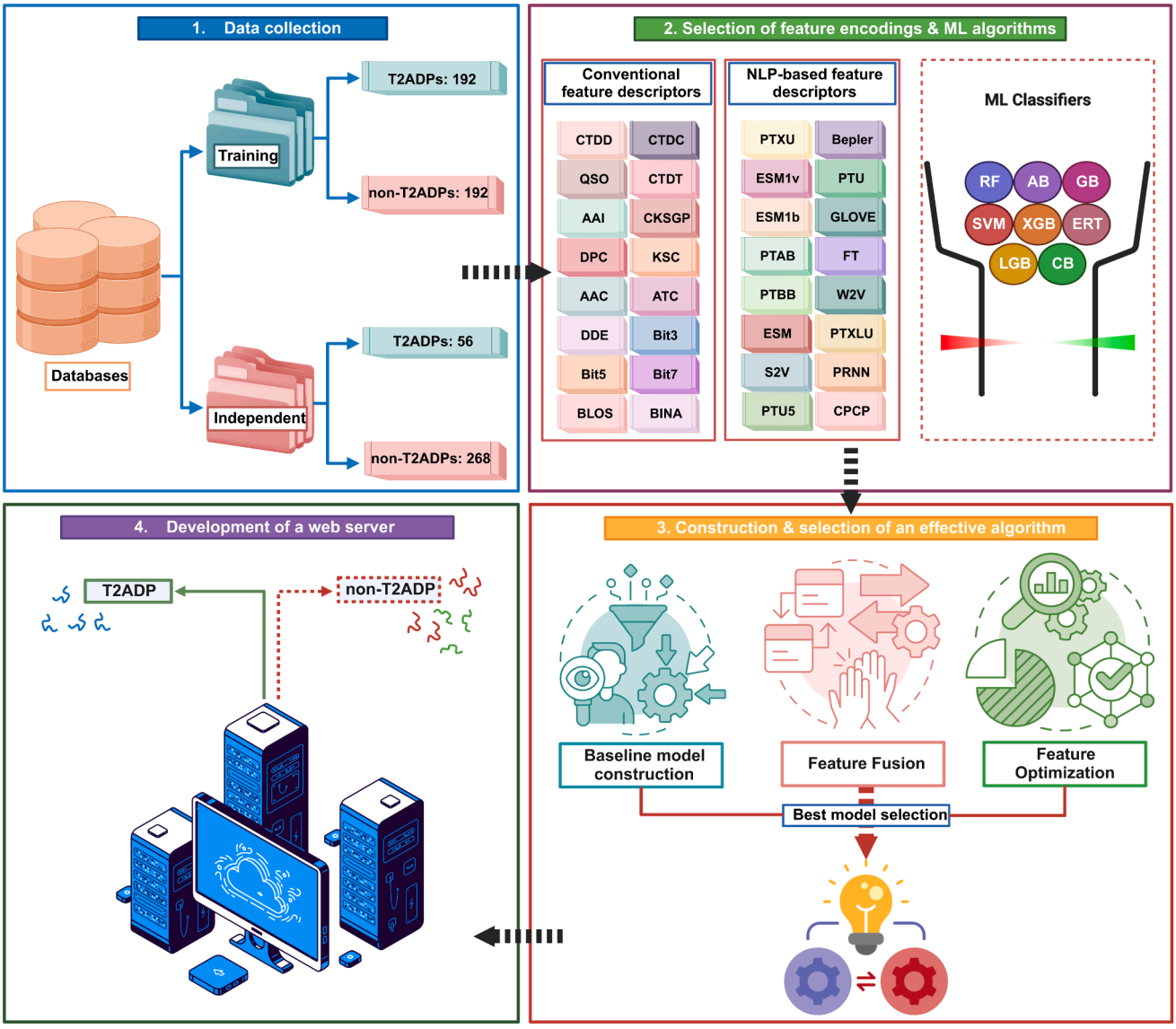

Fig. 1. Schematic workflow of AntiT2ADP-Pred. This workflow outlines the four key steps in developing AntiT2ADP-Pred: (1) Construction of a non-redundant dataset; (2) Building 256 baseline models using 32 feature descriptors and eight ML classifiers; (3) Assessing top models through three strategies: single feature, feature fusion, and integrative framework. Following thorough testing, the optimal model, AntiT2ADP-Pred, was created using a feature fusion approach; and (4) Development of a user-accessible web server.

to its benefits, such as minimizing bias, providing a more accurate estimation of performance, improving model selection, and offering robustness against outliers. To optimize each classifier's hyperparameters, we employed grid search, which systematically examines a range of hyperparameter values to identify the best settings for each model, as supported by previous research [31–36]. Baseline models were subsequently established using the median parameters obtained from the 10-randomized 10-fold CV results. Ultimately, we developed a total of 256 baseline models (32 feature descriptors × 8 classifiers) on the training dataset and evaluated their transferability using an independent test set.

2.5. Feature fusion

This approach utilizes the complementary strengths of multiple feature sets, enabling the model to capture a wider range of information, which may enhance classification accuracy and improve generalization to new data. By integrating diverse feature types, feature fusion addresses the limitations of individual descriptors, leading to stronger overall model performance. The top 10 features including DDE, FT, S2V, W2V, CPCP, ESM1v, CTDD, ESM1b, QSO, and ATC showing strong predictive capability for T2ADPs were identified based on the performance assessment of our baseline models. To assess the benefits of feature fusion, we started by creating composite feature sets that combined the first two features, then the first three, and so on, until we included all ten features. These sets were designated as FF2, FF3, FF4, FF5, FF6, FF7, FF8, FF9, and FF10, respectively. For each of these hybrid feature sets, we trained models utilizing all eight ML classifiers and optimized their hyperparameters through a comprehensive 10-randomized 10-fold CV process.

2.6. Feature optimization

Selecting the optimal feature set is crucial for developing effective and efficient predictive models. To tackle this challenge, we implemented a feature optimization strategy based on the top 10 feature encodings. ML algorithms often produce feature importance scores (FIS) to rank features, but relying solely on these scores may not accurately

reflect the true order of importance. To improve ranking precision, we applied the Iscore method [37] to select optimal models across eight different ML classifiers. Each classifier independently calculates FIS, but because some classifiers produce scores outside the 0–1 range, we normalized the FIS values across classifiers. We then calculated an average score for each feature, referred to as the Iscore. Features were then ranked based on their Iscore values, leading to the selection of the top 2000 features for further analysis. We then created ten distinct feature sets, incrementally increasing the number of top features, including Top200, Top400, Top600, Top800, Top1000, Top1200, Top1400, Top1600, Top1800, and Top2000. Each feature set was used to train eight different classifiers using a 10-randomized 10-fold CV approach and assessed their performance on independent dataset.

2.7. Performance evaluation metrics

We evaluated model performance using standard metrics, including sensitivity (Sn), specificity (Sp), area under the receiver operating characteristic curve (AUC-ROC), accuracy (ACC), and Matthews' correlation coefficient (MCC) [38–46].

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

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

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

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

where true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) represent the classification outcomes. Additionally, ROC curves and AUC values were employed to evaluate the overall performance.

3. Results and Discussion

3.1. Compositional and positional preference analysis

We conducted a compositional analysis on the combined dataset (including both training and independent datasets). Amino acid composition (AAC) analysis revealed that certain non-polar residues such as Leu, Gly, Phe, Met, Pro, Trp, and Val, and polar residues including Gln and Tyr, were particularly abundant in T2ADPs. Notably, the high prevalence of Pro in T2ADPs aligns with experimental findings that Pro-rich, rationally designed peptides exhibit strong antidiabetic properties [47,48]. Conversely, in non-T2ADPs, charged residues (Asp, Glu, His, Lys, and Arg), uncharged polar residues (Ala, Cys, Ser, and Thr), and the non-polar residue (Ile) were more prominent (Welch's *t*-test; $P \leq 0.05$) (Fig. 2A). Furthermore, dipeptide composition (DPC) analyses revealed that out of 400 dipeptides analyzed, 84 were statistically significant ($P \leq 0.01$) when comparing T2ADPs and non-T2ADPs. This indicates that nearly 21 % of all dipeptides showed differential representation between the two groups. The 15 most abundant dipeptides in T2ADPs include PF, GP, LP, PQ, FP, PL, PV, YP, PP, PG, VP, MP, IP, PI, and AG (Fig. 2B). A notable pattern is that many of these dipeptides feature Pro as one of the amino acids, highlighting the importance of this residue in T2ADPs. In contrast, non-T2ADPs showed 69 dipeptides as abundant, consisting mostly of charged or uncharged polar residues. The top 10 dipeptides in this group are RR, KK, RK, RV, AR, AS, RL, KR, AK, and GR, reflecting a composition dominated by Arg and Lys, which are positively charged residues. These significant compositional differences could enhance predictive model performance, so we incorporated them as input features for our ML models.

To examine the positional tendencies of each residue, we generated a sequence logo displaying the first five N-terminal and last five C-terminal residues for both T2ADPs and non-T2ADPs (Fig. 2B), using two-sample logo (TSL) analysis [49]. To confirm statistical significance, we scaled the height of the sequence logos (*t*-test, $P < 0.05$). The analysis revealed that Pro residues were significantly enriched at both terminal ends of T2ADPs. Other amino acids showing high occurrence in T2ADPs at the N-terminal included Leu and Trp at position 1, Met and Tyr at position 2, Glu and Trp at position 4, and Glu at position 5. At the C-terminal of T2ADPs, Val and Trp were enriched at position 7, Gln and Gly at position 8, Glu at position 9, and Phe, Leu, and Trp at position 10. In contrast, non-T2ADPs showed lower occurrence of positively charged amino acids such as Lys and Arg at both terminal ends. The negatively charged Asp residue was significantly underrepresented at positions 1 and 2 in non-T2ADPs. Additionally, amino acids Gly, Ile, Thr, Trp, Ala, Ser, and Cys were found to be generally underrepresented in non-T2ADPs. These findings indicate that certain residues, especially Pro, Leu, and Phe are favored in T2ADPs, corroborating both the AAC and DPC analysis and experimental data [47]. Understanding the specific positional preferences of residues in T2ADPs provides valuable insights for researchers designing de novo ADPs. This information allows for targeted amino acid substitutions at specific positions to enhance peptide efficacy, making this positional preference analysis a powerful tool for peptide optimization in antidiabetic applications.

3.2. Comparative analysis of conventional- and NLP-based feature encodings on different ML classifiers

We analyzed the distinctive capability of 16 NLP-based and 16 conventional-based feature encodings by feeding them into eight different ML classifiers. To evaluate predictive performance, we used 10-randomized 10-fold CV across the 32 feature encodings and eight classifiers. This resulted in 256 baseline (single feature-based) models developed using the training dataset (Table S1). Fig. 3A and B represent the overall performance of the classifiers, measured by MCC, using the 16 conventional and 16 NLP-based feature encodings, respectively. These models were further assessed on independent datasets, as shown in Fig. 3C and D (see also Table S1). Among all the models, the XGB model with DDE features showed the best performance on the training dataset, achieving an AUC of 0.991, MCC of 0.935, ACC of 0.966, Sn of 0.953, and Sp of 0.979. On the independent dataset, this model achieved AUC, MCC, ACC, Sn, and Sp values of 0.978, 0.802, 0.938, 0.911, and 0.944, respectively.

To assess the effectiveness of each feature in differentiating between T2ADPs and non-T2ADPs, we calculated the average training performance of eight classifiers using MCC evaluation metrics. Fig. 3A shows that the top ten feature encodings including DDE, FT, S2V, W2V, CPCP, ESM1v, CTDD, ESM1b, QSO, and ATC exhibit superior discriminatory power in distinguishing T2ADPs from non-T2ADPs, with an average MCC ranging from 88.60 % to 91.12 %. In contrast, the five position-specific based conventional feature descriptors (BINA, Bit3, Bit5, Bit7, and BLOS) show limited capability in distinguishing between the two classes, resulting in an average MCC of less than 50 %. The other conventional-based descriptors (composition) exhibit superior discrimination, with an average MCC falling between 75.86 % to 91.12 %. Meanwhile, NLP-based feature encodings show comparable performance to composition-based conventional feature descriptors, with an average MCC values ranging from 76.27 % to 90.84 %. In contrast, several feature descriptors did not demonstrate superior performance on the independent dataset, with some feature encodings showing inconsistent or suboptimal results (Table S1 and Fig. 3C and D). To address this inconsistency and underperformance, we aimed to improve the model's robustness by exploring alternative approaches. Specifically, we investigated two computational frameworks: feature fusion and feature optimization, to further improve model performance.

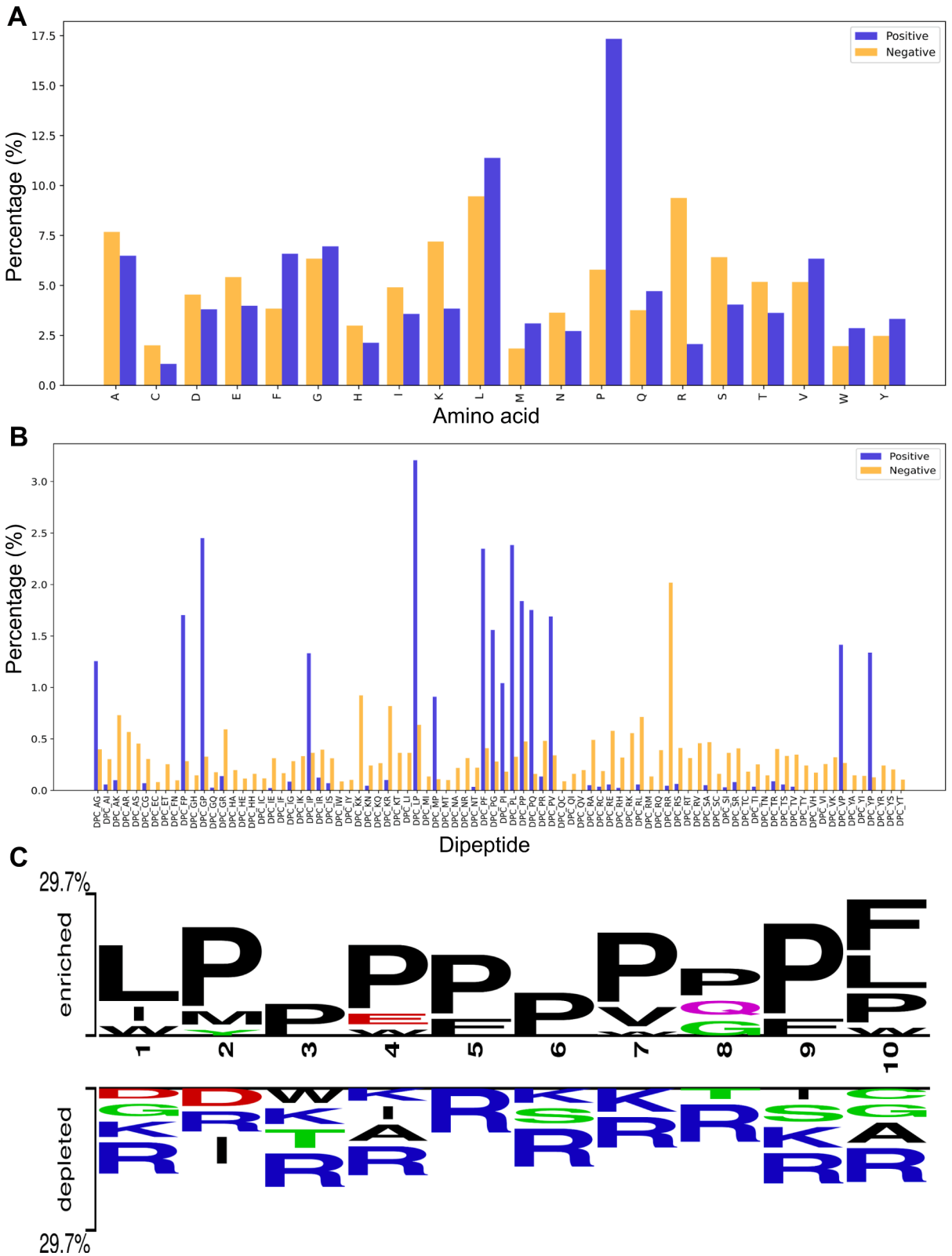

Fig. 2. Compositional and positional preference analysis. (A) and (B) represent the amino acid and dipeptide preferences between T2ADPs and non-T2ADPs. (C) Depicts positional conservation of first five residues at N-terminal end and last five residues at the C-terminal end between T2ADPs and non-T2ADPs, respectively, generated using two sample logos.

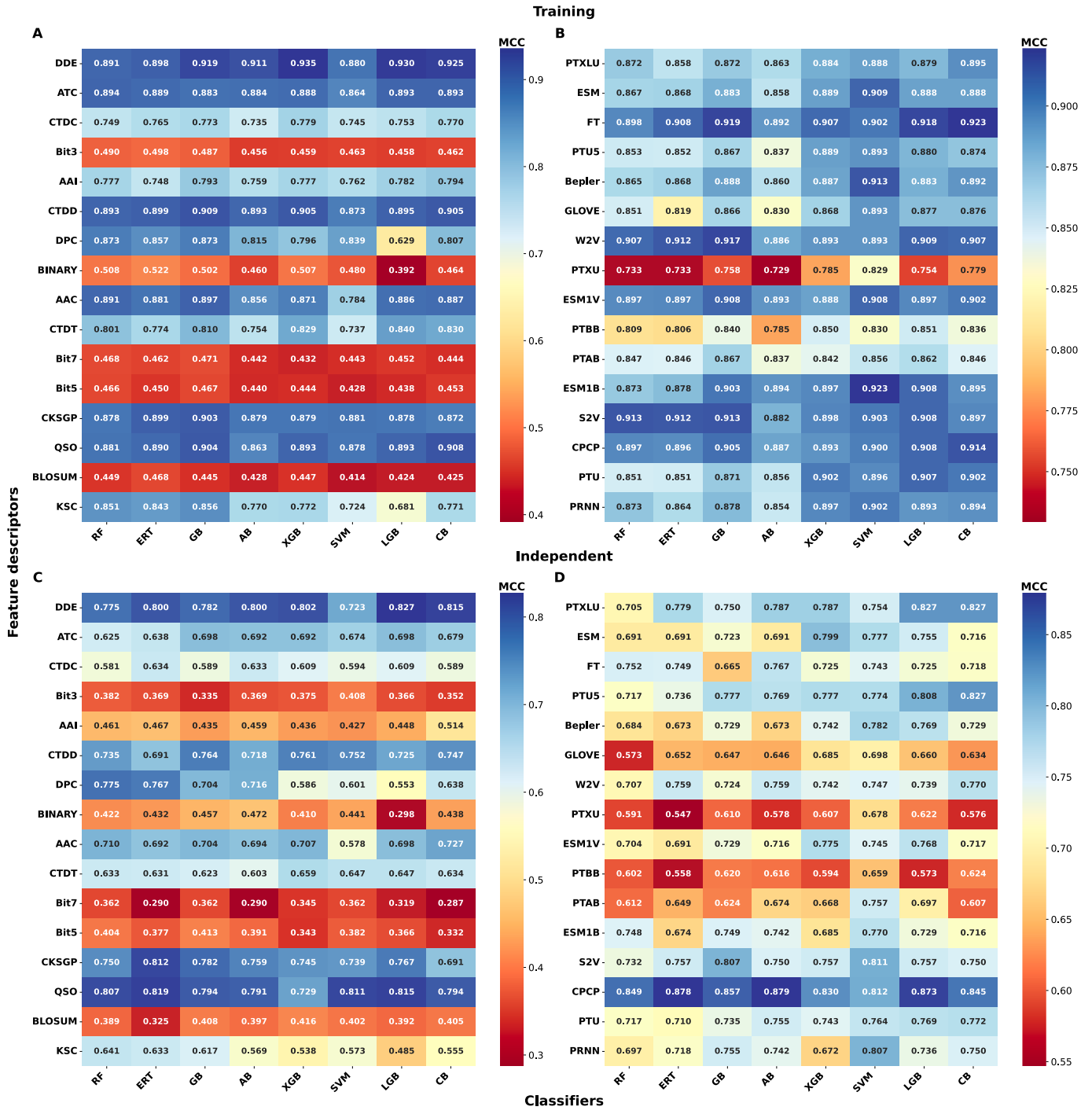

Fig. 3. Performance comparison of baseline models for each feature descriptor. The evaluation of baseline models for each descriptor during training is presented using Matthews' correlation coefficient (MCC). This includes the MCC metrics for single-feature models based on (A) conventional feature descriptors and (B) NLP-based feature descriptors across all eight ML classifiers. Additionally, the MCC performance evaluation for the independent testing is displayed for (C) conventional feature descriptors and (D) NLP-based feature descriptors across all eight ML classifiers.

3.3. Construction of AntiT2DMP-Pred using feature fusion

We implemented feature fusion, an approach that combines two or more distinct features in a linear fashion to potentially enhance the model's predictive capability. Considering that the top ten feature encodings such as DDE, FT, S2V, W2V, CPCP, ESM1v, CTDD, ESM1b, QSO, and ATC demonstrate strong discriminative power in differentiating between T2ADPs and non-T2ADPs, we investigated their combined potential. We trained eight different ML classifiers on various sets of hybrid features, ranging from the top 2 to the top 10 encodings (Top2 to Top10), and assessed their performance on an independent dataset (see Fig. 4A and B). As illustrated in Table S2, feature fusion (hybrid) models based on boosting algorithms, such as CB, GB, XGB, and LGB, consistently delivered superior performance across various feature fusion (FF2 – FF10) combinations. Notably, the Top4 CB-based feature fusion model stood out with its outstanding results on the training dataset, achieving an AUC of 0.992, MCC of 0.944, ACC of 0.971, Sn of 0.953, and Sp of 0.999 (Fig. 5A). When tested on the independent dataset, the model continued to demonstrate strong performance, with AUC, MCC, ACC, Sn, and Sp values of 0.970, 0.851, 0.957, 0.893, and

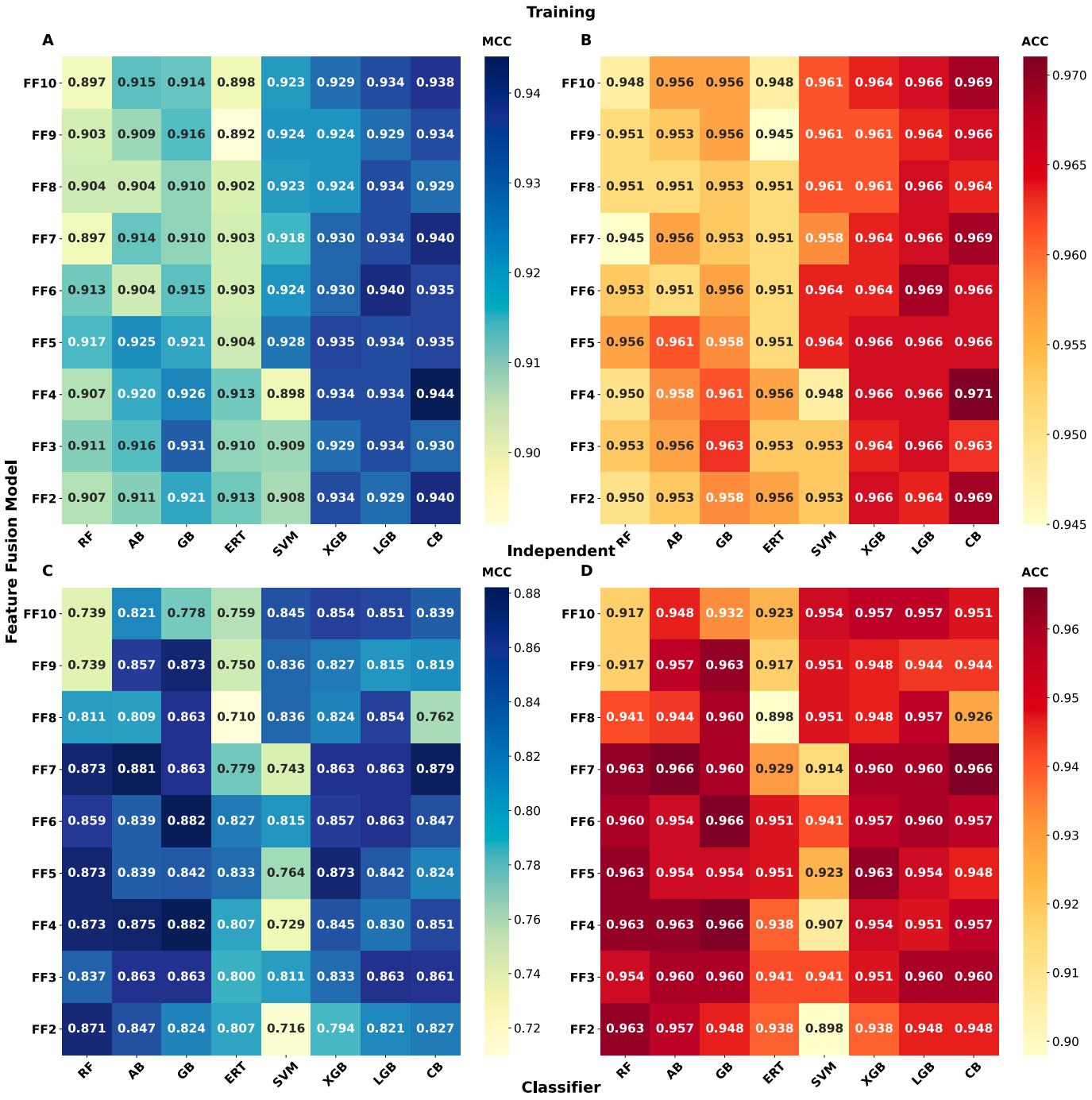

Fig. 4. Performance evaluation of AntiT2ADP-Pred with feature fusion models (FF2 to FF10) developed through a feature fusion approach. The models were evaluated using Matthews' correlation coefficient (MCC) and accuracy (ACC) metrics during training (A and B) and independent testing (C and D) across all ML classifiers.

0.970 (Fig. 5B), respectively.

Compared to the top-performing baseline model, the Top4 CB-based feature fusion model exhibited a notable increase in MCC and ACC by 0.9 % and 0.5 %, respectively, on the training dataset. Furthermore, when evaluated on the independent dataset, the Top4-based model showed even more significant improvements, achieving gains of 4.9 % in MCC and 1.9 % in ACC, highlighting its robustness and effectiveness in handling unseen data. These results emphasize the model's ability to generalize well and outperform other approaches across different datasets and feature fusion strategies. Given its consistent and exceptional performance across both the training and independent datasets, the Top4 CB-based feature fusion model significantly outperformed other models and established itself as the top-performing approach. As a result, this model was selected as the final model for our study and formally named it as AntiT2DMP-Pred.

3.4. *Effect of feature optimization on model robustness*

While feature fusion models generally outperformed the top baseline models, some inconsistencies were noted between the training and testing datasets. Specifically, certain feature fusion models demonstrated superior performance on the training dataset but exhibited diminished or inconsistent results when evaluated on the testing dataset. This discrepancy may be attributed to issues such as feature redundancy

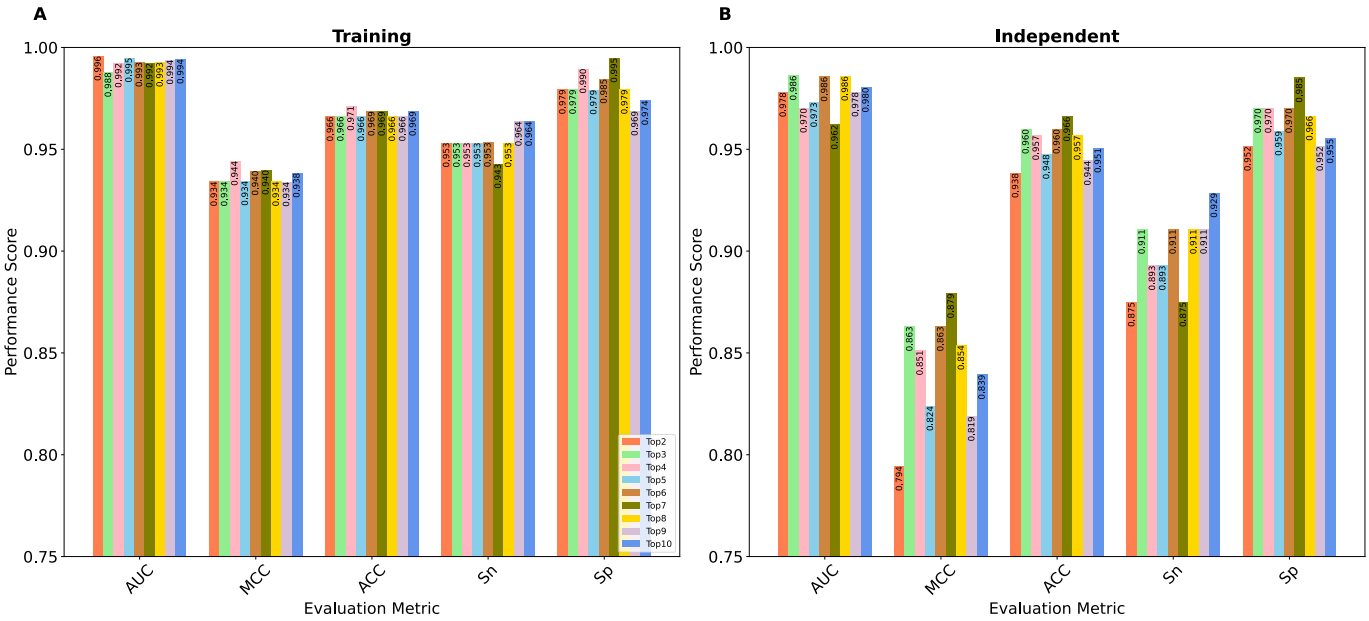

Fig. 5. Performance evaluation of AntiT2ADP-Pred (Top4) alongside the top-performing eight feature fusion models (Top2, Top3, Top5-Top10), developed using a feature fusion approach. The assessment was conducted during (A) training and (B) independent testing. The models were evaluated using various metrics, including AUC, MCC, ACC, sensitivity (Sn), and specificity (Sp).

or inconsistency, which can lead to overfitting and hinder the model's ability to generalize effectively to unseen data. To optimize feature selection for predictive modeling, we implemented a strategy based on the top 10 feature encodings. While ML algorithms provide feature importance scores (FIS) to rank features, relying solely on these can be imprecise. To improve accuracy, we applied the Iscore method [37] across eight ML classifiers. Each classifier's FIS was normalized to a 0–1 range, then averaged to calculate an Iscore for each feature. This ranking identified the top 2000 features for further analysis. We then created ten feature sets, incrementally increasing from the top 200 to the top 2000 features, in steps of 200. Each set trained eight classifiers using a randomized 10-fold CV approach, with performance evaluated on an independent dataset.

As illustrated in Fig. 6A, the Top200 LGB-based model demonstrated outstanding performance on the training dataset, achieving an AUC of 0.998, MCC of 0.949, ACC of 0.974, Sn of 0.959, and Sp of 0.990. However, when evaluated on the independent dataset, the model's performance declined to more moderate levels, with AUC, MCC, ACC, Sn, and Sp values of 0.979, 0.824, 0.948, 0.893, and 0.959, respectively (Fig. 6B). In comparison to the AntiT2DMP-Pred model, the Top200 LGB-based optimized model achieved slight improvements in MCC and ACC on the training dataset, increasing by 0.05 % and 0.03 %, respectively. However, the performance on the independent dataset showed a noticeable decline, with MCC and ACC dropping by 2.7 % and 1.0 %, respectively. This suggests that the Top200 LGB-based model may be more prone to overfitting and struggles to generalize effectively to

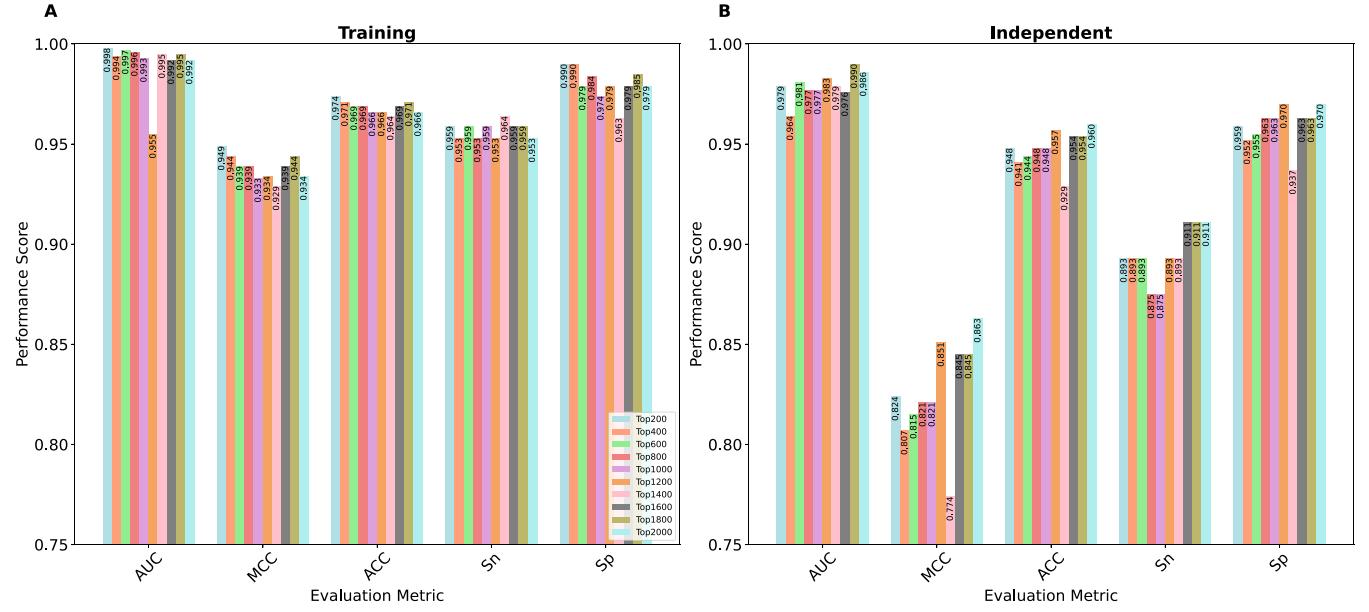

Fig. 6. Performance evaluation of the top-performing feature optimization models (Top200-Top2000) developed through a feature optimization approach. The assessment was conducted during (A) training and (B) independent testing. The models were evaluated using various metrics, including AUC, MCC, ACC, sensitivity (Sn), and specificity (Sp).

unseen data. Moreover, several other feature-optimized models failed to achieve superior performance on the independent dataset. Despite showing promising results during training, many of these models exhibited inconsistencies or suboptimal performance when applied to the independent data (Table S3). This highlights the challenge of selecting feature sets that balance both training and independent dataset performance, emphasizing the need for robust optimization strategies to ensure model generalizability and reliability. The Top4 CB-based feature fusion model a.k.a. AntiT2DMP-Pred model demonstrated consistent and exceptional performance across all datasets, outperforming the other models. As a result, AntiT2DMP-Pred was selected as the final approach for our study.

3.5. Comparison of AntiT2DMP-Pred with the leading baseline and feature-optimized models

As illustrated in Fig. 7, the AntiT2DMP-Pred model outperformed both the leading baseline and feature-optimized models across multiple performance metrics, demonstrating its robustness and superior predictive capability. On the training dataset, AntiT2DMP-Pred achieved a higher overall ACC and exhibited greater consistency in results, with notable improvements in metrics such as MCC and ACC compared to the top baseline models. Although the Top200 LGB-based feature-optimized model displayed excellent results on the training dataset reaching an MCC of 0.949 and ACC of 0.974, its performance gains were relatively marginal when compared to AntiT2DMP-Pred. Specifically, AntiT2DMP-Pred demonstrated MCC and ACC values of 0.953 and 0.976, respectively, showing slight improvements over the Top200 LGB-based model (Fig. 7A).

More importantly, AntiT2DMP-Pred maintained consistent performance and demonstrated strong generalization capabilities across both the training and independent datasets. When evaluated on the independent dataset, AntiT2DMP-Pred achieved an MCC of 0.851 and ACC of 0.957 (Fig. 7B). In contrast, the top baseline model showed a considerable drop in performance on the independent dataset, with an MCC of 0.803 and ACC of 0.940, indicating that single feature-based baseline models lacked robustness and struggled to generalize effectively. Similarly, several feature-optimized models, including the Top200 LGB-based model, experienced notable declines in performance metrics on the independent dataset, with MCC and ACC dropping to 0.824 and 0.948, respectively. This decline suggests that these models may have encountered issues with feature redundancy or inconsistency, affecting their ability to generalize beyond the training data.

Overall, the AntiT2DMP-Pred model demonstrated superior stability and effectiveness, achieving balanced performance across both the training and independent datasets. In contrast, the top baseline and feature-optimized models, while competitive on the training dataset, often displayed inconsistencies and reduced robustness when applied to independent data. This underscores the advantages of AntiT2DMP-Pred in real-world applications, where model generalizability and reliability are critical.

3.6. Webserver construction

Previously, we have made all our peptide prediction tools publicly accessible through web servers, providing researchers with a valuable resource for conducting experimental analyses [2,45,50–59]. By offering these tools online, we aim to support the scientific community by simplifying access to computational models and enabling researchers to perform peptide predictions efficiently. To facilitate the identification of T2ADPs, we have created an online web server called AntiT2DMP-Pred, which is freely available at https://balalab-skku.org/AntiT2DMP-Pred. Users can input their sequences in FASTA format by either pasting them into the designated text field or uploading them through the file upload option. The server then processes the input and delivers predictions in a user-friendly table format with four columns: serial number, FASTA ID, predicted class (T2ADP or non-T2ADP), and the probability score for T2ADP prediction. The probability score (PS) ranges from 0 to 1, where values nearer to 1 suggest a higher probability of the sequence being classified as T2ADP, thus providing an intuitive measure of prediction confidence. For comprehensive instructions on using the AntiT2DMP-Pred web server, users are encouraged to refer to the help page available on the website.

4. Conclusions

The computational prediction of ADPs plays a crucial role in the discovery of new and potent therapeutic peptides for treating diabetes. By leveraging computational approaches, researchers can streamline peptide-based drug discovery, reducing the time and cost associated

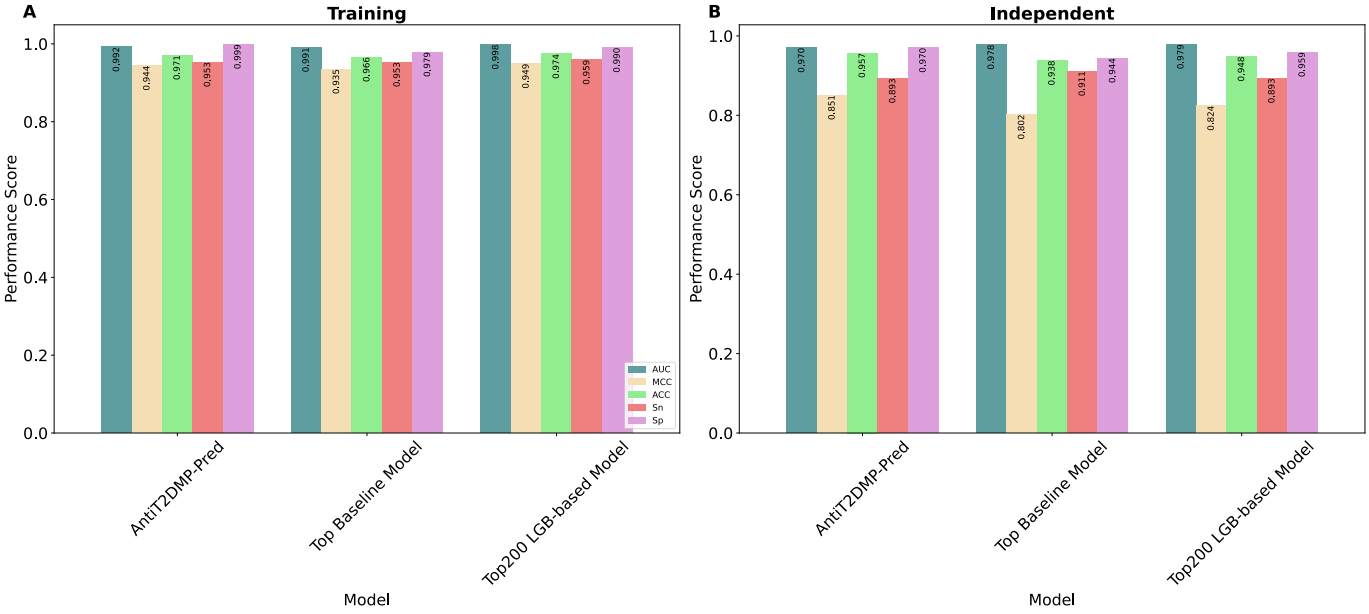

Fig. 7. A direct performance comparison of AntiT2ADP-Pred with the leading baseline and feature optimization models during (A) training and (B) independent testing. The evaluation utilized metrics such as AUC, MCC, ACC, sensitivity (Sn), and specificity (Sp).

with experimental procedures. Although ML techniques are actively being explored for identifying potential ADPs [2,15], this area of research is still evolving, driven by ongoing advancements in data analysis and algorithm development. To date, there are only two publicly available ML-based prediction methods for the general identification of ADPs including AntiDMPpred [15] and ADP-Fuse [2]. These tools provide valuable frameworks for predicting ADPs, however, the limited number of established methods highlights the need for continued innovation and expansion of computational models to better serve the research community.

AntiT2DMP-Pred serves as the first method that has been exclusively developed for the identification of T2ADPs. By integrating advanced ML algorithms and fusion of top feature descriptors, AntiT2DMP-Pred offers high accuracy and robustness in distinguishing between T2ADPs and non-T2ADPs. The comparative analysis with baseline and feature-optimized models demonstrates the superior generalization capability and consistency of AntiT2DMP-Pred, making it a valuable resource for bioinformatics and biomedical research. The freely accessible web server further broadens the scope of application, allowing researchers worldwide to leverage this tool in their experimental studies and contribute to the understanding of T2ADP-related mechanisms. As the field progresses, there is significant potential for the development of more sophisticated models that incorporate diverse data types and feature engineering techniques, thereby enhancing the accuracy and applicability of ADP predictions. Such advancements will enable the systematic identification of therapeutic peptides with high efficacy and minimal side effects, ultimately contributing to the development of novel peptide-based treatments for diabetes and related metabolic disorders.

Author contributions

SB collected and processed the datasets. SB implemented the algorithms and developed the prediction models. SB analyzed the results. BM created the back-end and front-end user interface of the web server. SB, BM, and GL performed the writing, reviewing and draft preparation of the manuscript. BM and GL conceived and coordinated the project. All authors have read and approved the final manuscript.

Funding

This work was supported by grants from the National Research Foundation (NRF), funded by the Ministry of Science and ICT (MSIT) in Korea (RS-2022-NR075341, 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ymeth.2025.01.003.

Data availability

Data will be made available on request.

References

[1] W.Q. Al-Bukhaiti, S. Al-Dalali, H. Li, L. Yao, S.M. Abed, L. Zhao, S.X. Qiu, Identification and in vitro Characterization of Novel Antidiabetic Peptides Released Enzymatically from Peanut Protein, Plant Foods Hum. Nutr. 79 (1) (2024) 66–72.
[2] 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.
[3] U. Galicia-Garcia, A. Benito-Vicente, S. Jebari, A. Larrea-Sebal, H. Siddiqi, K. B. Uribe, H. Ostolaza, C. Martin, Pathophysiology of Type 2 Diabetes Mellitus, Int. J. Mol. Sci. 21 (17) (2020).
[4] N.H. Cho, J.E. Shaw, S. Karuranga, Y. Huang, J.D. da Rocha Fernandes, A.W. Ohlrogge, B. Malanda, IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045, Diabetes Res Clin Pract 138 (2018) 271-281.
[5] V.A. Mustad, D.T.T. Huynh, J.M. Lopez-Pedrosa, C. Campoy, R. Rueda, The Role of Dietary Carbohydrates in Gestational Diabetes, Nutrients 12 (2) (2020).
[6] G. Grunberger, Should Side Effects Influence the Selection of Antidiabetic Therapies in Type 2 Diabetes? Curr. Diab. Rep. 17 (4) (2017) 21.
[7] I.A. Harsch, R.H. Kaestner, P.C. Konturek, Hypoglycemic side effects of sulfonylureas and repaglinide in ageing patients - knowledge and self-management, J. Physiol. Pharmacol. 69 (4) (2018).
[8] M. Zhou, G. Ren, B. Zhang, F. Ma, J. Fan, Z. Qiu, Screening and identification of a novel antidiabetic peptide from collagen hydrolysates of Chinese giant salamander skin: network pharmacology, inhibition kinetics and protection of IR-HepG2 cells, Food Funct. 13 (6) (2022) 3329–3342.
[9] Y. Hou, Z. Wu, Z. Dai, G. Wang, G. Wu, Protein hydrolysates in animal nutrition: Industrial production, bioactive peptides, and functional significance, J. Anim. Sci. Biotechnol. 8 (2017) 24.
[10] B.A. Kehinde, P. Sharma, Recently isolated antidiabetic hydrolysates and peptides from multiple food sources: a review, Crit. Rev. Food Sci. Nutr. 60 (2) (2020) 322–340.
[11] Y. Zhang, R. Chen, X. Chen, Z. Zeng, H. Ma, S. Chen, Dipeptidyl Peptidase IV-Inhibitory Peptides Derived from Silver Carp (Hypophthalmichthys molitrix Val.) Proteins, J. Agric. Food Chem. 64 (4) (2016) 831–839.
[12] Y. Zhang, N. Wang, W. Wang, J. Wang, Z. Zhu, X. Li, Molecular mechanisms of novel peptides from silkworm pupae that inhibit alpha-glucosidase, Peptides 76 (2016) 45–50.
[13] A. Zambrowicz, M. Pokora, B. Setner, A. Dabrowska, M. Szoltysik, K. Babij, Z. Szewczuk, T. Trziszka, G. Lubec, J. Chrzanowska, Multifunctional peptides derived from an egg yolk protein hydrolysate: isolation and characterization, Amino Acids 47 (2) (2015) 369–380.
[14] P. Wan, B. Cai, H. Chen, D. Chen, X. Zhao, H. Yuan, J. Huang, X. Chen, L. Luo, J. Pan, Antidiabetic effects of protein hydrolysates from Trachinotus ovatus and identification and screening of peptides with alpha-amylase and DPP-IV inhibitory activities, Curr. Res. Food Sci. 6 (2023) 100446.
[15] X. Chen, J. Huang, B. He, AntiDMPpred: a web service for identifying anti-diabetic peptides, PeerJ 10 (2022) e13581.
[16] A. Qureshi, N. Thakur, H. Tandon, M. Kumar, AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses, Nucleic Acids Res. 42 (Database issue) (2014) D1147.
[17] X. Kang, F. Dong, C. Shi, S. Liu, J. Sun, J. Chen, H. Li, H. Xu, X. Lao, H. Zheng, DRAMP 2.0, an updated data repository of antimicrobial peptides, Sci. Data 6 (1) (2019) 148.
[18] G. Shi, X. Kang, F. Dong, Y. Liu, N. Zhu, Y. Hu, H. Xu, X. Lao, H. Zheng, DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides, Nucleic Acids Res 50(D1) (2022) D488-D496.
[19] Z. Chen, P. Zhao, F. Li, T.T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D.R. Powell, T. Akutsu, G.I. Webb, K.C. Chou, A.I. Smith, R.J. Daly, J. Li, J. Song, iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and Protein Sequence Data, Brief Bioinform 21 (3) (2020) 1047–1057.
[20] Z. Yan, F. Ge, Y. Liu, Y. Zhang, F. Li, J. Song, D.J. Yu, TransEFVP: A Two-Stage Approach for the Prediction of Human Pathogenic Variants Based on Protein Sequence Embedding Fusion, J. Chem. Inf. Model. 64 (4) (2024) 1407–1418.
[21] Z. Chen, X. Liu, P. Zhao, C. Li, Y. Wang, F. Li, T. Akutsu, C. Bain, R.B. Gasser, J. Li, Z. Yang, X. Gao, L. Kurgan, J. Song, iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets, Nucleic Acids Res 50(W1) (2022) W434-W447.
[22] C. Dallago, K. Schutze, M. Heinzinger, T. Olenyi, M. Littmann, A.X. Lu, K.K. Yang, S. Min, S. Yoon, J.T. Morton, B. Rost, Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets, Curr Protoc 1 (5) (2021) e113.
[23] L.J.M.l. Breiman, Random forests, 45 (2001) 5-32.
[24] F. Yoav, R.E.J.C. Schapire, A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting, 10 (1.56) (1995) 9855.
[25] J.H. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 5 (2001) 1189–1232.
[26] P. Geurts, D. Ernst, L.J.M.l. Wehenkel, Extremely randomized trees, 63 (2006) 3-42.
[27] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[28] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: a highly efficient gradient boosting decision tree, Nips 17 (2017) 3149–3157.
[29] C.J.M.L. Cortes, Support-Vector Networks, (1995).
[30] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, Nips 18 (2018) 6639–6649.

[31] H.W. Park, T. Pitti, T. Madhavan, Y.-J. Jeon, B.J.C. Manavalan, S.B. Journal, MLACP 2.0: An updated machine learning tool for anticancer peptide prediction, 20 (2022) 4473-4480.
[32] 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.
[33] B. Manavalan, S. Basith, T.H. Shin, G. Lee, Computational prediction of species-specific yeast DNA replication origin via iterative feature representation, Brief Bioinform 22 (4) (2021).
[34] B. Manavalan, M.C. Patra, MLCPP 2.0: An Updated Cell-penetrating Peptides and Their Uptake Efficiency Predictor, J Mol Biol 434 (11) (2022) 167604.
[35] 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 (11) (2020) 3350–3356.
[36] 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.
[37] N.T. Pham, L.T. Phan, J. Seo, Y. Kim, M. Song, S. Lee, Y.-J. Jeon, B. Manavalan, Advancing the accuracy of SARS-CoV-2 phosphorylation site detection via meta-learning approach, Brief. Bioinform. 25 (1) (2023).
[38] 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, Brief Bioinform 25 (1) (2023).
[39] 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.
[40] X. Zou, L. Ren, P. Cai, Y. Zhang, H. Ding, K. Deng, X. Yu, H. Lin, C. Huang, Accurately identifying hemagglutinin using sequence information and machine learning methods, Front Med (Lausanne) 10 (2023) 1281880.
[41] H. Zulfiqar, Z. Guo, R.M. Ahmad, Z. Ahmed, P. Cai, X. Chen, Y. Zhang, H. Lin, Z. Shi, Deep-STP: a deep learning-based approach to predict snake toxin proteins by using word embeddings, Front Med (Lausanne) 10 (2023) 1291352.
[42] B. Shaherin, S. Vinoth Kumar, M. Balachandran, L. Gwang, mHPpred: Accurate identification of peptide hormones using multi-view feature learning, Computers in Biology and Medicine 183 (2024) 109297.
[43] B. Shaherin, P. Nhat Truong, M. Balachandran, L. Gwang, SEP-AlgPro: An efficient allergen prediction tool utilizing traditional machine learning and deep learning techniques with protein language model features, International Journal of Biological Macromolecules 273 (2024) 133085.
[44] 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 (11) (2022) 167549.
[45] B. Manavalan, S. Basith, G. Lee, Comparative analysis of machine learning-based approaches for identifying therapeutic peptides targeting SARS-CoV-2, Brief. Bioinform. 23 (1) (2022).
[46] S. Basith, G. Lee, B. Manavalan, STALLION: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction, Brief. Bioinform. 23 (1) (2022).
[47] B. Vivesh, S. Kaur, S. Jaglan, Y. Rani, P. Batra, Singh, Proline based rationally designed peptide esters against dipeptidyl peptidase-4: Highly potent anti-diabetic agents, Bioorg. Med. Chem. Lett. 76 (2022) 129018.
[48] F. Rivero-Pino, F.J. Espejo-Carpio, E.M. Guadix, Antidiabetic Food-Derived Peptides for Functional Feeding: Production, Functionality and In Vivo Evidences, Foods 9 (8) (2020).
[49] V. Vacic, L.M. Iakoucheva, P. Radivojac, Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments, Bioinformatics 22 (12) (2006) 1536–1537.
[50] S. Basith, N.T. Pham, B. Manavalan, G. Lee, SEP-AlgPro: An efficient allergen prediction tool utilizing traditional machine learning and deep learning techniques with protein language model features, Int. J. Biol. Macromol. 273 (Pt 2) (2024) 133085.
[51] 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, Brief. Bioinform. 22 (6) (2021).
[52] S. Basith, B. Manavalan, T.H. Shin, D.Y. Lee, G. Lee, Evolution of Machine Learning Algorithms in the Prediction and Design of Anticancer Peptides, Curr. Protein Pept. Sci. 21 (12) (2020) 1242–1250.
[53] 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 (4) (2020) 1276–1314.
[54] 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 (11) (2020) 3350–3356.
[55] 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 (16) (2019) 2757–2765.
[56] B. Manavalan, S. Basith, T.H. Shin, L. Wei, G. Lee, AtbPpred: A Robust Sequence-Based Prediction of Anti-Tubercular Peptides Using Extremely Randomized Trees, Comput Struct, Biotechnol. J. 17 (2019) 972–981.
[57] S. Basith, B. Manavalan, T.H. Shin, G. Lee, iGHBP: Computational identification of growth hormone binding proteins from sequences using extremely randomised tree, Comput Struct, Biotechnol. J. 16 (2018) 412–420.
[58] 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.
[59] B. Manavalan, S. Basith, T.H. Shin, S. Choi, M.O. Kim, G. Lee, MLACP: machine-learning-based prediction of anticancer peptides, Oncotarget 8 (44) (2017) 77121–77136.