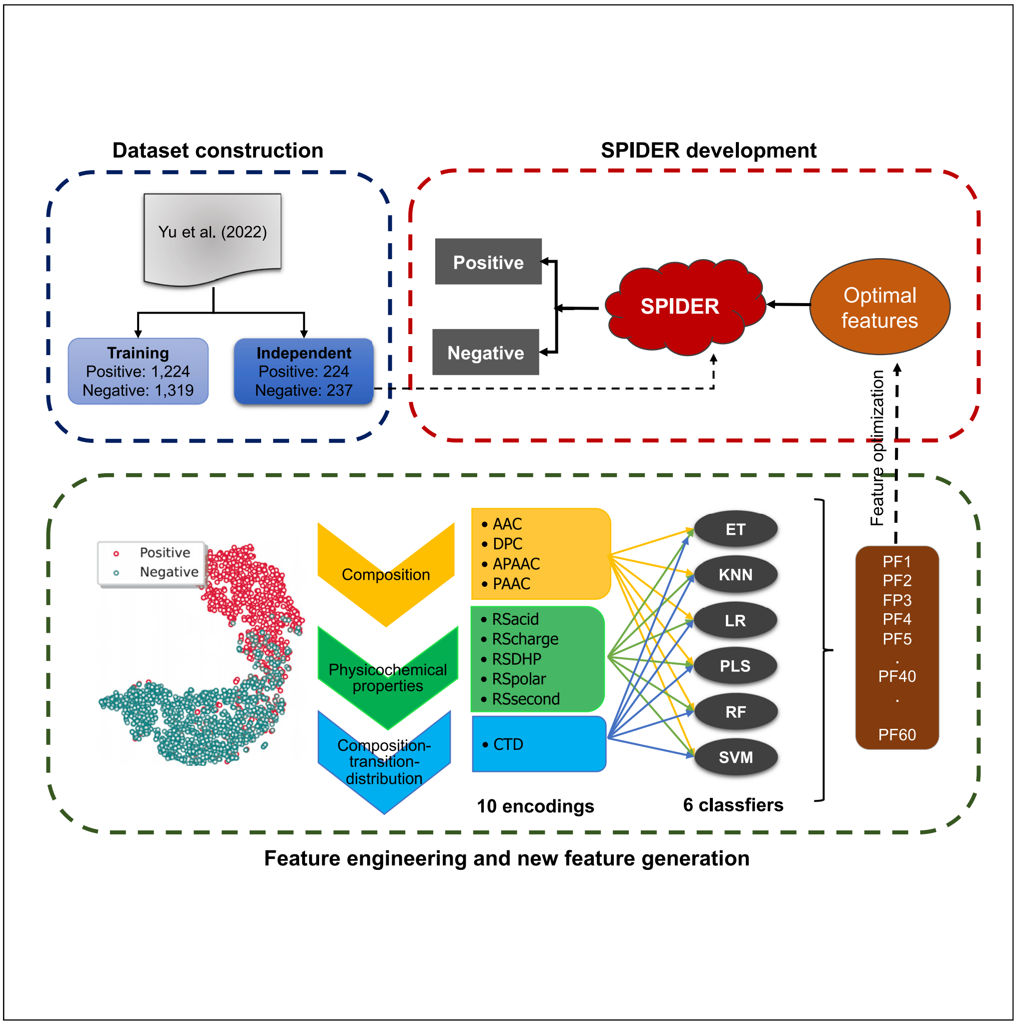
iScience

CellPress
OPEN ACCESS

Article

Computational prediction and interpretation of druggable proteins using a stacked ensemble-learning framework



Phasit Charoenkwan, Nalini Schaduangrat, Pietro Lio', Mohammad Ali Moni, Watshara Shoombuatong, Balachandran Manavalan

watshara.sho@mahidol.ac.th (W.S.)
bala2022@skku.edu (B.M.)

Highlights

Computational models can expedite the identification of potential druggable proteins

SPIDER represents the first stacked model proposed for druggable protein prediction

SPIDER enables more precise prediction of druggable proteins than existing methods

The SPIDER web server is available at http://pmlabstack.pythonanywhere.com/SPIDER.

Charoenkwan et al., iScience *25*, 104883
September 16, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.104883

iScience

CellPress
OPEN ACCESS

Article

Computational prediction and interpretation of druggable proteins using a stacked ensemble-learning framework

Phasit Charoenkwan,[1] Nalini Schaduangrat,[2] Pietro Lio',[3] Mohammad Ali Moni,[4] Watshara Shoombuatong,[2,*] and Balachandran Manavalan[5,6,*]

SUMMARY

Discovery of potential drugs requires rapid and precise identification of drug targets. Although traditional experimental methodologies can accurately identify drug targets, they are time-consuming and inappropriate for high-throughput screening. Computational approaches based on machine learning (ML) algorithms can expedite the prediction of druggable proteins; however, the performance of the existing computational methods remains unsatisfactory. This study proposes a computational tool, SPIDER, to enhance the accurate prediction of druggable proteins. SPIDER employs various feature descriptors pertaining to several aspects, including physicochemical properties, compositional information, and composition-transition-distribution information, coupled with well-known ML algorithms to facilitate the construction of the final meta-predictor. The experimental results showed that SPIDER enabled more precise and robust prediction of druggable proteins than the baseline models and current existing methods in terms of the independent test dataset. An online web server was established and made freely available online.

INTRODUCTION

A druggable protein refers to a protein that can bind to small drug-like molecules with a high affinity and produce desirable therapeutic effects (Liu and Altman, 2014). Druggable proteins are usually members of large protein families that have been successfully identified as drug targets (Owens, 2007). Failure of projects in the drug discovery field is usually due to the target being undruggable, as estimated in approximately 60% of all cases (Sakharkar et al., 2007). As such, the druggability of a protein is crucial for the progression of a drug discovery project, wherein the accurate identification of drug targets is necessary (Overington et al., 2006). Experimental methods require the analysis of the three-dimensional structure of a protein, which results in a long development cycle (Sakharkar et al., 2007). Traditional experimental methods can precisely identify the drug targets; however, these methods are laborious and challenging for high-throughput applications. Computational methods based solely on the primary sequences of drugs can complement experimental methods to expedite the characterization and prediction of druggable proteins. Owing to the vast number of novel proteins generated via next-generation sequencing, the possibility of identifying candidate druggable proteins that have not yet been characterized is immense. Hence, the precise and quick identification of druggable proteins from a vast pool of sequenced proteins is highly desirable for the development of new drugs (Lindsay, 2005).

Drug target prediction is complemented by numerous computational tools. For instance, Dezső and Ceccarelli developed a random forest (RF)-based method for selecting and prioritizing drug targets. In their study, the predictive model was trained using different feature descriptors and achieved an area under the receiver operating curve (AUC) of 0.89 in terms of the independent test dataset (Dezső and Ceccarelli, 2020). In addition, existing data-driven approaches can predict the drug similarity (Ma'ayan et al., 2014), drug-target interactions (Fakhraei et al., 2014; Perlman et al., 2011), and similarities between drugs and potential predicted targets (Wang et al., 2013). Detailed information on these data-driven approaches is available in the articles by Dezső and Ceccarelli (2020) and Gong et al. (2021). Several computational methods based on machine learning (ML) techniques, such as DrugMiner (Jamali et al., 2016), Sun's method (Sun et al., 2018), GA-bagging-SVM (Lin et al., 2019), DrugHybrid_BS (Gong et al., 2021), Yu's method

[1] Modern Management and Information Technology, College of Arts, Media and Technology, Chiang Mai University, Chiang Mai 50200, Thailand

[2] Center of Data Mining and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bangkok 10700, Thailand

[3] Department of Computer Science and Technology, University of Cambridge, Cambridge CB3 0FD, UK

[4] Artificial Intelligence & Digital Health, School of Health and Rehabilitation Sciences, Faculty of Health and Behavioural Sciences, The University of Queensland, St Lucia, QLD 4072, Australia

[5] Computational Biology and Bioinformatics Laboratory, Department of Integrative Biotechnology, College of Biotechnology and Bioengineering, Sungkyunkwan University, Suwon 16419, Gyeonggi-do, Republic of Korea

[6] Lead contact

*Correspondence:
watshara.sho@mahidol.ac.th (W.S.),
bala2022@skku.edu (B.M.)

https://doi.org/10.1016/j.isci.2022.104883

Table 1. Summary of existing methods and tools for prediction of druggable proteins

Method (Year)	Classifier[a]	Features[b]	Evaluation strategy[c]	Web server availability
DrugMiner (2016) (Jamali et al., 2016)	NN	AAC, DPC, PCP	5CV	Yes
Sun's method (2018) (Sun et al., 2018)	NN	CTD	5CV/IND	No
GA-Bagging-SVM (2019) (Lin et al., 2019)	SVM	DPC, RC, PAAC	5CV	No
DrugHybrid_BS (2021) (Gong et al., 2021)	SVM	CC, GAAC, monoDIKgap	5CV	No
Yu's method (2022) (Yu et al., 2022)	CNN-RNN	Dictionary, DPC, TPC, CTD	5CV/IND	No
XGB-;DrugPred (2022) (Sikander et al., 2022)	XGB	GDPC, S-PseAAC, RAAA	10CV	No
SPIDER (This study)	SVM	AAC, APAAC, DPC, CTD, PAAC, RC	10CV/IND	Yes

[a] CNN-RNN: hybrid model integrating convolutional recurrent neural networks and deep neural networks, NN: neural networks, SVM: support vector machine.
[b] AAC: amino acid composition, APAAC: amphiphilic pseudo-amino acid composition, CC: Cross Covariance, CTD: Composition-Transition-Distribution, DPC: dipeptide composition, DPS: dipeptide propensity score; GAAC: grouped amino acid composition, PCP: physicochemical properties, PACC: pseudo amino acid composition, TPC: tripeptide composition.
[c] 5CV: 5-fold cross-validation test, 10CV: 10-fold cross-validation test IND: independent test.

(Yu et al., 2022), and XGB-DrugPred (Sikander et al., 2022), have been designed for the *in silico* prediction of druggable proteins based on their protein sequence information, as summarized in Table 1.

In 2016, Jamali et al. developed DrugMiner (Jamali et al., 2016), the first computational method in this field, based on their own dataset comprising 1224 druggable and 1319 non-druggable proteins. DrugMiner was created using a neural network algorithm in conjunction with various types of feature descriptors. Furthermore, Lin et al. created a GA-bagging-SVM (Lin et al., 2019) by integrating various support vector machine (SVM)-based classifiers and a genetic algorithm (GA) through the bagging ensemble learning strategy. In the GA-bagging-SVM, Lin et al. employed three feature encodings to represent the druggable proteins, dipeptide composition (DPC), pseudo amino acid composition (PAAC), and reduced sequences (RS), which encompass the secondary structure, DHP, acidity, polarity, and charge (referred to herein as RSsecond, RSDHP, RSacid, RSpolar, and RScharge, respectively). Recently, Gong et al. (2021) developed DrugHybrid_BS, a bagging ensemble learning model combined with monoDiKGap, cross-covariance, and grouped amino acid composition. DrugHybrid_BS can provide a reasonably high predictive performance with an accuracy (ACC) of 0.970 and an AUC of 0.992. Recently, Yu et al. (2022) created hybrid convolutional recurrent neural networks (CNN-RNNs), which utilized both dictionary and sequence encoding schemes to enhance the prediction performance. Yu et al. first established an independent test dataset, which contained 224 druggable and 237 non-druggable proteins. This method provided an ACC of 0.898 and a Matthew's correlation coefficient (MCC) of 0.799 for the independent test dataset.

All aforementioned methods have facilitated the identification of druggable proteins and promoted the progress in this field. However, certain concerns still need to be addressed. First, most of the existing methods, except Yu's method (Yu et al., 2022), were not performed on an independent test dataset; thus, their prediction performance may fail in terms of generalizability. Second, there is no comprehensive analysis or evaluation of conventional feature encodings and ML algorithms for druggable proteins. Third, all existing methods are considered as black-box models; as such, it is difficult to provide a straightforward interpretation of the functional mechanisms of druggable proteins. Finally, all existing methods, except that of DrugMiner (Jamali et al., 2016), were not deployed as web servers. Therefore, they can only be used by experimental scientists.

Considering the above-mentioned limitations, herein, a new computational tool, named SPIDER (Stacked PredIctor of DruggablE pRoteins), is presented to improve the prediction accuracy of druggable proteins and enhance the most important features contributing to druggable protein prediction (see Figure 1). The significance and major advantages of SPIDER are as follows: (i) SPIDER represents the first stacked

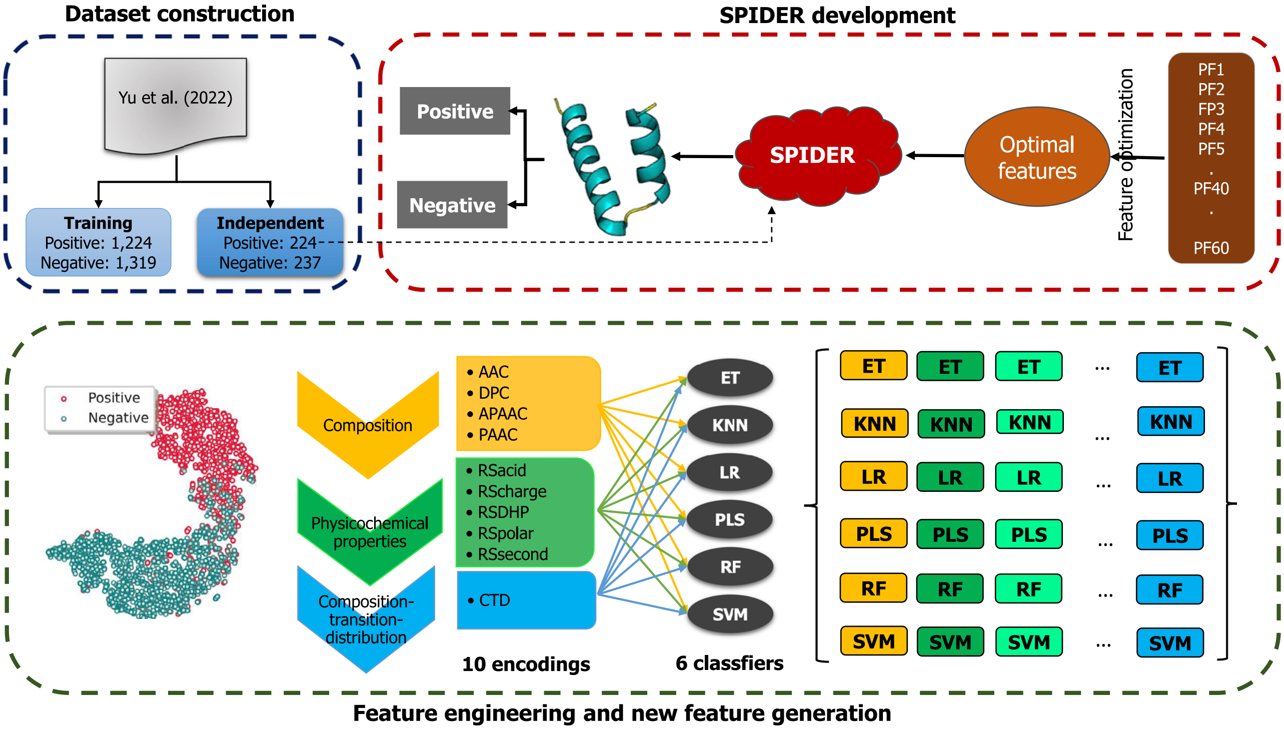


Figure 1. Schematic flowchart of the development of the SPIDER

There are four major steps, including dataset construction, feature engineering, new feature generation, and meta-predictor development.

ensemble learning approach proposed for druggable protein prediction. Specifically, SPIDER was trained and constructed by integrating *m* = 10 selected ML classifiers to facilitate the stable and accurate prediction of druggable proteins; (ii) we comprehensively investigated and assessed the predictive ability of various types of feature encodings coupled with popular ML algorithms in the prediction of druggable proteins. SPIDER was found to be more effective and outperformed several ML classifiers and existing methods for this prediction problem in terms of an independent test dataset; (iii) we employed an interpretable Shapley Additive exPlanations (SHAP) method to shed light on the impact of each feature on the output of SPIDER. Finally, to facilitate community-wide efforts in the prediction of druggable proteins, an online web server based on SPIDER was created and is easily accessible at http://pmlabstack.pythonanywhere.com/SPIDER.

RESULTS AND DISCUSSION

Performance of different feature-encoding schemes and ML algorithms

In this study, we comprehensively analyzed and assessed the predictive ability of various baseline models trained with ten feature encodings and six ML algorithms to distinguish druggable proteins from non-druggable ones. Comprehensive information regarding the feature encodings and ML algorithms is presented in Table 2 and S1. Both 10-fold cross-validation and independent tests were implemented on the training and independent test datasets to assess the performance of each baseline model, as summarized in Figure 2 and Tables S2–S5. As described in the SPIDER framework section, the baseline model with the highest MCC in the training dataset is regarded as the most efficient.

To analyze the overall effect of each feature encoding on the prediction of druggable proteins, we computed the average cross-validation performance for each feature encoding over six different ML algorithms. Among the ten feature encodings, the top five feature encodings comprising the highest performance corresponded to RSpolar, RSDHP, RSsecond, RSacid, and RScharge, with average MCC values of 0.727, 0.721, 0.716, 0.713, and 0.712, respectively (Table S4). Similarly, the top-three feature ML algorithms with the highest performance contained SVM, RF, and extremely randomized trees (ET), with corresponding average MCC values of 0.762, 0.747, and 0.743, respectively (Table S5). Interestingly, the SVM, RF, and

Table 2. Summary of ten different sequence-based feature descriptors along with their corresponding description and dimension

Order	Descriptors	Description	Dimension	Reference
1	AAC	Frequency of 20 amino acids	20	(Charoenkwan et al., 2021b, 2022)
2	APAAC	Amphiphilic pseudo-amino acid composition	22	(Chou, 2001, 2005)
3	CTD	Composition, transition, and distribution	273	(Li et al., 2006)
4	DPC	Frequency of 400 dipeptides	400	(Chen et al., 2016; Lin and Chen, 2011)
5	PAAC	Pseudo amino acid composition	21	(Chou, 2001, 2005)
6	RSacid	Reduced amino acid sequences according to acidity	32	(Xu et al., 2017)
7	RScharge	Reduced amino acid sequences according to charge	50	(Xu et al., 2017)
8	RSDHP	Reduced amino acid sequences according to DHP	32	(Xu et al., 2017)
9	RSpolar	Reduced amino acid sequences according to polarity	32	(Xu et al., 2017)
10	RSsecond	Reduced amino acid sequences according to secondary structure	40	(Xu et al., 2017)

ET models developed using RSpolar, RScharge, and RSpolar achieved the highest MCC values of 0.796, 0.778, and 0.769, respectively. This observation indicates that the feature group of RS could be more beneficial for druggable protein prediction. Among the 60 baseline models, the highest MCC of 0.796 was achieved by the SVM-RSpolar model, while the second and third highest MCC values of 0.792 and 0.780 were achieved by the SVM-RSDHP and SVM-Rsecond models, respectively. This indicated that the SVM-RSpolar model could be considered the most efficient one for the prediction of druggable proteins. Regarding the independent test results, the best-performing model provided an MCC of 0.770, an ACC of 0.883, and an AUC of 0.936. In contrast, the highest independent MCC of 0.808 was obtained using the PLS-AAC model. Overall, our comprehensive analysis suggests that single-feature-based models might fail in terms of generalizability and stability in this prediction problem. As such, we applied a stacked ensemble learning methodology to generate a model with the strongest stability and generalization ability in terms of both cross-validation and independent tests.

Construction of SPIDER

Next, we developed an ensemble model that integrates several ML classifiers using the stacking approach. To this end, we employed both 60-D and *m*-D feature vectors to develop mSVM predictors. As described in the SPIDER framework section, the GA in combination with the self-assessment-report (GA-SAR) approach was employed to optimize the 60-D feature vector by selecting *m* informative probabilistic features (PFs). After applying the GA-SAR approach, the experimental results indicated that the best number of informative PFs was $m = 10$. Specifically, $m = 10$ informative PFs were derived from ten baseline models, namely SVM-AAC, LR-DPC, ET-CTD, RF-PAAC, LR-APAAC, SVM-RSacid, SVM-RSpolar, LR-RSsecond, PLS-RScharge, and ET-RSDHP.

Table 3 provides information pertaining to the performance evaluation of the two new feature vectors. As shown in Table 3, the 10-D feature vector (referred to herein as the optimal feature vector) was found to provide an enhancement, as judged by ACC, MCC, sensitivity (Sn), and specificity (Sp), not only in the training dataset but also in the independent dataset. Remarkably, the ACC, MCC, and Sn of the optimal feature vector in the independent test dataset were 0.907, 0.816, and 0.857, respectively, which were 1.74%, 3.37%, and 2.68% higher than the compared feature vectors, respectively. For convenience of discussion, we denote the mSVM predictor trained with the optimal feature vector as SPIDER. We also compared the performance of SPIDER with that of a popular ensemble approach (voting strategy) in the independent test dataset. As shown in Table S6, the ACC, MCC, Sn, and Sp of SPIDER outperformed the voting strategy by 1.52%, 2.97%, 2.23%, and 0.84%, respectively.

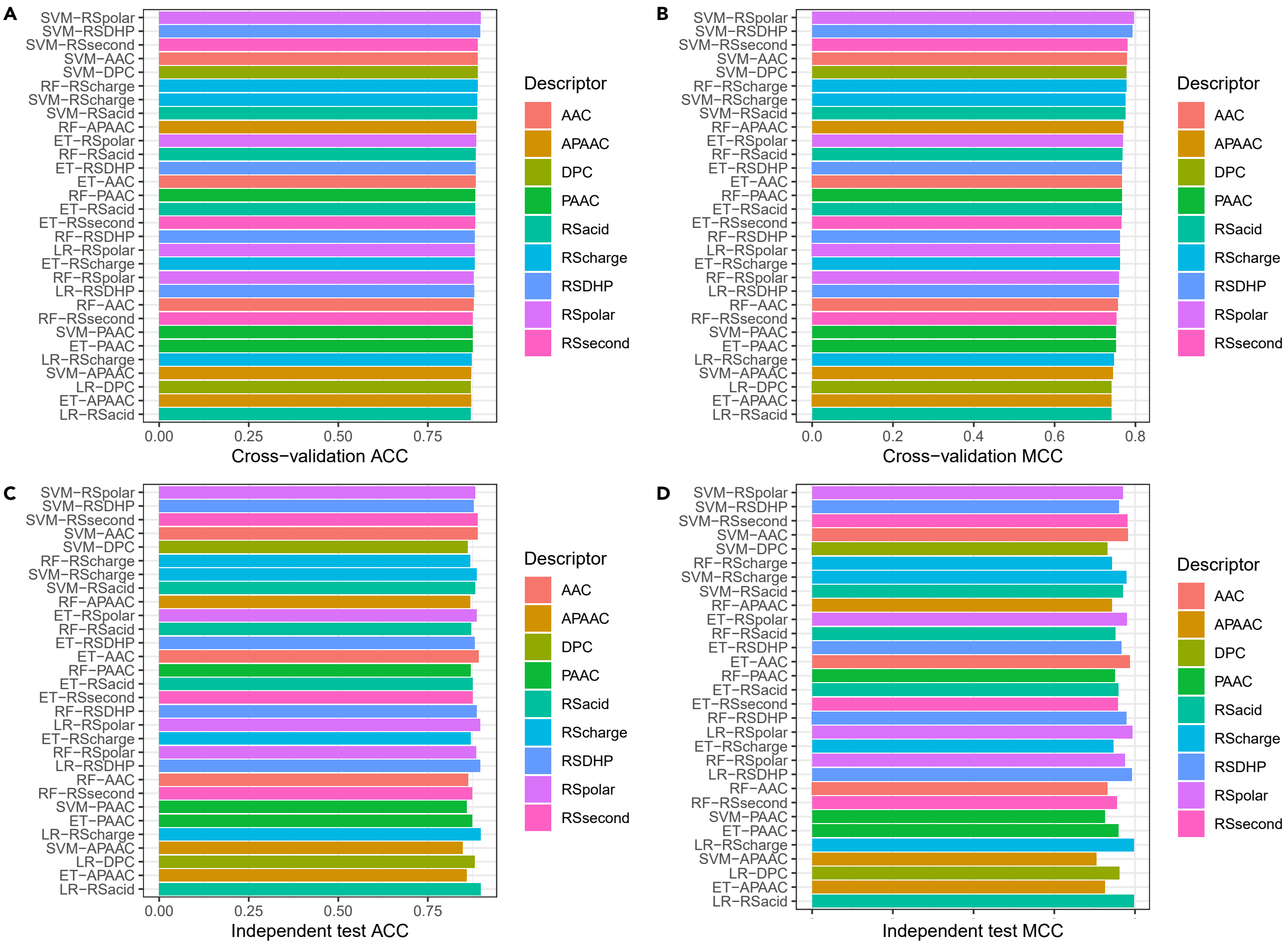


Figure 2. Performance evaluations of top 30 baseline models

(A and B) Cross-validation ACC and MCC of top 30 baseline models.

(C and D) Independent test ACC and MCC of top 30 baseline models.

Performance comparison between SPIDER and single-feature-based models

To elucidate the advantage of the stacking approach, we compared the performance of SPIDER with that of single-feature-based models. Figure 2 shows the five top-ranked baseline models, as indicated by MCC—SVM-RSpolar, SVM-RSDHP, SVM-RSsecond, SVM-AAC, and RF-RScharge—with corresponding MCC values of 0.796, 0.792, 0.780, 0.779, and 0.778, respectively. Thus, the performance of SPIDER was evaluated against the top five baseline models. The performance results are summarized in Figure 3 and Table 4. As summarized in Table 4, SPIDER clearly outperformed the top five baseline models in the training dataset and the independent dataset in terms of most of the performance metrics, with the exception of AUC. Specifically, SPIDER achieved enhanced performance in comparison to the best-performing baseline model (SVM-RSpolar) in the independent dataset in terms of ACC (0.907 vs. 0.883), Sn (0.857 vs. 0.821), Sp (0.954 vs. 0.941), and MCC (0.816 vs. 0.770). The above-mentioned results demonstrate that SPIDER

Table 3. Cross-validation results for the control and optimal model

Evaluation strategy	Model	Number of feature	ACC	Sn	Sp	MCC	AUC
Cross-validation	Control	60	0.909	0.888	0.929	0.819	0.955
	Optimal	10	0.919	0.895	0.942	0.839	0.950
Independent test	Control	60	0.889	0.830	0.945	0.783	0.934
	Optimal	10	0.907	0.857	0.954	0.816	0.902

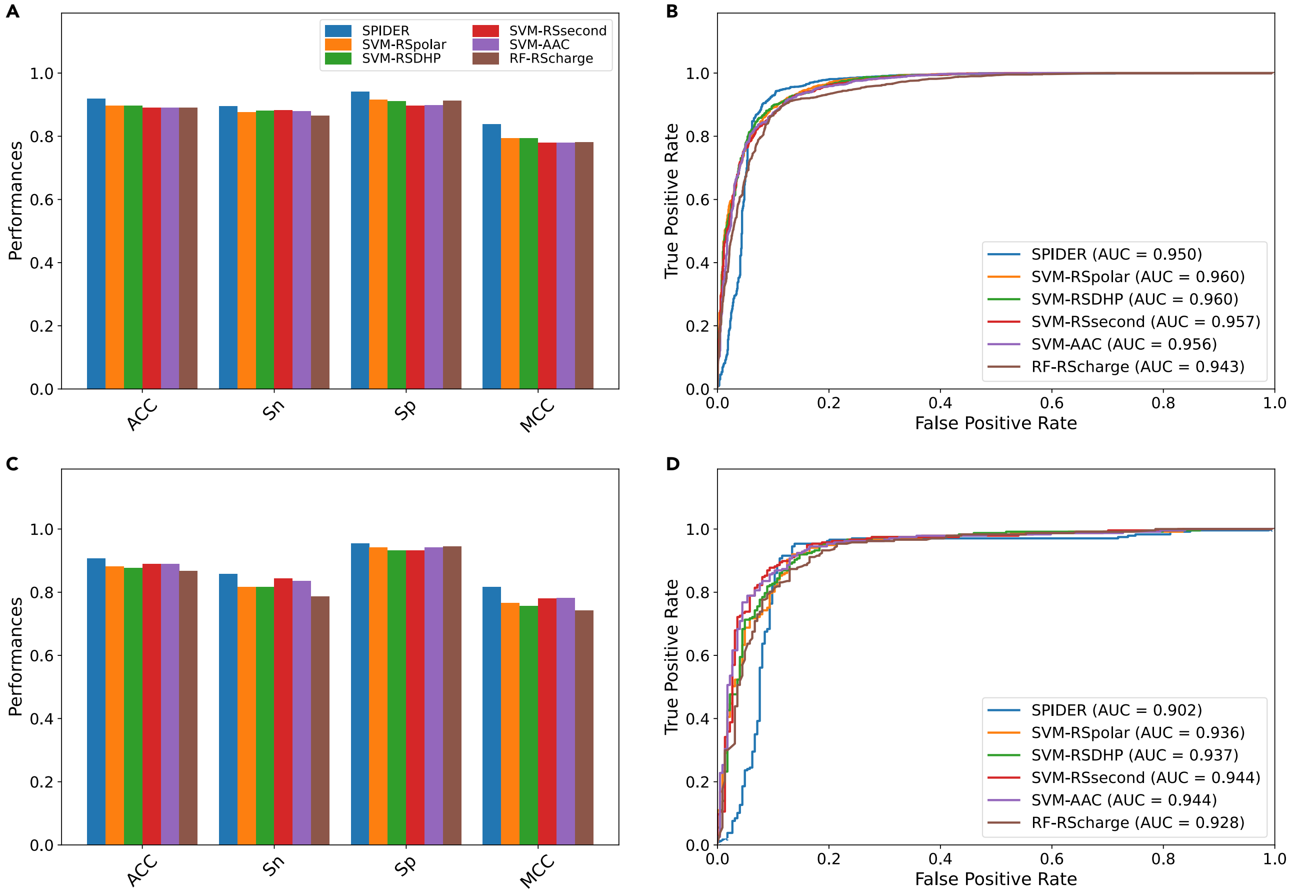


Figure 3. Predictive performance of various models

Performance comparison of SPIDER with the top five baseline models on the training (A–B) and independent test (C–D) datasets. Prediction results of SPIDER and the top five baseline models in terms of ACC, Sn, Sp, and MCC (A, C). ROC curves and AUC values of SPIDER and the top five baseline models (B–D).

achieved improved performance and stability compared with several single-feature-based models in both the training and independent test datasets.

To further reveal the improved performance of SPIDER, the distribution of the 2D feature space from the top three informative feature descriptors (RSpolar, RSDHP, and RSsecond), all features, the 60-D feature vector, and the optimal feature vector were visualized using the t-distributed stochastic neighbor embedding (t-SNE) (Van Der Maaten, 2014; Van der Maaten and Hinton, 2008) framework, wherein the red and green dots indicate druggable and non-druggable proteins, respectively (Figure 4). As shown in Figures 4A–4D, the red and green dots derived from the four t-SNE plots were mixed together, indicating that these feature descriptors have limited discriminative power for identifying druggable proteins. However, we noticed that the 60-D and optimal feature vectors showed a sharp distinction between the distribution of red and green dots (Figure 4F). Altogether, the stacking strategy used in SPIDER seems to be an effective and useful approach for improving prediction performance and model generalizability.

Performance comparison between SPIDER and state-of-the-art methods

Here, the performance of SPIDER was compared with that of state-of-the-art methods. Table 1 provides details of various ML-based methods that have been designed based on sequence information, namely DrugMiner (Jamali et al., 2016), Sun's method (Sun et al., 2018), GA-Bagging-SVM (Lin et al., 2019), DrugHybrid_BS (Gong et al., 2021), Yu's method (Yu et al., 2022), and XGB-DrugPred (Sikander et al.,

Table 4. Performance comparison of SPIDER and top five baseline models on the training and independent test datasets

Evaluation strategy	Method	ACC	Sn	Sp	MCC	AUC
Cross-validation	SPIDER	0.919	0.895	0.942	0.839	0.950
	SVM-RSpolar	0.898	0.885	0.911	0.796	0.960
	SVM-RSDHP	0.896	0.884	0.908	0.792	0.960
	SVM-RSsecond	0.890	0.885	0.895	0.780	0.957
	SVM-AAC	0.890	0.882	0.897	0.779	0.956
	RF-Scharge	0.889	0.862	0.914	0.778	0.943
Independent test	SPIDER	0.907	0.857	0.954	0.816	0.902
	SVM-RSpolar	0.883	0.821	0.941	0.770	0.936
	SVM-RSDHP	0.879	0.821	0.932	0.760	0.937
	SVM-RSsecond	0.889	0.844	0.932	0.781	0.944
	SVM-AAC	0.889	0.835	0.941	0.782	0.944
	RF-Scharge	0.868	0.786	0.945	0.743	0.928

2022), to determine the druggability of proteins. Among these existing methods, Yu's method (Yu et al., 2022) was the only one that was evaluated on both the training (1,224 druggable and 1,319 non-druggable proteins) and independent test (224 druggable and 237 non-druggable proteins) datasets. To perform a

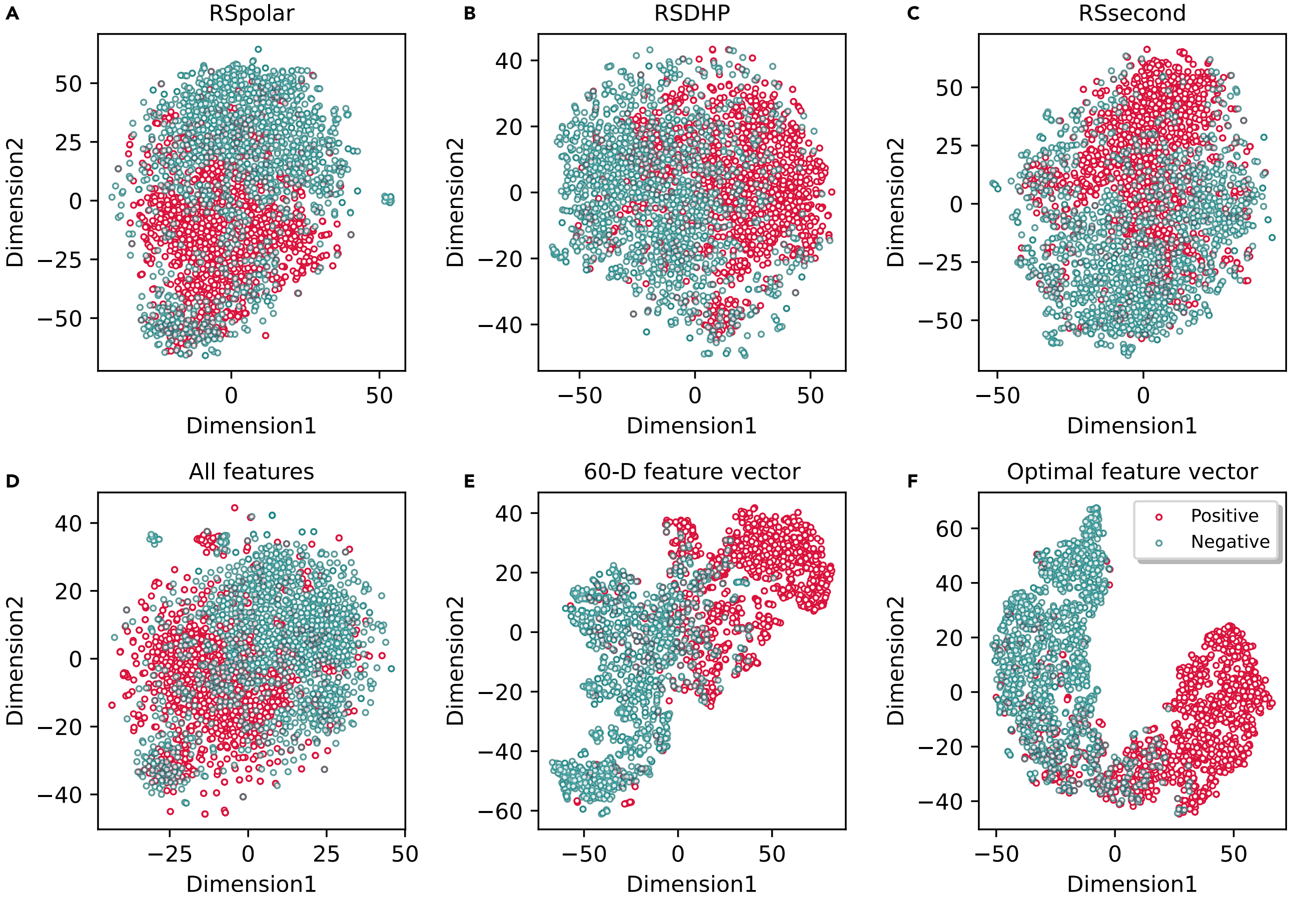


Figure 4. t-distributed stochastic neighbor embedding (t-SNE) distribution of positive and negative samples on the training dataset
(A) RSpolar, (B) RSDHP, (C) RSsecond, (D) All features, (E) 60-D feature vector, and (F) Optimal feature vector.

Table 5. Performance comparison of SPIDER and the state-of-the-art method

Evaluation strategy	Method	ACC	Sn	MCC	F-score	PRE
Cross-validation	Yu's method[a]	0.900	0.890	0.800	0.896	0.905
	SPIDER	0.919	0.895	0.839	0.914	0.895
Independent test	Yu's method[a]	0.898	0.848	0.799	0.889	0.936
	SPIDER	0.907	0.857	0.816	0.899	0.857

[a] Results were reported from the work of Yu's method (Yu et al., 2022).

comprehensive comparison, we compared the performance of SPIDER with that of Yu's method. The results of the comparison of the two methods are listed in Table 5. As can be observed, SPIDER attained the highest performance in terms of ACC, MCC, Sn, and F-values on the training dataset, which were 1.94%, 3.92%, 0.53%, and 1.80% higher than those obtained using Yu's method, respectively. Furthermore, the independent test results demonstrated that SPIDER achieved better performance, achieving an ACC of 0.907, Sn of 0.857, MCC of 0.816, and F-value of 0.899. Taken together, these results demonstrate that SPIDER is an accurate prediction model with efficient generalization ability compared with the available methods.

Mechanistic interpretation of SPIDER

As mentioned above, we applied the GA-SAR algorithm to select *m* important features to generate the optimal feature vector. However, the relationship between these features remains unknown. To address this problem, we used the SHAP framework not only to assess the value of each feature but also to shed light on the output of the model, which plays a crucial role in many bioinformatic applications (Li et al., 2021a, 2021b). As previously stated, SPIDER was constructed using a combination of ten selected baseline models—SVM-AAC, LR-DPC, ET-CTD, RF-PAAC, LR-APAAC, SVM-RSacid, SVM-RSpolar, LR-RSsecond, PLS-RScharge, and ET-RSDHP. SHAP positive and negative values indicate the prediction of druggable and non-druggable proteins, respectively. As illustrated in Figure 5, the top five informative features with the highest SHAP values were SVM-RSpolar, LR-DPC, LR-RSsecond, SVM-AAC, and PLS-RScharge. Interestingly, most of the top five informative features (except PLS-RScharge) had positive SHAP values. Taking SVM-RSpolar as an example, for a given uncharacterized protein sequence *P*, if the value of SVM-RSpolar is very high, then *P* is predicted as a druggable protein; otherwise, *P* is predicted as a non-druggable protein.

To further reveal the influence of the optimal feature vector on the functioning of SPIDER, the performance of SPIDER was compared with that of a model lacking the optimal feature vector. Detailed comparison results of the two models on the training and independent test datasets are presented in Figure 6 and

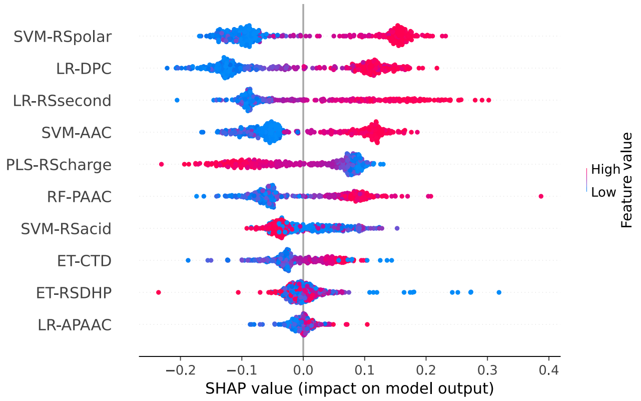


Figure 5. Ten important features of SPIDER ranked by SHAP values

SHAP values represent the directionality of the ten important features, where positive and negative SHAP values represent druggable protein and non-druggable protein predictions, respectively.

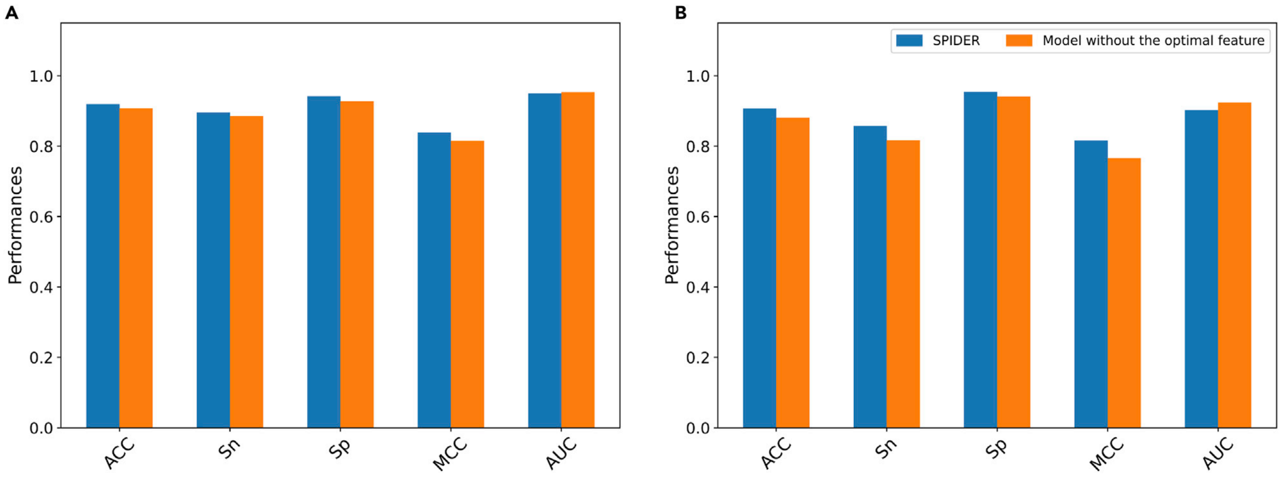


Figure 6. Predictive performance of various models

Performance comparison of SPIDER with the model without the optimal feature vector, as assessed by 10-fold cross-validation (A) and independent test (B).

Table S7. The comparison outcomes clearly indicated that SPIDER achieved an overall better performance than the compared model with regard to all performance metrics, with the exception of AUC. Specifically, the ACC, Sn, Sp, and MCC of SPIDER in the independent test dataset were 2.60%, 4.02%, 1.27%, and 5.05% higher, respectively, than those of the model lacking the optimal feature vector. These results demonstrate that *m* = 10 selected informative PFs are vital in capturing the key information pertaining to druggable proteins, thus contributing to the improvement in performance.

Utilization of the SPIDER webserver

Publicly accessible web servers are more beneficial for experimental researchers to identify their desired samples rather than developing their own internal prediction models. Therefore, we developed an online webserver for SPIDER, which is freely available at http://pmlabstack.pythonanywhere.com/SPIDER, to aid the broader research community in the identification of druggable protein candidates from large-scale proteins. In addition, we have provided stepwise instructions on the usage of the SPIDER webserver, which can be accessed using the "About" tab of the webserver.

Conclusion

This study presents SPIDER, an innovative stacked ensemble learning framework established for the precise prediction of druggable proteins. In SPIDER, we utilized ten distinctive feature descriptors based on various features, including physicochemical properties, composition-transition-distribution information, and compositional information. These feature descriptors, in conjunction with popular ML algorithms, have been used to develop numerous baseline models. Ultimately, *m* = 10 selected baseline models derived from the GA-SAR method were integrated to create the final meta-predictor in this study. Comparative experimental results showed that SPIDER was more efficient for druggable protein predictions compared to its baseline models in terms of cross-validation and independent tests. Moreover, SPIDER achieved better performance than the existing method, Yu's method, with an ACC of 0.907, Sn of 0.857, and MCC of 0.816, in terms of the independent test dataset. In addition, the SHAP algorithm was applied to determine the impact of each baseline model on the output provided by SPIDER. Finally, to aid highly efficient prediction of druggable proteins, we created an accessible webserver based on SPIDER that is readily available at http://pmlabstack.pythonanywhere.com/SPIDER. We believe that SPIDER will be a useful tool for the screening and identification of potential druggable proteins and to expedite their application in the drug discovery and development process.

Limitations of the study

Overall, the computational tool proposed in this study could enable a more precise and robust prediction of druggable proteins as compared to the current existing methods. In the meanwhile, we employed the

SHAP approach to elucidate the effect of each feature on the prediction of druggable proteins. However, there is still ample room for improving the prediction performance. Recently, several computational frameworks have been developed and reported, such as a flexible deep learning (DL)-based approach (Liang et al., 2022), DL-based hybrid approach (Hasan et al., 2022; Xie et al., 2021), and multilayer ensemble learning frameworks (Shoombuatong et al., 2022). In consideration of the effectiveness of these frameworks, in the future, we plan on integrating the appropriate computational methodologies for further enhancement of the prediction performance of druggable proteins.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ Benchmark datasets
  - ○ Feature engineering
  - ○ Feature selection based on GA-SAR
  - ○ SPIDER framework
  - ○ Performance evaluation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104883.

ACKNOWLEDGMENTS

This work was fully supported by College of Arts, Media and Technology, Chiang Mai University and partially supported by Chiang Mai University and Mahidol University. Furthermore, this work was supported by the National Research Foundation of Korea (NRF) and funded by the Korean government (MSIT) (2021R1A2C1014338), as well as by the SungKyunKwan University and the BK21 FOUR (Graduate School Innovation) funded by the Ministry of Education (MOE, Korea) and NRF. Many thanks to the Information Technology Service Center (ITSC) of Chiang Mai University for providing computational resources.

AUTHOR CONTRIBUTIONS

W.S. and B.M. conceived the idea and supervised the research. P.C. established ML models, performed the experiments, and developed the web server. W.S. and B.M. collected the data, analyzed the experiments, and revised the manuscript. M.A.M., P.L., W.S., and N.S. wrote the initial draft of the manuscript. All authors reviewed and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Received: April 24, 2022
Revised: July 8, 2022
Accepted: August 2, 2022
Published: September 16, 2022

REFERENCES

Azadpour, M., McKay, C.M., and Smith, R.L. (2014). Estimating confidence intervals for information transfer analysis of confusion matrices. J. Acoust. Soc. Am. *135*, EL140–EL146.

Bakheet, T.M., and Doig, A.J. (2009). Properties and identification of human protein drug targets. Bioinformatics *25*, 451–457.

Cao, Z., Pan, X., Yang, Y., Huang, Y., and Shen, H.-B. (2018). The lncLocator: a subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. Bioinformatics *34*, 2185–2194.

Charoenkwan, P., Chiangjong, W., Lee, V.S., Nantasenamat, C., Hasan, M.M., and Shoombuatong, W. (2021a). Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. Sci. Rep. *11*, 3017.

Charoenkwan, P., Chiangjong, W., Nantasenamat, C., Hasan, M.M., Manavalan, B., and Shoombuatong, W. (2021b). StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. Brief. Bioinform. *22*, bbab172.

Charoenkwan, P., Nantasenamat, C., Hasan, M.M., Moni, M.A., Lio', P., Manavalan, B., and Shoombuatong, W. (2022). StackDPPIV: a novel computational approach for accurate prediction of dipeptidyl peptidase IV (DPP-IV) inhibitory peptides. Methods *204*, 189–198.

Charoenkwan, P., Nantasenamat, C., Hasan, M.M., Moni, M.A., Manavalan, B., and Shoombuatong, W. (2021c). UMPred-FRL: a new approach for accurate prediction of umami peptides using feature representation learning. Int. J. Mol. Sci. *22*, 13124.

Charoenkwan, P., Nantasenamat, C., Hasan, M.M., and Shoombuatong, W. (2020). Meta-iPVP: a sequence-based meta-predictor for improving the prediction of phage virion proteins using effective feature representation. J. Comput. Aided Mol. Des. *34*, 1105–1116.

Charoenkwan, P., Schaduangrat, N., Nantasenamat, C., Piacham, T., and Shoombuatong, W. (2019). Int. J. Mol. Sci. *21*, 75.

Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K.-C. (2016). iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget *7*, 16895–16909.

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T.T., Wang, Y., Webb, G.I., Smith, A.I., Daly, R.J., Chou, K.-C., and Song, J. (2018). iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics *34*, 2499–2502.

Chou, K.-C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics *21*, 10–19.

Chou, K.C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins *43*, 246–255.

Dao, F.-Y., Lv, H., Zhang, D., Zhang, Z.-M., Liu, L., and Lin, H. (2021a). DeepYY1: a deep learning approach to identify YY1-mediated chromatin loops. Brief. Bioinform. *22*, bbaa356.

Dao, F.-Y., Lv, H., Zulfiqar, H., Yang, H., Su, W., Gao, H., Ding, H., and Lin, H. (2021b). A computational platform to identify origins of replication sites in eukaryotes. Brief. Bioinform. *22*, 1940–1950.

Dezső, Z., and Ceccarelli, M. (2020). Machine learning prediction of oncology drug targets based on protein and network properties. BMC Bioinf. *21*, 104–112.

Fakhraei, S., Huang, B., Raschid, L., and Getoor, L. (2014). Network-based drug-target interaction prediction with probabilistic soft logic. IEEE/ACM Trans. Comput. Biol. Bioinform. *11*, 775–787.

Fu, X., Cai, L., Zeng, X., and Zou, Q. (2020). StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. Bioinformatics *36*, 3028–3034.

Gong, Y., Liao, B., Wang, P., and Zou, Q. (2021). DrugHybrid_BS: using hybrid feature combined with bagging-SVM to predict potentially druggable proteins. Front. Pharmacol. *12*, 771808.

Hasan, M.M., Tsukiyama, S., Cho, J.Y., Kurata, H., Alam, M.A., Liu, X., Manavalan, B., and Deng, H.-W. (2022). Deepm5C: a deep-learning-based hybrid framework for identifying human RNA N5-methylcytosine sites using a stacking strategy. Mol. Ther. *30*, 2856–2867.

Ho, S.-Y., Chen, J.-H., and Huang, M.-H. (2004). Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. IEEE Trans. Syst. Man Cybern. B Cybern. *34*, 609–620.

Jamali, A.A., Ferdousi, R., Razzaghi, S., Li, J., Safdari, R., and Ebrahimie, E. (2016). DrugMiner: comparative analysis of machine learning algorithms for prediction of potential druggable proteins. Drug Discov. Today *21*, 718–724.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. *42*, D1091–D1097.

Li, F., Chen, J., Ge, Z., Wen, Y., Yue, Y., Hayashida, M., Baggag, A., Bensmail, H., and Song, J. (2021a). Computational prediction and interpretation of both general and specific types of promoters in Escherichia coli by exploiting a stacked ensemble-learning framework. Brief. Bioinform. *22*, 2126–2140.

Li, F., Guo, X., Jin, P., Chen, J., Xiang, D., Song, J., and Coin, L.J.M. (2021b). Porpoise: a new approach for accurate prediction of RNA pseudouridine sites. Brief. Bioinform. *22*, bbab245.

Li, Q., and Lai, L. (2007). Prediction of potential drug targets based on simple sequence properties. BMC Bioinf. *8*, 353.

Li, Z.-R., Lin, H.H., Han, L.Y., Jiang, L., Chen, X., and Chen, Y.Z. (2006). PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res. *34*, W32–W37.

Liang, Y., Wu, Y., Zhang, Z., Liu, N., Peng, J., and Tang, J. (2022). Hyb4mC: a hybrid DNA2vec-based model for DNA N4-methylcytosine sites prediction. BMC Bioinf. *23*, 258.

Lin, H., and Chen, W. (2011). Prediction of thermophilic proteins using feature selection technique. J. Microbiol. Methods *84*, 67–70.

Lin, J., Chen, H., Li, S., Liu, Y., Li, X., and Yu, B. (2019). Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier. Artif. Intell. Med. *98*, 35–47.

Lindsay, M.A. (2005). Finding new drug targets in the 21st century. Drug Discov. Today *10*, 1683–1687.

Liu, T., and Altman, R.B. (2014). Identifying druggable targets by protein microenvironments matching: application to transcription factors. CPT Pharmacometrics Syst. Pharmacol. *3*, e93–e99.

Lv, H., Dao, F.-Y., Guan, Z.-X., Yang, H., Li, Y.-W., and Lin, H. (2021a). Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method. Brief. Bioinform. *22*, bbaa255.

Lv, H., Dao, F.-Y., Zulfiqar, H., Su, W., Ding, H., Liu, L., and Lin, H. (2021b). A sequence-based deep learning approach to predict CTCF-mediated chromatin loop. Brief. Bioinform. *22*, bbab031.

Ma'ayan, A., Rouillard, A.D., Clark, N.R., Wang, Z., Duan, Q., and Kou, Y. (2014). Lean Big Data integration in systems biology and systems pharmacology. Trends Pharmacol. Sci. *35*, 450–460.

Mishra, A., Pokhrel, P., and Hoque, M.T. (2019). StackDPPred: a stacking based prediction of DNA-binding protein from sequence. Bioinformatics *35*, 433–441.

Overington, J.P., Al-Lazikani, B., and Hopkins, A.L. (2006). How many drug targets are there? Nat. Rev. Drug Discov. *5*, 993–996.

Owens, J. (2007). Determining druggability. Nat. Rev. Drug Discov. *6*, 187.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: machine learning in Python. J. Machine Learning Res. *12*, 2825–2830.

Perlman, L., Gottlieb, A., Atias, N., Ruppin, E., and Sharan, R. (2011). Combining drug and gene similarity measures for drug-target elucidation. J. Comput. Biol. *18*, 133–145.

Qiang, X., Zhou, C., Ye, X., Du, P.-f., Su, R., and Wei, L. (2020). CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning. Briefings Bioinf. *21*, 11–23.

Rao, B., Zhou, C., Zhang, G., Su, R., and Wei, L. (2020). ACPred-Fuse: fusing multi-view information improves the prediction of anticancer peptides. Brief. Bioinform. *21*, 1846–1855.

Sakharkar, M.K., Sakharkar, K.R., and Pervaiz, S. (2007). Druggability of human disease genes. Int. J. Biochem. Cell Biol. *39*, 1156–1164.

Shoombuatong, W., Basith, S., Pitti, T., Lee, G., and Manavalan, B. (2022). THRONE: a new approach for accurate prediction of human RNA N7-methylguanosine sites. J. Mol. Biol. *434*, 167549.

Sikander, R., Ghulam, A., and Ali, F. (2022). XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. Sci. Rep. *12*, 5505.

Sun, T., Lai, L., and Pei, J. (2018). Analysis of protein features and machine learning algorithms for prediction of druggable proteins. Quant. Biol. *6*, 334–343.

Van Der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. J. Mach. Learn. Res. *15*, 3221–3245.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

Wang, D., Zhang, Z., Jiang, Y., Mao, Z., Wang, D., Lin, H., and Xu, D. (2021). DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism. Nucleic Acids Res. *49*, e46.

Wang, K., Sun, J., Zhou, S., Wan, C., Qin, S., Li, C., He, L., and Yang, L. (2013). Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. PLoS Comput. Biol. *9*, e1003315.

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics *34*, 4007–4016.

Wolpert, D.H. (1992). Stacked generalization. Neural Network. *5*, 241–259.

Xie, R., Li, J., Wang, J., Dai, W., Leier, A., Marquez-Lago, T.T., Akutsu, T., Lithgow, T., Song, J., and Zhang, Y. (2021). DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy. Brief. Bioinform. *22*, bbaa125.

Xu, C., Ge, L., Zhang, Y., Dehmer, M., and Gutman, I. (2017). Computational prediction of therapeutic peptides based on graph index. J. Biomed. Inform. *75*, 63–69.

Yu, L., Xue, L., Liu, F., Li, Y., Jing, R., and Luo, J. (2022). The applications of deep learning algorithms on in silico druggable proteins identification. J. Adv. Res. https://doi.org/10.1016/j.jare.2022.01.009.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
iFeature	(Chen et al., 2018)	https://github.com/Superzchen/iFeature/
Reduced Sequences	(Lin et al., 2019)	https://github.com/QUST-AIBBDRC/GA-Bagging-SVM
Python package Scikit-learn v0.24.1	(Pedregosa et al., 2011)	https://scikit-learn.org/stable/
SPIDER	This paper	https://github.com/plenoi/SPIDER

RESOURCE AVAILABILITY

Lead contact

Further information regarding the methods and the dataset should be directed to and will be fulfilled by the lead contact, Professor Balachandran Manavalan (bala2022@skku.edu).

Materials availability

This study did not generate new reagents.

Data and code availability

All the dataset used in this study are available at http://pmlabstack.pythonanywhere.com/SPIDER. The source code for the SPIDER has been deposited at https://github.com/plenoi/SPIDER. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Benchmark datasets

We used the same training dataset derived from a study by Jamali et al. (2016) to generate and optimize our proposed models. The dataset comprised 1,224 druggable and 1,319 non-druggable proteins, representing positive and negative samples, respectively. Specifically, compilation of the positive samples was performed using the DrugBank database (Law et al., 2014), while Swiss-Prot was employed for the negative samples using the methods described by Li et al. (Li and Lai, 2007) and Bakheet et al. (Bakheet and Doig, 2009). Yu et al. (2022) recently established the first independent test dataset for this prediction problem. This independent test dataset contained 224 druggable and 237 non-druggable proteins. Additional details regarding the construction of the training and independent test datasets can be found in the articles by Jamali et al. (2016) and Yu et al. (2022), respectively. These datasets are available at http://pmlabstack.pythonanywhere.com/dataset_SPIDER.

Feature engineering

To obtain key information on druggable proteins, we utilized ten different feature-encoding schemes based on sequence information, namely PAAC, DPC, RSsecond, RSDHP, RSacid, RSpolar, RScharge, amino acid composition (AAC), amphiphilic pseudo-amino acid composition (APAAC), and composition-transition-distribution (CTD), indicating different aspects, including physicochemical properties, composition-transition-distribution information, and compositional information. The ten sequence-based feature encodings are sorted into three key groups as follows: (i) the first group consists of CTD-based features (Charoenkwan et al., 2021b, 2022); (ii) the second group consists of composition-based features (AAC and DPC (Rao et al., 2020; Wei et al., 2018)); and (iii) the third group consists of physicochemical property-based features (APAAC, PAAC, RSsecond, RSDHP, RSacid, RSpolar, and RScharge (Lin et al., 2019; Xu et al., 2017)). In this study, the aforementioned feature encodings were retrieved using the iFeature package (Chen et al., 2018). Comprehensive information regarding the feature encoding is presented in Table 2.

Feature selection based on GA-SAR

To enhance the predictive capability of the proposed models, we employed the GA-SAR approach to determine the model parameters and optimize the selection of informative features. This method was initially introduced by our group for the prediction of quorum-sensing peptides (Charoenkwan et al., 2019). This feature selection method has been successfully used in several bioinformatic applications (Charoenkwan et al., 2020, 2021a, 2021c, 2022). In brief, GA-SAR creates a profile that is employed to assess the importance of a feature. Note that the most important feature shows the highest correlation between the feature and output variable (Charoenkwan et al., 2019; Ho et al., 2004). In the GA-SAR algorithm, the chromosome contains binary and parametric genes, which are created for two main purposes: feature selection and ML parameter optimization. For ease of discussion, Gene and Chrom were used to represent the genes and chromosomes, respectively. Features with increased frequency are deemed significant for the prediction of druggable proteins. The implementation of GA-SAR algorithm to identify important features involves the following steps: (i) Randomly create 50 Chroms containing assigned values of binary Genes as means to get the number of features (*m*) equal to the selected number; (ii) Evaluate the performance for each Chrom in terms of the 10-fold cross-validation test; (iii) Construct a mating pool by performing a tournament selection; (iv) Perform a 20-point crossover on the selected parents; (v) Identify the optimal feature set by employing the SAR mutation operator; and (vi) Stop if the termination condition is reached; otherwise go to Step (ii). Additional information regarding the GA-SAR approach is available in the article by Charoenkwan et al. (2019).

SPIDER framework

In this study, we employed the stacking approach to establish SPIDER to improve the prediction of druggable proteins. This approach represents an efficient learning technique based on the ensemble method, which incorporates the individual abilities of various ML classifiers to create a single stable model (Cao et al., 2018; Fu et al., 2020; Mishra et al., 2019; Wolpert, 1992). To date, various stacking-based computational approaches have achieved improved performance compared with their baseline models (Charoenkwan et al., 2021b, 2021c; Li et al., 2021a, 2021b; Qiang et al., 2020; Xie et al., 2021). In particular, the construction process of the SPIDER includes three major steps, as summarized in Figure 1. Briefly, several baseline models were created and used to generate PFs. Finally, informative PFs were selected and employed in meta-predictor construction. Further details of the SPIDER framework are provided in the following paragraphs.

First, we created 60 baseline models developed using six different ML algorithms, SVM, RF, logistic regression (LR), k-nearest neighbor (KNN), ET, and partial least squares (PLS), in conjunction with ten widely used feature encodings, CTD, AAC, DPC, APAAC, PAAC, RSsecond, RSDHP, RSacid, RSpolar, and RScharge. We then systematically assessed the implementation of these six ML algorithms and ten feature encodings in the prediction of druggable proteins using the training and independent test datasets. Notably, the baseline model yielding the highest MCC in terms of the training dataset was deemed as the best-performing model. The Scikit-learn v0.24.1 package (Pedregosa et al., 2011) was utilized for the development and optimization of all baseline models, and the search range is documented in Table S1.

As each baseline model provided probabilistic information, we used this as the second step. Specifically, this information is the prediction confidence that the implied protein is druggable. Herein, the predicted confidence was considered PF, where the value of PF ranged from 0 to 1. As a result, for a given protein sequence *P*, we obtained 60 PFs generated by all 60 baseline models, which can be defined as follows:

$$\mathrm{P} = \left[PF(\mathrm{M}_1,\ F_1),\ldots.,PF\left(\mathrm{M}_i,\ F_j\right),\ldots,PF(\mathrm{M}_6,\ F_{10})\right)\right]^T \qquad \text{(Equation 1)}$$

where $P(\mathrm{M}_i,\ F_j)$ represents the PF generated by the baseline model trained using the i^{th} ML algorithm coupled with the j^{th} feature descriptor. Finally, *P* is converted into a 60-dimensional (60-D) feature vector.

In the third step, we used the 60-D feature vector to construct the meta-predictor based on the SVM algorithm (mSVM) using the Scikit-learn v0.24.1 package. To improve the performance of the mSVM predictor, we employed a GA-SAR approach (Charoenkwan et al., 2019). This allowed us to determine *m* informative PFs from 60 PFs, where *m* is in the range from 5 to 20. Herein, Chrom consisted of n = 60 binary genes (bg_i) to select *m* informative PFs ($m < n$) and 3-bit genes to optimize the parameters of the mSVM predictor (Table S1). If $bg_i = 1$, the i^{th} PF is used for the construction of the mSVM predictor; otherwise, the i^{th} PF is omitted from the optimal feature vector. Finally, the feature vector that achieves the highest cross-validation MCC is deemed to be the ideal one and is applied for the final meta-predictor construction.

Performance evaluation

Seven widely used performance metrics, MCC, ACC, AUC, F-value, precision (PRE), Sn, and Sp, were applied to the two-class prediction problem (Azadpour et al., 2014; Charoenkwan et al., 2021b) as follows:

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})} \tag{Equation 2}$$

$$\mathrm{F}-\mathrm{value} = 2\times\frac{\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}\times\mathrm{FN}} \tag{Equation 3}$$

$$\mathrm{PRE} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})} \tag{Equation 4}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \tag{Equation 5}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FP})} \tag{Equation 6}$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{Equation 7}$$

Specifically, TP and TN represent the numbers of correctly predicted true druggable and true non-druggable proteins, respectively. Furthermore, FP and FN indicate the number of non-druggable proteins that are predicted to be druggable proteins and the number of druggable proteins that are predicted to be non-druggable proteins, respectively (Dao et al., 2021a, 2021b; Lv et al., 2021a, 2021b; Wang et al., 2021).