ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

TCellPredX: A Novel Approach for Accurate Prediction of Hepatitis C Virus Linear T Cell Epitopes

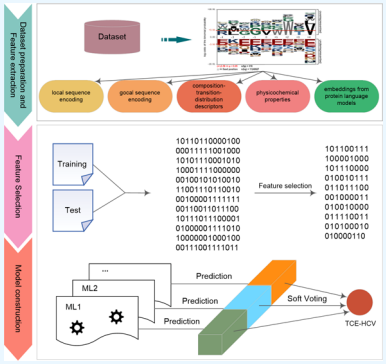
Fang Ge,* Hao-Yang Li, Ming Zhang, Muhammad Arif, and Tanvir Alam*

Cite This: *ACS Omega* 2024, 9, 51494−51507

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Hepatitis C Virus (HCV) is a bloodborne RNA virus that leads to severe liver diseases, and currently, no effective prophylactic biologics are available to prevent its transmission. The prevention of HCV is closely related to the major histocompatibility complex (MHC). Linear antigenic peptides of HCV, known as T cell epitopes (TCEs), are crucial in the presentation process by MHC molecules to T cells, playing a key role in immune responses. Therefore, the rapid and accurate identification of these TCE−HCVs is essential for advancing vaccine development. Herein, we propose TCellPredX, a novel integrated predictor for TCE−HCV identification. TCellPredX leverages five distinct feature encoding schemes, including local and global sequence encodings, composition-transition-distribution descriptors, physicochemical properties, and embeddings from two protein language models, which are processed through 12 machine learning algorithms. Our results indicate that feature fusion significantly enhances predictive accuracy. Moreover, the maximal relevance minimal redundancy feature selection method is particularly effective in isolating informative features, ensuring the model's use of the most informative data. Additionally, ensemble models, especially when combined with an averaged voting strategy, demonstrate superior stability and accuracy compared to individual classifiers, effectively reducing noise and enhancing model robustness. TCellPredX achieves notable accuracies of 0.900 and 0.897 in 10-fold cross-validation and independent test, respectively. Furthermore, TCellPredX's high accuracy is validated on experimentally verified peptide sequences documented for their potential benefits in vaccine development. Overall, TCellPredX can offer a robust tool for the precise identification of TCE−HCV, potentially serving as a cornerstone for future epitope research and advancing HCV vaccines development.



1. INTRODUCTION

Hepatitis C virus (HCV), classified under the genus Hepacivirus within the Flaviviridae family, is a single-stranded, positive-sense, enveloped RNA virus.[1,2] Due to the lack of proofreading function in its RNA-dependent RNA polymerase, HCV rapidly evolves into millions of quasispecies.[3] In recent years, there have been substantial advances in the pharmacological treatment of HCV, particularly with the development of "direct-acting antivirals" that target the viral replication machinery, leading to significant breakthroughs in HCV treatment. However, despite these advancements, nearly 80% of HCV-infected individuals worldwide remain undiagnosed and without access to affordable treatment,[4] making the World Health Organization's goal of eliminating HCV infections by 2030 increasingly challenging. Therefore, developing an effective vaccine remains the most viable strategy for controlling HCV-related diseases.[5] Traditional HCV vaccine designs have shown limited success, proving to be costly and inadequate in addressing pathogens with high antigenic diversity.[6]

The nucleotide sequence of the Hepatitis C virus varies by over 30% between different genotypes and over 15% between subtypes.[7] These genetic differences are particularly pronounced across various geographic regions, making it imperative for vaccine design to either induce broad immune responses or target highly conserved regions of the viral genome.[8] Numerous studies have shown that immune responses targeting these conserved regions may be effective against multiple HCV strains, providing broader protection. Although initial progress has been made in developing vaccines that induce neutralizing antibodies against the envelope glycoproteins E1/E2, the heterogeneity of HCV significantly impacts the broad-spectrum protective efficacy of the vaccine, posing a substantial challenge for vaccine development.[9] In contrast, vaccines focusing on specific antigenic peptide segments presented by major histocompatibility complex (MHC) molecules and their impact on T cells show great potential.

Within the immune response, $CD8^+$ T cells are the principal effectors in controlling viral infections, whereas $CD4^+$ T cells

Received: September 23, 2024
Revised: November 29, 2024
Accepted: December 4, 2024
Published: December 16, 2024

are essential for sustaining $CD8^+$ T cell functionality and preventing viral escape mutations within $CD8^+$ T cell epitopes.[10] The HCV has developed multiple strategies to disrupt the antigen presentation process, thereby evading the host's immune system, particularly T cell-mediated responses.[11] MHC molecules are crucial for presenting antigens to T cells, thereby initiating and orchestrating immune reactions.[12] MHC molecules are categorized into two classes based on their antigen presentation pathways: MHC class I and MHC class II. MHC class I molecules present endogenous antigen peptides, activating $CD8^+$ T cells, which leads to the elimination of infected cells.[13] In contrast, MHC class II molecules present exogenous antigen peptides to $CD4^+$ T cells, activating helper T cells that support B cell antibody production and establish long-term immune memory.[14]

Accurately identifying linear T cell epitopes of the Hepatitis C virus can enhance the binding efficiency of MHC molecules, facilitating effective antigen presentation, which is crucial for activating $CD8^+$ T cells and helper T cells, thereby stimulating, regulating, and forming memory within the immune system. Although significant progress has been made in the bioinformatics research of T cell epitopes TCE−HCV, clinical trials are needed to validate these findings. In this context, our goal is to accurately identify TCE−HCV using only sequence information, thereby improving the antigen presentation efficiency of MHC molecules and providing new directions for vaccine development.

Machine learning methods have been extensively employed in the prediction of linear T-cell epitopes for HCV. Phasit et al. introduced the TROLLOPE method, which combines 12 features with 12 machine learning models to produce 144 base classifiers that generate a feature set called APF, subsequently refined using the GA-SAR feature selection algorithm and applied to build the final classifier via PLS.[15] Despite its innovative approach, TROLLOPE[15] faces significant limitations in predictive accuracy, largely due to its lack of comprehensive feature fusion and model integration. This shortcoming affects both its overall performance and stability. Moreover, TROLLOPE's reliance on a single model, without testing the benefits of model fusion, raises concerns about its robustness and generalizability.

In response to these challenges, we developed TCellPredX, a novel ensemble model that integrates 13 features and 12 models using a soft voting strategy. TCellPredX optimally combines feature sets with top-performing models and further refines them through the mRMR feature selection method. Our experimental results clearly demonstrate that TCellPredX not only surpasses TROLLOPE in predictive accuracy but also offers enhanced precision in TCE−HCV predictions, thereby providing new opportunities for advancing HCV vaccine development.

2. MATERIAL AND METHODS

2.1. Benchmark Data Set. The data set utilized in this study is derived from the work of Charoenkwan et al.,[15] who extracted both positive and negative samples related to HCV (ID 11103) from the Immune Epitope Database version 2.26 (www.iedb.org).[16] Specifically, the data set comprises peptide sequences primarily associated with T-cell assays conducted in humans, mice, and nonhuman primates.[17] These peptides are categorized as either TCE−HCV or non-TCE−HCV.[15] The original data set contained 711 positive and 790 negative samples. To enhance data quality and reduce redundancy, Charoenkwan et al.[15] performed preprocessing, eliminating redundant samples. This process resulted in a final data set comprising of 446 TCE−HCV and 525 non-TCE−HCV. All selected peptide sequences are between 8 and 10 amino acid (AA) to ensure consistency and reliability of the data.

2.2. Feature Representation. HCV epitopes exhibit a range of characteristics, including sequence patterns, structural features, and physicochemical properties. To comprehensively capture these attributes, we incorporated multiple feature groups: local sequence encoding methods (AAC, PAAC, APAAC), global sequence encoding techniques (DDE, DPC, TPC), and physicochemical descriptors (PCP, AAI). These features enhance the model's ability to discern subtle yet significant differences among HCV epitopes. Additionally, embeddings from protein language models such as ESM and PortT5 were employed to capture contextual sequence information by integrating evolutionary and structural insights. This approach was motivated by the necessity to represent higher-order dependencies within protein sequences, which are challenging to capture using simpler encoding methods alone.

Converting peptide sequences into numerical vector representations is a critical step in this experiment.[18−20] To accomplish this, we utilized a diverse array of 13 features, incorporating traditional sequence encodings, composition-transition-distribution descriptors, physicochemical properties, and advanced protein language models, as detailed below. Initially, local sequence encoding methods, such as AAC, transform sequences into 20-dimensional vectors by calculating AA frequency in protein sequence.[21] Building on this, PAAC introduces additional sequence-order information,[22] while its variant, APAAC, captures hydrophilicity−hydrophobicity distribution patterns along peptide chain.[23] On the other hand, global sequence encoding is exemplified by DDE, which calculates the deviation between observed dipeptide frequencies and their expected values, generating a corresponding feature vector.[24] Furthermore, DPC records the frequency of all possible dipeptides within the sequence,[25] and TPC extends this to tripeptides, resulting in an 8000-dimensional vector reflecting position-specific occurrence frequencies[26] For composition-transition-distribution descriptors, CTDC encodes various physicochemical properties.[27] CTDD subsequently maps the positional distribution of specific amino acids within the sequence, for example, within the first 10% or last 20%.[28] Following this, CTDT characterizes the frequency of amino acid pair transitions between different groups.[29] Physicochemical properties are further captured through the AA index, which generates indices representing attributes like hydrophobicity and molecular volume. These indices are then employed in the PCP to encode the characteristics of amino acids.[29] Lastly, we integrated embeddings from protein language models. Specifically, the ESM-2 model, a large-scale pretrained protein language model, produces 1280-dimensional vectors that encapsulate deep contextual and structural information from protein sequences.[30] Similarly, ProtT5 converts protein sequences into textual representations, leveraging the T5 architecture to extract sophisticated features from these sequences.[31] Collectively, this comprehensive feature representation framework ensures robust and precise modeling for the experimental tasks.

2.3. Feature Normalization and Selection. To address the variability in feature ranges, we standardized the data, which is essential for improving interpretability and accuracy, especially in classifiers sensitive to feature scales like k-nearest

https://doi.org/10.1021/acsomega.4c08715
ACS Omega 2024, 9, 51494−51507

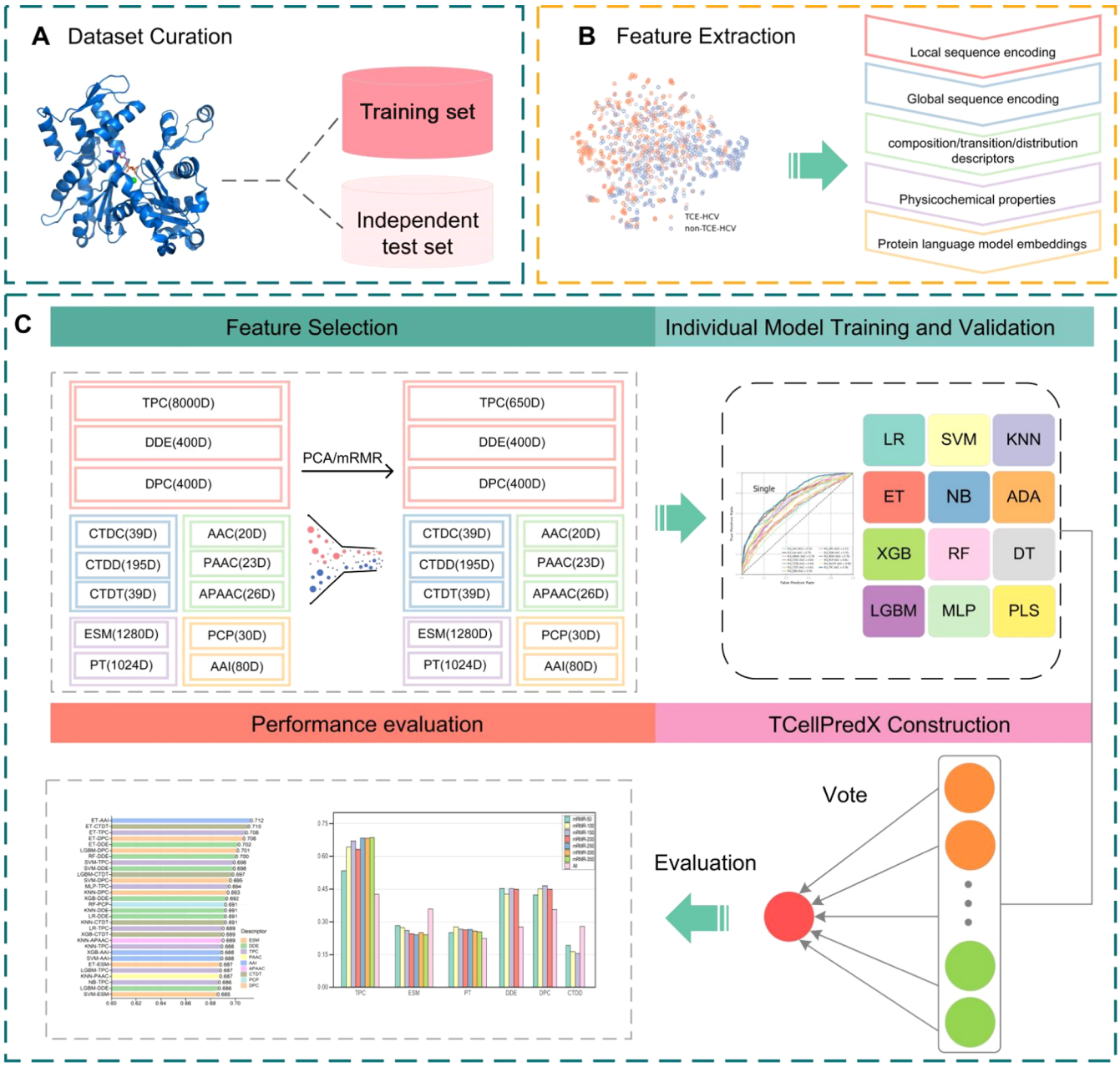


Figure 1. Overall workflow of TCellPredX. (A) Data set curation, (B) Feature extraction, (C) Feature selection, individual model training and testing, TCellPredX construction, and performance evaluation.

neighbors and support vector machines. The standardization formula applied is $x' = x - \mu/\sqrt{std}$, x and x' are the original and standardized data, respectively, μ and std denote the mean and standard deviation. Recognizing the importance of model interpretability, we applied the max-relevance-min-redundancy (mRMR) technique following feature fusion. This approach allowed us to retain the most informative features while reducing redundancy, thereby ensuring a streamlined feature set without compromising accuracy. Moreover, feature selection is crucial for reducing dimensionality and enhancing model performance. We employed two key techniques: PCA and mRMR. PCA involves standardizing the data, calculating the covariance matrix, performing eigenvalue decomposition, and mapping the data onto the chosen principal components.[32] In contrast, mRMR focuses on selecting features most relevant to target while minimizing redundancy among them, making it widely applicable across different domains.[33] Detailed methodologies for feature extraction and selection are provided in Supporting Text S1.

2.4. Performance Evaluation Strategies and Metrics. The data set was first split into: 80% (training) and 20% (independent test). The training data is then used for model validation through 10-fold cross-validation, repeating ten times to ensure that each subset is used for testing, providing a robust measure of the model's reliability.[34] Subsequently, the remaining 20% data is used for independent test to objectively assess the model's performance and generalization capability.[35] To evaluate the model's performance, we referred to four tools, PSRQSP,[36] DrugormerDTI,[37] FRTpred,[38] and PRR-HyPred[39] and selected five commonly used evaluation metrics: Matthews correlation coefficient (MCC), specificity (SP), sensitivity (SN), accuracy (ACC), and area under the curve (AUC).

https://doi.org/10.1021/acsomega.4c08715
ACS Omega 2024, 9, 51494−51507

These metrics collectively assess different aspects of model performance.[40,41]

$$SN = \frac{TP}{TP + FN} \quad (1)$$

$$SP = \frac{TN}{TN + FP} \quad (2)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP \times FN) \times (TN + FP) \times (TN + FN)}} \quad (4)$$

where TP represents the number of correctly predicted TCE−HCV samples (true positives), TN is the number of correctly predicted non-TCE−HCV samples (true negatives), FP denotes the number of non-TCE−HCV samples incorrectly predicted as TCE−HCV (false positives), and FN is the number of TCE−HCV samples incorrectly predicted as non-TCE−HCV (false negatives).

3. PROPOSED MODEL TCELLPREDX

3.1. Workflow of TCellPredX. The development and evaluation of TCellPredX (Figure 1) involve six key steps. First, during data set curation, we utilized a data set from ref 15, which include both TCE−HCV and non-TCE−HCV, and was divided into training and independent test sets to ensure robust model generalizability. Next, in feature extraction, we extracted 13 features categorized into five distinct types—local and global sequence encoding, physicochemical descriptors, composition-transition-distribution descriptors, and embeddings from protein language models, thereby providing a comprehensive, multidimensional representation for distinguishing between TCE−HCV and non-TCE−HCV. Subsequently, in feature selection, we employed PCA and mRMR to identify the most informative features, effectively reducing redundancy and enhancing both training efficiency and model performance.

Following this, during individual model training and testing, we trained and tested 12 classifiers, including LR, SVM, KNN, ET, NB, ADA, XGB, RF, DT, LGBM, MLP, and PLS, with performance assessed via 10-fold cross-validation and independent test. Then, in TCellPredX construction, we adopted an ensemble classification strategy, incorporating a soft voting mechanism to integrate predictions from multiple models, thereby enhancing robustness and predictive accuracy. Finally, in performance evaluation, we assessed the models using key metrics such as SN, SP, ACC, AUC, and MCC, ensuring a thorough evaluation of model performance and generalization capabilities.

3.2. Parameter Settings. To optimize model performance, we meticulously configured key parameters across different models and conducted parameter optimization using grid search. For partial least squares (PLS), we systematically varied the number of components between 1 and 20. Additionally, we assessed the impact of scaling by toggling it on and off (true/false) to gauge its influence on performance. In logistic regression (LR), the "max_iter" parameter was set to 5000 to ensure sufficient iterations for convergence, particularly given the complexity of the feature space. For MLP, a thorough exploration of the architecture was performed, testing configurations with one to three hidden layers. The size of these layers was varied between 20 and 500 neurons to find the optimal balance between model capacity and the risk of overfitting.

4. RESULTS AND DISCUSSIONS

4.1. Model Performance Assessment through 10-Fold Cross-Validation and Independent Test. As outlined in Section 2.4, the data set was divided into training data for 10-fold cross-validation and 20% for independent testing. Each classifier and feature encoding were subjected to both 10-fold cross-validation and independent test, with the average results from ten iterations recorded. These results are presented in Tables 1 and 2, Figures 2 and 3.

4.1.1. Comprehensive Evaluation of Classifier Performance and Computational Efficiency. We derived 156 base classifiers from cross-validation results and conducted a comprehensive analysis to identify the most effective models. As shown in Figure 2, the ACC and AUC values for the top 30

Table 1. Performance Comparison of Classifiers on the Training Data via 10-Fold Cross-Validation

feature	classifier	ACC	SN	SP	MCC	AUC
AAC	ET	0.678	0.654	0.699	0.351	0.764
	KNN	0.693	0.656	0.731	0.387	0.752
	XGBoost	0.662	0.638	0.685	0.322	0.729
AAI	SVM	0.683	0.664	0.701	0.361	0.745
	LGBM	0.693	0.664	0.722	0.385	0.741
	XGBoost	0.673	0.640	0.704	0.343	0.74
APAAC	ET	0.678	0.642	0.715	0.356	0.762
	KNN	0.687	0.648	0.730	0.377	0.742
	SVM	0.696	0.684	0.707	0.387	0.724
CTDC	ET	0.675	0.653	0.695	0.346	0.757
	LGBM	0.680	0.655	0.701	0.354	0.735
	XGBoost	0.680	0.654	0.706	0.358	0.733
CTDD	LGBM	0.715	0.703	0.729	0.427	0.760
	ET	0.691	0.683	0.699	0.375	0.753
	XGBoost	0.679	0.658	0.698	0.353	0.743
CTDT	ET	0.698	0.680	0.716	0.393	0.779
	KNN	0.697	0.666	0.728	0.394	0.759
	LGBM	0.680	0.658	0.700	0.356	0.749
DDE	SVM	0.718	0.727	0.715	0.431	0.785
	KNN	0.719	0.706	0.732	0.434	0.773
	MLP	0.695	0.698	0.711	0.392	0.765
DPC	SVM	0.720	0.725	0.723	0.437	0.780
	KNN	0.710	0.696	0.724	0.416	0.770
	ET	0.697	0.695	0.702	0.388	0.764
PAAC	ET	0.679	0.658	0.701	0.356	0.750
	KNN	0.692	0.649	0.743	0.391	0.736
	PLS	0.655	0.642	0.664	0.299	0.725
PCP	ET	0.664	0.650	0.677	0.321	0.729
	RF	0.668	0.702	0.654	0.330	0.714
	LGBM	0.668	0.684	0.665	0.330	0.708
TPC	PLS	0.715	0.720	0.716	0.426	0.794
	LR	0.719	0.700	0.739	0.435	0.787
	SVM	0.719	0.707	0.734	0.435	0.786
ESM	ET	0.701	0.688	0.714	0.397	0.764
	SVM	0.706	0.693	0.718	0.407	0.762
	LGBM	0.696	0.671	0.719	0.389	0.753
PortT5	SVM	0.697	0.778	0.670	0.402	0.758
	LGBM	0.691	0.672	0.710	0.378	0.745
	ADA	0.683	0.659	0.706	0.362	0.736

https://doi.org/10.1021/acsomega.4c08715
ACS Omega 2024, 9, 51494−51507

Table 2. Performance Comparison of Various Classifiers on the Training Data via 10-Fold Cross-Validation (Based on Group Features)[a]

feature	classifier	ACC	SN	SP	MCC	AUC
Group_1	ET	0.691	0.673	0.709	0.378	0.781
	LGBM	0.68	0.652	0.71	0.36	0.752
	XGBoost	0.683	0.65	0.715	0.364	0.745
Group_2	ET	0.707	0.697	0.717	0.410	0.767
	LGBM	0.702	0.675	0.731	0.403	0.756
	RF	0.673	0.677	0.671	0.336	0.752
Group_3	ET	0.702	0.675	0.733	0.406	0.794
	LGBM	0.722	0.701	0.742	0.44	0.786
	SVM	0.724	0.716	0.739	0.447	0.785
Group_4	LGBM	0.717	0.699	0.734	0.43	0.795
	ET	0.711	0.694	0.728	0.419	0.793
	ADA	0.705	0.698	0.714	0.406	0.778
Group_5	ET	0.684	0.677	0.691	0.36	0.771
	LGBM	0.677	0.658	0.695	0.348	0.755
	SVM	0.689	0.678	0.701	0.374	0.754

[a]Note: Group_1, local sequence encoding (AAC, PAAC, APAAC); Group_2, composition-transition-distribution descriptors (CTDC, CTDD, CTDT); Group_3, global sequence encoding (DDE, DPC, TPC); Group_4, physicochemical properties (PCP, AAI); Group_5, embeddings from protein language models (ESM, PortT5).

base classifiers were evaluated through both cross-validation and independent testing. Among the various feature encodings, KNN, SVM, and ET consistently demonstrated superior performance. Notably, under the TPC encoding, PLS achieves an AUC of 0.794, while SVM-TPC and LR-TPC reach AUC values of 0.786 and 0.787, respectively (Table 1). These results are consistent with the findings of Phasit Charoenkwan et al.,[15] who reported that SVM−TPC and LR-TPC are among the top performers out of all 144 base classifiers, each achieving AUC values greater than 0.780.

To ensure a balanced and thorough comparison of the 12 classifiers, we also assessed their performance in terms of MCC, computational efficiency, and ACC. Tables 1 and 2 present detailed performance metrics for each classifier across different feature encodings. The ACC results indicate that the NB and ET performed exceptionally well in both 10-fold cross-validation and independent test. For instance, ET-DPC achieves an 0.697 in 10-fold cross-validation and ACC of 0.706 in independent test, showing minimal variance and suggesting stable and robust performance in predicting HCV linear T-cell epitopes. As an ensemble classifier, ET is particularly effective at handling data impacted by noise, high dimensionality, and highly correlated features, all without overfitting.

Regarding MCC values, SVM and KNN outperform other classifiers across most feature encodings. For example, SVM-DDE and KNN-DDE achieve MCC values of 0.431 and 0.434, respectively, which are higher than those of other classifiers. Although PLS and LR show average performance across other feature encodings, they perform comparably to SVM and KNN under the TPC encoding. Specifically, PLS-TPC and LR-TPC achieve MCC values of 0.426 and 0.435, respectively (Table 1). Figure 3 illustrates the MCC values for the top six machine learning models across different feature encodings.

In terms of computational efficiency, we employed mRMR to reduce the 8000D TPC feature encoding to 650D before training six models (i.e., ET, KNN, NB, SVM, LR, and PLS), while recording their cross-validation and parameter optimization times (Figure 4). The results indicate that KNN and SVM require significantly more time for parameter optimization compared to other classifiers. Specifically, SVM need 89.98 min, whereas ET requires only 2.01 min. PLS have the shortest total time, at just 2.421 min (Figure 4A). Figure 4B,C further reveal that ET consumed the most computational resources during training, while SVM have the highest computational cost during parameter optimization. Overall, PLS emerges as a promising model due to its low computational cost and strong performance.

4.1.2. Optimal Feature Combination Strategies for Enhancing HCV Linear T-Cell Epitope Prediction. The AUC values of the top 30 base classifiers (Figure 2C,D) reveal that among the six classifiers achieving an AUC of 0.78 or higher during cross-validation, four were with the TPC features, with three ranking in the top three. This highlights TPC as one of the most effective feature encodings. Specifically, SVM-TPC achieves an AUC of 0.786, LR−TPC reaches 0.787, and PLS−TPC attains 0.794 (Figure 2C). Additionally, other global sequence encoding features, such as DDE and DPC, also perform well across various metrics, further demonstrating the strengths of global sequence features. For example, SVM−DDE achieves an AUC of 0.785, and KNN−DDE reaches 0.773 (Table 1). In contrast, local sequence encodings perform well only with the ET model, showing mediocre results with other classifiers. For instance, ET−AAC achieves an AUC of 0.764, with no other classifier exceeding 0.76 using this encoding (Table 1). Despite generally weaker performance when used individually, composition-transition-distribution descriptors, protein language models, and physicochemical descriptors exhibit notable improvements when physicochemical descriptors are combined. For instance, ET−PCP yields an AUC of 0.729, whereas integrating features as in LGBM-Group_4 raises the AUC to 0.795 (Figure 2). This demonstrates that while physicochemical features like AAI and PCP may perform moderately on their own, their combination can substantially enhance predictive accuracy.

The key advantage of feature integration lies in its comprehensive utilization of information across different levels, providing a holistic representation of HCV T cell epitopes. Accurate prediction of these epitopes requires not only consideration of the local arrangement and sequence information on amino acids but also a deeper understanding of their biological functions and interactions. By amalgamating multiple feature sets, the model constructs a comprehensive representation of T cell epitopes across various layers, which is crucial for improving predictive performance.

In the context of HCV vaccine development, feature fusion serves as a potent strategy by integrating diverse layers of biological information to enhance the accuracy and reliability of T cell epitope predictions. This approach effectively addresses the complexity of HCV, whose genome varies significantly among different genotypes and subtypes, complicating the design of a universally effective vaccine. Consequently, feature fusion plays a central role in enhancing both the predictive capability and biological relevance of T cell epitope prediction models for HCV vaccine development. By combining multiple sources of information—from sequence data to structural insights—feature fusion enables the identification of epitopes with improved binding affinity, stability, and cross-genotype coverage. This approach not only elevates the performance of prediction models but also

https://doi.org/10.1021/acsomega.4c08715
ACS Omega 2024, 9, 51494−51507

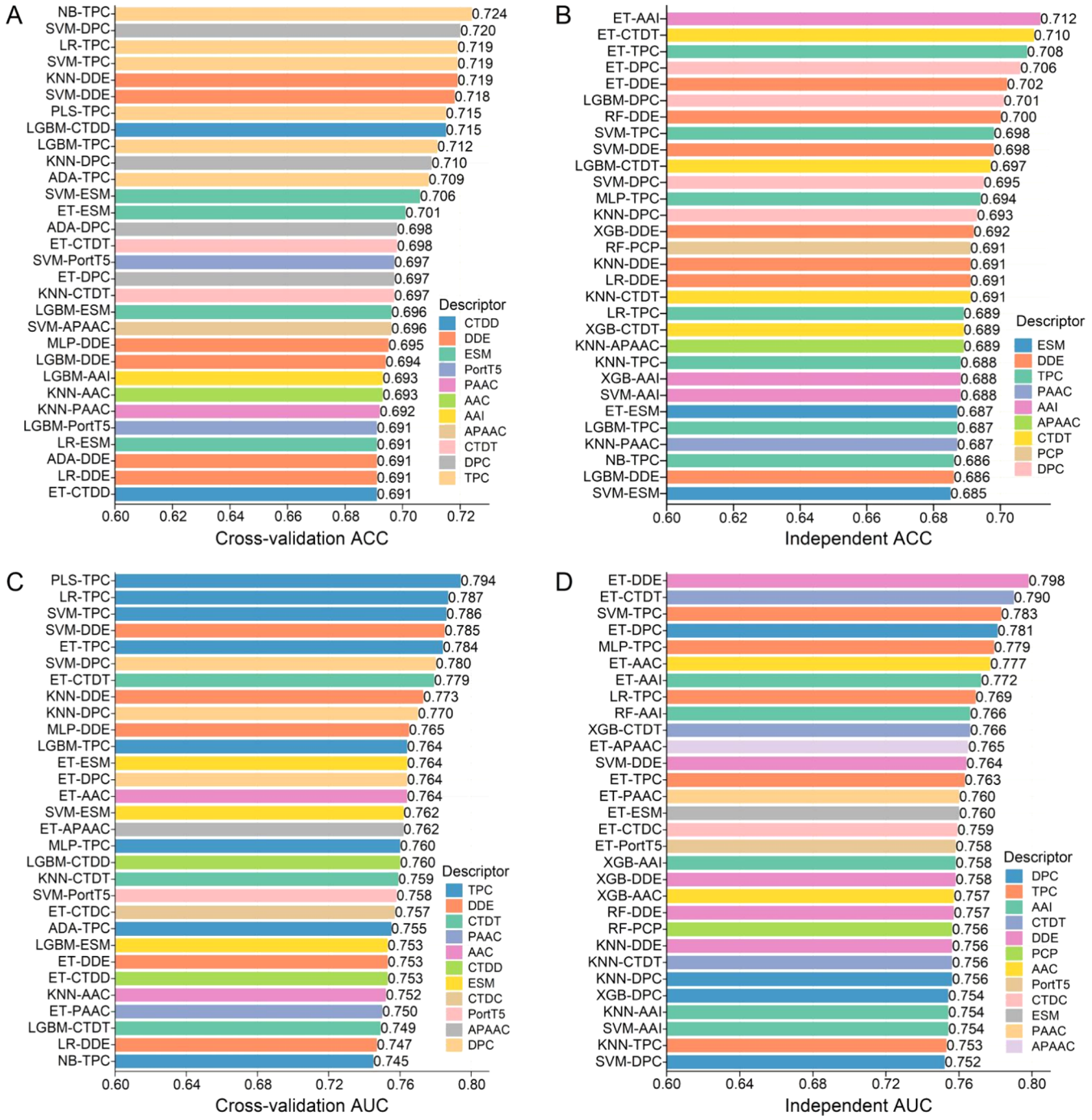


Figure 2. ACC and AUC values of the top 30 base classifiers, evaluated through 10-fold cross-validation and independent testing. (A, B) The ACC values for these classifiers in both 10-fold cross-validation and independent tests. (C, D) The AUC values for the same classifiers under the same evaluation conditions.

holds the potential to drive the development of more effective, broad-spectrum vaccines, thereby making a significant impact on HCV prevention.

Consequently, we are motivated to explore further feature combinations, such as integrating TPC with Group_1. Figure 5 presents the ROC curves for six classifiers (i.e., ET, NB, LR, SVM, PLS, and KNN) across different features. It is evident that Group_3 and Group_4 outperform others, while Group_1, Group_2, and Group_5 demonstrate relatively weaker performance.

While previous studies have demonstrated that combining physicochemical descriptors can improve AUC values, the overall performance across the five feature groups (Table 2) suggests that simple feature fusion does not significantly enhance model performance or fully exploit the features' potential. Notably, Group_3 shows relatively strong results compared to other groups, but the performance gains over the single TPC feature are minimal and, in some cases, even decline in specific metrics. For instance, the AUC for SVM–TPC is 0.786, whereas the SVM combination in Group_3 achieves an AUC of 0.785. This indicates that TPC remains

https://doi.org/10.1021/acsomega.4c08715
ACS Omega 2024, 9, 51494−51507

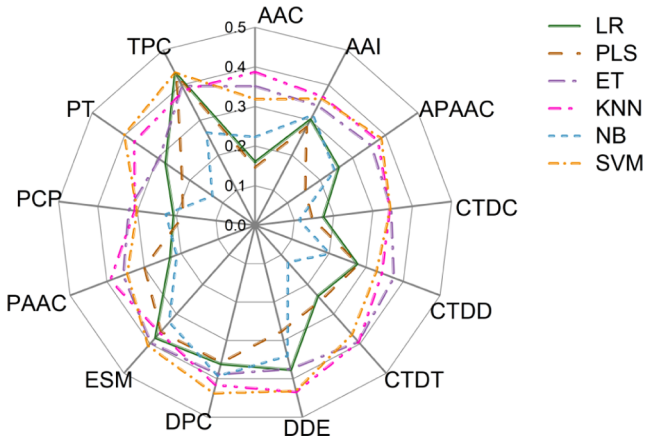


Figure 3. MCC values of the top six classifiers on the training data with different features via 10-fold cross-validation.

the dominant feature even after fusion, with the final predictive outcome largely reliant on it. Consequently, we opted to combine the strongest TPC feature with each of the other four feature groups separately, employing mRMR for feature selection.

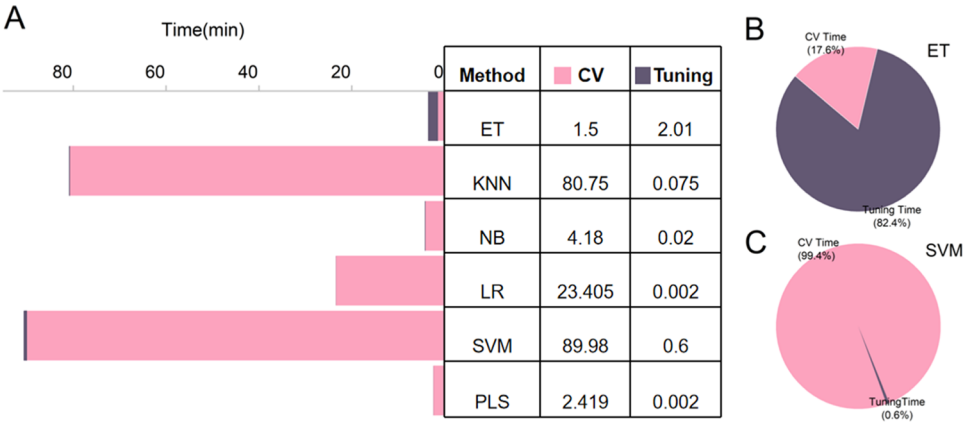
The analysis and comparison of classifier performance in Tables 1 and 2 lead to several key insights: First, the performance difference between single features and feature combinations shows that feature fusion does not substantially improve model performance, particularly for TPC feature, where the enhancement is limited. Second, the optimal strategy involves combining the strongest TPC feature with other features and applying mRMR for feature selection, which maximizes the exploitation of feature potential and further enhances model performance. Finally, the importance of feature selection lies in mRMR's ability to effectively remove redundant information, especially in high-dimensional feature, thereby improving classifier performance.

4.2. Impact of Dimensionality Reduction on Model Performance. The initial dimensions of the DDE, ESM, PortT5, and TPC features are 400, 1280, 1024, and 8000, respectively. These high-dimensional features not only consumed substantial memory during classifier training but also significantly increased the time required for parameter optimization and model training. To address these challenges, we employed PCA and mRMR for feature selection, aiming to simplify the model, reduce computational costs, and eliminate noise and redundant information. We reduced the feature dimensions to 50, 100, 150, 200, 250, 300, and 350, and compared these dimensions with the original features.

Figure 6 illustrates that PCA is less effective at retaining key information, especially for TPC and CTDD features, leading to inferior performance compared to the original features. For instance, the MCC value for TPC is substantially lower with PCA (Figure 6B). In contrast, mRMR outperforms the original features, particularly for TPC, where feature selection enhances performance by approximately 30% (Figure 6A). This highlights mRMR's capability to effectively select informative subsets from the 8000-dimensional TPC features, reducing noise and redundancy for improved classifier modeling. Figure 6A shows mRMR performance across various feature dimensions, Figure 6B presents PCA results, and Figure 6C compares the performance of feature subsets selected by mRMR and PCA.

Based on these comparisons, we draw the following conclusions: First, PCA is inadequate in preserving key information when dealing with high-dimensional features such as TPC and CTDD, resulting in diminished performance. Second, mRMR generally outperform the original features across most encodings, particularly in handling high-dimensional features like TPC, significantly enhancing classifier performance. Finally, mRMR proves effective in selecting the most informative subset from high-dimensional features, removing noise and redundancy, simplifying the model, and reducing computational costs. Therefore, mRMR is a superior tool for feature selection, especially when dealing with high-dimensional data, as it significantly improves both classifier performance and efficiency.

4.3. Refining Predictive Models with Soft Voting-Based Ensemble Learning. Our extensive cross-validation results demonstrated that the ensemble model utilizing soft voting consistently delivered robust and enhanced performance. The complementary strengths of the individual classifiers augmented the model's predictive capabilities, significantly improving performance without introducing unnecessary complexity. In this section, we evaluated optimal strategies



Method	CV	Tuning
ET	1.5	2.01
KNN	80.75	0.075
NB	4.18	0.02
LR	23.405	0.002
SVM	89.98	0.6
PLS	2.419	0.002

Figure 4. Comparison of CV and tuning times across different classifiers in 10-fold cross-validation. (A): Compares the training times for each classifier using TPC features, reduced to 650 dimensions with mRMR, including both parameter tuning and CV. (B, C): the proportional breakdown of parameter tuning and CV times for ET and SVM, respectively. Note: All experiments were performed on a Windows 10 operating system using Python version 3.8. The hardware configuration comprised an Intel(R) Core (TM) i5-7300U CPU and 8 GB of memory.

https://doi.org/10.1021/acsomega.4c08715
ACS Omega 2024, 9, 51494−51507

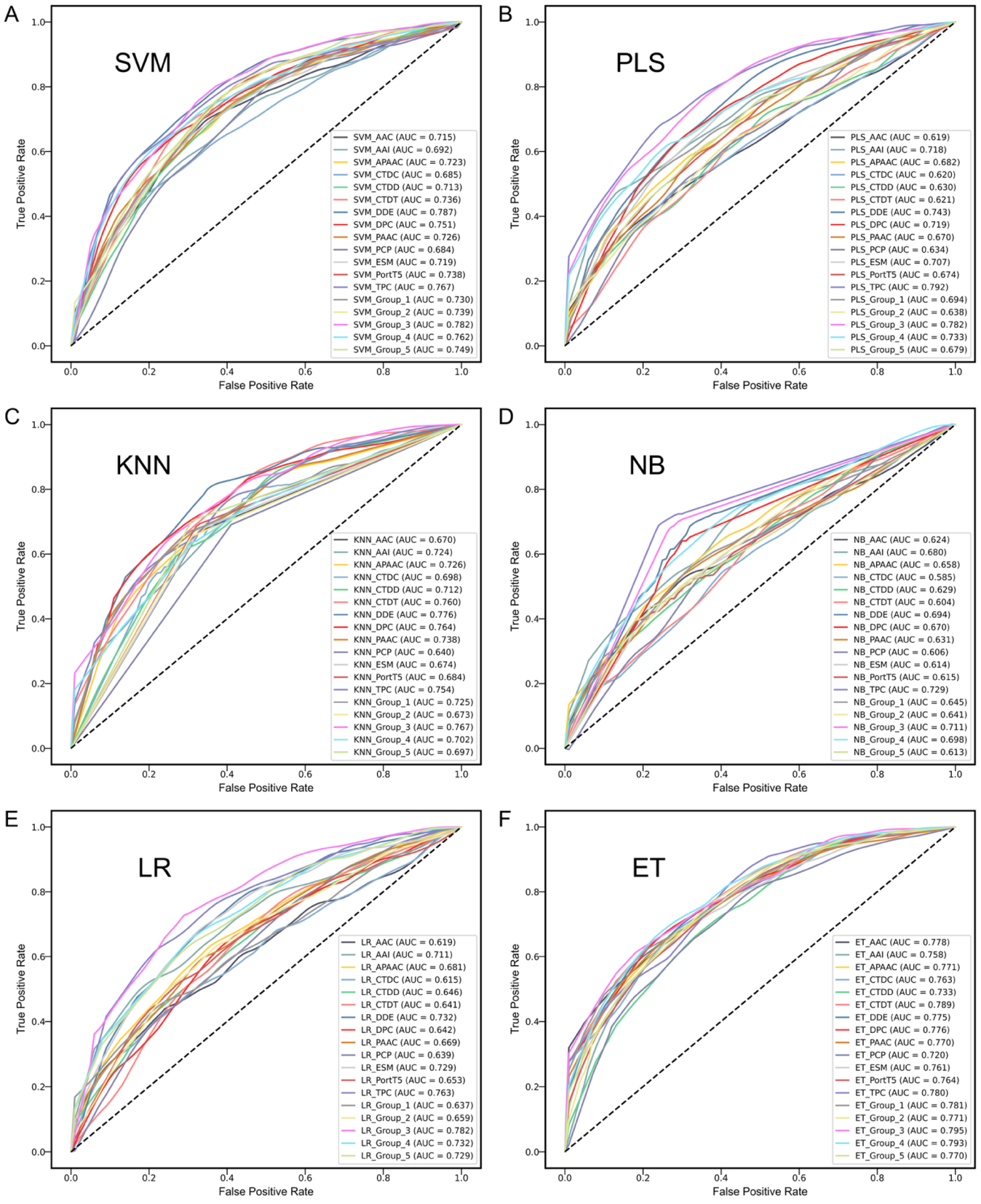


Figure 5. ROC curves of six classifiers with different features. (A) SVM, (B) PLS, (C) KNN, (D) NB, (E) LR, and (F) ET. Note: Group_1, local sequence encoding (AAC, PAAC, and APAAC); Group_2, composition/transition/distribution descriptors (CTDC, CTDD, and CTDT); Group_3, global sequence encoding (DDE, DPC, and TPC); Group_4, physicochemical properties (PCP, AAI); Group_5, embeddings from protein language models (ESM, PortT5).

https://doi.org/10.1021/acsomega.4c08715
ACS Omega 2024, 9, 51494−51507

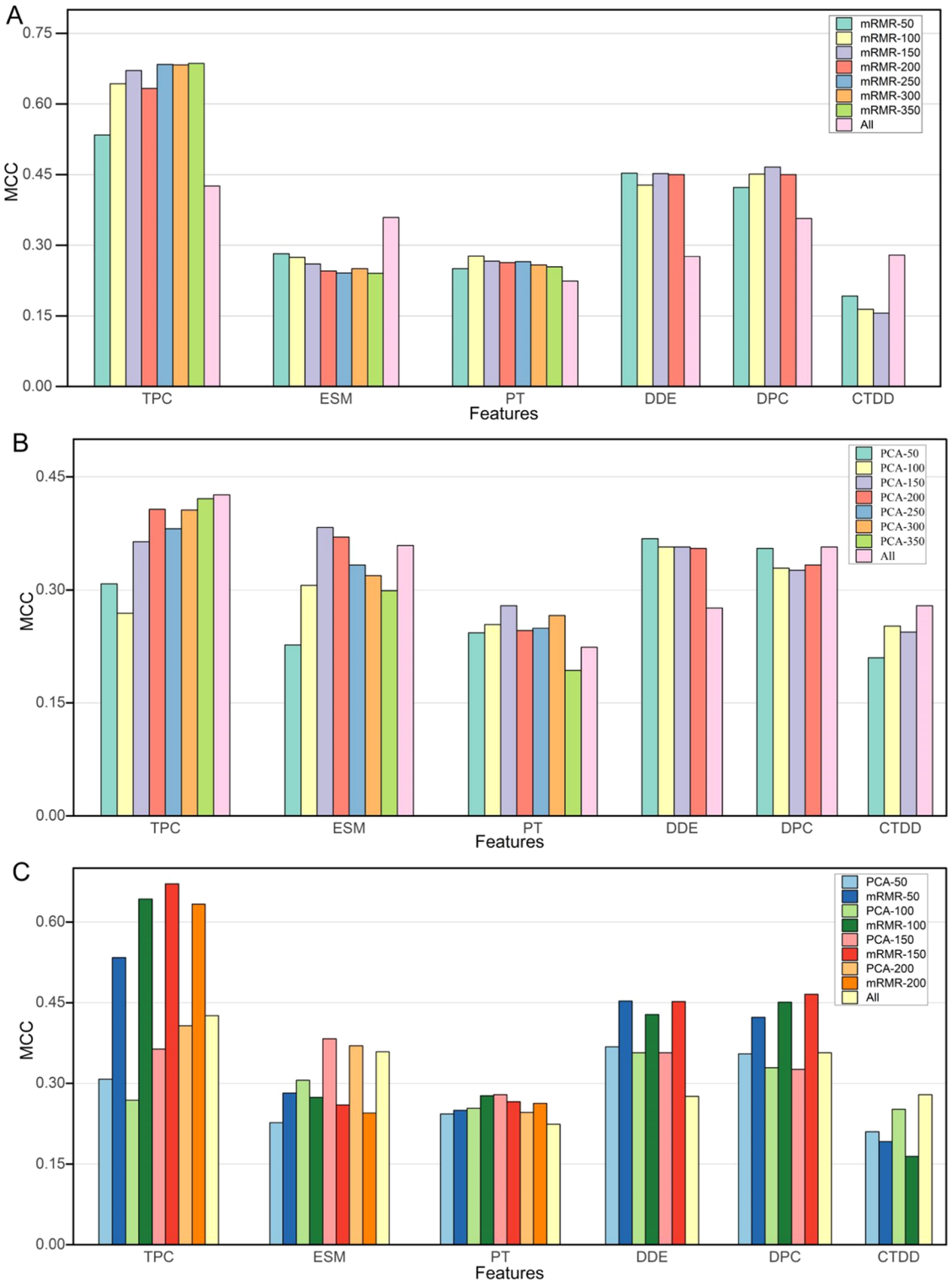


Figure 6. Feature selection using PCA and mRMR. (A) Comparison of performance across different dimensions selected by mRMR for six features. (B) Performance comparison of PCA-selected feature vectors across different dimensions. (C) Comparative analysis of the performance of feature subsets selected by mRMR and PCA.

for feature fusion and model integration by analyzing cross-validation results after feature selection. As detailed in Table 3, the 8000-dimensional TPC feature emerged as the most effective single feature, achieving the highest AUC of 0.794, which underscores its superior predictive power.

Moreover, the combined feature sets {1, 2, 3, 9} and {4, 5, 6, 9} demonstrate strong performances, with AUC of 0.938 and 0.932, respectively. Although the feature set {10, 11, 9} slightly surpass these with an AUC of 0.943, its MCC is less stable, suggesting potential variability in model reliability. To further investigate the performance of various classifiers, Table 4 compares both single and ensemble models. The ensemble model {a, b, c}, which integrates LR, SVM, and PLS classifiers, exhibited the best overall performance, achieving an AUC of

https://doi.org/10.1021/acsomega.4c08715
ACS Omega 2024, 9, 51494−51507

Table 3. Performance Evaluation of Single and Combined Features Using PLS on Training Data via 10-Fold Cross-Validation

feature[a]	feature dimension	ACC	SN	SP	MCC	AUC
1. AAC	20	0.579	0.550	0.599	0.146	0.620
2. PAAC	23	0.655	0.642	0.664	0.299	0.725
3. APAAC	26	0.583	0.554	0.605	0.154	0.632
4. CTDC	39	0.579	0.549	0.600	0.145	0.622
5. CTDD	195	0.642	0.627	0.658	0.279	0.669
6. CTDT	39	0.629	0.600	0.654	0.252	0.642
7.DDE	400	0.639	0.607	0.670	0.276	0.660
8. DPC	400	0.682	0.673	0.691	0.357	0.736
9. TPC	8000	0.715	0.720	0.716	0.426	0.794
10. AAI	80	0.643	0.614	0.667	0.280	0.701
11. PCP	30	0.598	0.580	0.610	0.183	0.637
12. ESM	1280	0.680	0.654	0.708	0.359	0.740
13. PortT5	1024	0.615	0.584	0.642	0.224	0.660
{1, 2, 3}	69	0.631	0.605	0.653	0.255	0.696
{4, 5, 6}	273	0.606	0.577	0.631	0.206	0.638
{7, 8, 9}	8800	0.702	0.684	0.722	0.402	0.779
{10, 11}	110	0.664	0.640	0.687	0.324	0.735
{12, 13}	2304	0.629	0.596	0.663	0.257	0.682
{1, 2, 3, 9}	500	0.844	0.932	0.798	0.698	0.938
{4, 5, 6, 9}	500	0.845	0.917	0.806	0.697	0.932
{10, 11, 9}	600	0.843	0.895	0.814	0.689	0.943
{12, 13, 9}	500	0.782	0.806	0.772	0.565	0.871
{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	1500	0.800	0.802	0.803	0.599	0.891

[a]Note: The "feature dimension" column indicates the number of dimensions after mRMR feature selection. The "feature" represents either a single feature or a combination of multiple features. For example, "1. AAC" indicates a model trained using AAC encoding, while {1, 2, 3} refers to a feature set that combines AAC, PAAC, and APAAC encodings.

Table 4. Performance Evaluation of Classifiers on the Training Data via 10-Fold Cross-Validation

classifier[a]	ACC	SN	SP	MCC	AUC
a.LR	0.836	0.935	0.788	0.685	0.935
b.SVM	0.878	0.892	0.869	0.755	0.883
c.PLS	0.832	0.934	0.784	0.675	0.942
d.ET	0.781	0.824	0.756	0.562	0.859
e.NB	0.847	0.975	0.787	0.713	0.846
f.KNN	0.796	0.815	0.785	0.590	0.790
{a, b}	0.892	0.925	0.870	0.785	0.945
{a, b, c}	0.900	0.946	0.871	0.803	0.953
{a, b, c, d}	0.860	0.937	0.816	0.727	0.945
{a, b, c, d, e}	0.853	0.963	0.799	0.721	0.947
{a, b, c, d, e, f}	0.854	0.955	0.803	0.722	0.938

[a]Note: Each term in the "classifier" represents either a single classifier or an ensemble classifier with group feature {1, 2, 3, 9}. Ensemble classifier {a, b, c} refers to an ensemble classifier integrating with "a.LR", "b.SVM", and "c.PLS".

0.953 and an MCC of 0.803. This comparison highlights that ensemble models, particularly those utilizing soft voting strategies, can significantly enhance predictive accuracy and robustness compared to individual classifiers.

Additionally, Figure 7 presents the ROC curves for both single and combined features based on PLS. When evaluating the performance of individual models using different features (Figure 7A), the single TPC feature emerges with the highest AUC of 0.78, outperforming other features. Subsequently, the impact of feature combinations on model performance is assessed (Figure 7B). Notably, Group_3 exhibits strong performance among the five groups. However, the improvement over the single TPC feature is minimal. Further analysis of Table 2 indicates that the AUC for Group_3 is 0.794, identical to that of the single TPC feature, confirming that TPC remains the dominant feature even after fusion.

To further enhance predictive performance, the strongest TPC feature is combined with each of the other four feature groups separately, followed by mRMR. The experimental results show that the combined feature sets {1, 2, 3, 9} and {4, 5, 6, 9} perform optimally, achieving AUCs of 0.938 and 0.932, respectively. Although the {10, 11, 9} combination reaches an AUC of 0.943, its MCC exhibits greater variability, indicating instability (Table 3). In contrast, the {1, 2, 3, 9} combination exhibits strong performance with excellent generalization capabilities. Lastly, the performance of ensemble classifiers is further investigated. Table 4 reveals that the ensemble model {a, b, c} achieves the highest performance in terms of ACC, MCC, and AUC. By integrating the ensemble model {a, b, c} with the {1, 2, 3, 9} feature combination, the constructed TCellPredX demonstrate outstanding predictive performance, significantly improving prediction accuracy.

In conclusion, combining features alone provides limited improvement in model performance, particularly when TPC is the predominant feature. However, when TPC is integrated with additional features and optimized through mRMR feature selection, the full potential of these features is realized, leading to notable performance gains. Additionally, mRMR effectively reduces redundancy, particularly in high-dimensional scenarios, thereby further enhancing classifier performance.

4.4. Comparative Analysis of TCellPredX against Existing Methods. To assess TCellPredX, we conducted a comparative analysis against the TROLLOPE and other classifiers, including weighted voting and Boosting algorithms. In weighted voting approach, prediction probabilities from PLS, LR, and SVM were combined, while the boosting algorithm utilized LR as a weak classifier. We also compared

https://doi.org/10.1021/acsomega.4c08715
ACS Omega 2024, 9, 51494−51507

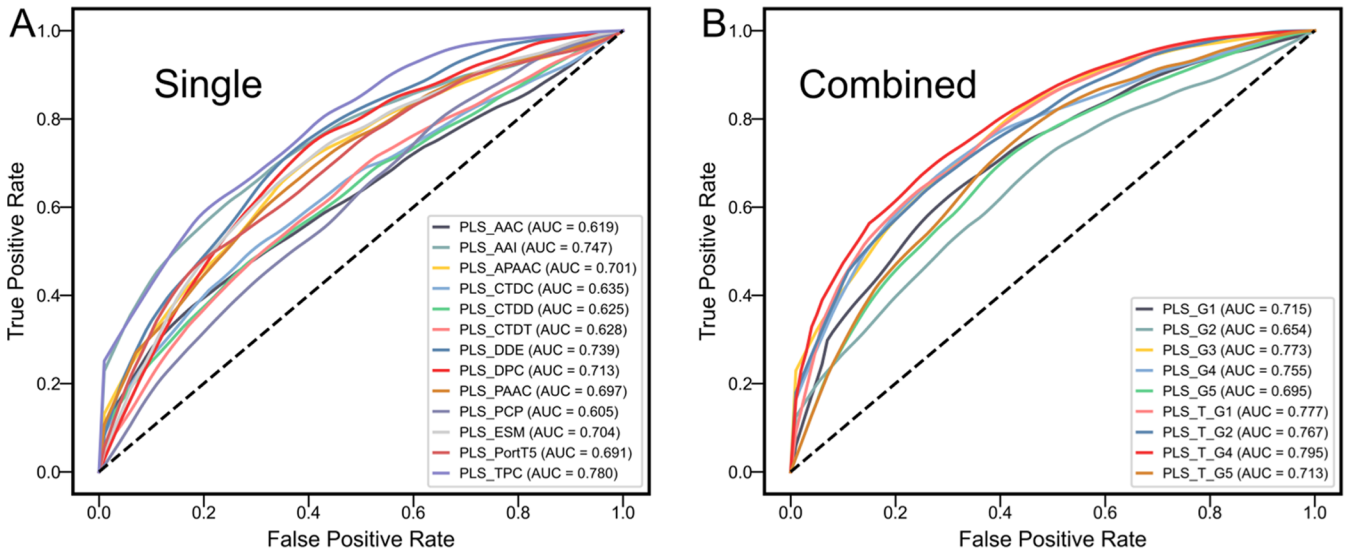


Figure 7. ROC curves for single and combined feature sets using PLS. (A) ROC curves for single features. (B) ROC curves for combined features. Note: the label "PLS_[Feature]" corresponds to independent test results for PLS using the specified features, while "PLS_G1-G5" illustrates the results for feature Groups 1 through 5. Additionally, "PLS_T_G1", "PLS_T_G2", "PLS_T_G4", and "PLS_T_G5" present the outcomes when TPC is combined with Group_1, Group_2, Group_4, and Group_5, respectively.

Table 5. Comparison Results of TCellPredX with Existing Methods[a]

evaluation strategies	method	ACC	SN	SP	MCC	AUC
10-fold cross-validation	TCellPredX	0.900	0.946	0.871	0.803	0.953
	TROLLOPE	0.745	0.686	0.795	0.487	0.808
	Boosting	0.803	0.938	0.747	0.628	0.925
	Weighted voting	0.871	0.953	0.826	0.751	0.955
independent test	TCellPredX	0.897	0.946	0.866	0.796	0.951
	TROLLOPE	0.747	0.742	0.752	0.493	0.827
	Boosting	0.795	0.940	0.739	0.612	0.923
	Weighted voting	0.865	0.944	0.822	0.736	0.956

[a]Note: TCellPredX integrates the TPC, AAC, PAAC, and APAAC features, reduces them to 500D using mRMR, and then employs PLS, LR, and SVM classifiers within an average voting ensemble. "Weighted voting" combines the prediction probabilities of PLS, LR, and SVM, while "Boosting" uses LR as a weak classifier to construct an ensemble model..

the top 5 classifiers with TCellPredX, as depicted in Supporting Figures S1 and S2.

As demonstrated in Table 5, TCellPredX outperformed both TROLLOPE and other machine learning methods, particularly in 10-fold cross-validation. TCellPredX achieves ACC, SN, SP, and AUC values of 0.900, 0.946, 0.871, and 0.953, respectively, significantly surpassing TROLLOPE's scores of 0.745, 0.686, 0.795, and 0.808. Additionally, the MCC improves from 0.487 to 0.803, marking a 31.6% increase. While TCellPredX is slightly outperformed in certain metrics by the weighted average voting method, its superior MCC indicates greater stability and predictive capability. While in the independent test, TCellPredX demonstrates outstanding performance, with ACC, SN, SP, and AUC values reaching 0.897, 0.946, 0.866, and 0.951, respectively, all surpassing TROLLOPE's corresponding values of 0.747, 0.742, 0.752, and 0.827. The MCC also increases significantly from 0.493 to 0.796. Compared to other methods, TCellPredX shows a marked improvement in MCC over both Boosting and weighted average voting algorithms, further highlighting its robust predictive accuracy.

Overall, TCellPredX consistently outperformed TROLLOPE and other ensemble methods across all metrics in both 10-fold cross-validation and independent test, with particularly strong results in MCC. By integrating TPC with other features and applying mRMR for feature selection, TCellPredX significantly enhanced predictive performance, reinforcing its superiority in identifying HCV linear T-cell epitopes.

4.5. Enhancing Epitope Identification in HCV through Feature Integration and t-SNE Analysis. We present a detailed analysis of linear T-cell epitopes in HCV. Experimental findings revealed that although the physicochemical descriptors AAI and PCP performed moderately when used individually, their combined application markedly enhanced feature effectiveness. To further validate the utility of these features, we employed t-SNE to examine the spatial distribution of four feature sets: Group_4, TPC, DDE, and DPC. As shown in Figure 8, the t-SNE distributions indicate that AAI and PCP are particularly effective in distinguishing between TCE−HCV and non-TCE−HCV, corroborating previous studies that identified AAI and PCP as critical for analyzing and characterizing various protein functions.[42]

Figure 8A illustrates the t-SNE distribution of Group_4 features, while Figure 8B−D show the distributions for TPC, DDE, and DPC, respectively. The data points corresponding to Group_4 features are more densely clustered, resulting in a better distinction between TCE−HCV and non-TCE−HCV

https://doi.org/10.1021/acsomega.4c08715
ACS Omega 2024, 9, 51494−51507

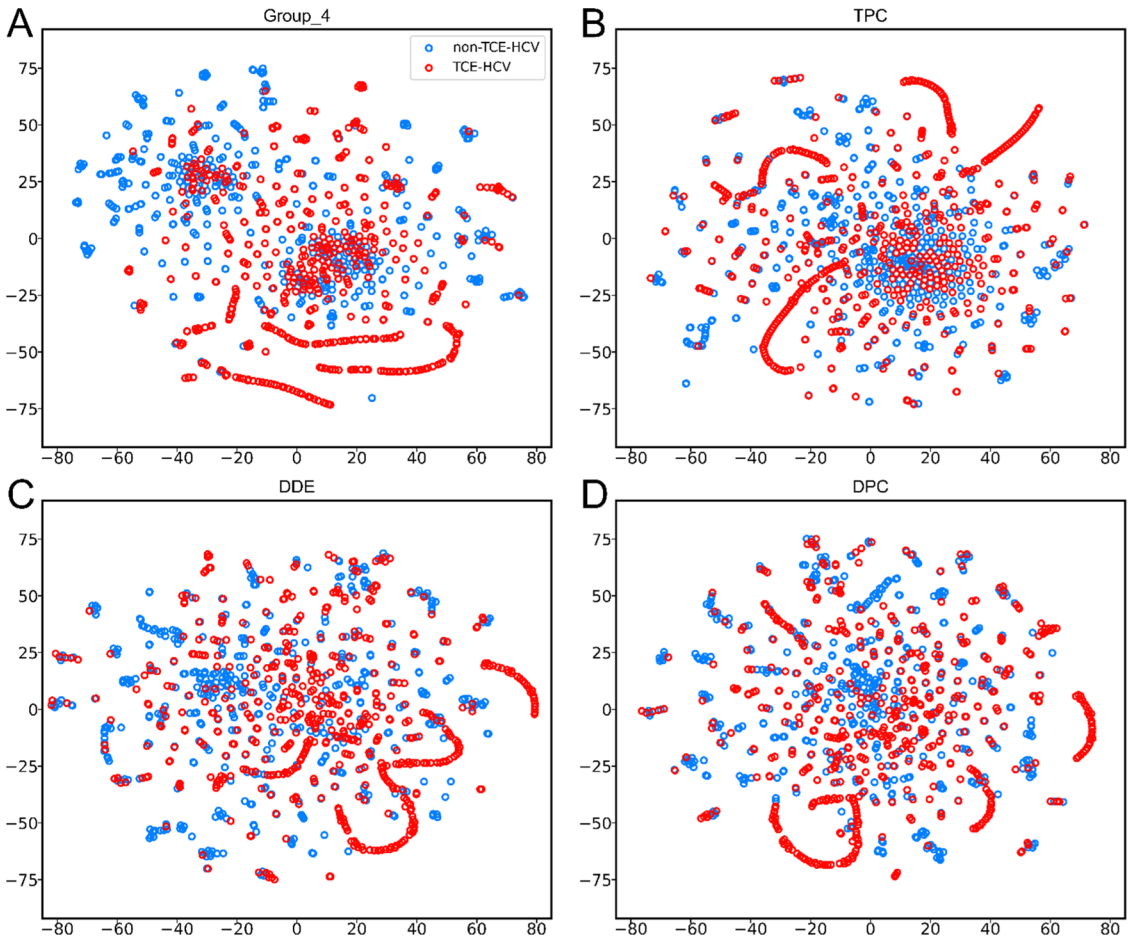


Figure 8. t-SNE visualization. (A−D) show the t-SNE plots for Group_4, TPC, DDE, and DPC, respectively.

samples. This suggests that Group_4 features have a high discriminative ability in identifying HCV linear T-cell epitopes.

Moreover, the physicochemical properties of TCE−HCV are critical in the binding between MHC molecules and T-cell epitopes. For a T-cell epitope to be presented in complex with MHC class I molecules, it must exhibit sufficient affinity and stability. Once presented, the complex must remain on the cell surface for an adequate duration. If the affinity is low, the peptide may not effectively bind to MHC-I molecules, resulting in inefficient antigen presentation.[43,44] MHC class II molecules possess an open binding groove that can accommodate peptides of varying lengths. While MHC class I molecules typically bind peptides ranging from 8 to 14 amino acids, MHC class II molecules can accommodate longer peptides. Numerous studies have shown that, within a certain range, extending peptide length may enhance affinity for MHC class II molecules, with optimal peptide-MHC affinity often occurring around 18−20 amino acids.[45,46] Therefore, focusing on the length of linear T-cell epitopes of the Hepatitis C virus is significant for peptide-MHC interactions, offering new avenues for vaccine design and development.

In summary, the visual analysis provided by t-SNE reinforces the effectiveness of AAI and PCP features in identifying HCV linear T-cell epitopes. These findings not only offer theoretical support for HCV vaccine development but also highlight the potential of combined features in bioinformatics applications.

4.6. Case Study. To validate the clinical applicability of TCellPredX, we conducted predictive analyses on peptide sequences documented in the literature as having potential benefits for vaccine development (see Figure 9). In the study by Timothy Donnison et al., a mouse model was utilized to develop a vaccine capable of inducing virus-specific B cell and T cell responses.[47] Their research identified the epitopes listed in Figure 9 as playing a critical role in vaccine development. Figure 9 presents the prediction results for six peptide sequences using TCellPredX, TROLLOPE, and five other

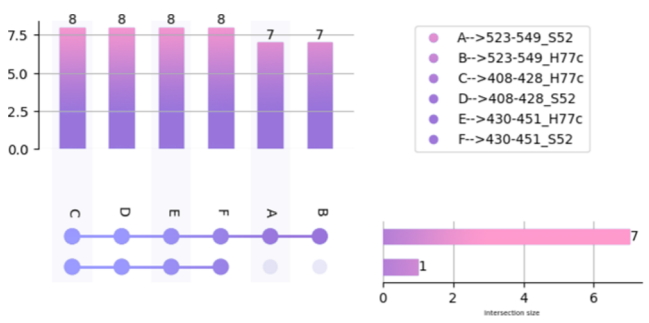


Figure 9. Upset plot comparing the prediction accuracy of TCellPredX with other classifiers for vaccine-relevant HCV epitopes. Note: 408−428_H77c, KQNIQLINTNGSWHINSTALN; 408−428_S52, KQKLQLVNTNGSWHINSTALN; 430−451_H77c, NESLNTGWLAGLFYQHKFNSSG; 430−451_S52, NESINTGFIAGLFYYHKFNSTG; 523−549_H77c, GAPTYSWGANDTDVFVLNNTRPPLGNW; 523−549_S52, GRPTYNWGENETDVFLLESLRPPSGRW.

https://doi.org/10.1021/acsomega.4c08715
ACS Omega 2024, 9, 51494−51507

top classifiers (LR−TPC, SVM−TPC, ET−TPC, PLS−TPC, KNN−DDE). Notably, TCellPredX accurately identified all six peptide sequences as TCE−HCV, fully aligning with the actual experimental outcomes. This result underscores the high efficiency and reliability of TCellPredX in identifying TCE−HCV.

In contrast, KNN−DDE incorrectly classified the ID:523−549_H77c and ID:523−549_S52 peptide sequences as non-TCE−HCV, highlighting the superior accuracy of TCellPredX. Although other classifiers such as TROLLOPE, LR−TPC, SVM−TPC, ET−TPC, and PLS−TPC correctly predicted TCE−HCV in most cases, the errors made by KNN−DDE emphasize TCellPredX's robustness as a more reliable tool. Collectively, these findings strongly support the potential of TCellPredX in vaccine development, demonstrating its high-precision predictive capability as a valuable asset for formulating vaccine development strategies and conducting clinical trials.

5. CONCLUSIONS

Rapid and accurate identification of linear T-cell epitopes in HCV is crucial for understanding antigen presentation by MHC molecules in the immune process and has a significant impact on advancing vaccine development. In this study, we investigated 13 diverse feature encoding schemes, including local and global sequence encodings, physicochemical properties, composition-transition-distribution descriptors, and embeddings from two protein language models. These features were integrated with 12 machine learning classifiers to construct TCellPredX, aimed at assessing its effectiveness in predicting HCV linear T-cell epitopes. Our findings revealed that PLS, SVM, KNN, LR, ET, and NB models consistently outperformed others, with PLS emerging as the most effective model. Furthermore, feature fusion techniques outperformed single-feature approaches, with TPC identified as the most influential feature among the 12 individual encodings, significantly enhancing the performance of integrated models. Additional analysis showed that the mRMR feature selection method was more effective than PCA in eliminating redundant information and noise.

Despite TCellPredX's strong performance, it has several limitations. First, the mRMR feature selection method may discard valuable features, potentially affecting the model's ability to capture relevant T-cell epitope patterns. Second, the model's complexity reduces interpretability, limiting its practical use for vaccine developers who need insights into individual feature contributions. Furthermore, TCellPredX is specifically tailored for HCV and may require significant adjustments, along with pathogen-specific data sets, to generalize to other pathogens—particularly when facing scalability challenges with larger data sets. Moreover, the limited availability of high-quality data sets for less-studied pathogens constrains the model's generalization across diseases. To address these issues, future research should prioritize: (1) enhancing interpretability through explainable AI techniques, (2) exploring feature selection methods that retain more informative features, (3) expanding the model's applicability to other pathogens, (4) incorporating detailed physicochemical properties, and (5) improving scalability through efficient algorithms.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c08715.

Detailed description of feature representation (Text S1); confusion matrix of TCellPredX and the top 5 classifiers on the training data via 10-fold cross-validation (Figure S1); confusion matrix of TCellPredX and the top 5 classifiers via independent test (Figure S2) (PDF)

AUTHOR INFORMATION

Corresponding Authors

Fang Ge − *State Key Laboratory of Organic Electronics and Information Displays, Institute of Advanced Materials (IAM), Nanjing University of Posts and Telecommunications, Nanjing 210023, China*; Email: gfang0616@njupt.edu.cn

Tanvir Alam − *College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar*; orcid.org/0000-0001-7033-3693; Email: talam@hbku.edu.qa

Authors

Hao-Yang Li − *School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212100, China*

Ming Zhang − *School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212100, China*; orcid.org/0009-0004-3197-3523

Muhammad Arif − *College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c08715

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY223062). Open access publication of this article was supported by the College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar.

REFERENCES

(1) Manne, V.; Ryan, J.; Wong, J.; et al. Hepatitis C vaccination: where we are and where we need to be. *Pathogens* **2021**, *10* (12), 1619.

(2) Arca-Lafuente, S.; Martínez-Román, P.; Mate-Cano, I.; et al. Nanotechnology: A reality for diagnosis of HCV infectious disease. *J. Infect.* **2020**, *80* (1), 8−15.

(3) Messina, J. P.; Humphreys, I.; Flaxman, A.; et al. Global distribution and prevalence of hepatitis C virus genotypes. *Hepatology* **2015**, *61* (1), 77−87.

(4) mondiale de la Santé, O.; World Health Organization. Meeting of the Immunization and Vaccine-related Implementation Research Advisory Committee (IVIR-AC), February 2023−Réunion du Comité consultatif sur la vaccination et la recherche sur la mise en œuvre des vaccins (IVIR-AC), février 2023. *Weekly Epidemiol. Rec. = Relevé épidémiologique hebdomadaire* **2024**, *98* (13), 127−139.

(5) Forns, X.; Bukh, J.; Purcell, R. H. The challenge of developing a vaccine against hepatitis C virus. *J. Hepatol.* **2002**, *37* (5), 684−695.

(6) Wolf, M. C.; Freiberg, A. N.; Zhang, T.; et al. A broad-spectrum antiviral targeting entry of enveloped viruses. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (7), 3157−3162.

(7) Bailey, J. R.; Barnes, E.; Cox, A. L. Approaches, Progress, and Challenges to Hepatitis C Vaccine Development. *Gastroenterology* **2019**, *156* (2), 418−430.
(8) Grebely, J.; Prins, M.; Hellard, M.; et al. Hepatitis C virus clearance, reinfection, and persistence, with insights from studies of injecting drug users: towards a vaccine. *Lancet Infect. Dis.* **2012**, *12* (5), 408−414.
(9) Man John Law, L.; Landi, A.; Magee, W. C.; et al. Progress towards a hepatitis C virus vaccine. *Emerging Microbes Infect.* **2013**, *2* (1), 1−7.
(10) Thimme, R. T cell immunity to hepatitis C virus: Lessons for a prophylactic vaccine. *J. Hepatol.* **2021**, *74* (1), 220−229.
(11) Herzer, K.; Falk, C. S.; Encke, J.; et al. Upregulation of Major Histocompatibility Complex Class I on Liver Cells by Hepatitis C Virus Core Protein via p53 and TAP1 Impairs Natural Killer Cell Cytotoxicity. *J. Virol.* **2003**, *77* (15), 8299−8309.
(12) McKiernan, S. M.; Hagan, R.; Curry, M.; et al. Distinct MHC class I and II alleles are associated with hepatitis C viral clearance, originating from a single source. *Hepatology* **2004**, *40* (1), 108−114.
(13) Perot, B. P.; Ingersoll, M. A.; Albert, M. L. The impact of macroautophagy on CD 8+ T-cell-mediated antiviral immunity. *Immunol. Rev.* **2013**, *255* (1), 40−56.
(14) Kedzierska, K.; Koutsakos, M. The ABC of major histocompatibility complexes and T cell receptors in health and disease. *Viral Immunol.* **2020**, *33* (3), 160−178.
(15) Charoenkwan, P.; Waramit, S.; Chumnanpuen, P.; et al. TROLLOPE: A novel sequence-based stacked approach for the accelerated discovery of linear T-cell epitopes of hepatitis C virus. *PLoS One* **2023**, *18* (8), No. e0290538.
(16) Vita, R.; Zarebski, L.; Greenbaum, J. A.; et al. The Immune Epitope Database 2.0. *Nucleic Acids Res.* **2010**, *38* (suppl_1), D854−D862.
(17) Yu, K. K. Q.; Wilburn, D. B.; Hackney, J. A.; et al. Conservation of molecular and cellular phenotypes of invariant NKT cells between humans and non-human primates. *Immunogenetics* **2019**, *71*, 465−478.
(18) Ge, F.; Arif, M.; Yan, Z.; et al. MMPatho: Leveraging Multilevel Consensus and Evolutionary Information for Enhanced Missense Mutation Pathogenic Prediction. *J. Chem. Inf. Model.* **2023**, *63* (22), 7239−7257, DOI: 10.1021/acs.jcim.3c00950.
(19) Ge, F.; Zhu, Y. H.; Xu, J.; et al. MutTMPredictor: robust and accurate cascade XGBoost classifier for prediction of mutations in transmembrane proteins. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 6400−6416.
(20) Ge, F.; Zhang, Y.; Xu, J.; et al. Prediction of disease-associated nsSNPs by integrating multi-scale ResNet models with deep feature fusion. *Briefings Bioinf.* **2022**, *23* (1), No. bbab530.
(21) Wang, R.; Wang, Z.; Wang, H.; et al. Characterization and identification of lysine crotonylation sites based on machine learning method on both plant and mammalian. *Sci. Rep.* **2020**, *10* (1), No. 20447.
(22) Li, F.; Zhu, F.; Ling, X.; et al. Protein interaction network reconstruction through ensemble deep learning with attention mechanism. *Front. Bioeng. Biotechnol.* **2020**, *8*, 390.
(23) Chou, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21* (1), 10−19.
(24) Shi, H.; Zhang, S. Accurate prediction of anti-hypertensive peptides based on convolutional neural network and gated recurrent unit. *Interdiscip. Sci.: Comput. Life Sci.* **2022**, *14* (4), 879−894.
(25) Ahmad, S.; Charoenkwan, P.; Quinn, J. M. W.; et al. SCORPION is a stacking-based ensemble learning framework for accurate prediction of phage virion proteins. *Sci. Rep.* **2022**, *12* (1), No. 4106.
(26) Ferdous, S. M.; Mugdha, S. B. S.; Dehzangi, I. New Multi-View Feature Learning Method for Accurate Antifungal Peptide Detection. *Algorithms* **2024**, *17* (6), 247.
(27) Dubchak, I.; Muchnik, I.; Holbrook, S. R.; et al. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92* (19), 8700−8704.
(28) Yuan, S.-S.; Gao, D.; Xie, X. Q.; et al. IBPred: A sequence-based predictor for identifying ion binding protein in phage. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 4942−4951.
(29) Chao, W.; Quan, Z. A machine learning method for differentiating and predicting human-infective coronavirus based on physicochemical features and composition of the spike protein. *Chin. J. Electron.* **2021**, *30* (5), 815−823.
(30) Yang, W.; Liu, C.; Li, Z. Lightweight Fine-tuning a Pretrained Protein Language Model for Protein Secondary Structure Prediction *bioRxiv* 2023, p 2023-03.
(31) Zhang, M.; Gong, C.; Ge, F.; et al. FCMSTrans: Accurate Prediction of Disease-Associated nsSNPs by Utilizing Multiscale Convolution and Deep Feature Combination within a Transformer Framework. *J. Chem. Inf. Model.* **2024**, *64* (4), 1394−1406.
(32) Verbanck, M.; Josse, J.; Husson, F. Regularised PCA to denoise and visualise data. *Stat. Comput.* **2015**, *25* (2), 471−486.
(33) Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27* (8), 1226−1238.
(34) Ogundunmade, T. P.; Adepoju, A. A.; Allam, A. Stock price forecasting: Machine learning models with K-fold and repeated cross validation approaches *Mod. Econ. Manag.* 2022; Vol. *1* DOI: 10.53964/mem.2022001.
(35) Ershoff, B. D.; Lee, C. K.; Wray, C. L.; et al. Training and Validation of Deep Neural Networks for the Prediction of 90-Day Post-Liver Transplant Mortality Using UNOS Registry Data. *Transplant. Proc.* **2020**, *52* (1), 246−258.
(36) Charoenkwan, P.; Chumnanpuen, P.; Schaduangrat, N.; et al. PSRQSP: An effective approach for the interpretable prediction of quorum sensing peptide using propensity score representation learning. *Comput. Biol. Med.* **2023**, *158*, No. 106784.
(37) Hu, J.; Yu, W.; Pang, C.; et al. DrugormerDTI: Drug Graphormer for drug−target interaction prediction. *Comput. Biol. Med.* **2023**, *161*, No. 106946.
(38) Manavalan, B.; Lee, J. FRTpred: A novel approach for accurate prediction of protein folding rate and type. *Comput. Biol. Med.* **2022**, *149*, No. 105911.
(39) Firoz, A.; Malik, A.; Ali, H. M.; et al. PRR-HyPred: A two-layer hybrid framework to predict pattern recognition receptors and their families by employing sequence encoded optimal features. *Int. J. Biol. Macromol.* **2023**, *234*, No. 123622.
(40) Azadpour, M.; McKay, C. M.; Smith, R. L. Estimating confidence intervals for information transfer analysis of confusion matrices. *J. Acoust. Soc. Am.* **2014**, *135* (3), EL140−EL146.
(41) Nakashima, H.; Nishikawa, K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.* **1994**, *238* (1), 54−61.
(42) Huang, H.-L.et al. Prediction and Analysis of Protein Solubility Using a Novel Scoring Card Method with Dipeptide Composition. In *BMC Bioinformatics*; BioMed Central, 2012; Vol. *13*, pp 1−14.
(43) Rasmussen, M.; Fenoy, E.; Harndahl, M.; et al. Pan-specific prediction of peptide−MHC class I complex stability, a correlate of T cell immunogenicity. *J. Immunol.* **2016**, *197* (4), 1517−1524.
(44) Schaap-Johansen, A.-L.; Vujović, M.; Borch, A.; et al. T cell epitope prediction and its application to immunotherapy. *Front. Immunol.* **2021**, *12*, No. 712488.
(45) Miles, J. J.; Elhassen, D.; Borg, N. A.; et al. CTL recognition of a bulged viral peptide involves biased TCR selection. *J. Immunol.* **2005**, *175* (6), 3826−3834.
(46) O'Brien, C.; Flower, D. R.; Feighery, C. Peptide length significantly influences in vitro affinity for MHC class II molecules. *Immunome Res.* **2008**, *4* (1), 6.
(47) Donnison, T.; McGregor, J.; Chinnakannan, S.; et al. A pan-genotype hepatitis C virus viral vector vaccine generates T cells and neutralizing antibodies in mice. *Hepatology* **2022**, *76* (4), 1190−1202.