ASFP (AI-based Scoring Function Platform): a web server for the development of customized scoring functions

Xujun Zhang
Zhejiang University https://orcid.org/0000-0002-5633-1517
Chao Shen
Zhejiang University https://orcid.org/0000-0003-2783-5529
Zhe Wang
Zhejiang University https://orcid.org/0000-0003-4102-1353
Gaoqi Weng
Zhejiang University https://orcid.org/0000-0001-8476-7548
Qing Ye
Zhejiang University https://orcid.org/0000-0003-3927-1919
Gaoang Wang
Zhejiang University https://orcid.org/0000-0002-0184-9562
Qiaojun He
Zhejiang University https://orcid.org/0000-0002-0104-6989
Bo Yang
Zhejiang University https://orcid.org/0000-0003-3524-4307
Dongsheng Cao
Central South University https://orcid.org/0000-0003-3604-3785
Tingjun Hou (tingjunhou@zju.edu.cn)
Zhejiang University https://orcid.org/0000-0001-7227-2580

Software

Keywords: Scoring functions, Descriptors, Machine learning, Virtual screening

Posted Date: October 27th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-96877/v1

Version of Record: A version of this preprint was published on February 4th, 2021. See the published version at https://doi.org/10.1186/s13321-021-00486-3.

Abstract

Virtual screening (VS) based on molecular docking has emerged as one of the mainstream technologies of drug discovery due to its low cost and high efficiency. However, the scoring functions (SFs) implemented in most docking programs are not always accurate enough and how to improve their prediction accuracy is still a big challenge. Here, we propose an integrated platform called ASFP, a web server for the development of customized SFs for structure-based VS. There are three main modules in ASFP: 1) the descriptor generation module that can generate up to 3437 descriptors for the modelling of protein-ligand interactions; 2) the AI-based SF construction module that can establish target-specific SFs based on the pre-generated descriptors through three machine learning (ML) techniques; 3) the online prediction module that provides some well-constructed target-specific SFs for VS and a generic SF for binding affinity prediction. Our methodology has been validated on several benchmark datasets. The target-specific SFs can achieve an average ROC AUC of 0.841 towards 32 targets and the generic SF can achieve the Pearson correlation coefficient of 0.81 on the PDBbind version 2016 core set. To sum up, the ASFP server is a powerful tool for structure-based VS and binding affinity prediction. Availability and Implementation: ASFP web server is freely available at http://cadd.zju.edu.cn/asfp/.

Introduction

As one of the core technologies in virtual screening (VS), molecular docking has been extensively used to screen small molecule libraries for lead discovery.[1] A protein-ligand docking algorithm consists of two basic components: a search algorithm to generate a large number of potential ligand binding poses within the binding site and a scoring function (SF) to evaluate the binding strength for a particular pose. In general, most SFs implemented in docking programs cannot give a reliable prediction to the relative binding affinity of a set of compounds.[2] Therefore, how to improve the accuracy of SFs still remains a big challenge.

Traditional SFs can be roughly classified into three categories:1) force field-based SFs, 2) knowledge-based SFs and 3) empirical SFs. Unlike traditional SFs, MLSFs do not have particular theory-motivated functional forms, and they are developed by learning from very large volumes of protein-ligand structural and interaction data through ML algorithms, such as random forest (RF), support vector machine (SVM), artificial neural network (ANN), gradient boosting decision tree (GBDT), etc.[3–7] Consequently, MLSFs have the capability to capture the non-linear relationship between protein-ligand interaction features and binding affinities that are difficult to be characterized by classical SFs, thus yielding better binding affinity predictions. However, in order to develop a MLSF, we need to generate a set of features to characterize protein-ligand interactions, and furthermore we need to be familiar with ML algorithms, which may be a difficult task for non-experts.

Here, we developed the ASFP server that can be used to develop customized MLSFs for structure-based VS and provide a generic MLSF for binding affinity prediction. The ASFP server has three basic modules: descriptor generation, AI-based SF construction and online prediction. In the descriptor generation

module, 15 computational tools (only 9 tools are available due to license restriction) are embedded into the module for the characterization of ligand, protein binding pocket and protein-ligand interaction information, and up to 3437 descriptors can be generated. The AI-based SF construction module can be used to develop customized SFs with easy operation. In the online prediction module, 15 well-validated target-specific classification models for VS and a generic regression model for binding affinity prediction are provided for users. All the above modules in the ASFP server are automated and the results are presented interactively through a user-friendly interface.

Implementation

The implementation of ASFP consists of two parts: the model construction and validation and the development of the web server that purposes in ML-based SF construction.

Model construction

Benchmark. The benchmark dataset I (Dataset I), which contains the kinase subset and the diverse subset in the Directory of Useful Decoys-Enhanced (DUD-E) benchmark, was used to train and assess the MLSFs. The kinase subset contains the inhibitors and decoys generated by DUDE for 26 kinases, and the diverse subset contains the inhibitors and decoys for seven representative targets in the entire DUDE set. The basic information of Dataset I is shown in **Table S1**.

The benchmark dataset □ (Dataset □) extracted from the PDBbind database (version 2016)[8] was used to train and evaluate the SVM regression model for binding affinity prediction. There are 4057 protein-ligand complexes in the "refined set" and 290 complexes in the "core set" of PDBbind version 2016.

Evaluation criteria. In this study, six evaluation criteria were utilized to assess the performance of the models. Among them, F1 score, Cohen's kappa, Matthews correlation coefficient (MCC), the area under the receiver operating characteristic curve (ROC AUC) and the enrichment factor (EF) at 1% were used to evaluate the performance of target-specific models while the Pearson correlation coefficient (R_p) was calculated to assess the performance of the SVM regression model. The details of the metrics can be found in Supplementary material.

Preparation. The protein targets were prepared by using the *Structure Preparation wizard* in Schrodinger version 2018, which added hydrogen atoms, repaired the side-chains of the imperfect residues using Prime, and optimized the steric hindrance of side-chains. The protonation states of the proteins were determined by using PROPKA and the het groups were preprocessed by Epik. The ligands were prepared using the *ligprep* module, which added hydrogen atoms, ionized the structures using Epik, desalted, generated tautomers and stereoisomers. In the preparation process, the default settings were used.

Docking. The grids were firstly generated by using the *Receptor Grid Generation* utility with the size of binding box set to 10 Å × 10 Å × 10 Å centered on the co-crystallized ligand. Then, the Glide docking

program with the SP scoring mode was used to dock the prepared ligands into the prepared proteins. For every ligand, only the pose with the highest docking score will be retained.

Descriptors generation. After molecular docking, the structural files of Dataset I and Dataset were retained for descriptors generation. In this study, a total of 15 descriptors calculation tools of various types were included in computing descriptors (Table 1). Considering some of the tools were restricted by license, two schemes were employed to generate the descriptors to establish MLSFs. First, all the SFs (excluding fingerprints and dpocket) supported by the computational tools in Table 1 were used to generate descriptors (ALL descriptors). Second, all the SFs supported by the computational tools without licenses restrictions in Table 1 (i.e. AffiScore version 3.0, AutoDock version 6.8, DSX version 0.9, GalaxyDockBP2, NNScore version 2.01 and SMoG2016) were used to generate descriptors (FREE descriptors). Both descriptors were implemented in the generic SF construction while only FREE descriptors were utilized to build target-specific classification models due to the huge computational cost.

Table 1
The basic information of the computational tools supported by the descriptor generation module.

Computational tools	Type of descriptors	No.	Types
AffiScore[1]	Energy terms	13	Empirical
ASP[1]	Energy terms	5	Knowledge Empirical
AutoDock	Energy terms	6	Force field
ChemPLP	Energy terms	11	Empirical
ChemScore	Energy terms	10	Empirical
DPOCKET	Pocket descriptors	49	-
DSX	Energy terms	1	Knowledge
RDKit	ECFP fingerprint	2048	-
GalaxyDockBP2	Energy terms	11	Empirical
Glide SP	Energy terms	17	Empirical
Glide XP	Energy terms	27	Empirical
GoldScore	Energy terms	6	Force field
NNscore	Energy terms	348	ML
PaDEL	Pubchem fingerprint	881	-
SMoG2016	Energy terms	5	Knowledge Empirical

[1]Computational tools without license restriction are marked in bold.

Modeling. For the construction of target-specific MLSFs, the dataset for each target in Dataset I was split into the training set and test set with the ratio of 3:1, and preprocessed to scale the data and remove duplicated features. Then, three ML algorithms, including Support Vector Machine (SVM), Random Forest (RF) and eXtreme Gradient Boosting (XGboost), were used to develop the MLSF for each target, and the hyperparameters were optimized with the hyperopt package. The performance of each model was assessed by a ten-fold cross-validation on the training set and the actual prediction on the test set. To develop the generic SVM regression model for binding affinity prediction, the PDBbind version 2016 'refined set' (excluding the PDBbind version 2016 'core set') was used as the training set and the PDBbind version 2016 'core set' was used as the test set.

Web API

Descriptors generation. With respect to the characterization of protein-ligand interactions, energy terms and knowledge-based pairwise potentials extracted from existing SFs are popular representation methods. These energy components correlated with the binding affinity of protein-ligand complexes can be used as the input for the development of MLSFs. Therefore, 12 scoring programs were integrated into this module and the scoring components from the output of the SFs implemented in these computational tools can be generated automatically. Besides, two computational tools, i.e., RDkit and PaDEL, were integrated into this module to calculate the Extended-connectivity fingerprint (ECFP) and Pubchem fingerprint, respectively, to characterize the structural features of small molecules. Furthermore, the function of fpocket was supported by this module to calculate 49 descriptors to characterize the structural information of protein pockets. It should be noted that the protein-ligand complexes should be docked before submitted to the server and the descriptors for small molecules may not be recommended for the development of MLSFs. The information of the 15 computational tools supported by ASFP are listed in Table 1. Because some computational tools implemented by ASFP are commercial, and therefore their functions are disabled. Based on the descriptors generated by this module, users can further construct a customized SF through a ML algorithm.

AI-Based Scoring Functions Construction. As one of the modules implemented in the server, the AI-based SF construction is designed for building customizing target-specific MLSFs. After submission, the workflow is summarized in Fig. 1. In this module, the 384 descriptors computed and extracted from the SFs implemented in 6 freely available computational tools (AffiScore version 3.0, AutoDock version 6.8, DSX version 0.9, GalaxyDockBP2, NNScore version 2.01 and SMoG2016) can be used for training SFs. First, the whole dataset uploaded by the user is divided into the training set and the test set according to the user's input. Then, the dataset is preprocessed (standardization, removing features with low variance, and tree-based feature selection) using *sklearn*. For the sake of computational efficiency, three popular ML algorithms (RF, SVM and XGBoost) are provided. Users can choose a ML algorithm for training and set some options about hyperparameter optimization (which hyperparameter to be optimized, the hyperparameter range and the optimization times). Finally, according to the user's input, the server uses *hyperopt* to find the optimal hyperparameter combinations and chooses the corresponding ML algorithm for training (10-fold cross validation) and prediction, and then outputs the results with a PDF file.

Online Prediction. On the base of the model performance, 15 well-constructed customized SFs with research-worthy targets and the generic regression SF for binding affinity prediction were retained to form the third module, Online prediction. The detailed information of the models is provided in Table 2.

Table 2
The information of the 15 targets with well-established classification models.

Target	Data source	ML algorithm	ROC_AUC on test set
abl1	DUD-E Kinase subset	SVM	0.848
akt2			0.859
csf1r			0.902
egfr			0.894
igf1r			0.846
jak2			0.921
kpcb			0.890
mapk2			0.876
mk01			0.838
src			0.852
tgfr1			0.965
wee1			0.965
akt1	DUD-E Diverse subset		0.850
cxcr4			0.942
hivpr			0.947

The ASFP server based on a high-level Python web framework of Django is deployed on a Linux server of an Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20 GHz CPUs with 28 cores and 64 GB of memory. Several SFs programs like *autodock* [9] were integrated to automate the calculation process. The overall workflow implemented in the ASFP server is shown in Supplementary Figure S1, and the manual of ASFP can be downloaded from the website (http://cadd.zju.edu.cn/asfp/).

Results

As is shown in the Fig. 2A and 2B, target-specific SFs constructed by ASFP outperformed the docking method, Glide SP, achieving an average ROC AUC of 0.841 towards 32 targets on the DUDE dataset. As

for binding affinity prediction, the generic SF can achieve the Pearson correlation coefficient of 0.81 on the PDBbind version 2016 core set [8], which is comparable to the state-of-the-art regression MLSFs (Fig. 2C). The average speed of modeling is 10 ligand per minute which is influenced by the ligand size and the computational capacities.

Discussion

All the three modules of ASFP required protein and ligand files uploaded and users can not only get satisfactory results as described in this paper by easily click the 'Run' button using default settings but also be allowed to submit jobs with their own settings. As shown above, the ML-based SFs constructed by ASFP outperform the classic SF (Glide sp) and can be built easily through the ASFP server. Therefore, our ASFP server is a powerful tool that can calculate descriptors for modeling and construct ML-based SFs for virtual screening and binding affinity prediction.

To illustrate the practicability of the ASFP server, if one wants to construct an ML-based SF to find ligands targeting at Tyrosine-protein kinase ABL (abl1), one can use the AI-Based Scoring Functions Construction module with the input files including a ligand file in the MOL2 format containing 50 active molecules, a decoy file in the MOL2 format containing 150 molecules, a test file in the MOL2 format containing 100 molecules and a protein file in the PDB format (PDB ID: 2HZI10). Upload the files and submit the job with the default hyperparameters settings. As shown in Figure 3, the ASFP server succeeds in generating descriptors and constructing a customized MLSF. The returned PDF file shows that the SF successfully identifies 23 inhibitors from 100 molecules (25 inhibitors).

Conclusions

Here, we present a user-friendly ASFP server for customizing SFs for structure-based VS. We have validated our methodology on several benchmark datasets, and the target-specific SFs constructed by ASFP achieved an average ROC AUC of 0.841 towards 32 targets on the DUDE dataset and the generic SF can achieve the Pearson correlation coefficient of 0.81 on the PDBbind version 2016 core set, suggesting that the ASFP server is a useful and effective tool for MLSF construction. The combination of 15 computational descriptor generation tools, *sklearn* and *hyperopt* makes it very convenient to calculate different types of descriptors and construct customized MLSFs. The ASFP server is an on-going project and further developments will be focused on the integration of more descriptor generation tools, the development of an automatic modelling pipeline using deep learning algorithms (e.g. 3D-convolutional neural networks) and the acceleration in computational speed with the help of more computing resources.

Declarations

Availability and requirements

- **Project name**: ASFP (AI-based Scoring Function Platform)
- **Project home page**: http://cadd.zju.edu.cn/asfp/
- **Operating system(s)**: Platform independent
- **Programming language**: Python
- **Other requirements**: Mozilla Firefox or Google Chrome is recommended
- **License**: MIT.

Any restrictions to use by non-academics: no.

Availability of data and materials

The web server is available at http://cadd.zju.edu.cn/asfp/.

The ASFP manual is available at http://cadd.zju.edu.cn/asfp/extract/download/?name=h.

The data and source code are available at https://github.com/5AGE-zhang/ASFP .

Funding Sources

This study was supported by the Key R&D Program of Zhejiang Province (2020C03010), Zhejiang Provincial Natural Science Foundation of China (LZ19H300001), and National Natural Science Foundation of China (21575128, 81773632).

Conflict of Interest: none declared.

Authors' contributions

XZ and CS developed the web application, analyzed the data, and wrote the manuscript; ZW, GW, QY, GW and QH evaluated and interpreted the results and wrote the manuscript; BY, DC and TH conceived and supervised the project, interpreted the results, and wrote the manuscript.

Competing interests

The authors declare that they have no competing interests.

Additional files

Supplementary file SI.docx.

References

1. Chen, Y.; Shoichet, B. K., Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nature Chemical Biology* **2009**, 5, 358-364.

2. Wang, Z.; Sun, H.; Yao, X.; Li, D.; Xu, L.; Li, Y.; Tian, S.; Hou, T., Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics* **2016**, 18, 12964-12975.
3. Shen, C.; Ding, J.; Wang, Z.; Cao, D.; Ding, X.; Hou, T., From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *WIREs Computational Molecular Science* **2020**, 10, e1429.
4. Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J., Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews-Computational Molecular Science* **2015**, 5, 405-424.
5. Durrant, J. D.; McCammon, J. A., NNScore 2.0: A Neural-Network Receptor-Ligand Scoring Function. *Journal of Chemical Information and Modeling* **2011**, 51, 2897-2903.
6. Trott, O.; Olson, A. J., Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of Computational Chemistry* **2010**, 31, 455-461.
7. Pereira, J. C.; Caffarena, E. R.; dos Santos, C. N., Boosting Docking-Based Virtual Screening with Deep Learning. *Journal of Chemical Information and Modeling* **2016**, 56, 2495-2506.
8. Li, Y.; Liu, Z.; Li, J.; Han, L.; Liu, J.; Zhao, Z.; Wang, R., Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *Journal of Chemical Information and Modeling* **2014**, 54, 1700-+.
9. Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J., Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **1998**, 19, 1639-1662.
10. Cowan-Jacob, S. W.; Fendrich, G.; Floersheimer, A.; Furet, P.; Liebetanz, J.; Rummel, G.; Rheinberger, P.; Centeleghe, M.; Fabbro, D.; Manley, P. W., Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. *Acta Crystallographica Section D-Biological Crystallography* **2007**, 63, 80-93.

Figures

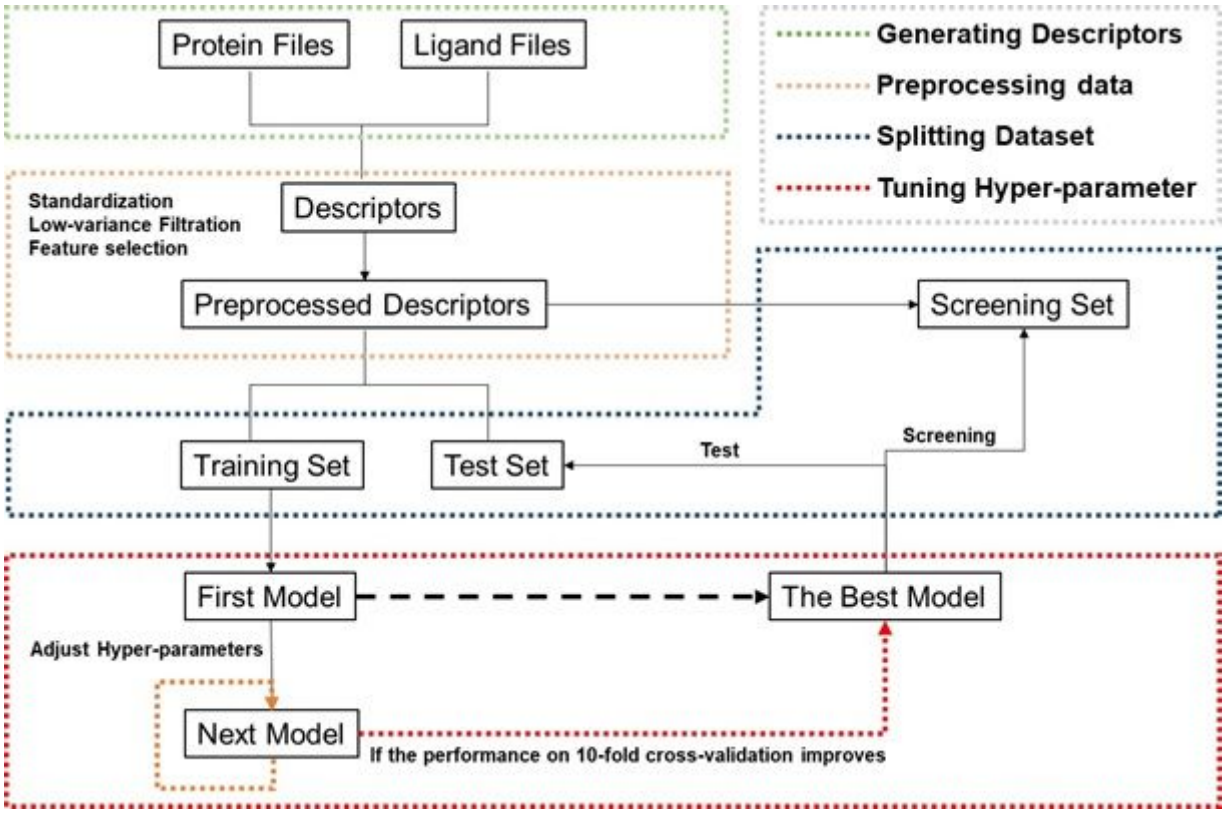

Figure 1

The workflow of the ASFP server for the AI-based scoring function construction.

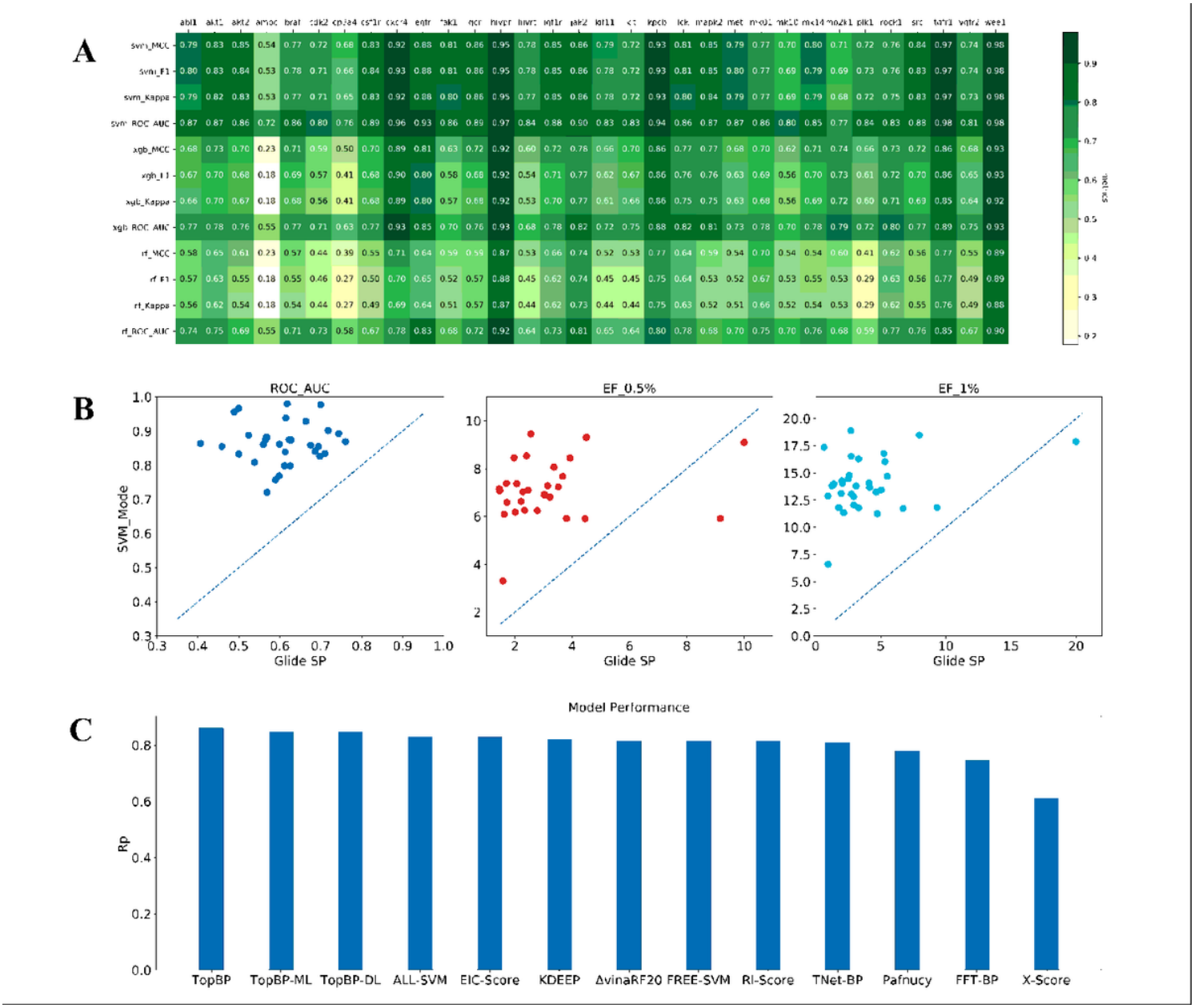

Figure 2

(A) The performance of the customizing SFs built by 3 ML algorithms (SVM, XGBoost and RF) on the Dataset □ under 4 metrics (MCC, F1, Kappa and ROC AUC) (B) Comparison of the screening power of customizing SFs and Glide SP based on the ROC AUC, EF at 0.5% level and EF at 1% level (C) The Scoring power of the generic SF implemented on the Online Prediction module. ALL-SVM represents the model learning from features part of which are generated from commercial scoring programs while FREE-SVM utilizing the descriptors calculated from free tools.

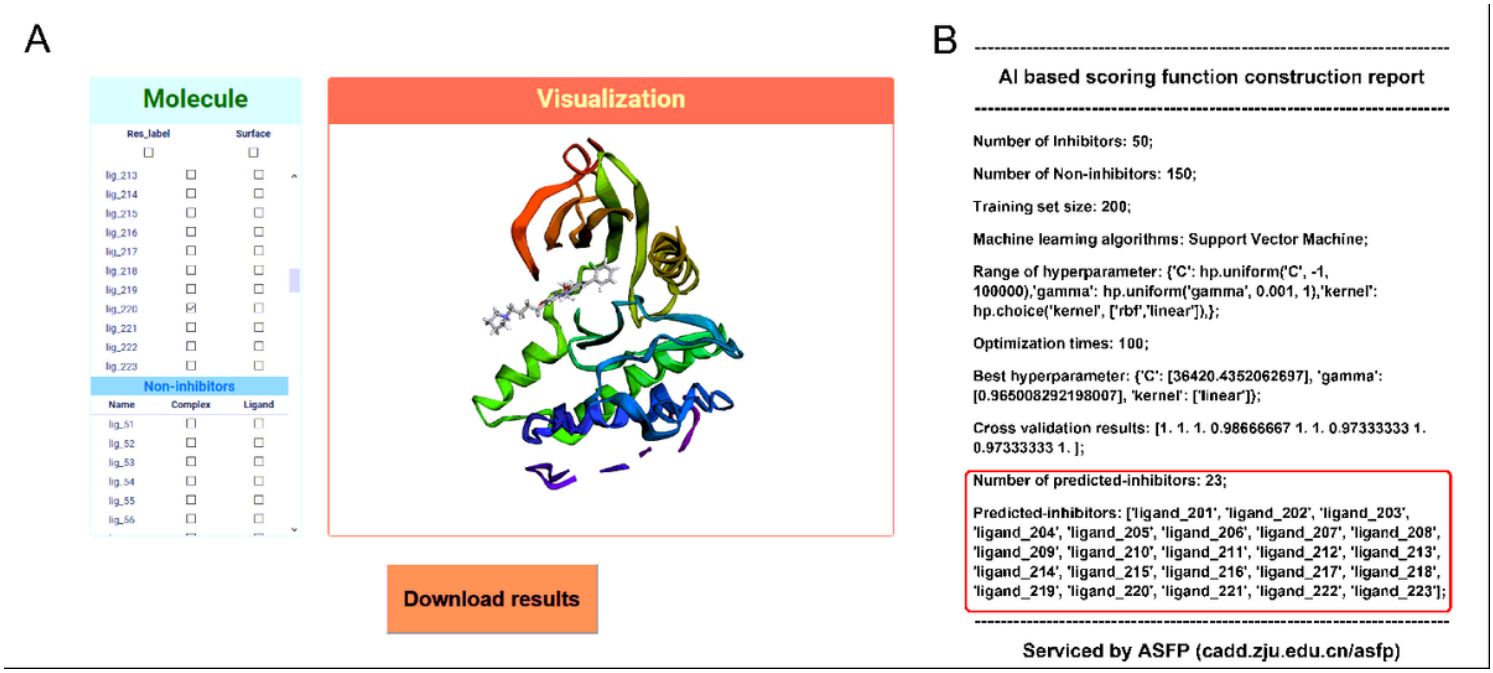

Figure 3

The AI-based scoring function construction result of the example (target: abl1). (A) The Visualization page of the results. (B) The prediction results in the report PDF file.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ASFPmanual.pdf
- SI.docx