Universitat
Pompeu Fabra
Barcelona

MACHINE LEARNING-BASED ALGORITHM FOR LIGAND BINDING SITE PREDICTION OF PROTEINS

Paula Mayo and Mariona Isern

Pompeu Fabra University, Barcelona

11 April 2023

1. Introduction

1.1. Background and significance of protein ligand binding site research

Proteins are some of the most important elements for life. They are not only crucial cellular components, but they also participate in various critical activities and processes in the life cycle of organisms, which can achieve or help to achieve important biological functions. Proteins are not isolated in living organisms, they need to interact with other biomolecules or ions such as nucleic acids, inorganic or organic small molecules and metal ions to perform their functions (1). These molecules and ions are called ligands. Particularly, intermolecular interactions between proteins and ligands occur via amino acid residues at specific positions in the protein, usually located in pocket-like regions. These specific key amino acid residues are called ligand binding sites (LBSs). Since, as we previously mentioned, proteins carry out their functions through interactions with ligands, accurately identifying the LBSs have attracted much attention in protein functional annotation, as well as in the fields of molecular docking, drug-target interactions, compound design, and even molecular dynamics (2–6). Thus, this project aims to develop a machine learning-based (ML) algorithm for LBSs prediction, using sequence and structural information. The source code and the dataset made to train and test the model are available at https://github.com/pmayog/PREDICT-RF.

1.2. Methods for detecting protein ligand binding sites

In principle, binding sites at protein surfaces can be detected experimentally. Hajduk and co-workers used heteronuclear-NMR-based screening to identify and characterize hot spots on protein surfaces (7). X-ray crystallography and tethering technologies are other approaches used (8). However, beyond time-consuming biological experiments, computational methods for the detection and characterisation of protein ligand-binding sites have increasingly become an area of interest now that large amounts of protein structural information are becoming available prior to any knowledge of protein function (9).

In the last twenty years, under the promotion of CASP and other research goals, researchers have made great progress in the field of LBS predictions. In this sense, many protein LBS prediction algorithms have been developed: geometric algorithms such as POCKET (10), SURFNET (11), or APROPOS (12), which are based on analyzing geometrical features like cleft or cavity, energetics-based methods (13–16), as well as machine learning and deep-learning related algorithms (14,17–19).

1.3. Implementation of machine learning approaches to predict protein ligand binding sites

The main purpose of ML algorithms is to reveal hidden information in data. Datasets are fed into the algorithm for ML models to understand how to perform different tasks. The model sees and learns from the training data automatically in order to be able to recognize the underlying, hidden relationships and patterns in the data or to predict a certain outcome. The validation dataset, on the other hand, is a different dataset that is often used during training to assess how well the model is performing and used to tune the hyperparameter of the model. After the model has been fully trained and validated, one can run assessment metrics on an independent test dataset not used in training to monitor the performance of the model predictions (20). In this sense, regarding the prediction of LBSs, ML follows the approach of learning the relation between different characteristics of the protein residues and their activity (i.e. if they bind or not to the ligand) from the known structures of protein–ligand complex pairs to derive statistical models for predicting the unknown LBSs of proteins.

1.3.1. Benchmark Dataset

PDBind represents one of the main prevailing dataset in the field of protein-ligand interactions that serve as a source of training, testing and validating data for ML methods. The database was created in 2004 by Wang et al. (21) with the aim of yield a comprehensive collection of experimentally measured binding affinity data for all biomolecular complexes deposited in the Protein Data Bank (PDB), so it provides an essential linkage between the energetic and structural information of those complexes (22). The dataset contains a broad set of protein-ligand complexes selected from the PDB database together with other types of biomolecular complexes. Being updated annually, the latest release (i.e. version 2020) contains binding data (K_d , K_i and IC_{50} values) for 19.443 protein–ligand, 2852 protein-protein, 1052 protein-nucleic acid and 149 nucleic acid-ligand complexes.

A high-quality subset of the complexes selected out of the general set is compiled into a set referred to as the “refined set”. Each protein-ligand complex in the refined set must meet the following requirements: (i) it must have a stable experimental binding affinity value such as K_d or K_i; complexes for which only IC_{50} values are known are excluded due to the measurement’s dependence on binding assay conditions; (ii) only one ligand is bound to the protein; (iii) protein and ligand molecules must be non-covalently bound to each other; (iv) the resolution of the crystal structure of the complex does not exceed 2.5 Angstrom (Å); and (v) for compatibility with molecular modeling software, neither the residues in the protein binding pocket nor the ligand are allowed to have organic elements other than C, N, O, P, S, F, Cl, Br, I, and H (23).

In addition, to make the refined set readily readable by most molecular modeling software, the original PDB structural file of each complex is split into a protein file and a ligand file. Both partners in each complex are processed properly and saved in standard formats (such as PDB, Mol2 and SDF). All these processed structural files together with the binding data can be downloaded as a package from PDBbind-CN Web site (24).

1.3.2. Main important features

The essence of protein LBSs resides in the geometry of the binding site and in the chemical composition of the ligand-contact surface. The LBSs geometries are diverse but not unlimited, represented by around 1000 pocket shapes. However, similar pockets could accommodate very different ligand scaffolds, and similar ligands could bind to pockets with very different geometries (25). The complexity in the chemical composition of the LBSs linings is not less daunting: statistical propensities for amino acids are not strikingly different between the LBSs and the general protein surfaces exposed to solvent (26). For these reasons, it is not easy to establish specific characteristics of protein residues and ligands that allow to perfectly discriminate between LBSs and non-LBSs or, at least, properties which may be responsible for protein-ligand interactions are not fully understood nowadays. Therefore, how to apply feasible features in protein interaction prediction remains an open problem.

Nevertheless, a large number of ML approaches to predict LBSs have achieved successful results (27) by combining different residue characteristics, even though prediction efficiency can still be improved. Some of the most included features can be classified into sequence or structural characteristics:

1.3.2.1. Sequence-based features

Hydrophobicity

Protein folds in an aqueous environment to acquire their native conformation, the one with the minimum free energy. The most determining property that affects protein folding is the hydrophobic effect: when a protein folds, solvation water molecules are released, increasing the entropy of the system, which decreases the free energy of it. As for the interface prediction, the hydrophobic effect is often a major contributor to stabilize protein complexes (28).

Gallet et al. proposed a fast method to predict protein interaction sites by analyzing hydrophobicity distribution (29). This work suggested that interface residues can be identified by using the hydrophobicity and the hydrophobic moment. Nevertheless, the role of hydrophilic and hydrophobic amino acids in ligand-binding sites is a complex issue and is still an active area of research. While some publications show that the binding residues are more hydrophobic than the other residues on the protein surface (30), other authors conclude that hydrophobic atoms have no apparent preference for interfaces (31) and even others discovered that hydrophilic amino acids are more likely to be ligand-binding (20).

Therefore, it seems that both hydrophilic and hydrophobic amino acids can be involved in ligand binding, and the specific amino acids involved can depend on the ligand and the protein in question.

Amino acid charges

Amino acid charges also play an important role in protein-ligand interactions since they can affect the electrostatic interactions between the protein and its ligands, which can impact the binding affinity and specificity.

Several studies have used amino acid charges to predict ligand binding residues. For example, a study by Tuo Zhang et al. (32) used machine learning models to predict ligand binding residues based on various protein features, including amino acid charges. Similarly, a study by Chen, P et al. developed a method for predicting ligand binding residues based on sequence information alone, including amino acid charges and other physicochemical properties (33). Finally, another study used amino acid charges and other features to predict protein-protein interaction sites. The authors found that charged amino acids seem to have a preference for non-interfaces (31).

Polarity

The polarity of amino acids can have an impact on the prediction of ligand-binding residues because it affects the types of interactions that can occur between amino acid residues and ligands.

Polar residues are typically hydrophilic and can form hydrogen bonds with other polar or charged molecules, including ligands. Thus, polar residues can often be found at the binding sites of proteins that interact with polar or charged ligands, such as enzymes that bind to substrates through hydrogen bonds. In this sense, looking at the distribution of amino acid residues, some articles found that polar residues were more abundant in interfaces (34–39).

In contrast, nonpolar residues, which are typically hydrophobic, tend to be buried in the interior of proteins away from water and polar or charged molecules. However, some nonpolar residues, such as tryptophan and phenylalanine, can also play important roles in ligand binding by forming hydrophobic interactions with ligands.

Isoelectric point

The isoelectric point of an amino acid is the pH at which it carries no net charge. It is an important feature when creating a LBS predictor because the charge of amino acids can influence the way they interact with other molecules.

As previously mentioned, many ligand-binding sites in proteins involve interactions between charged and polar groups on the protein and the ligand. The charge of the amino acids in the binding site can affect the strength and specificity of these interactions. For example, if the binding site contains acidic amino acids, it may interact more strongly with basic ligands, and if it contains basic amino acids, it may interact more strongly with acidic ligands.

The isoelectric point is also important because it can help to determine the overall charge of the protein at a given pH. Proteins can have different charges depending on the pH of their environment, and this can affect their interactions with other molecules. By incorporating the isoelectric point of amino acids into a ligand-binding site predictor, the algorithm can better account for the charge and pH-dependent properties of the protein, which can improve its accuracy in predicting ligand-binding sites.

Overall, the isoelectric point of amino acids is an important feature because it can help to capture the charge-dependent properties of the protein, which can influence its interactions with ligands. Thus, some authors have taken this feature into account in order to develop accurate ML methods for predicting the protein-ligand binding sites (30,40).

Evolutionary conservation

It has been long suggested that functional residues tend to be evolutionarily conserved. Consequently, it also has been suggested this to be true for binding surfaces (41,42). In fact, several publication results have shown a higher degree of conservation of the residues in LBSs compared to the rest of the protein's surface and that residue conservation is strongly correlated with ligand binding and catalytic sites (43).

Therefore, evolutionary conservation is an un-ignorable feature in ligand binding site prediction (44). One of the plausible methods to acquire this information is to compute the Position-Specific Scoring Matrix (PSSM), since it is the probability of mutating to each type of amino acid at each position. By analyzing the entropy of amino acids in a protein sequence, researchers can also gain insights into which positions are likely to be functionally important or evolutionarily conserved. This can be understood because the entropy of amino acids refers to the variability or diversity of amino acids at a particular position within a protein sequence across different species. Evolutionarily conserved positions in a protein sequence tend to have a lower entropy, meaning that fewer amino acids are observed at that position across different species. Conversely, positions that are less conserved tend to have a higher entropy, meaning that a wider range of amino acids are observed at that position across different species.

1.3.2.2. Structure-based features

Solvent accessible surface area (SASA)

Solvent Accessible Surface Area (SASA) [or briefly Accessible Surface Area (ASA)] of proteins is related to the spatial arrangement and packing of residues during the protein folding process. Residues that are exposed to the solvent are more likely to be involved in ligand-binding interactions, so SASA is another effective feature for protein–ligand binding site prediction (45).

Secondary structure

Regarding the secondary structure, there seems to be controversy on the preference of these in the binding sites. While some publications reported the preference of β-strands and relatively long non-structured chains for interfaces and the disfavor of α-helices (31), others conclude the reverse preferences for β-strands and α-helices (34). This may be the result of the different databases used.

1.3.3. Machine learning classification algorithms

There are two main issues that make the prediction of LBSs a very difficult task. The first one, that we already commented on, concerns the lack of understanding about the biological properties that are responsible for protein–ligand interactions, which leads to the difficulty of extracting informative features common to all the binding sites. The second one regards the fact that the number of interacting sites of a protein is much smaller than that of non-interacting sites, which leads to a very challenging problem, the so-called imbalanced data classification problem (46).

One drawback of imbalanced datasets is that it may introduce a bias towards the most represented class, leading to poor performance for the minority class. Another consequence is that some commonly used evaluation measures, such as accuracy, are not appropriate because they favour the most frequent class. Instead, several metrics that can provide better insight are precision, recall (i.e. sensitivity), F1 score and AUC (area under the ROC curve), commonly used in the Information Retrieval sciences (47).

Among the many classic ML algorithms, the naive Bayesian algorithm needs to calculate the prior probability and does not apply to data with a correlation between samples. Although the logistic regression is simple to implement, its accuracy is poor because it tends to under-fit characteristics. Besides, although the KNN algorithm is fast and has low training costs, the classification effect is poor under the sample imbalance situation. In the case of the support vector machine (SVM), it stands out from many traditional ML algorithms by virtue of its high classification accuracy, strong generalization ability, and excellent classification ability for high-dimensional small sample data. As a result, it has become a popular ML method in the field of LBS predictions (48). Moreover, eXtreme Gradient Boosting (XGBoost) is also a common ML algorithm used for this classification task (49). However, the issue of imbalanced datasets can be a challenge for SVM and XGBoost, as well as other classification methods. Nevertheless, Random Forest (RM) algorithm has been shown to perform well on imbalanced datasets. Moreover, RF can handle high-dimensional data and is robust to noise and outliers, making it a suitable choice for LBS prediction (20). For example, LigandRFs (33), developed by Chen et al, is a sequence-based method for identifying protein–ligand binding residues with a RF-based classifier. Besides LigandRFs, RF algorithm was also implemented by Qiu and Wang's method (50), Bordner (51), PRANK (52) and UTProt Galaxy (53).

In order to handle imbalanced data and thus improve the classification performance, one approach is to use cost-sensitive learning, where different misclassification costs are assigned to different classes. Another approach is to use different sampling techniques to balance the class distribution, such as oversampling the minority class or undersampling the majority class. Moreover, some publications have used an algorithm called Synthetic Minority Oversampling Technique (SMOTE), which basically populates the minority class with synthetic observations (46,54).

Overall, different algorithms are powerful for LBS prediction. It is important to experiment with many of them and choose the one that yields the best performance.

2. Material and methods

2.1. Dataset construction

The dataset was constructed based on 268 biologically relevant protein-ligand complexes (214 used as a training and 54 used as a test set) from the refined set of the PDBbind database. The PDB codes of all of them are shown on the supplementary material (Supplementary Table 1).

In accordance with PrankWeb database (55), for each residue on the pocket, if the distance between any atom of the residue and any atom of the ligand was less than 4 angstroms, this residue was tagged as binding; otherwise, the residue was tagged as a non-binding candidate.

2.2. Feature extraction

From each protein we extracted, based on the literature, the main important features used to build a predictor that can distinguish LBSs from non-LBSs.

On one hand, regarding the sequence-based ones, we obtained from the AAindex1 database (56) the hydrophobicity, the mean polarity, positive and negative charges and the isoelectric point of each amino acid of the protein (with accession numbers PRAM900101, RADA880108, FAUJ880111, FAUJ880112 and ZIMJ680104, respectively). Concerning the evolution information, we obtained the PSSM of each sequence using PSI-BLAST in the database of Swiss-Prot with five iterations and extracted entropy values from it.

On the other hand, two structure-based features were also extracted. The secondary structures of proteins were calculated by the DSSP algorithm, which is the standard method for assigning secondary structure to the amino acids given the atomic-resolution coordinates of a protein or, in other words, given the 3D structure of a protein. Although DSSP uses an eight-state secondary structure representation for the class label of each amino acid, it was reduced to 3-states: helix (H*),* strand (E) and coil (C). Lastly, in order to compute the solvent accessible surface area (SASA) of residues we took advantage of the Biopython package which calculates SASAs using the Shrake-Rupley algorithm.

2.3. Feature selection

In order to determine the most important features from those measured, the spearman correlation between the dependent variable (i.e. 'binding' or 'non-binding') and the independent ones (i.e. the different computed features) was assessed. Also the spearman correlation between the different independent variables was computed in order to avoid possible multicollinearity problems. Correlations were represented with a table and a heatmap, respectively.

2.4. Machine learning classification algorithms

Regarding the classification algorithms, we tried the more used ones in the field of LBS predictions, which are RF, XGBoost and SVC.

As the actual acquired protein–ligand binding site data show many fewer binding residues than non-binding residues, to effectively utilize the extracted features and to deal with the imbalanced data classification problems, different balancing techniques were tried in each algorithm used: undersampling, SMOTE and a combination of both.

Moreover, as the classification predictions result in a high percentage of false positives (FP), an increase of the threshold to the positive class probabilities was applied. Increasing the threshold means that fewer samples are being classified as positive, which reduces the number of false positives. This is because when the threshold is increased, a higher level of confidence in the model's prediction before classifying an instance as positive is required. Instances that are closer to the decision boundary (i.e. have probabilities close to 0.5) are more likely to be misclassified, so by setting a higher threshold we avoid classifying these instances as positive and potentially reduce the number of FP. However, increasing the threshold may also increase the number of false negatives (FN) and, consequently, decrease the overall number of positive predictions, which means that we may miss some true positive (TP) instances that would have been correctly classified with a lower threshold. As a result, the recall (i.e. true positive rate) increases because more true positives are correctly identified, but the precision (i.e. positive predictive value) decreases. Therefore, it is important to balance the trade-off between FP and FN or between the precision and the recall. In fact, this trade-off is a common issue in classification tasks. Focusing on LBSs prediction, we decided to increase the threshold from 0.5 (default) to 0.7 since successful results have been obtained using this threshold (57,58) and, at the same time, we also obtained well-balanced precision and recall values with our data by using this threshold.

Finally, a comparison between all the models built with the different algorithms and balancing techniques was made in order to determine the best one. Since the accuracy is insufficient for performance evaluation in unbalanced learning, the performance of classification was evaluated on the precision, recall, F1 score, and with a confusion matrix. The overall prediction quality of the binary models was evaluated using the area under the receiver operating characteristic curve (AUC).

3. Results

3.1. Dataset construction

The dataset contains 80.720 rows, that is the number of residues of all the proteins together, and 29 columns, which are the predictor features and the target variable (i.e. activity).

About the target variable, it is a binary variable with the number "0" representing a non-binding residue and the number "1" representing a binding one. The graph below shows the counts of each label of the variable. Specifically, there are 76.253 non-binding and 4.467 binding residues.

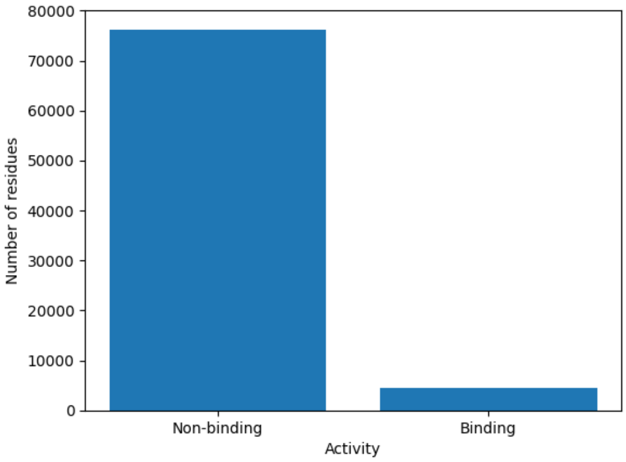


Figure 1. Counts of the residues belonging to each label (i.e. binding and non-binding) of the target (i.e. activity) variable.

Concerning the predictor variables, all of them were converted to numeric. Secondary structure and PSSM features are integers, while the entropy, hydrophobicity, polarity, amino acid charges, the isoelectric point and SASA are float variables.

Finally, no missing data was found in the dataset, so it did not need extensive preprocessing.

3.2. Feature selection

On one hand, regarding the correlation between the dependent variable (i.e. activity) and the independent ones (i.e. the different computed features), no strong or medium correlations were observed. Nevertheless, a low positive correlation between the activity variable and the entropy, as well as a low negative correlation between the activity variable and the PSSM feature of gluramine, threonine, glutamate, lysine, alanine and proline residues were detected.

On the other hand, about the correlation between the different independent variables, a very strong correlation (coefficient value of -0.91) was observed between the hydrophobicity and the polarity variables. Both variables were also correlated with amino acid charges (medium correlations) and the majority of PSSM features (from weak to medium correlations). Of the amino acid charges we can highlight a correlation with coefficient values above 0.55 with the isoelectric point (positive charge was positively correlated and negative charge was negatively correlated). Also, SASA feature showed a correlation of 0.44 with positive charge of residues. Regarding the correlations of the PSSM feature for the different residues, coefficients range from almost 0 to 0.88. Entropy was also correlated (medium correlations) with some of the PSSM features. Finally, no strong or medium correlations were observed between the secondary structure and any of the other features.

Despite the fact that some of the characteristics presented a correlation close to 0 with the dependent variable and, on the other hand, some explanatory variables were highly correlated, we decided to not drop any of the features to be able to further analyze the importance of each one in the final model.

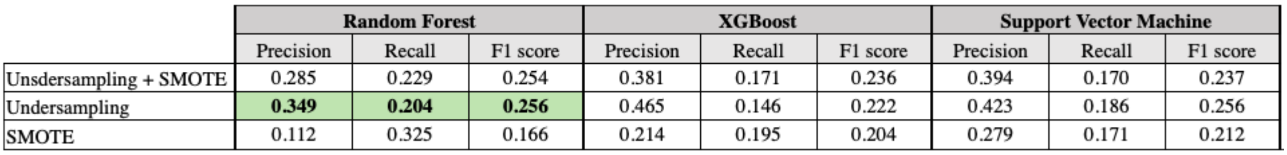
The correlation coefficients between the dependent variable and the independent ones can be seen on Table 2 of the supplementary material. In the same way, correlations between all pairs of features are shown on the Figure 1 of the supplementary material.

3.3. Machine learning classification algorithms

We split the original data into three sets: the training, the validation and the testing. The training set is used to train all models. Then, the validation set is used to assess the quality of the model prediction once the classifier is trained. Finally, once the better-performing model is chosen, the testing set is used to assess the quality of the model prediction but now only with the selected model.

The data was splitted in order that the test data represents a 20% and the training a 80%. The ratio of the split can vary depending on the size of the dataset and the complexity of the model. However, a 80/20 split is a common practice in many machine learning applications. We chose that ratio because by reserving 80% of the data for training, we have enough data to train the model effectively. If we allocate too much data for testing, we may not have enough data to train the model correctly. In the same way, by using 20% of the data for testing, we can reliably evaluate the performance of the model. If we allocate too little data for testing, the evaluation may not be representative of the model's true performance. Moreover, by reserving a significant portion of the data for testing, we can prevent overfitting. Overfitting occurs when a model is trained on the training data so well that it becomes too specialized to the training data and performs poorly on new data. By testing the model on a separate set of data, we can ensure that the model has not overfit to the training data.

Three algorithms (RF, XGBoost and SVC) were trained and evaluated with each of the balancing techniques, obtaining nine different model performances whose evaluation is summarized in the table below using three different metrics.



	Random Forest			XGBoost			Support Vector Machine		
	Precision	Recall	F1 score	Precision	Recall	F1 score	Precision	Recall	F1 score
Unsdersampling + SMOTE	0.285	0.229	0.254	0.381	0.171	0.236	0.394	0.170	0.237
Undersampling	**0.349**	**0.204**	**0.256**	0.465	0.146	0.222	0.423	0.186	0.256
SMOTE	0.112	0.325	0.166	0.214	0.195	0.204	0.279	0.171	0.212

Table 1. Evaluation using three different metrics (precision, recall and F1 score) of the models built with different algorithms (Random Forest, XGBoost and Support Vector Machine) and balancing techniques (undersampling, SMOTE and the combination of both).

Of these models, the RF algorithm with undersampling technique (which results in 2.914 samples in each class) was the one selected to construct the definitive prediction model since precision and recall did not differ much from each other and were quite high in comparison with the rest. Also the F1 score was the one with the highest value.

In order to understand the contribution of each feature to make accurate predictions, the importance of these was analyzed. As shown in the graph below, entropy was the most important one, followed by the SASA. Then PSSMs were also quite important. The isoelectric point, the polarity and the hydrophobicity were almost equally important. Finally, amino acid charges were the less important features.

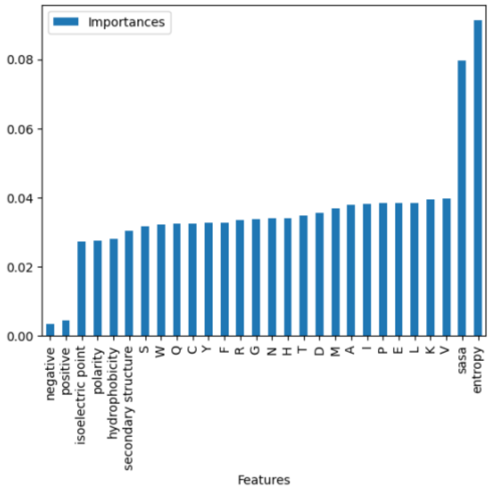


Figure 2. Random Forest importance of all the different features.

3.4. Analysis of some examples

So as to test PREDICT-RF, some PDBs were used as input to analyze the predicted residues compared to binding-site residues found in the PDB's sequence section.

3.4.1. *1hii.pdb*

Title: Comparative Analysis Of The X-Ray Structures Of Hiv-1 And Hiv-2 Proteases In Complex With Cgp 53820, A Novel Pseudosymmetric Inhibitor (59).

Type: Hydrolase (aspartic proteinase)

Molecule: HIV-2 protease

- **Chains**: A, B
- **Sequence length**: 99

Ligand: C20 and SO4

Reference Binding sites from PDB sequence section:

PDB's reference binding sites	Chain A	Chain B
P1, Binding site SO4	Not predicted	Not predicted
Q2, Binding site SO4	Not predicted	Not predicted
R8, Binding site	**Predicted**	**Predicted**
L23, Active site	**Predicted**	**Predicted**
G27, Binding site	**Predicted**	**Predicted**
A28, Binding site	**Predicted**	**Predicted**
D29, Binding site	**Predicted**	**Predicted**
D30, Binding site	Not predicted	Not predicted
I32, Binding site	**Predicted**	**Predicted**
V47, Binding site	Not predicted	Not predicted
G48, Binding site	Not predicted	Not predicted
G49, Binding site	**Predicted**	**Predicted**
I50, Binding site	Not predicted	Not predicted
K69, Binding site	Not predicted	Not predicted
T74, Binding site	Not predicted	Not predicted
T80, Binding site	Not predicted	Not predicted
P81, Binding site	Not predicted	Not predicted

I82, Binding site	Not predicted	Not predicted
I84, Binding site	**Predicted**	**Predicted**
N88, Binding site	**Predicted**	**Predicted**

Table 2. *1hii.pdb* chain A and chain B predicted and not predicted binding site residues.

Other predicted residues:

- **Chain A: PRO 9**, ILE 15, **ASP 25**, **THR 26**, **SER 31**, **GLY 86**, **ARG 87** and **LEU 90**.
- **Chain B: PRO 9**, ILE 15, **ASP 25**, **THR 26**, **SER 31**, **GLY 86** and **ARG 87** and **LEU 90**.

As shown in **Table 2**, 9 out of 20 residues correspond to different binding sites found in PDB sequence section. However, other 8 residues not found in the reference have been predicted in both chains. Despite the false positive results, analyzing these residue positions, the vast majority of them (remarked in black) are three residues far from a binding site, at most. For instance, residues located in positions 25, 26 and 31, are surrounded by binding sites.

1hii_binding_residues.cmd visualization in Chimera:

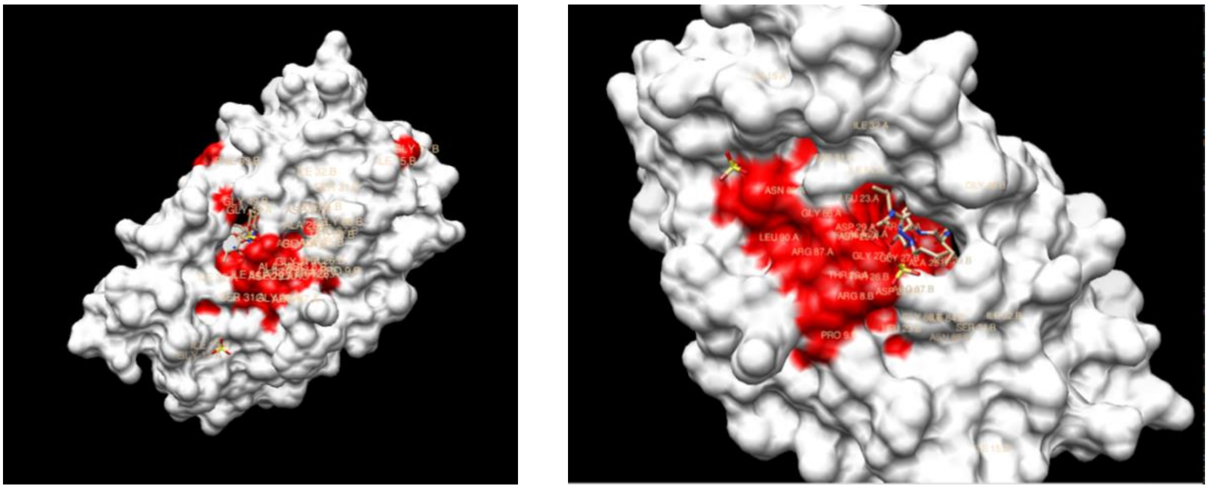


Figure 3. *1hii.pdb* predicted residues visualized in Chimera.

Lastly, despite the false positive ratio, when visualizing the predicted residues in Chimera, red color-labeled residues seem to accumulate close to the ligand.

3.4.2. *3npc.pdb*

Title: X-ray crystal structure of JNK2 complexed with the p38alpha inhibitor BIRB796: Insights into the rational design of DFG-out binding MAP kinase inhibitors (60).

Type: Transferase/transferase inhibitor

Molecule: Mitogen-activated protein kinase 9

- **Chains**: A, B
- **Sequence length**: 365

Ligand: B96

Reference Binding sites from PDB sequence section:

PDB's reference binding sites	Chain A	Chain B
32-40, Binding site	**Partially predicted (32, 33, 35, 37, 38, 40)**	**Partially predicted (32, 33 , 35, 37, 38, 40)**
A53, Binding site	**Predicted**	**Predicted**
K55, Binding site	**Predicted**	**Predicted**
R69, Binding site	Not predicted	Not predicted
E73, Binding site	**Predicted**	**Predicted**
L76, Binding site	Not predicted	Not predicted
I86, Binding site	Not predicted	Not predicted
L106, Binding site	Not predicted	Not predicted
M108, Binding site	**Predicted**	**Predicted**
L110, Binding site	Not predicted	Not predicted
M111, Binding site	Not predicted	Not predicted
L142, Binding site	Not predicted	Not predicted
D151, Active site	**Predicted**	**Predicted**
L168, Binding site	**Predicted**	**Predicted**
D169, Binding site	**Predicted**	**Predicted**
F170, Binding site	**Predicted**	**Predicted**

Table 3. *3npc.pdb* chain A and chain B predicted and not predicted binding site residues.

Other predicted residues:

- **Chain A:** TYR 26, **ARG 72**, **LEU 77**, HIS 82, **ASN 84**, **TYR 105**, **GLU 109**, LEU 115, GLN 134, **HIS 143**, **HIS 149**, **ARG 150**, **LEU 152**, **LYS 153**, ASN 156, ILE 157, VAL 158, LEU 165, **LYS**

166, **GLY 171**, ALA 173, THR 188, TYR 191, PRO 194, GLU 195, TYR 202, ASP 207, SER 210, GLY 212, PHE 225, VAL 237, LEU 302, ARG 309, SER 311, HIS 318, TYR 320, TYR 325, TYR 338, GLY 359.

- **Chain B**: PHE 10, TYR 26, **ARG 72**, **LEU 77,** HIS 82, **ASN 84, TYR 105, GLU 109**, LEU 115, GLN 134, **HIS 143**, **HIS 149**, **ARG 150**, **LEU 152**, **LYS 153**, ASN 156, ILE 157, VAL 158, LEU 165, **LYS 166**, **GLY 171,** ALA 173, THR 188, TYR 191, PRO 194, GLU 195, TYR 202, ASP 207, SER 210, GLY 212, PHE 225, VAL 237, LEU 302, ARG 309, HIS 318, TYR 320, TYR 325, TYR 338.

As summarized in **Table 3**, 14 out of 24 residues correspond to different binding sites found in the reference. However, 39 extra residues in chain A and 38 in chain B have been predicted. Looking at their positions, indicated residues (in black) are between one and two residues apart from a binding residue.

3npc_binding_residues.cmd visualization in Chimera:

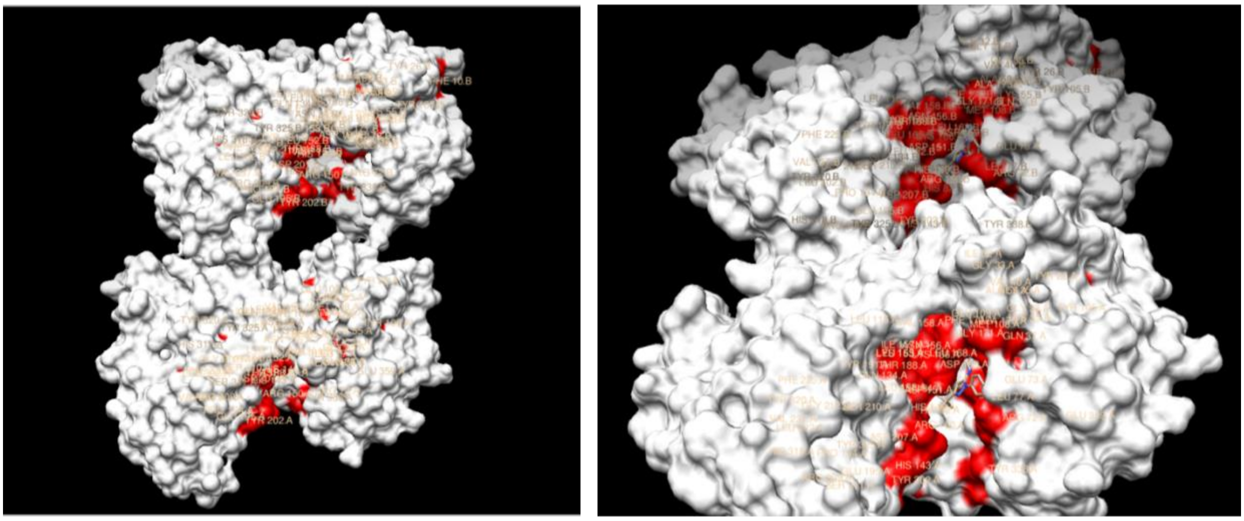


Figure 4. *3npc.pdb* predicted residues visualized in Chimera.

Finally, when looking at **Figure 4**, some predicted residues are located in the pocket region, where the ligand is situated, but also other residues are distributed in other regions far from this pocket.

3.4.3. *2wa3.pdb*

Title: Factor Inhibiting Hif-1 Alpha With 2-(3-Hydroxyphenyl)-2-Oxoacetic Acid (61).

Type: Oxidoreductase

Molecule: Hypoxia-inducible factor 1-alpha inhibitor

- **Chains**: A
- **Sequence length**: 349

Ligand: A29 and FE2

Reference Binding sites from PDB sequence section:

PDB's reference binding sites	Chain A
Y145, Binding site	**Predicted**
D152, Binding site	Not predicted
Q181, Binding site	Not predicted
L182, Binding site	Not predicted
T183, Binding site	Not predicted
L188, Binding site	Not predicted
T196, Binding site	Not predicted
H199, Binding site	**Predicted**
D201, Binding site	**Predicted**
E202, Binding site	Not predicted
Q203, Binding site	Not predicted
N205, Binding site	Not predicted
F207, Binding site	Not predicted
K214, Binding site	**Predicted**
R238, Binding site	Not predicted
Q239, Binding site	Not predicted
H279, Binding site	**Predicted**
I281, Binding site	Not predicted
N294, Binding site	**Predicted**
W296, Binding site	Not predicted
A300, Binding site	Not predicted
N321, Binding site	Not predicted

Table 4. *2wa3.pdb* chain A predicted and not predicted binding site residues.

Other predicted residues:

- **Chain A:** TRP 27, TYR 71, PHE 128, **GLY 190**, **PHE 206, GLY 212**, GLY 268, **TRP 277**.

Regarding results from **Table 4**, 6 out of 22 residues correspond to a binding site indicated in PDB sequence section. Moreover, 9 residues that do not correspond to any binding site have also been predicted. In comparison to other tested proteins, few non-binding but predicted residues are closely located to binding sites.

2wa3_binding_residues.cmd visualization in Chimera:

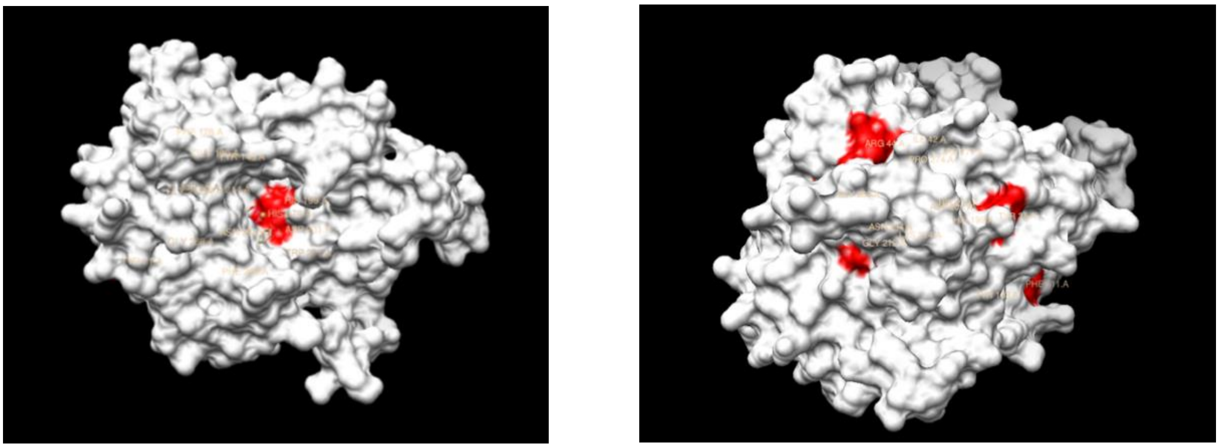


Figure 5. *2wa3.pdb* predicted residues visualized in Chimera.

Regarding **Figure 5**, it can be seen that some residues seem to be located in the pocket region, where the ligand is found, but there are also other residues in red that are far away from this area and seem to be dispersed.

3.4.4. *1rqp.pdb*

Title: Crystal structure and mechanism of a bacterial fluorinating enzyme (62).

Type: Transferase

Molecule: 5'-fluoro-5'-deoxyadenosine synthase

- **Chains**: A, B, C
- **Sequence length**: 299

Ligand: SAM

Reference Binding sites from PDB sequence section:

PDB's reference binding sites	Chain A	Chain B	Chain C
D16, Binding site	**Predicted**	**Predicted**	**Predicted**
D21, Binding site	Not predicted	Not predicted	Not predicted
D22, Binding site	Not predicted	Not predicted	Not predicted
S23, Binding site	Not predicted	Not predicted	Not predicted
W50, Binding site	Not predicted	Not predicted	Not predicted
Y77, Binding site	Not predicted	Not predicted	Not predicted
P78, Binding site	Not predicted	Not predicted	Not predicted
T80, Binding site	**Predicted**	Not predicted	Not predicted
T155, Binding site	Not predicted	**Predicted**	**Predicted**
F156, Binding site	**Predicted**	**Predicted**	**Predicted**
S158, Binding site	Not predicted	Not predicted	Not predicted
D210, Binding site	**Predicted**	Not predicted	Not predicted
F213, Binding site	**Predicted**	**Predicted**	**Predicted**
N215, Binding site	**Predicted**	**Predicted**	**Predicted**
W217, Binding site	Not predicted	Not predicted	Not predicted
F254, Binding site	Not predicted	Not predicted	Not predicted
S269, Binding site	Not predicted	Not predicted	Not predicted
R270, Binding site	Not predicted	Not predicted	Not predicted
R277, Binding site	Not predicted	Not predicted	Not predicted
N278, Binding site	**Predicted**	**Predicted**	**Predicted**
A279, Binding site	Not predicted	Not predicted	Not predicted

Table 5. *1rqp.pdb* chain A, chain B and chain C predicted and not predicted binding site residues.

Other predicted residues:

- **Chain A:** VAL 40, VAL 41, ASP 42, VAL 43, HIS 45, **VAL 52**, PHE 65, THR 82, ASN 124, **VAL 216, TYR 266**, VAL 296, ALA 298.
- **Chain B:** ARG 8, VAL 40, VAL 41, ASP 42, VAL 43, HIS 45, **VAL 52**, PHE 65, ASN 124, **VAL 216, TYR 266**, VAL 296, ALA 298.
- **Chain C:** VAL 41, VAL 43, HIS 45, **VAL 52**, PHE 65, THR 82, ASN 124, ILE 209, **VAL 216, TYR 266**, VAL 296, ALA 298.

Results summarized in **Table 5** show that 7 out of 21 in chain A and 6 out of 21 in chains B and C were correctly predicted. With respect to predicted residues not found in the reference, chains A and B presented 13 extra residues, and chain C 12 extra residues, with 3 residues each being close to a binding site.

1rqp_binding_residues.cmd visualization in Chimera:

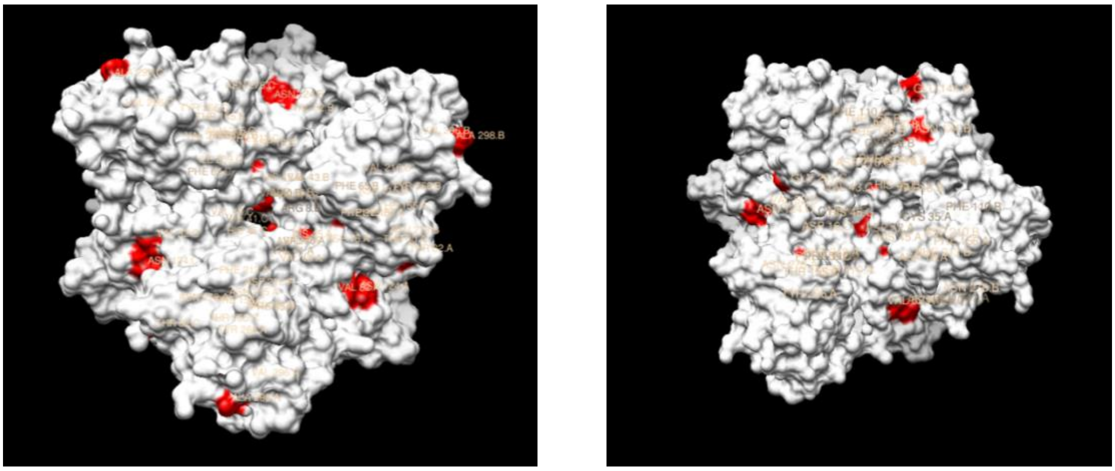


Figure 6. *1rqp.pdb* predicted residues visualized in Chimera.

Regarding **Figure 6,** it can be seen that some residues are dispersed in the protein. However, the ligands are situated in the inside of the protein, where some predicted residues are located.

4. Discussion

The main objective of this work was to develop a ML algorithm to predict the ligand binding sites of proteins.

To achieve it, 28 characteristics were computed for each residue of 268 proteins included in the refined set of the PDBind database. These characteristics can be classified into sequence-based features and structure-based features. Regarding the first group, hydrophobicity, negative and positive charges, mean polarity, the isoelectric point, the PSSM and the entropy of amino acids were extracted. With respect to the structure-based features, secondary structure and solvent accessible surface area were calculated. Afterwards, a data frame containing these characteristics (as columns) for each residue (as rows) of all the proteins was constructed in order to be able to train, test and validate the model. The last column refers to the target variable and, therefore, defines the type of residue, that is, whether or not it binds to the ligand. In order to determine that, for each residue on the pocket, if the distance between any atom of the residue and any atom of the ligand was less than 4 angstroms, this residue was tagged as binding; otherwise, the residue was tagged as a non-binding candidate.

The most used algorithms (i.e. Random Forest, XGBoost Classifier and Support Vector Classifier) in LBSs prediction, as well as different balancing techniques (i.e. undersampling, SMOTE and undersampling + SMOTE) were tried to construct the ML model. Finally, the RF algorithm with undersampling outperformed the rest when it comes to the F1 score, which was one of the metrics used to evaluate them. Moreover, precision and sensitivity results were similar to each other. Our results align with previous studies that have also used this algorithm for their ML model (33,50–53). Thus, our program uses sequence-based features, structure-based features and RF with undersampling to build prediction models.

Nevertheless, a set of limitations should be considered. In the first place, none of the included features showed a strong or moderate correlation with the target variable, which suggests the need to discover new properties with better prediction performance. The ones that were more correlated with the target variable were the entropy and the PSSM. The fact that the variables that were more correlated with the target variable were both sequence-based features can be seen as an advantage, particularly in cases where the 3D structure of the input proteins is unknown.

Secondly, a distance-based approach to discern between binding and non-binding residues have been used. Despite it is a commonly used method, other factors such as SASA or hydrogen bonding interactions could also have been considered in order to be able to increase the accuracy and the reliability of predictions of ligand binding residues.

SASA measures the exposed surface area of a residue, which can be used to estimate the extent to which a residue is accessible to solvent molecules and potentially to ligands. Residues with high SASA values may be more likely to be involved in binding interactions, as they are more exposed to the solvent and thus more likely to come into contact with ligands. In addition, hydrogen bonding interactions between residues and ligands can also be used to identify potential binding sites since hydrogen bonding interactions play a key role in stabilizing protein-ligand complexes.

On the third place, it is important to note that our program does not contemplate information about the local environment around each residue. Thus, using a windows-based approach would have been a good practice since it allows the model to capture this missed information and, therefore, the model can take into account the influence of nearby residues on the likelihood of ligand binding of the target residue. This is important because ligand binding sites are often composed of a network of residues that work together to stabilize the ligand and facilitate binding interactions. Moreover, using a window-based approach can help to reduce the noise in the input data and increase the robustness of the predictor.

Taking into account all the above mentioned limitations and, in addition, the difficulty itself of obtaining a model that achieves excellent predictions considering the high complexity that ligand binding sites can have, our model results in a precision of 35% and a sensitivity of 20%. Although the percentages may not appear promising, our model has yielded favorable outcomes in numerous analyses of various protein-ligand binding sites. Notably, it has accurately predicted binding residues in many instances, despite some occurrences of both false positives and false negatives. Moreover, some of these false positive residues tend to locate near binding sites, which seem to be part of protein pockets, although these are not considered ligand binding residues.

Regarding its applicability, some issues have arised concerning PDBs variability used during the program testing. As any set specifying PDB characteristics was found, testing phase was performed with manual-searched PDBs that included ligands and were preferably simple enzymes. Thus, not many different results were obtained with respect to its precision and sensitivity. For this reason, using a well-curated and classified set including different types of molecules could reveal more information about its applicability.

5. Conclusion

The current era is marked by advanced ML techniques, rapid growth of public data and increase in computing power. Taking advantage of these developments, large amounts of projects with the aim of predicting ligand binding sites of proteins are currently being published. Despite many excellent computational methods are available, prediction efficiency can still be improved

In this sense, this paper proposes a new ML-based method for predicting LBSs of proteins by using both sequence and structure information. Features extracted and RF algorithm with undersampling were used to construct the prediction model.

We were able to successfully predict the binding residues with a precision of 35%. This success rate was achieved using only 28 RF attributes, so prediction performance should be improved when more properties that distinguish between binding residues and the rest of the protein surface become available.

REFERENCES

1. Chen K, Mizianty MJ, Kurgan L. ATPsite: sequence-based prediction of ATP-binding residues. Proteome Sci. 2011 Dec 14;9(S1):S4.
2. Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biol. 2011 Dec 28;9(1):71.
3. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics. 2018 Sep 1;34(17):i821–9.
4. Ballester PJ, Mitchell JBO. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics. 2010 May 1;26(9):1169–75.
5. Seco J, Luque FJ, Barril X. Binding Site Detection and Druggability Index from First Principles. J Med Chem. 2009 Apr 23;52(8):2363–71.
6. Heo L, Shin WH, Lee MS, Seok C. GalaxySite: ligand-binding-site prediction by using molecular docking. Nucleic Acids Res. 2014 Jul 1;42(W1):W210–4.
7. Hajduk PJ, Huth JR, Fesik SW. Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. J Med Chem. 2005 Apr 1;48(7):2518–25.
8. Vajda S, Guarnieri F. Characterization of protein-ligand interaction sites using experimental and computational methods. Curr Opin Drug Discov Devel. 2006 May;9(3):354–62.
9. Campbell SJ, Gold ND, Jackson RM, Westhead DR. Ligand binding: functional site location, similarity and docking. Curr Opin Struct Biol. 2003 Jun;13(3):389–95.
10. Levitt DG, Banaszak LJ. POCKET: A computer graphies method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph. 1992 Dec;10(4):229–34.
11. Laskowski RA. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph. 1995 Oct;13(5):323–30.
12. Peters KP, Fauck J, Frömmel C. The Automatic Search for Ligand Binding Sites in Proteins of Known Three-dimensional Structure Using only Geometric Criteria. J Mol Biol. 1996 Feb;256(1):201–13.
13. Ghersi D, Sanchez R. EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. Bioinformatics. 2009 Dec 1;25(23):3185–6.
14. Hernandez M, Ghersi D, Sanchez R. SITEHOUND-web: a server for ligand binding site identification in protein structures. Nucleic Acids Res. 2009 Jul 1;37(Web Server):W413–6.
15. Ngan CH, Hall DR, Zerbe B, Grove LE, Kozakov D, Vajda S. FTSite: high accuracy detection of ligand binding sites on unbound protein structures. Bioinformatics. 2012 Jan 15;28(2):286–7.
16. Xie ZR, Hwang M. Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. Bioinformatics. 2012 Jun 15;28(12):1579–85.
17. Yang J, Roy A, Zhang Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics. 2013 Oct 15;29(20):2588–95.

18. Dou Y, Wang J, Yang J, Zhang C. L1pred: A Sequence-Based Prediction Tool for Catalytic Residues in Enzymes with the L1-logreg Classifier. PLoS One. 2012 Apr 27;7(4):e35666.

19. Wass MN, Kelley LA, Sternberg MJE. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res. 2010 Jul 1;38(suppl_2):W469–73.

20. Dhakal A, McKay C, Tanner JJ, Cheng J. Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. Brief Bioinform. 2022 Jan 17;23(1).

21. Wang R, Fang X, Lu Y, Yang CY, Wang S. The PDBbind Database: Methodologies and Updates. J Med Chem. 2005 Jun 1;48(12):4111–9.

22. Prof. Renxiao Wang. PDB bind database [Internet]. Available from: http://www.pdbbind.org.cn/index.php

23. Ashtawy HM, Mahapatra NR. A Comparative Assessment of Predictive Accuracies of Conventional and Machine Learning Scoring Functions for Protein-Ligand Binding Affinity Prediction. IEEE/ACM Trans Comput Biol Bioinform. 2015 Mar 1;12(2):335–47.

24. Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, et al. PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics. 2015 Feb 1;31(3):405–12.

25. Gao M, Skolnick J. A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. PLoS Comput Biol. 2013 Oct 24;9(10):e1003302.

26. Khazanov NA, Carlson HA. Exploring the Composition of Protein-Ligand Binding Sites on a Large Scale. PLoS Comput Biol. 2013 Nov 21;9(11):e1003321.

27. Yang C, Chen EA, Zhang Y. Protein–Ligand Docking in the Machine-Learning Era. Molecules. 2022 Jul 18;27(14):4568.

28. Glaser F, Steinberg DM, Vakser IA, Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins. 2001 May 1;43(2):89–102.

29. Bradford JR, Westhead DR. Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics. 2005 Apr 15;21(8):1487–94.

30. Wang K, Gao J, Shen S, Tuszynski JA, Ruan J, Hu G. An Accurate Method for Prediction of Protein-Ligand Binding Site on Protein Surface Using SVM and Statistical Depth Function. Biomed Res Int. 2013;2013:1–7.

31. Neuvirth H, Raz R, Schreiber G. ProMate: A Structure Based Prediction Program to Identify the Location of Protein–Protein Binding Sites. J Mol Biol. 2004 Apr;338(1):181–99.

32. Zhang T, Zhang H, Chen K, Shen S, Ruan J, Kurgan L. Accurate sequence-based prediction of catalytic residues. Bioinformatics. 2008 Oct 15;24(20):2329–38.

33. Chen P, Huang JZ, Gao X. LigandRFs: random forest ensemble to identify ligand-binding residues from sequence information alone. BMC Bioinformatics. 2014 Dec 3;15(S15):S4.

34. Jones S, Thornton JM. Protein-protein interactions: A review of protein dimer structures. Prog Biophys Mol Biol. 1995;63(1):31–65.

35. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches 1 1Edited by G.Von Heijne. J Mol Biol. 1997 Sep;272(1):121–32.

36. Aloy P, Querol E, Aviles FX, Sternberg MJE. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol. 2001 Aug;311(2):395–408.

37. Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, et al. An Accurate, Sensitive, and Scalable Method to Identify Functional Sites in Protein Structures. J Mol Biol. 2003 Feb;326(1):255–61.

38. Albeck S, Unger R, Schreiber G. Evaluation of direct and cooperative contributions towards the strength of buried hydrogen bonds and salt bridges. J Mol Biol. 2000 May;298(3):503–20.

39. Bright JN, Woolf TB, Hoh JH. Predicting properties of intrinsically unstructured proteins. Prog Biophys Mol Biol. 2001;76(3):131–73.

40. Li L, Koh CC, Reker D, Brown JB, Wang H, Lee NK, et al. Predicting protein-ligand interactions based on bow-pharmacological space and Bayesian additive regression trees. Sci Rep. 2019 May 22;9(1):7703.

41. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, et al. ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. Bioinformatics. 2003 Jan 1;19(1):163–4.

42. Xie L, Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments. Proceedings of the National Academy of Sciences. 2008 Apr 8;105(14):5441–6.

43. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics. 2007 Aug;23(15):1875–82.

44. Dai T, Liu Q, Gao J, Cao Z, Zhu R. A new protein-ligand binding sites prediction method based on the integration of protein sequence conservation information. BMC Bioinformatics. 2011 Dec 14;12(S14):S9.

45. Hu X, Feng Z, Zhang X, Liu L, Wang S. The Identification of Metal Ion Ligand-Binding Residues by Adding the Reclassified Relative Solvent Accessibility. Front Genet. 2020 Mar 19;11.

46. Chen X wen, Jeong JC. Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics. 2009 Mar 1;25(5):585–91.

47. Šikić M, Tomić S, Vlahoviček K. Prediction of Protein–Protein Interaction Sites in Sequences and 3D Structures by Random Forests. PLoS Comput Biol. 2009 Jan 30;5(1):e1000278.

48. Zhao J, Cao Y, Zhang L. Exploring the computational methods for protein-ligand binding site prediction. Comput Struct Biotechnol J. 2020;18:417–26.

49. Zhao Z, Xu Y, Zhao Y. SXGBsite: Prediction of Protein–Ligand Binding Sites Using Sequence Information and Extreme Gradient Boosting. Genes (Basel). 2019 Nov 22;10(12):965.

50. Qiu Z, Wang X. Improved Prediction of Protein Ligand-Binding Sites Using Random Forests. Protein Pept Lett. 2011 Dec 1;18(12):1212–8.

51. Bordner AJ. Predicting small ligand binding sites in proteins using backbone structure. Bioinformatics. 2008 Dec 15;24(24):2865–71.

52. Krivák R, Hoksza D. Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features. J Cheminform. 2015 Dec 1;7(1):12.

53. Komiyama Y, Banno M, Ueki K, Saad G, Shimizu K. Automatic generation of bioinformatics tools for predicting protein–ligand binding sites. Bioinformatics. 2016 Mar 15;32(6):901–7.

54. Chen C, LA, & BL. Using Random Forest to Learn Imbalanced Data. University of California, Berkeley. 2004;

55. Jendele L, Krivak R, Skoda P, Novotny M, Hoksza D. PrankWeb: a web server for ligand binding site prediction and visualization. Nucleic Acids Res. 2019 Jul 2;47(W1):W345–9.

56. AAindex [Internet]. Available from: https://www.genome.jp/aaindex/

57. Qiu Z, Wang X. Improved Prediction of Protein Ligand-Binding Sites Using Random Forests. Protein Pept Lett. 2011 Dec 1;18(12):1212–8.

58. Suresh MX, Gromiha MM, Suwa M. Development of a Machine Learning Method to Predict Membrane Protein-Ligand Binding Residues Using Basic Sequence Information. Adv Bioinformatics. 2015 Jan 31;2015:1–7.

59. Priestle J, Fässler A, Rösel J, Tintelnot-Blomley M, Strop P, Grütter M. Comparative analysis of the X-ray structures of HIV-1 and HIV-2 proteases in complex with CGP 53820, a novel pseudosymmetric inhibitor. Structure. 1995 Apr;3(4):381–9.

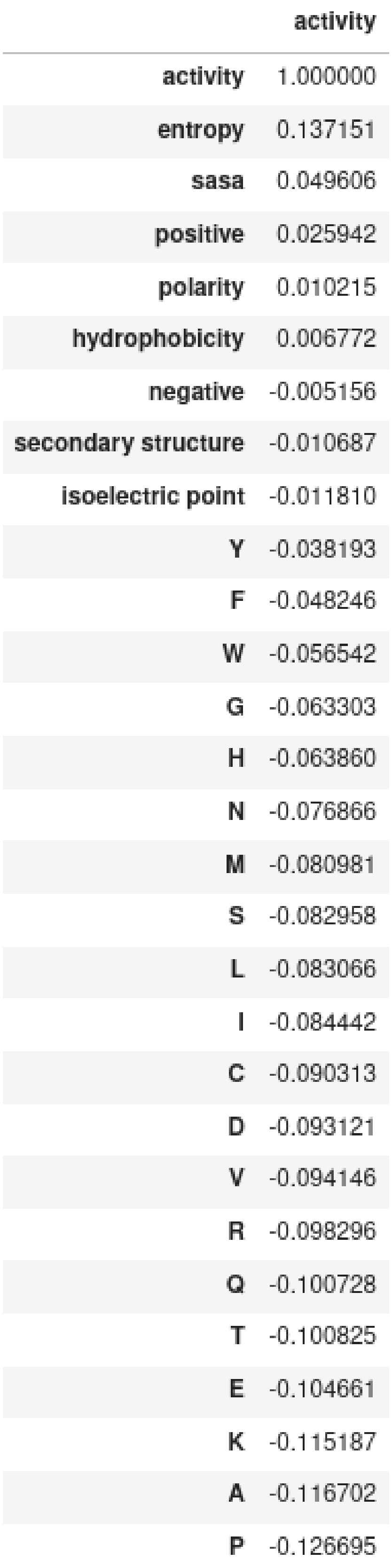
60. Kuglstatter A GMTSVASDBJBM. X-ray crystal structure of JNK2 complexed with the p38alpha inhibitor BIRB796: Insights into the rational design of DFG-out binding MAP kinase inhibitors. Bioorg Med Chem Lett. 2010;

61. Conejo-Garcia A, McDonough MA, Loenarz C, McNeill LA, Hewitson KS, Ge W, et al. Structural basis for binding of cyclic 2-oxoglutarate analogues to factor-inhibiting hypoxia-inducible factor. Bioorg Med Chem Lett. 2010 Oct;20(20):6125–8.

62. Dong C, Huang F, Deng H, Schaffrath C, Spencer JB, O'Hagan D, et al. Crystal structure and mechanism of a bacterial fluorinating enzyme. Nature. 2004 Feb;427(6974):561–5.

PDB codes of training set and test set
2q88 3cm2 1ua4 1a99 2gz2 1upf 1m1b 1a4r 1yqy 2v2h 2jiw 2oxy 5yas 1m0n 2h6t 1wm1 1k1y 1jzs 2vw5 2vt3 1nki 2cbj 1h0a 2o8h 1hlk 3pce 1t7d 2pqc 2aj8 2mas 1olx 2evl 1moq 3eqr 1j36 3e5u 2zcs 2bvd 1gyx 3d7k 2qbu 1s5z 1dy4 2fqt 1ogz 2jkp 2v95 1ajp 2i3h 1duv 2gsu 1uwf 1szd 2hhn 1alw 2gvj 3b4p 1br6 1m2x 1q5k 1k9s 1nw7 2a5b 1gai 1xk9 10gs 1oba 1bai 2v54 3c2r 1t5f 1j4r 1ikt 1hi5 2gyi 2hzl 1drv 1ws1 1ax0 3b2q 1r9l 6std 1bzy 2cht 2pu2 2afw 1nc1 1i5r 1q54 1rql 2fqy 2bt9 2gst 1r1j 1ogd 2q8z 3eeb 1n4k 2i2c 2ews 2r75 1c3x 3ckb 1fzq 2hjb 3coz 1m5w 2i80 1jqy 1rd4 1pbq 1vyg 1lyb 2bfr 3bfu 1njs 3f8f 1dzk 1l83 1nje 2e94 1uho 1k27 2vfk 1rpj 1swr 1q1g 2qm9 2epn 1efy 1tkb 1lrh 2vuk 2i4z 1ez9 2pql 1kc7 2c80 3bxh 1uou 1xk5 1p19 1dqn 1v1j 1grp 1h6h 1e3v 1zhy 3cd7 1b3l 2doo 2amt 2j78 1jcx 3ebl 1q91 1elr 1ew8 1pfu 3cj5 1jlr 6rnt 2hxm 1koj 1mai 5tmp 2gvv 2r0h 2rio 2v8w 3c2u 2aac 1wcq 1ws4 2byr 2pwd 2ha3 1oar 1atl 2qpu 1n4h 1yvm 1ydk 2qrl 1x8j 1qji 2za0 1f4f 1pzp 2csn 1jq8 1ro6 3cke 1m83 1gwv 1ado 1pkx 1fh7 1kjr 2fxv 2rk8 3b3c 1m48 1lbf 1ec9 1n0s 1n51 3cd5 1a94 3brn 2rcn 2vj8 2yz3 1jak 3e5a 1fiv 2d1o 3d52 1nli 1u1w 1qan 1kmy 1tmn 2sim 1uj5 2p4s 2bet 1pa9 1kyv 1fcy 3d0e 2fu8 1nf8 1kdk 1pvn 1ork 1e6q 1ui0 1s89 1qy2 1pgp 1y3p 2vyt 1e2k 2ogy 3czv 1b55 1wur 2d0k 1hee 1z4o 1fao 1xgi 2oi2 1hyo 1ppi 2ewb 1wn6 1ctt 2rcb 1bq4 1m7y 2am4 2q6f 2fxu 2b4l 1pdz 2glp

Table 1. PDB codes of training set and test set.



	activity
activity	1.000000
entropy	0.137151
sasa	0.049606
positive	0.025942
polarity	0.010215
hydrophobicity	0.006772
negative	-0.005156
secondary structure	-0.010687
isoelectric point	-0.011810
Y	-0.038193
F	-0.048246
W	-0.056542
G	-0.063303
H	-0.063860
N	-0.076866
M	-0.080981
S	-0.082958
L	-0.083066
I	-0.084442
C	-0.090313
D	-0.093121
V	-0.094146
R	-0.098296
Q	-0.100728
T	-0.100825
E	-0.104661
K	-0.115187
A	-0.116702
P	-0.126695

Table 2. Correlation coefficients between the dependent variable (i.e. activity) and the independent ones (i.e. the different features). The last 20 rows refer to the PSSM of each residue.

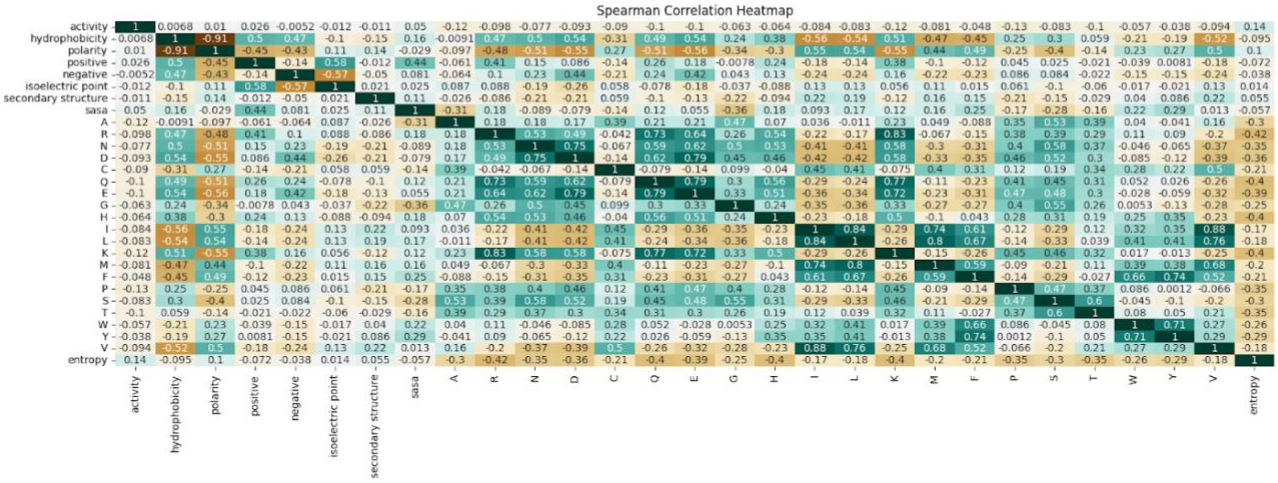


Figure 1. Correlation coefficients between all pairs of variables. Amino acid letters refer to the PSSM of each residue.