

RESEARCH

Open Access



# TIPred: a novel stacked ensemble approach for the accelerated discovery of tyrosinase inhibitory peptides

Phasit Charoenkwan<sup>1</sup>, Sasikarn Kongsompong<sup>2</sup>, Nalini Schaduangrat<sup>3</sup>, Pramote Chumnanpuen<sup>4,5\*</sup> and Watshara Shoombuatong<sup>3\*</sup>

\*Correspondence:  
pramote.c@ku.th; watshara.sho@mahidol.ac.th

<sup>1</sup> Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai 50200, Thailand

<sup>2</sup> Interdisciplinary Graduate Program in Bioscience, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand

<sup>3</sup> Center for Research Innovation and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

<sup>4</sup> Department of Zoology, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand

<sup>5</sup> Omics Center for Agriculture, Bioresources, Food, and Health, Kasetsart University (OmiKU), Bangkok 10900, Thailand

## Abstract

**Background:** Tyrosinase is an enzyme involved in melanin production in the skin. Several hyperpigmentation disorders involve the overproduction of melanin and instability of tyrosinase activity resulting in darker, discolored patches on the skin. Therefore, discovering tyrosinase inhibitory peptides (TIPs) is of great significance for basic research and clinical treatments. However, the identification of TIPs using experimental methods is generally cost-ineffective and time-consuming.

**Results:** Herein, a stacked ensemble learning approach, called TIPred, is proposed for the accurate and quick identification of TIPs by using sequence information. TIPred explored a comprehensive set of various baseline models derived from well-known machine learning (ML) algorithms and heterogeneous feature encoding schemes from multiple perspectives, such as chemical structure properties, physicochemical properties, and composition information. Subsequently, 130 baseline models were trained and optimized to create new probabilistic features. Finally, the feature selection approach was utilized to determine the optimal feature vector for developing TIPred. Both tenfold cross-validation and independent test methods were employed to assess the predictive capability of TIPred by using the stacking strategy. Experimental results showed that TIPred significantly outperformed the state-of-the-art method in terms of the independent test, with an accuracy of 0.923, MCC of 0.757 and an AUC of 0.977.

**Conclusions:** The proposed TIPred approach could be a valuable tool for rapidly discovering novel TIPs and effectively identifying potential TIP candidates for follow-up experimental validation. Moreover, an online webserver of TIPred is publicly available at <http://pmlabstack.pythonanywhere.com/TIPred>.

**Keywords:** Tyrosinase inhibitory peptides, Sequence analysis, Bioinformatics, Machine learning, Feature selection, Stacking strategy

## Background

Tyrosinase is a metalloenzyme that possesses a copper binding domain which is conserved across different organisms including fruits, vegetables, fungi, mammals, and insects that utilize it for cuticle sclerostasis and wound healing [2, 3]. The enzyme catalyzes



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

the transformation of tyrosine, an amino acid, into DOPA (dihydroxyphenylalanine), which is subsequently converted into melanin—the pigment responsible for determining skin, hair, and eye color [1, 4]. Moreover, this enzyme also plays a role in the biosynthesis of other pigments such as dopamine and norepinephrine [2]. Overproduction of melanin and instability of tyrosinase activity could cause several hyperpigmentation disorders, which are the conditions that result in excessive skin pigmentation and cause darker, discolored patches on the skin [3]. These disorders can be caused by a variety of factors, including sun exposure, hormonal changes, inflammation, genetics, and certain medications [4, 5]. Examples of common hyperpigmentation disorders include melasma, age spots, and post-inflammatory hyperpigmentation [4]. Treatment options for hyperpigmentation disorders include topical lightening agents, chemical peels, and laser therapy. In some cases, reducing exposure to triggers and protecting the skin from further sun damage can also help reduce the appearance of hyperpigmentation [4–6].

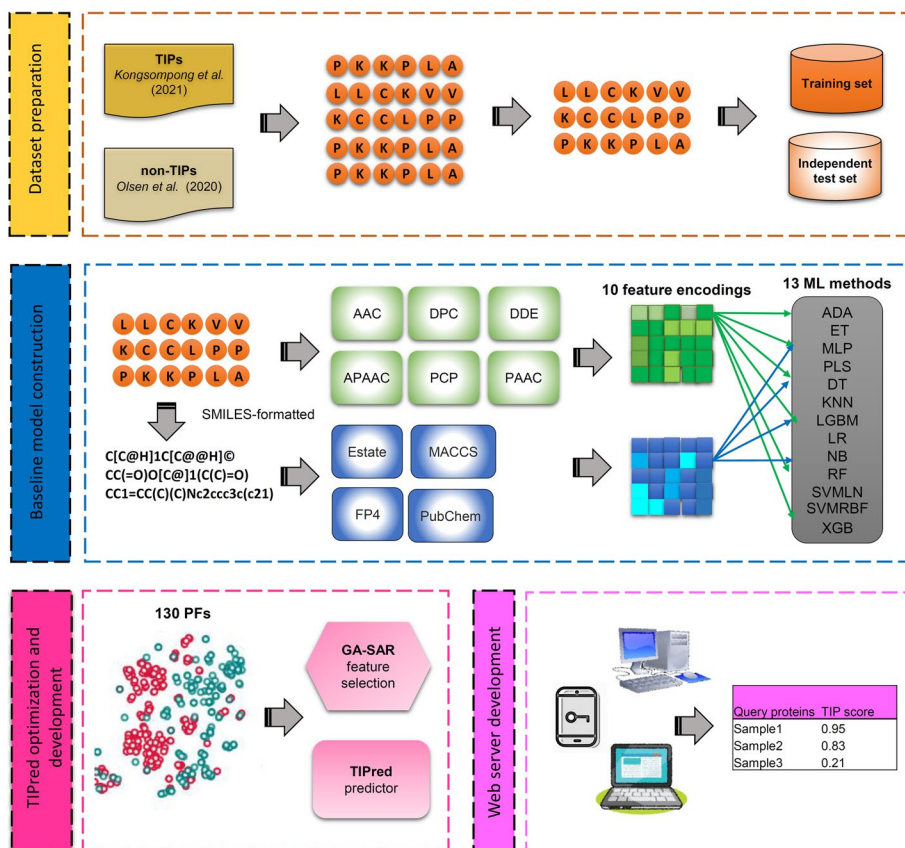
Substances that can hinder the function of the enzyme tyrosinase are known as tyrosinase inhibitors, and are frequently utilized in skin lightening products aimed at reducing the visibility of hyperpigmentation and dark spots on the skin [5]. Some natural tyrosinase inhibitors include kojic acid, arbutin, and licorice extract [7, 8]. On the other hand, chemical tyrosinase inhibitors include hydroquinone, azelaic acid, and glycolic acid [9, 10]. However, these chemical whitening agents may lead to various undesirable side effects such as skin irritation, allergic reactions, sensitivity to sunlight, discoloration, and exogenous ochronosis [11–14]. Tyrosinase inhibitory peptides (TIPs) refer to short chains of amino acids, usually comprised of 3–20 units, that are capable of impeding the activity of the tyrosinase enzyme. This enzyme is responsible for the synthesis of melanin, the pigment that determines skin color [17]. Recently, bioactive peptides have become an increasingly popular medicinal agent, and TIPs derived from food sources are highly favored due to their excellent biological safety and ease of absorption. These peptides are viewed as a promising alternative to chemical tyrosinase inhibitors like hydroquinone [17–19]. In addition to TIPs, amino acids released during digestion in the gastrointestinal tract can also be completely absorbed even without ingestion [15]. The clinical trials for novel TIPs derived from various animal and plant sources are currently underway [16–18].

Therefore, the identification of TIPs through the use of sequence information is crucial to accelerate their implementation in clinical settings. In this regard, machine learning (ML) techniques have been explored to facilitate the high-throughput discovery of new TIPs. Currently, only one computational method has been developed for the identification of TIPs. This method was introduced by Kongsompong et al. [19]. Specifically, this group employed two popular ML methods (random forest (RF) and k-nearest neighbour (KNN)) trained with three interpretable feature descriptors (amino acid composition (AAC), physicochemical properties (PCP), and dipeptide composition (DPC)). These KNN and RF classifiers were trained and evaluated on the dataset consisting of 133 TIPs and 13 non-TIPs. The performance of KNN and RF classifiers were 0.97 and 0.99, respectively, in terms of accuracy (ACC) on the independent test dataset. Although Kongsompong's method provides a high prediction performance, their method has a few flaws and needs to be improved. Firstly, the method was developed using a small number of negative samples. Hence, their

performance in non-TIP identification might not be satisfactory. Secondly, this study did not offer a comparative analysis of the impact of well-known feature encodings and ML algorithms on TIP prediction. Thirdly, this study did not provide a web server.

Considering these limitations, we introduce TIPred for the large-scale identification of TIPs by using only peptide sequence information. The design and development of TIPred is summarized in Fig. 1. Major contributions of this study are listed as follows:

- (i) To the best of our knowledge, TIPred is the first stacked ensemble approach developed for the identification and characterization of TIPs.
- (ii) TIPred employed different feature encoding schemes from multiple aspects, including, amino acid composition, chemical structure properties, physicochemical properties and pseudo-amino acid composition, integrated state-of-the-art ML classifiers to develop a more stable meta-model. In addition, we investigated the contributions of different types of feature encodings in TIP prediction.
- (iii) The independent test results indicated that TIPred achieved a better performance compared to the existing method and several conventional ML classifiers in terms of ACC (0.923), Matthew’s correlation coefficient (MCC) (0.744) and area under the receiver operating characteristics (ROC) curve (AUC) (0.964).



**Fig. 1** System flowchart of the proposed TIPred. The overall workflow for the development of TIPred contains four major steps: dataset preparation, baseline model construction, TIPred optimization, and web server development

- (iv) In TIPred, we utilized an interpretable Shapley Additive exPlanation (SHAP) approach to provide a better understanding of the functional mechanisms of TIPs.
- (v) TIPred-assisted virtual screening approach was introduced and used for the accelerated discovery of novel TIPs.

## Materials and methods

### Construction of training and independent datasets

In this study, the positive dataset containing 133 TIPs was derived from the study of Kongsompong et al. [19]. These TIPs are peptides that have been experimentally verified as having tyrosinase inhibitory properties. Until now, there has been no source of experimentally verified non-TIPs. Therefore, to establish a dependable dataset, the non-antioxidative peptides obtained from Olsen et al. [20] were utilized to create the negative dataset in this study. TIPs usually exhibit dual activities, they can inhibit tyrosinase through the photoaging system by binding to the catalytic domain, as well as directly scavenge free radicals as antioxidants [15, 18, 21–23]. Thus, we selected peptides that were empirically confirmed as non-effective in both categories of antioxidant properties (i.e., free radical scavengers and iron chelators) as negative samples. After removal of duplicated sequences, 287 non-TIPs were obtained. Finally, the benchmark dataset contained 133 TIPs and 287 non-TIPs. Among these, 106 TIPs and 230 non-TIPs were randomly selected to construct the training dataset (called TIP-TRN), the remaining peptides were used to construct the independent test dataset (called TIP-IND).

### Feature encoding methods

To encode the TIPs and non-TIPs, we utilized 10 different feature encoding schemes, which are categorized into multiple groups, namely amino acid composition (AAC, DPC, and DDE), pseudo-amino acid composition (APAAC and PAAC), physicochemical properties (PCP) and chemical structure properties (Estate, FP4, MACCS, and PubChem). Among these feature encoding schemes, AAC, APAAC, DDE, DPC, PAAC, and PCP, which are known as sequence-based feature descriptors, can be used to encode FASTA-formatted TIPs and non-TIPs into fixed-length feature vectors by considering the 20 standard amino acids along with the *iFeature* Python package [24–26]. By using the remaining feature encoding schemes, the FASTA-formatted TIPs and non-TIPs were converted into their corresponding chemical structures (SMILES format) by using the *RDKit* software [27]. Then, the Chemistry Development Kit (CDK) was used to encode SMILES-formatted TIPs and non-TIPs into fixed-length feature vectors [26, 28–31]. Details of all the 10 feature encodings are summarized in Table 1.

### Feature selection technique

Training a prediction model with high-dimensional input feature vectors can cause overfitting and underfitting issues. In this context, the feature selection method is needed to enhance the prediction performance and optimize computational time [28–30, 32]. In this study, we used our proposed genetic algorithm (GA-SAR) for constructing an optimal feature set containing  $m$  useful features [25, 26, 33]. In 2019, Charoenkwan et al. initially introduced this method for the interpretable identification of quorum sensing

**Table 1** Summary of ten different feature encodings along with their corresponding description and dimension

Order	Descriptors	Description	Dimension	References
1	AAC	Frequency of 20 amino acids	20	[59, 60]
2	APAAC	Amphiphilic pseudo-amino acid composition	22	[61]
3	DDE	Dipeptide deviation from expected mean	400	[24]
4	DPC	Frequency of 400 dipeptides	400	[59, 60]
5	PCP	Different biochemical and biophysical properties extracted from the AAindex database	11	[59, 60]
6	PAAC	Pseudo amino acid composition	21	[61]
7	Estate	Electrotopological state atom types	79	[29, 30, 62]
8	FP4	Presence of SMARTS patterns for functional groups	307	[29, 30, 63]
9	MACCS	Binary representation of chemical features defined by MACCS keys	166	[29, 30, 64]
10	Pubchem	Binary representation of substructures defined by PubChem	881	[29, 30, 65]

peptides [33]. Until now, the GA-SAR method has been applied for the prediction and characterization of many protein and peptide functions [25, 26, 34, 35]. In brief, the chromosomes of the GA-SAR consist of two main genes, namely binary gene and parametric gene. The chromosomes and gene of the GA-SAR are referred as GA-chrom and GA-gene herein, respectively. Herein, the parameters and their values for the GA-SAR consist of  $m_{start} = 5$ ,  $m_{end} = 20$ ,  $P_m = 0.05$ , and  $Pop = 50$ . Detailed report regarding this algorithm is provided in our previous studies [25, 26, 33] along with the Additional file 1 [25, 26, 33].

#### The architecture of the proposed model TIPred

Herein, TIPred was developed by using the stacking strategy. Stacking is well-known as a powerful ensemble learning approach that is able to automatically combine multiview information derived from different ML classifiers as means to create a more accurate and stable predictor [25, 26, 29, 30]. Thus, the development of our proposed TIPred involves two main steps, including (1) baseline model construction and (2) meta-model development.

In the first part, we encoded the TIP-TRN dataset using 10 types of feature encodings, including DPC, PAAC, PCP, AAC, DDE, APAAC, FP4, Estate, PubChem, and MACCS, in combination with 13 ML methods, including ADA, ET, MLP, PLS, DT, KNN, LGBM, LR, NB, RF, SVMLN, SVMRBF and XGB, for the baseline model development. Specifically, the baseline models were created based on a wide range of feature encodings from multiple perspectives, including amino acid composition, chemical structure properties, pseudo-amino acid composition, and physicochemical properties [36–40]. In total, 130 baseline models were trained and constructed by using the scikit-learn package (Table 2). In addition, we conducted a comprehensive assessment of all the 130 baseline models in TIP prediction by performing both cross-validation and independent tests. Herein, the best-performing baseline model was indicated by using the Matthew's Correlation Coefficient (MCC) on the TIP-TRN dataset.

In the second part, we generated a new probabilistic feature vector (PFV) of 130 dimension (130-D) by using 130 probabilistic features (PFs) derived from all the

**Table 2** Parameter search details used for the construction of nine ML-based classifiers

Method <sup>a</sup>	Parameters	Range of parameters
ADA	n_estimators	[20, 50, 100, 200, 500]
ET	n_estimators	[20, 50, 100, 200, 500]
LGBM	n_estimators	[20, 50, 100, 200, 500]
LR	Cost	[0.001, 0.01, 0.1, 1, 10, 100]
MLP	hidden_layer_sizes	[20, 50, 100, 200, 500]
RF	n_estimators	[20, 50, 100, 200, 500]
SVMLN	Cost	[2 <sup>0</sup> to 2 <sup>5</sup> ] in log <sub>2</sub> steps
SVMRBF	Cost	[2 <sup>-4</sup> to 2 <sup>4</sup> ] in log <sub>2</sub> steps
XGB	n_estimators	[20, 50, 100, 200, 500]

<sup>a</sup> ADA: AdaBoost, DT: decision tree, ET: extremely randomized trees, KNN: k-nearest neighbor, LGBM: light gradient boosting machine, LR: logistic regression, MLP: multilayer perceptron, NB: naive Bayes, PLS: partial least squares, RF: random forest, SVMRBF: support vector machine with radial basis function, SVMLN: support vector machine with linear kernels, XGB: extreme gradient boosting

130 baseline models and tenfold cross-validation scheme, where PFs were the predicted confidence of TIPs. Then, the 130-D feature vector was used to develop the meta-model based on PLS (called mPLS) [25]. Although the 130-D feature vector contains only 130 PFs, some of these PFs involve redundant and noisy information. Thus, the GA-SAR was used to establish an optimal feature set containing  $m$  useful PFs. Specifically, the GA-SAR's chromosome used herein involved  $n = 130$  features. As a result, the GA-chrom contains 130 binary GA-genes ( $f_i$ ). If the  $i^{th}$  PP is considered as a useful feature when  $f_i = 1$ ; otherwise, the  $i$ th feature is not considered. Finally, the feature set exhibiting the highest cross-validation MCC was deemed as the optimal one. Furthermore, additional evaluation metrics, including AUC, ACC, MCC, balanced accuracy (BACC), sensitivity (Sn), and specificity (Sp), were selected to evaluate the effectiveness of our proposed model. The descriptions of these evaluation metrics can be found in the Additional file 1 [40, 41].

### Screening novel TIPs

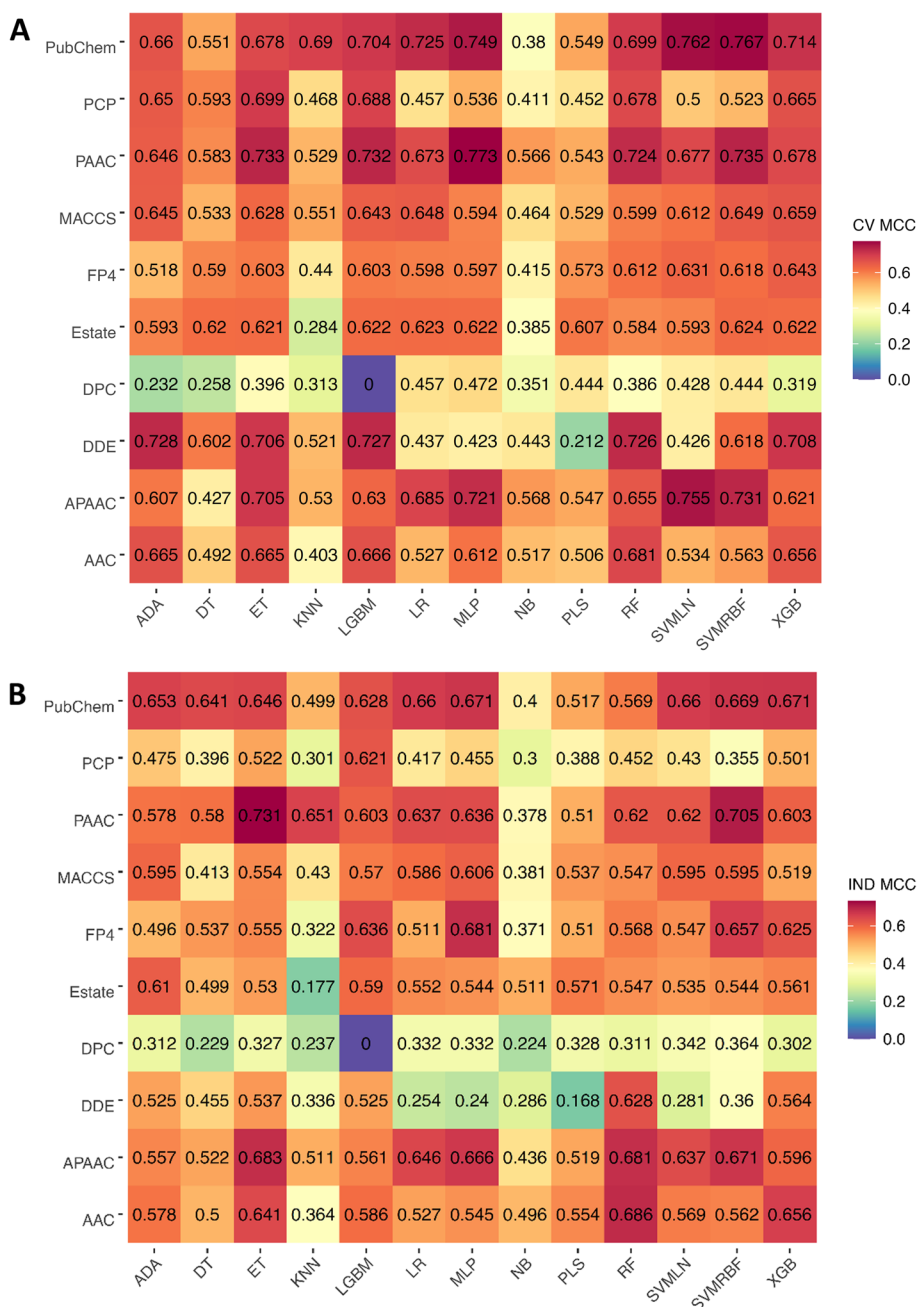
In this study, our proposed model was employed to perform a large-scale identification of TIPs from the putative hempseed (*Cannabis sativa*) trypsinized peptidome derived from a previous study [42]. Only 73 unique peptides with the proper amino acid length (10–57) were considered for our analysis. The putative peptides from *Cannabis sativa* seed having the highest probabilistic scores were deemed as candidate TIPs. After that, the molecular docking approach was used to assess the ability of the selected TIPs to bind to the active site of the tyrosinase enzyme. Specifically, the molecular docking between the selected TIPs and the polyphenol oxidase domain (chains A–D) of the crystal structure of mushroom tyrosinase from *Agaricus bisporus* (PDB: 2Y9X) was performed using two protein-peptide docking web servers, namely [43] GalaxyPepDock (<http://galaxy.seoklab.org/pepdock>) and HPEPDOCK (<http://huanglab.phys.hust.edu.cn/hpepdock/>). The GalaxyPepDock server was used to conduct the template-based molecular docking simulation, while the HPEPDOCK server was used to estimate the template-free (global) molecular docking scores.



## Results and discussion

### Investigation of the contribution of different machine learning methods and feature encodings

In this section, we investigated the contribution of different types of feature encodings in TIP prediction. Thus, all the 10 feature encodings were assessed pairwise using all the 13 ML methods in terms of tenfold cross-validation and independent tests. Figure 2 and Additional file 1: Tables S1-S3 detail the predictive performance of the 130 different ML classifiers. Additional file 1: Table S3 shows that the highest average MCC of 0.664 is



**Fig. 2** MCC values of 130 baseline models in terms of tenfold cross-validation (A) and independent (B) tests

achieved by using PubChem, while the second and third highest average MCC of 0.661 and 0.629 were achieved by using PAAC and APAAC, respectively. And, we noticed that there were ten PubChem-based (range 0.660–0.767), nine PAAC-based (range 0.646–0.773), nine APAAC-based (range: 0.607–0.755) classifiers with MCC greater than 0.6 (see Fig. 2). Interestingly, all the top ten ML classifiers were developed based on PubChem, PAAC, and APAAC, including MLP-PAAC, SVMRBF-PubChem, SVMLN-PubChem, SVMLN-APAAC, MLP-PubChem, SVMRBF-PAAC, ET-PAAC, LGBM-PAAC, and SVMRBF-APAAC. This demonstrates that these feature encodings could be beneficial for TIP prediction. Although it could be noticed that MLP-PAAC attained the highest performance in terms of ACC (0.882) and MCC (0.767) on the TIP-TRN dataset, this classifier failed to achieve a better performance on the TIP-IND dataset, with ACC of 0.870, MCC of 0.636, and AUC of 0.960. This evidence indicates that the performance of a single feature-based models is not stable on the TIP-IND dataset. To address this issue, we were motivated to generate a more comprehensive and reliable model by using the stacking strategy.

#### Performance evaluation of TIPred

Herein, we utilized the stacked strategy to create a stacked model by using PLS method in conjunction with the 130-D feature vector generated from multiple baseline models. To maximize the utility of the 130-D feature vector, this feature vector was optimized by using the GA-SAR as means to determine the optimal number ( $m$ ) of PFs. In this study, the optimal number of PFs was 11 (or the 11-D feature vector). Specifically, the 11-D feature vector was generated by the baseline models of ET-DDE, MLP-PubChem, XGB-PubChem, SVMRBF-APAAC, NB-DDE, ADA-FP4, NB-Estate, LR-MACCS, SVMRBF-FP4, MLP-FP4, and PLS-PubChem. The performance of the 130-D and 11-D feature vectors are recorded in Table 3. As seen in Table 3, MCC, ACC, Sn, and Sp of the 11-D feature vector are 0.920, 0.958, 0.945, and 0.973, which are 13.21, 6.58, 6.55, and 6.64%, respectively, higher than the 130-D feature vector in terms of the tenfold cross-validation test. Furthermore, in case of the independent test results, the 11-D feature vector still achieved the overall best performance compared to the 130-D feature vector. In this context, we utilized the 11-D feature vector to build our proposed model, TIPred.

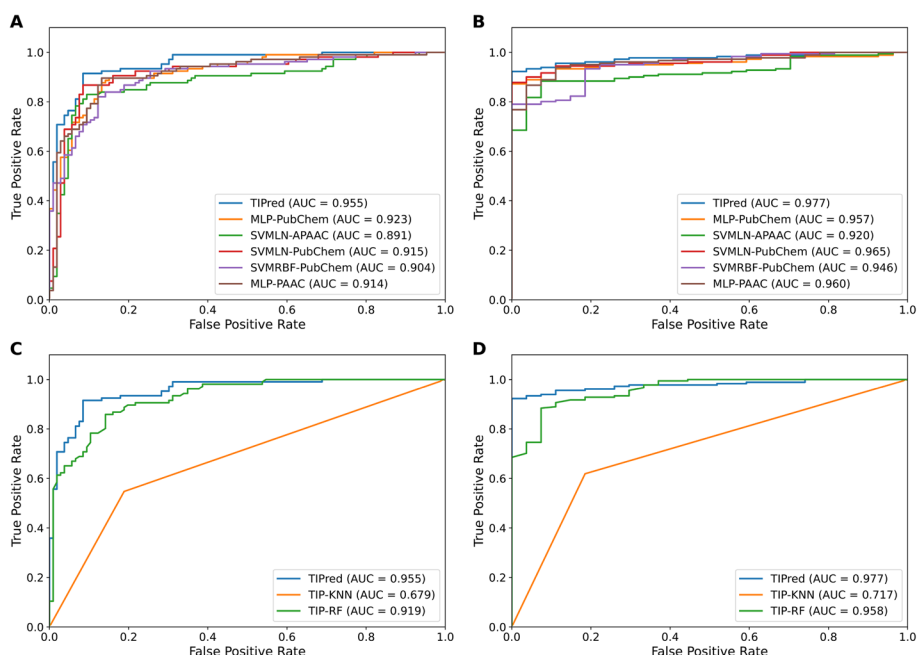
#### The Stacking model is capable of improving the predictive performance

In this section, we aim to highlight the improved performance provided by the stacking strategy, by comparing the performance of TIPred with BLAST-based predictors and the top five baseline models (i.e., MLP-PubChem, SVMLN-APAAC, SVMLN-PubChem,

**Table 3** Cross-validation and independent test results for the control and optimal model

Evaluation strategy	Feature	Number of feature	ACC	BACC	Sn	Sp	MCC	AUC
Cross-validation	AFV	130	0.869	0.869	0.870	0.868	0.741	0.942
	BFV	11	0.916	0.917	0.917	0.917	0.837	0.955
Independent test	AFV	130	0.909	0.948	1.000	0.895	0.725	0.989
	BFV	11	0.923	0.956	1.000	0.912	0.757	0.977





**Fig. 3** Performance comparison of TIPred with related methods in terms of tenfold cross-validation (A, C) and independent (B, D) tests. (A, B) ROC curves of TIPred and top five baseline models. (C, D) ROC curves of TIPred and existing methods

**Table 4** Performance comparison of TIPred and top five ML classifiers

Evaluation strategy	Method	ACC	BACC	Sn	Sp	MCC	AUC
Cross-validation	MLP-PubChem	0.873	0.872	0.850	0.895	0.749	0.923
	SVMMLN-APAAC	0.873	0.874	0.916	0.831	0.755	0.891
	SVMMLN-PubChem	0.878	0.877	0.887	0.866	0.762	0.915
	SVMRBF-PubChem	0.882	0.881	0.869	0.894	0.767	0.904
	MLP-PAAC	0.882	0.884	0.871	0.897	0.773	0.914
	TIPred	0.916	0.917	0.917	0.917	0.837	0.955
Independent test	MLP-PubChem	0.889	0.921	0.963	0.878	0.671	0.957
	SVMMLN-APAAC	0.880	0.899	0.926	0.873	0.637	0.920
	SVMMLN-PubChem	0.875	0.928	1.000	0.856	0.660	0.965
	SVMRBF-PubChem	0.880	0.931	1.000	0.862	0.669	0.946
	MLP-PAAC	0.870	0.910	0.963	0.856	0.636	0.960
		TIPred	0.923	0.956	1.000	0.912	0.757

SVMRBF-PubChem, and MLP-PAAC). Additional file 1: Table S4 presents the independent test results of BLAST-based predictors with various *E*-values. It is worth noting that the highest MCC of 0.406 was achieved using an *E*-value cutoff value of 0.1. However, the Sn of this optimal cutoff value was unsatisfactory (Sn of 0.185), while Sn of other cutoff values were in the range of 0.037–0.111. This demonstrated that the BLAST-based predictor was not capable of precisely identifying true TIPs. As can be seen from Fig. 3 and Table 4, TIPred outperformed the top five baseline models in terms of all performance metric on the TIP-IND dataset. Specifically, the BACC, Sn, Sp, and AUC of TIPred were 0.959, 1.000, 0.912, and 0.757, which were 4.61, 3.70, 5.52, and 12.06%,

respectively, higher than that of the best-performing baseline model (MLP-PAAC). This indicates that the stacked ensemble learning approach is indeed effective in improving the performance of TIP prediction.

#### TIPred outperforms the existing method

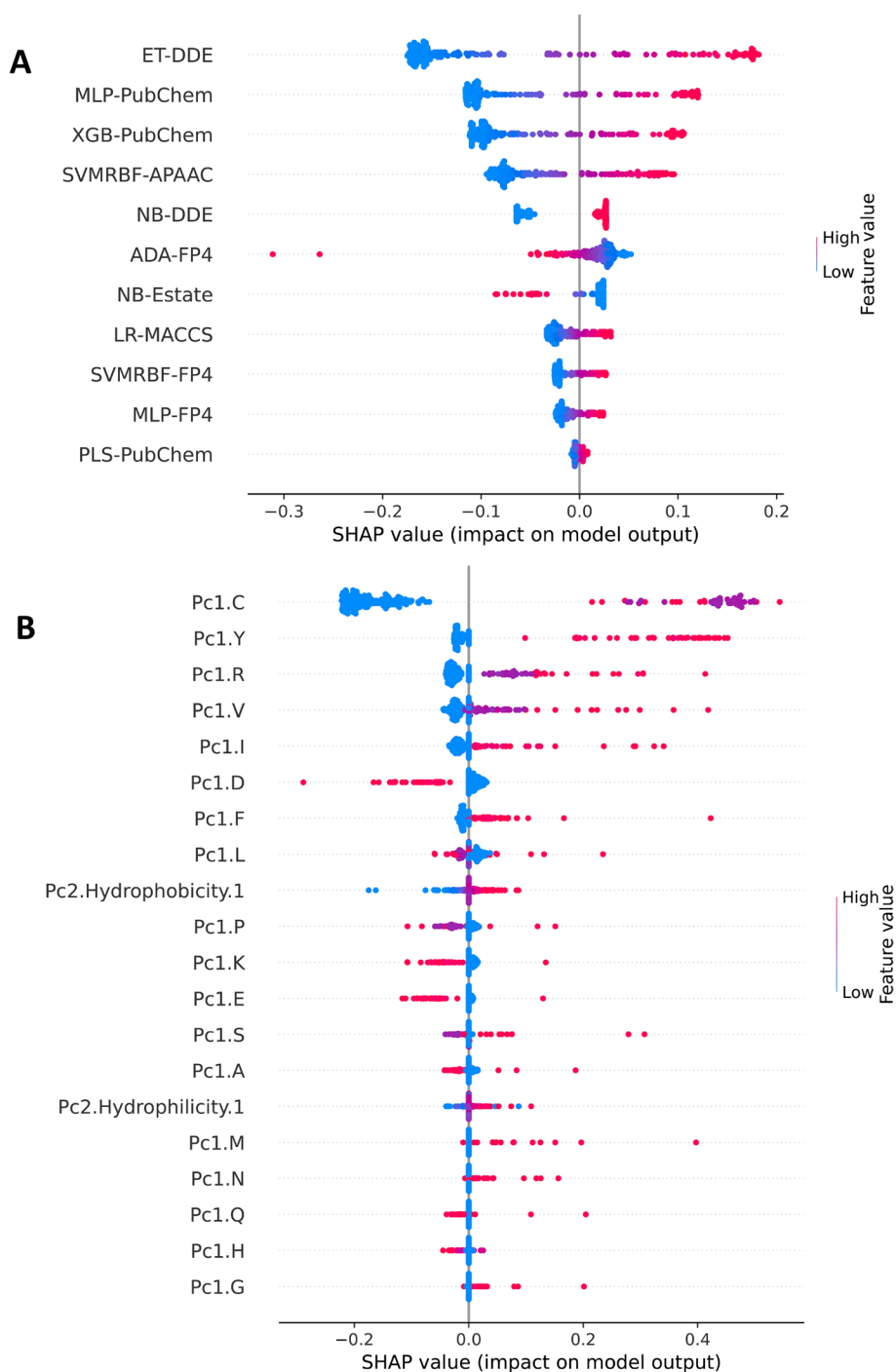
To demonstrate the effectiveness of the model, it is necessary to compare the proposed model TIPred with the existing method [19]. As the existing method did not provide a webserver, we implemented KNN-based and RF-based classifiers by strictly utilizing the same procedure reported in the study of Kongsompong et al. [19] using the same training dataset. Table 5 illustrates that our proposed model, TIPred, achieved the best performance as judged by five out of six evaluation metrics (i.e., ACC, BACC, Sn, Sp, MCC, and AUC) on both the TIP-TRN and TIP-IND datasets. Specifically, the BACC, Sn, and MCC values achieved by TIPred were 8.98, 18.52, and 11.33% higher than RF-based classifier in terms of the TIP-IND dataset. Altogether, based on both the cross-validation and independent test results, TIPred consistently demonstrated a stable performance, indicating its effectiveness and robustness. In addition, the higher values of Sp and MCC in terms of the TIP-IND dataset are sufficient to elucidate that TIPred could effectively reduce the number of false positives, which plays a crucial role for minimizing the experimental costs and burden.

#### Feature importance analysis

The SHAP framework is well-known as an interpretable and powerful framework used to provide information about how features can affect the output of the model. Therefore, we utilized this framework to analyze the prediction outputs of the proposed TIPred and its baseline models. Figure 4A and Additional file 1: Fig. S1A demonstrate the impact of the 11 PFs on the prediction of TIPred, where the positive and negative SHAP values indicate the probability that the prediction outputs are relatively positive and negative classes, respectively. We obtained the top-eight informative PFs with the highest SHAP values from eight baseline models of SVM-LN-PubChem, ET-DDE, SVM-RBF-APAAC, MLP-PubChem, LGBM-PubChem, XGB-MACCS, PLS-Estate, and PLS-APAAC (refer to Fig. 4A and Additional file 1: Table S5). Taking SVM-LN-PubChem as an example, peptide sequences with high PF values of SVM-LN-PubChem have a high probability of being TIPs. On the other hand, peptide sequences with high PF values of PLS-Estate have a high possibility of being non-TIPs. Among the top-eight informative PFs, SVM-RBF-APAAC was found to be the fourth-best informative PF. Figure 4B along with

**Table 5** Performance comparison of TIPred and the existing predictors

Evaluation strategy	Method	ACC	BACC	Sn	Sp	MCC	AUC
Cross-validation	TIP-KNN	0.680	0.679	0.811	0.547	0.383	0.679
	TIP-RF	0.845	0.845	0.821	0.869	0.695	0.919
	TIPred	0.916	0.917	0.917	0.917	0.837	0.955
Independent test	TIP-KNN	0.644	0.717	0.815	0.619	0.294	0.717
	TIP-RF	0.904	0.866	0.815	0.917	0.643	0.958
	TIPred	0.923	0.956	1.000	0.912	0.757	0.977



**Fig. 4** Feature importance from TIPred (A) and SVMRBF-APAAC (B) as ranked by SHAP values based on the training dataset. Color indicates the magnitude and direction of the contribution of features to TIPred and SVMRBF-APAAC for TIP prediction, where positive and negative SHAP values indicate the probability that the prediction outputs are positive and negative samples, respectively

Additional file 1: Fig. S1B and Table S6 display the impact of top 20 informative features on the prediction of SVMRBF-APAAC. Based on the SHAP values, we noticed that the ten top-ranked features consist of Cys, Tyr, Arg, Val, Ile, Asp, Phe, Leu, hydrophobicity,

and Pro. As shown in Fig. 4B, Cys, Tyr, Arg, Val, Ile, Phe, and hydrophobicity are abundant in TIPs compared to non-TIPs, while Asp, Leu, and Pro are abundant in non-TIPs compared to TIPs.

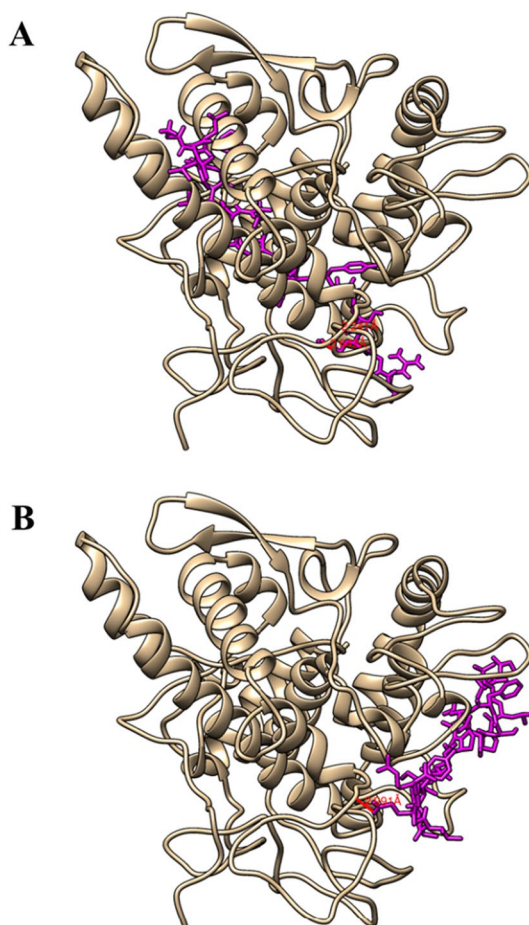
The amino acid composition findings are consistent with previous reports, indicating that TIPs tend to contain higher levels of certain amino acids, including Cys, Tyr, Arg, Phe, and Met [44–49]. Some of the well-known TIPs were rich in Ser, Trp, Arg, and sulfur-containing amino acids (Cys and Met). These characteristics are typical of well-known peptides that inhibit tyrosinase and chelate metals. The sulfur-containing amino acids, Cys and Met, have been found to be associated with tyrosinase inhibition and copper chelation in natural TIPs derived from hydrolyzed rice-bran-derived albumin [47]. Schurink et al. [42] observed that peptides with polar, uncharged amino acids, particularly Cys, are effective tyrosinase inhibitors that have a high reductive effect on the melanin synthesis pathway. According to research, Cys plays a role in limiting tyrosinase activity by converting o-quinone intermediates into stable colorless cysteine-quinone adducts, lowering polyphenol precursor levels and preventing the formation of polymeric melanin products. Cys-containing peptides can also compete with catalytically active copper ions, preventing them from binding to tyrosinase [48, 49].

Interestingly, hydrophobicity, a feature of physicochemical property has also been found in the ten top-ranked features together with other amino acids. There is also supporting evidence suggesting that hydrophobic amino acid residues can enhance tyrosinase inhibition. The hydrophobic nature of TIPs, including amino acids such as Phe, Trp, Met, and particularly Ala, has been found to play an important role in inhibiting melanogenesis [45]. Furthermore, it was observed that the aromatic amino acid Phe has the ability to stabilize free radicals through electron donation and maintain its antioxidant stability through its resonant structure [46–48]. The peptides containing the cationic amino acid Arg have been found to possess remarkable activity in chelating copper ions and generally exhibit excellent tyrosinase binding properties, owing to the presence of the guanidine group [42]. Interestingly, the dipeptide Arg-Lys, which has been reported as the active composition in the TIPs fraction of hydrolyzed rice albumin [47], was not among the top ten SHAP values identified in this study.

#### TIPred-assisted virtual screening for novel TIPs identification

Herein, we applied TIPred-assisted virtual screening approach for the identification of novel TIPs. To showcase TIPs prediction and screening, hempseed (*Cannabis sativa*) peptidome was the most suitable choice due to the fact that it is a high-protein plant source (20–25% content) that has been extensively utilized in the development of numerous products for the cosmetics, therapeutic, functional food, and nutraceutical industries [50, 51]. Hempseed protein hydrolysates have been recognized as a valuable source of bioactive peptides with various health-promoting effects [52]. As a result, hempseed peptides have garnered attention for their potential bioactive pharmaceutical properties, including antioxidant and tyrosinase inhibitory abilities. Additional file 1: Table S7 lists the probabilistic scores of the 73 putative peptides from *Cannabis sativa* seed. As mentioned above, we selected the candidate TIPs in terms of the probabilistic score and considered as potential TIPs. The top five TIPs with a probabilistic score of 1.000 were identified, including A-2 (ISSSTLALFAALMLVAHAVAFR),

E1–9 (YTIQQNGLHLPSYTNTPQLVYIVK), E2–12 (GLLLPSFLNAPMMFYVIQGR), E3–38 (NAMYPQYTMNAHNIIYAIR), and E3–6 (LTIQPNGLHLPSYTNQPLIHVIR). This suggests that these peptides are highly likely to have TIP activity and can be considered as potential TIPs for further validation. To further demonstrate the effectiveness of these top-five potential TIPs, we performed molecular docking between these TIPs and the polyphenol oxidase domain (chain A–D) of the crystal structure of mushroom tyrosinase by using GalaxyPepDock and HPEPDOCK (Fig. 5 and Additional file 1: Figures S2–S6). The calculated binding affinity between the TIP candidates and tyrosinase was represented by molar Gibbs free energy ( $\Delta G$ ) and the equilibrium dissociation constant (Kd). As shown in Table 6, the scores ( $\Delta G$ , Kd, and molecular docking score) of the top-five potential TIPs were  $-11.6$  to  $-9.4$  kcal/mol,  $3.1E-09$  to  $1.6E-07$ , and  $-201.2040$  to  $134.0631$  kJ/mol, respectively. Among the top-five potential TIPs, E2–12 outperformed other peptides in terms of molecular docking score. To be specific, the  $\Delta G$ , Kd, and molecular docking score of E2–12 were  $-10.1$ ,  $4.0E-08$ , and  $-201.2040$ , respectively.



**Fig. 5** Molecular docking of E2–12 to the tyrosinase (PDB: 2Y9X) generated from GalaxyPepDock (**A**) and HPEPDOCK (**B**). The tyrosinase is shaded in gold, while the peptide sequences and hydrogen bonds are shown in pink and red, respectively

**Table 6** Calculated binding affinity ( $\Delta G$ ), dissociation constant (Kd), and binding energy scores from the molecular docking results from HPEPDOCK of csTIPs and P4 to the tyrosinase (2Y9X chainD) based on the PROGIDY and PIMA web servers

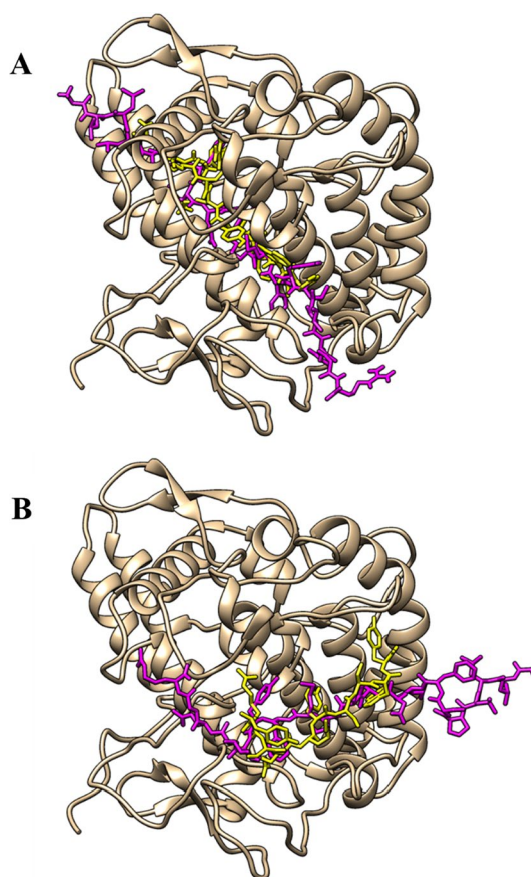
Name	$\Delta G$ (kcal/mol)	Kd at 25.0 °C	H-bond ener. (kJ/mol)	Elec. ener. (kJ/mol)	VDW. ener. (kJ/mol)	Molecular docking score (kJ/mol)
A-2	- 9.4	1.4E-07	- 5.4749	- 5.7960	145.3340	134.0631
E1-9	- 11.6	3.1E-09	- 3.1403	1.8997	- 90.3478	- 91.5884
E2-12	- 10.1	4.0E-08	- 12.6447	- 6.7877	- 181.7720	- 201.2040
E3-6	- 11.0	8.3E-09	- 4.5694	4.2906	- 141.8230	- 142.1020
E3-38	- 10.8	1.2E-08	- 13.4909	0.0000	- 48.8050	- 62.2959
P4	- 9.3	1.6E-07	- 34.2300	24.0025	- 9.2737	- 19.5013

To indicate the effectiveness of E2-12, we compared its performance with the commercial TIP, i.e., P4 (YRSRKSSWP) or decapeptide-12. There have been multiple studies that indicate P4 as the most well-known peptide that is currently being used as the primary active component in the Lumixyl™ skin brightening product [17, 53]. The common binding position of the highest molecular docking scored csTIP candidate (E2-12) and the positive control peptide (P4) was revealed by the comparative molecular docking on the crystal structure of tyrosinase (Fig. 6). The molecular docking result could confirm the overlapped interacting regions on the active site of tyrosinase structure on both protein-peptide docking methods (GalaxyPepDock and HPEPDOCK). Table 6 indicates that E2-12 exhibited a better potential inhibition of tyrosinase as compared to the commercial TIP P4 in terms of  $\Delta G$  (- 10.1 versus - 9.3), Kd (4.0E-08 versus 1.6E-07), and molecular docking score (- 201.2040 versus - 19.5013). Furthermore, all the hydrogen bonds between peptides and tyrosinase proposed by the molecular docking simulation are listed in Additional file 1: Table S8. According to the molecular docking simulation experiments, the distance in the hydrogen bond between E2-12 and tyrosinase structures was 1.5–3.3 Å, implying that E2-12 could be deemed as a moderately strong covalent interaction [54]. Similar to the analyzed results of abalone biomimetic peptides (hdTIPs) [43], the cationic amino acid residues (Arg20, 38, 95, 268, 321) of the catalytic domain on tyrosinase seems to be the key binding target of E2-12 and other hempseed TIPs. These results indicate that the peptide E2-12 as derived from this study could be a promising TIP. Altogether, the virtual screening result of TIPred is adequate to demonstrate that TIPred has the potential to be a useful and efficient tool for quickly screening and identifying promising TIPs.

## Conclusion

This study introduces a novel stacked ensemble approach (termed TIPred) for the accurate and high-throughput identification of TIPs. TIPred combines a comprehensive set of feature encoding schemes from multiple aspects, such as chemical structure properties, physicochemical properties, and composition information, with 13 well-known ML methods to create a more stable model. The experimental results of both the tenfold cross-validation and independent tests indicate the effectiveness of our stacked model TIPred, outperforming the existing method and several conventional ML classifiers. The improved performance of TIPred can be attributed to several factors: (1) The integration





**Fig. 6** Comparative molecular docking of the highest molecular docking scored csTIP candidate (E2-12) and the positive control peptide (P4) on the crystal structure of tyrosinase (PDB: 2Y9X) from different protein-peptide docking tools: GalaxyPepDock (**A**) and HPEPDOCK (**B**). The structure of the tyrosinase is shaded in gold, while E2-12 and P4 are shown in pink and yellow, respectively

of different feature encodings can provide more discriminative patterns; (2) The GA-SAR methods can determine the optimal number of features as a means of reducing the overfitting issue and improving the performance; and (3) The stacking strategy can effectively maximize the utilization of baseline models to obtain a more accurate TIP identification. Our new method is anticipated to contribute to community-wide efforts in screening and identifying potential TIP candidates for the treatment of skin pigmentation disorders and other clinical applications. Although TIPred has achieved better and more robust performance, it still has some limitations, which can be addressed in future work. One possible extension is to collect additional TIPs to develop a more comprehensive prediction model. Another extension could be the employment of well-known feature extractors, such as a bidirectional recurrent neural network (RNN) [55] and ProtBERT [56], to effectively capture the key information of TIPs. For the last extension, we can try to incorporate TIPred with recent innovative computational frameworks, such as an iterative feature representation algorithm [57] and deep learning (DL)-based framework [39, 58].

**Abbreviations**

DOPA	Dihydroxyphenylalanine
TIPs	Tyrosinase inhibitory peptides
RF	Random forest
KNN	K-nearest neighbour
AAC	Amino acid composition
PCPs	Physicochemical properties
DPC	Dipeptide composition
MCC	Matthew's correlation coefficient
ROC	Receiver operating characteristics
AUC	Area under the receiver operating characteristics curve
SHAP	Shapley Additive exPlanation
CDF	Chemistry development kit
ML	Machine learning
ET	Extremely randomized trees
SVM	Support vector machine
ACC	Accuracy
DT	Decision tree
LGBM	Light gradient boosting machine
LR	Logistic regression
MLP	Multilayer perceptron
NB	Naive Bayes
PLS	Partial least squares
SVMRBF	Support vector machine with radial basis function
SVMLN	Support vector machine with linear kernels
XGB	Extreme gradient boosting.
PF	Probabilistic feature
GA	Genetic algorithm
SAR	Self-assessment-report operation
Sp	Specificity
Sn	Sensitivity
TP	True positive
FP	False positive
TN	True negative
FN	False negative
APAAC	Amphiphilic pseudo-amino acid composition
DDE	Dipeptide deviation from expected mean
PAAC	Pseudo amino acid composition
ADA	AdaBoost
RNN	Recurrent neural network
DL	Deep learning

**Supplementary Information**

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-023-05463-1>.

**Additional file 1.** Supplementary Figures and Tables.

**Acknowledgements**

This study is supported by College of Arts, Media and Technology, Chiang Mai University and partially supported by Chiang Mai University. This study is also supported by International SciKU Branding (ISB), Faculty of Science, Kasetsart University. The authors thank the Information Technology Service Center (ITSC) of Chiang Mai University for supporting the computational resources.

**Author contributions**

PC: Designing the study, Methodology, Formal analysis, Software, Webserver development. SK: Formal analysis, Investigation, and Visualization. NS: Preparation of the manuscript. PM: Designing the study, Data curation, Formal analysis, Visualization, Investigation, Preparation of the manuscript, Revision of the manuscript. WS: Project administration, Supervision, Designing the study, Formal analysis, Visualization, Investigation, Preparation of the manuscript preparation, Revision of the manuscript.

**Funding**

This project is funded by National Research Council of Thailand (NRCT) and Mahidol University (N42A660380) and the Specific League Funds from Mahidol University.

**Availability of data materials**

All the data used in this study are available at <http://pmlabstack.pythonanywhere.com/TIPred>.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 30 March 2023 Accepted: 1 September 2023

Published online: 21 September 2023

## References

1. Casanola-Martin GM, et al. Tyrosinase enzyme: 1. An overview on a pharmacological target. *Curr Top Med Chem*. 2014;14(12):1494–501.
2. Kahn V. Effect of kojic acid on the oxidation of DL-DOPA, norepinephrine, and dopamine by mushroom tyrosinase. *Pigment Cell Res*. 1995;8(5):234–40.
3. Panzella L, Napolitano A. Natural and bioinspired phenolic compounds as tyrosinase inhibitors for the treatment of skin hyperpigmentation: recent advances. *Cosmetics*. 2019;6(4):57.
4. Pandya AG, Guevara IL. Disorders of hyperpigmentation. *Dermatol Clin*. 2000;18(1):91–8.
5. Briganti S, Camera E, Picardo M. Chemical and instrumental approaches to treat hyperpigmentation. *Pigment Cell Res*. 2003;16(2):101–10.
6. Perez-Bernal A, Munoz-Perez MA, Camacho F. Management of facial hyperpigmentation. *Am J Clin Dermatol*. 2000;1:261–8.
7. Leyden J, Shergill B, Micali G, Downie J, Wallo W. Natural options for the management of hyperpigmentation. *J Eur Acad Dermatol Venereol*. 2011;25(10):1140–5.
8. Kim Y-J, Uyama H. Tyrosinase inhibitors from natural and synthetic sources: structure, inhibition mechanism and perspective for the future. *Cell Mol Life Sci CMLS*. 2005;62:1707–23.
9. Chang T-S. An updated review of tyrosinase inhibitors. *Int J Mol Sci*. 2009;10(6):2440–75.
10. Zolghadri S, et al. A comprehensive review on tyrosinase inhibitors. *J Enzyme Inhib Med Chem*. 2019;34(1):279–309.
11. Migas P, Krauze-Baranowska M. The significance of arbutin and its derivatives in therapy and cosmetics. *Phytochem Lett*. 2015;13:35–40.
12. Hermawan M, Simbolon Sitohang I, Sirait SP. Exogenous ochronosis: Screening by dermoscopy and histopathological confirmation. *J Nat Sc Biol Med*. 2019;10:163–5.
13. Diven DG, Smith EB, Pupo RA, Lee M. Hydroquinone-induced localized exogenous ochronosis treated with dermabrasion and CO<sub>2</sub> laser. *J Dermatol Surg Oncol*. 1990;16(11):1018–22.
14. Arndt KA, Fitzpatrick TB. Topical use of hydroquinone as a depigmenting agent. *JAMA*. 1965;194(9):965–7.
15. Song Y, Chen S, Li L, Zeng Y, Hu X. The hypopigmentation mechanism of tyrosinase inhibitory peptides derived from food proteins: an overview. *Molecules*. 2022;27(9):2710.
16. Deng Y, et al. Skin-care functions of peptides prepared from Chinese quince seed protein: sequences analysis, tyrosinase inhibition and molecular docking study. *Ind Crops Prod*. 2020;148:112331.
17. Ochiai A, et al. New tyrosinase inhibitory decapeptide: molecular insights into the role of tyrosine residues. *J Biosci Bioeng*. 2016;121(6):607–13.
18. Addar L, Bensouici C, Zennia SSA, Haroun SB, Mati A. Antioxidant, tyrosinase and urease inhibitory activities of camel  $\alpha$ S-casein and its hydrolysate fractions. *Small Rumin Res*. 2019;173:30–5.
19. Kongsompong S, E-kobon T, Chumnanpuen P. K-nearest neighbor and random forest-based prediction of putative tyrosinase inhibitory peptides of abalone *Haliotis diversicolor*. *Molecules*. 2021;26(12):3671.
20. Olsen TH, et al. AnOxPePred: using deep learning for the prediction of antioxidative properties of peptides. *Sci Rep*. 2020;10(1):21471.
21. Yang J-K, et al.  $\beta$ -Lactoglobulin peptide fragments conjugated with caffeic acid displaying dual activities for tyrosinase inhibition and antioxidant effect. *Bioconjug Chem*. 2018;29(4):1000–5.
22. Yap P-G, Gan C-Y. Multifunctional tyrosinase inhibitor peptides with copper chelating, UV-absorption and antioxidant activities: kinetic and docking studies. *Foods*. 2021;10(3):675.
23. Luisi G, Stefanucci A, Zengin G, Dimmito MP, Mollica A. Anti-oxidant and tyrosinase inhibitory in vitro activity of amino acids and small peptides: new hints for the multifaceted treatment of neurologic and metabolic disfunctions. *Antioxidants*. 2018;8(1):7.
24. Chen Z, et al. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*. 2018;34(14):2499–502.
25. Charoenkwan P, Chiangjong W, Nantasenamat C, Hasan MM, Manavalan B, Shoombuatong W. StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Brief Bioinform*. 2021;22(6):bbab172.
26. Charoenkwan P, Nantasenamat C, Hasan MM, Moni MA, Manavalan B, Shoombuatong W. StackDPPiV: a novel computational approach for accurate prediction of dipeptidyl peptidase IV (DPP-IV) inhibitory peptides. *Methods*. 2022;204:189–98.
27. Landrum G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. London: Academic Press; 2013.
28. Willighagen EL, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform*. 2017;9(1):1–19.

29. Malik AA, Chotpatiwetchkul W, Phanus-Umporn C, Nantasenamat C, Charoenkwan P, Shoombuatong W. StackHCV: a web-based integrative machine-learning framework for large-scale identification of hepatitis C virus NS5B inhibitors. *J Comput Aided Mol Design*. 2021;35(10):1037–53.
30. Schaduangrat N, Anuwongcharoen N, Moni MA, Lio P, Charoenkwan P, Shoombuatong W. StackPR is a new computational approach for large-scale identification of progesterone receptor antagonists using the stacking strategy. *Sci Rep*. 2022;12(1):1–16.
31. Wei L, Ye X, Xue Y, Sakurai T, Wei L. ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief Bioinform*. 2021;22(5):bbab041.
32. Chen W, Lv H, Nie F, Lin H. i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics*. 2019;35(16):2796–800.
33. Charoenkwan P, Schaduangrat N, Nantasenamat C, Piacham T, Shoombuatong W. iQSP: a sequence-based tool for the prediction and analysis of quorum sensing peptides using informative physicochemical properties. *Int J Mol Sci*. 2019;21(1):75.
34. Charoenkwan P, et al. AMYPred-FRL is a novel approach for accurate prediction of amyloid proteins by using feature representation learning. *Sci Rep*. 2022;12(1):1–14.
35. Charoenkwan P, Schaduangrat N, Moni MA, Manavalan B, Shoombuatong W. SAPPHERE: a stacking-based ensemble learning framework for accurate prediction of thermophilic proteins. *Comput Biol Med*. 2022;146:105704.
36. Qiang X, Zhou C, Ye X, Du P-F, Su R, Wei L. CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. *Brief Bioinform*. 2020;21(1):11–23.
37. Rao B, Zhou C, Zhang G, Su R, Wei L. ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides. *Brief Bioinform*. 2020;21(5):1846–55.
38. Zhu L, Wang X, Li F, Song J. PreAcres: a machine learning framework for identifying anti-CRISPR proteins. *BMC Bioinform*. 2022;23(1):444.
39. Xie R, et al. DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy. *Brief Bioinform*. 2021;22(3):bbaa125.
40. Hasan MM, Schaduangrat N, Basith S, Lee G, Shoombuatong W, Manavalan B. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics*. 2020;36(11):3350–6.
41. Basith S, Manavalan B, Hwan-Shin T, Lee G. Machine intelligence in peptide therapeutics: a next-generation tool for rapid disease screening. *Med Res Rev*. 2020;40(4):1276–314.
42. Prasertsuk K, et al. Computer-aided screening for potential coronavirus 3-chymotrypsin-like protease (3CLpro) inhibitory peptides from putative hemp seed trypsinized peptidome. *Molecules*. 2022;28(1):50.
43. Kongsompong S, E-kobon T, Taengphan W, Sangkhawasi M, Khongkow M, Chumnanpuen P. Computer-aided virtual screening and in vitro validation of biomimetic tyrosinase inhibitory peptides from abalone peptidome. *Int J Mol Sci*. 2023;24(4):3154.
44. Schurink M, van Berkel WJ, Wichers HJ, Boeriu CG. Novel peptides with tyrosinase inhibitory activity. *Peptides*. 2007;28(3):485–95.
45. Ishikawa M, Kawase I, Ishii F. Combination of amino acids reduces pigmentation in B16F0 melanoma cells. *Biol Pharm Bull*. 2007;30(4):677–81.
46. Rajapakse N, Mendis E, Byun H-G, Kim S-K. Purification and in vitro antioxidative effects of giant squid muscle peptides on free radical-mediated oxidative systems. *J Nutr Biochem*. 2005;16(9):562–9.
47. Rajapakse N, Mendis E, Jung W-K, Je J-Y, Kim S-K. Purification of a radical scavenging peptide from fermented mussel sauce and its antioxidant properties. *Food Res Int*. 2005;38(2):175–82.
48. Zhou C, Yu X, Qin X, Ma H, Yagoub AEA, Hu J. Hydrolysis of rapeseed meal protein under simulated duodenum digestion: Kinetic modeling and antioxidant activity. *LWT Food Sci Technol*. 2016;68:523–31.
49. Kubglomsong S, Theerakulkait C, Reed RL, Yang L, Maier CS, Stevens JF. Isolation and identification of tyrosinase-inhibitory and copper-chelating peptides from hydrolyzed rice-bran-derived albumin. *J Agric Food Chem*. 2018;66(31):8346–54.
50. Oomah BD, Busson M, Godfrey DV, Drover JC. Characteristics of hemp (*Cannabis sativa* L.) seed oil. *Food chemistry*. 2002;76(1):33–43.
51. Girgih AT, Udenigwe CC, Aluko RE. In vitro antioxidant properties of hemp seed (*Cannabis sativa* L.) protein hydrolysate fractions. *J Am Oil Chem Soc*. 2011;88(3):381–9.
52. Santos-Sánchez G, et al. Hempseed (*Cannabis sativa*) protein hydrolysates: A valuable source of bioactive peptides with pleiotropic health-promoting effects. *Trends Food Sci Technol*. 2022;127:303–18.
53. Reddy B, Jow T, Hantash BM. Bioactive oligopeptides in dermatology: part I. *Exp Dermatol*. 2012;21(8):563–8.
54. Jeffrey GA, Jeffrey GA. An introduction to hydrogen bonding. New York: Oxford University Press; 1997.
55. Xuan W, Liu N, Huang N, Li Y, Wang J. Clpred: a sequence-based protein crystallization predictor using blstm neural network. *Bioinformatics*. 2020;36(Supplement\_2):i709–17.
56. Wang X, Han L, Wang R, Chen H. DaDL-SChlo: protein subchloroplast localization prediction based on generative adversarial networks and pre-trained protein language model. *Brief Bioinform*. 2023;24(3):bbad083.
57. Wei L, et al. Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics*. 2019;35(23):4930–7.
58. Li F, et al. Digerati—a multipath parallel hybrid deep learning framework for the identification of mycobacterial PE/PPE proteins. *Comput Biol Med*. 2023;21:107155.
59. Charoenkwan P, Nantasenamat C, Hasan MM, Moni MA, Manavalan B, Shoombuatong W. UMPred-FRL: a new approach for accurate prediction of umami peptides using feature representation learning. *Int J Mol Sci*. 2021;22(23):13124.
60. Charoenkwan P, Nantasenamat C, Hasan MM, Moni MA, Lio P, Shoombuatong W. iBitter-fuse: a novel sequence-based bitter peptide predictor by fusing multi-view features. *Int J Mol Sci*. 2021;22(16):8958.
61. Chou K-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*. 2005;21(1):10–9.

62. Hall LH, Kier LB. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J Chem Inf Comput Sci*. 1995;35(6):1039–45.
63. Laggner C. SMARTS patterns for functional group classification. 2005.
64. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci*. 2002;42(6):1273–80.
65. Kim S, et al. PubChem substance and compound databases. *Nucleic Acids Res*. 2016;44(D1):D1202–13.

### **Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

