Predicting Protein Binding Sites: Machine Learning Approach

Introduction to Python - Structural Bioinformatics

Documentation - Tutorial

Msc in Bioinformatics for Health Sciences

Universitat Pompeu Fabra
Barcelona

Alex Ascunce París
Paula Delgado Manzano
María Chacón Ortiz

INDEX

1. Introduction

Proteins, essential biomolecules composed of amino acids, play diverse and vital roles in living organisms. Each amino acid has its own particular properties that contribute to the protein main function and structure. The folding process is mainly driven by the nucleation of hydrophobic residues, which will tend to shield from the solvent. The final outcome is the protein fold that has a specific function, which may involve the interaction with other molecules.

In this context, proteins have been found to form complexes by either a permanent or transient binding with other proteins or with other molecules such as receptor ligands, enzyme substrates, nucleic acids, ions, and other molecules of small or larger sizes. This binding interaction occurs in the so-called binding site, located on the protein surface. These protein interactions are pivotal for biological and metabolic pathways, which constitute essential roles for sustaining life. Understanding protein-ligand interactions is fundamental in molecular biology and drug discovery. In the same tone, predicting binding sites on protein surfaces is a key step towards understanding their function and designing drugs that target specific sites.

Protein binding sites exhibit several distinguishing features that differentiate them from the surrounding protein surface. They often contain cavities or pockets that can accommodate ligands, providing the necessary space and shape complementarity for selective binding interactions. These sites are enriched in specific amino acid residues with functional groups that facilitate binding interactions, such as hydrogen bonding, electrostatic interactions, and hydrophobic interactions. The arrangement and properties of these residues are crucial for ligand recognition and binding [1]. Moreover, binding sites typically display higher solvent accessibility and flexibility compared to other regions of the protein, allowing them to accommodate ligand binding and undergo conformational changes upon interaction. Residues crucial for ligand recognition and binding are often evolutionarily conserved across homologous proteins, reflecting the functional importance of these residues in maintaining the protein's biological activity [2]. Additionally, binding sites exhibit dynamic properties, such as flexibility, conformational changes, and allosteric regulation, contributing to their versatility in binding different ligands and adapting to various environmental conditions [3]. These structural, physicochemical, and dynamic features are essential for their ability to selectively recognize and bind to their cognate ligands, enabling proteins to carry out their biological functions.

In conclusion, protein binding sites are characterized by their specialized structure, specific amino acid composition, and dynamic properties, enabling them to selectively recognize ligands and perform key biological functions. These fundamental features ensure the efficacy of protein-ligand interactions and underscore the importance of understanding the complexity of these sites in the context of molecular biology and pharmaceutical research [4].

2. Current approaches

Nowadays it is possible to find both experimental and computational methods to study protein interactions. In contrast to labor-intensive experimental methods, these computational methods offer crucial advantages in predicting ligand binding sites. These approaches leverage sequence and structural data, and accuracy of predictions present good results in many cases. They can be classified as 3D structure-based (geometric), template similarity-based, traditional machine learning-based and deep learning-based methods [1].

2.1. Geometric approach

Ligand binding predominantly occurs within cavities or pockets on protein surfaces, as substantial affinity typically arises from sufficiently expansive interfaces. Consequently, the predominant approach for identifying binding sites involves searching for distinct geometric or energetic features within protein structures. This method is typically executed following two main steps. Firstly, spatial geometric analyses are conducted on proteins to identify surface pockets. Alternatively, probes are positioned on the protein surface, followed by the identification of cavities through the estimation of energy potentials between the probes and the surrounding regions [5].

The basic idea behind these prediction methods based on spatial geometry is to locate large or even the largest hollow on the protein structure. Nevertheless, the geometric approach depends on the state of the given protein 3D structure, as the pocket may not exist in the apo state but it may be induced by the interaction between protein and ligand (holo state). An example of this method is MSPocket, which identifies surface pocket regions according to the normal vector directions at the vertices on the surface [6].

2.2. Template-similarity based

This approach mainly focuses on protein homologs. It is based on the idea that there are proteins that share a common ancestor and therefore we can infer characteristics of a problem protein taking into account its homologs. This way it is possible to transfer structural or functional information from the homolog to the query. Template similarity-based binding site prediction methods can be classified into two different approaches:

- Structure template-based methods.
- Sequence template-based methods.

ConSurf is an example of sequence template-based, in which phylogenetic relationships between sequences are taken into account [7]. On the other hand, FINDSITE is an instance of structure template-based, which uses a PROSPECTOR 3 threading algorithm [8].

2.3. Machine learning

There has been continuous computational development finally giving rise to artificial intelligence-related tools that provide high accuracy and performance in numerous fields. Machine learning in the prediction of protein binding sites involves leveraging these computational techniques to predict those regions in a protein of interest in which the binding with other proteins or molecules may happen. In general terms, machine learning approaches in this context first aim to extract features from protein structures and sequences.

Moreover, a sufficiently large dataset would be provided in which we can find both known and not known binding sites. This dataset is required for the training and subsequent testing of the generated model, as the data is split into training and testing sets. Finally, it is possible to predict a given protein binding site (independent from the dataset used to produce the model). It is plausible to implement different machine learning algorithms for this task, such as Support Vector Machine (SVM) or Random Forests, being the latter the chosen one for this project [9].

2.4. Deep learning

Deep learning approach is a further developed machine learning method that is continuously growing. It aims to imitate the functioning of the human brain and human neural networks. Deep learning architectures show these neural networks with multiple layers of interconnected nodes (neurons). In this scenario, a common architecture is Convolutional Neural Networks (CNNs). CNN is also composed of multiple neurons and each of them does a part of the calculation based on a portion of the input, finally giving the corresponding output portion [4]. An example would be DeepBind, that specifically predicts characteristics of DNA or RNA binding proteins [10].

3. Our approach: prediction of protein binding sites

Here we present a machine learning methodology utilizing a Random Forest classifier for prediction of protein binding sites. Our approach involves extracting structural and sequence-based features from protein structure datasets, followed by data preprocessing to enhance classifier compatibility and to avoid potential biases. After training, the classifier can distinguish between binding and non-binding sites using the extracted features. This framework holds potential for enhancing our comprehension of protein-ligand interactions and streamlining drug discovery processes.

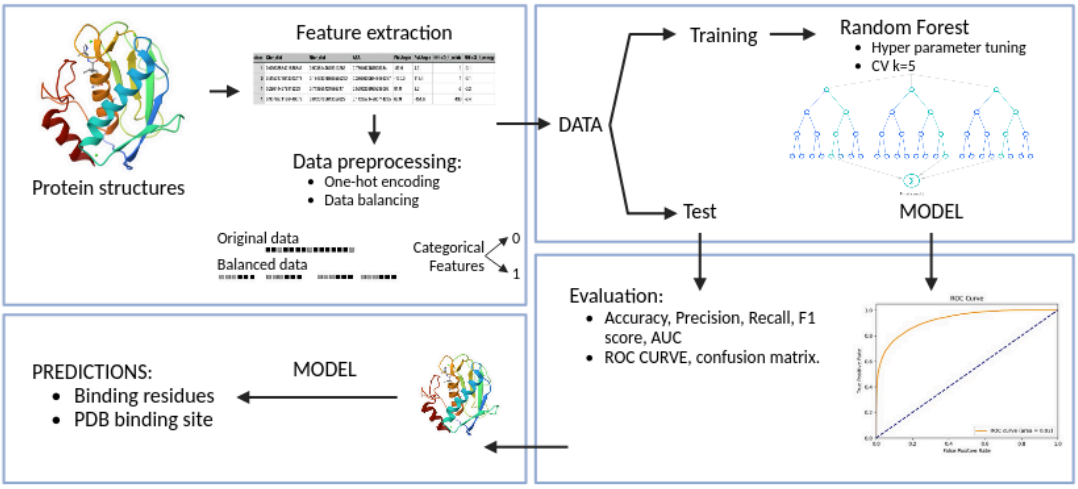


Figure 1. Workflow of the process segmented into four blocks: data acquisition and feature extraction; model generation and training; model evaluation and binding site prediction.

We have chosen this approach over others due to its demonstrated effectiveness in accurately predicting binding sites using both structural and sequence data. The Random Forest classifier offers advantages in handling complex datasets and has shown promising results in various biological prediction tasks. Additionally, our decision is influenced by the success of similar methodologies, such as GRaSP, which employs a graph-based residue neighborhood strategy to predict binding sites in protein interactions from which we have taken several ideas [11].

3.1. Data collection

First, we collected a dataset of protein structures with annotated binding sites comprising 1111 entries randomly selected from COACH and HOLO4k databases obtained from GRASP [11] to obtain a sufficient dataset size to build a robust model avoiding an excessive generation time. Using the provided dataset the model generation time extends approximately for one hour.

For each protein structure file (PDB file), an annotation file (CSV file) was provided, containing information about the corresponding binding sites. Then, various features were extracted from the protein structures to capture relevant information about residuals. These features capture diverse aspects such as spatial distribution, chemical properties, and structural characteristics of residues within the protein. Specifically, features were selected to capture details about the residue's position in the protein sequence, its relative distance from termini, and its surrounding environment. Additionally, structural attributes such as secondary structure elements, accessible surface area, and torsion angles were included to characterize the residue's conformational context. Furthermore, properties related to hydrogen bonding, hydrophobicity, aromaticity, and charge distribution were computed to provide insights into the residue's chemical composition and potential interaction patterns. This comprehensive feature set aimed to capture the multifaceted nature of residues within protein structures, enabling a holistic representation for predictive modeling (Table 1).

Feature	Description
Residue name	Name of each amino acid residue in the protein sequence.
Cter_dist	Fractional distance of the residue from the C-terminus.
Nter_dist	Fractional distance of the residue from the N-terminus.
Secondary structure (SS)	Local conformation of each residue.
Accessible surface area (ASA)	Surface area of each residue that is accessible to solvent.
Phi and Psi angles	Torsion angles describing the residual side-chain orientation.
NH->O_1_relidx	Relative index of the hydrogen bond formed between the nitrogen (N) of a residue and the oxygen (O) of the subsequent residue.
NH->O_1_energy	Energy associated with the hydrogen bond between the nitrogen (N) of a residue and the oxygen (O) of the subsequent residue.
O->NH_1_relidx	Relative index of the hydrogen bond formed between the oxygen (O) of a residue and the nitrogen (N) of the previous residue.
O->NH_1_energy	Energy associated with the hydrogen bond between the oxygen (O) of a residue and the nitrogen (N) of the previous residue.
NH->O_2_relidx	Relative index of the hydrogen bond formed between the nitrogen (N) of a residue and the oxygen (O) of the second subsequent residue.
NH->O_2_energy	Energy associated with the hydrogen bond between the nitrogen (N) of a residue and the oxygen (O) of the second subsequent residue.
DON	Hydrogen bond donor atoms within the residue.
HPB	Hydrophobicity of the residue.
ACP	Hydrogen bond acceptor atoms within the residue.
NEG	Negative charge present within the residue.
POS	Positive charge present within the residue.
ARM	Aromaticity of the residue.

Table 1. Features collected from PDB files for the Random Forest classifier

Given that the six last features (DON, HBP, ACP, NEG, POS, ARM) were computed by aggregating the characteristics of the atoms within each residue, their values would be identical for residues of the same type, providing no additional discriminative information beyond the residue name. Therefore, to enhance the discriminative power of these features, their values were weighted based on the properties of neighboring residues within a 6 Å radius. This distance was established based on the distance criteria of non covalent interactions [12, 13]. This approach allows for a more comprehensive understanding of the local environment surrounding each residue, thereby enriching the contextual information associated with each residue's properties.

3.2. Data preprocessing

After feature extraction, the dataset was balanced to address potential biases arising from the differences in the number of residues inside and outside binding sites. Given that binding sites are typically smaller in size compared to the overall protein structure, ensuring an equal representation of residues from both categories was essential. This balanced approach enabled the model to effectively discern between binding and non-binding regions, thereby improving its predictive accuracy. The extracted features were preprocessed to suit machine learning algorithms, with categorical features (residue name and secondary structure) encoded using one-hot encoding.

3.3. Random Forest

A Random Forest classifier was then trained to predict protein binding sites. The dataset was split into training and test sets (70-30%) for model generation and evaluation respectively. To optimize model performance, the max_features parameter was tuned using a randomized search, employing five-fold cross-validation and using the area under the curve of the ROC curve to analyze the performance.

3.4. Evaluation

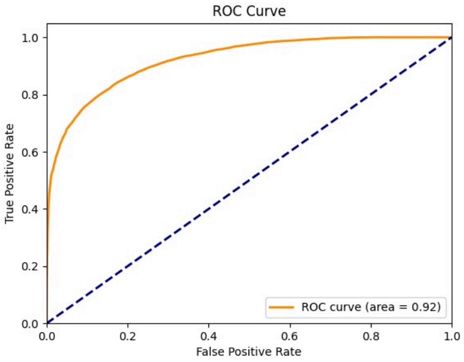
The trained model was then evaluated in the test set using various performance metrics including accuracy, AUC (Area Under the Curve), precision, recall, and F1-score (figure 2A). With an accuracy of 0.83, the model correctly predicts 83% of the instances, indicating a strong overall predictive capability. In terms of precision, the model correctly identifies 82% of the binding sites out of all those predicted to contain binding sites, indicating a relatively low false positive rate. Additionally, a recall of 0.85 implies that the model captures 85% of the actual binding site instances, reflecting its ability to identify a significant portion of binding sites. The F1-score, a harmonic mean of precision and recall, serves as a balanced measure of the model's performance. With an F1-score of 0.84, the model achieves a harmonious balance between precision and recall, indicating reliable predictive performance across the binding site prediction task.

Overall, these metrics suggest that the model exhibits strong predictive capabilities, effectively identifying potential binding sites while minimizing false predictions. However, there may still be opportunities to further improve its performance, particularly in scenarios where precision and recall need to be optimized simultaneously.

A

Metric	Value
Accuracy	0.83
AUC	0.92
Precision	0.82
Recall	0.85
F1-score	0.84

B



C

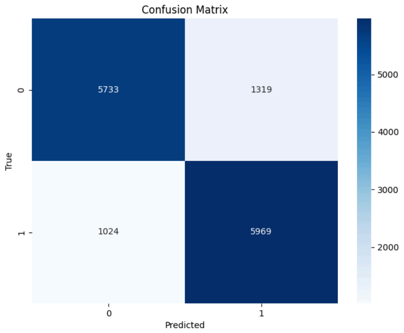


Figure 2. Evaluation Metrics for Model Performance (A) Table displaying the quality measures of the model. (B) Receiver operating characteristic (ROC) of the model (C) Confusion matrix showing the counts of true positive, true negative, false positive, and false negative predictions.

The ROC curve illustrates the trade-off between sensitivity (true positive rate) and specificity (false positive rate) across different threshold values, showcasing the model's ability to effectively differentiate between positive and negative instances. The area under the ROC curve (AUC) provides a quantitative measure of the model's overall performance, with an AUC of 0.92 suggesting that the model can effectively distinguish between sequences containing binding sites and those that do not, demonstrating its robust discriminative power.

The confusion matrix provides a detailed breakdown of the model's classification results, offering insights into its accuracy, precision, recall, and specificity. This allows for a comprehensive assessment of the model's performance across different classes. Overall, the combined analysis of the ROC curve and the confusion matrix indicates that the model demonstrates robust predictive capabilities, accurately classifying instances while effectively managing false positives and false negatives.

These findings provide valuable insights into the model's strengths and weaknesses, facilitating informed decision-making and potential enhancements to optimize its performance further.

3.5. Prediction

Finally, the trained model can be used to predict binding sites in new protein structures, facilitating the identification of potential drug targets or functional regions in proteins. To ensure the accuracy of predicted binding residues, a threshold probability of 0.8 is set. While this threshold helps reduce false positives, it might also result in an increase in false negatives by potentially missing some binding site residues. Despite this possibility, maintaining a stringent threshold in the prediction could be crucial for ensuring the precision and reliability of the identified binding residues.

4. Running the program

This section aims to provide a meaningful guide for installing and using our package with the necessary scripts and data implicated in the prediction of binding sites. A few examples will be also offered to show the overall performance of our method.

4.1. Program Requirements

This program has been implemented using Python3 as the primary programming language. The **installation of DSSP** (Define Secondary Structure of Proteins) software is mandatory for the well-functioning of the program, as it is used for the extraction of the secondary structure features related to the protein of interest. To install DSSP, use the following command, which relies on the conda package manager:

```
conda install -c salilab dssp
```

The bs_predictor.tar.gz file includes:

- **pdb_dir**: database of protein structures in pdb format.
- **labels_dir**: database of csv files with the residues labeled as binding and non binding of the corresponding pdb files provided.
- **model.rf:** model trained with pdb_dir and labels_dir data.
- **Python scripts:** accuracy.py, model_generator.py, main.py, extract_features.py, random_forest.py.
- **atom_types.csv:** file needed to compute properties of the residues within the proteins.
- **requirements.txt:** python packages needed to run the program.
- **examples:** predictions made in two example proteins described in section 5.
- **Documentation and tutorial file.**
- **README:** instructions to run the utilities of the program.

All python packages needed for the different scripts are specified in the requirements.txt file and can be installed using:

```
pip install -r requirements.txt
```

4.2. Main functionality - Prediction of binding sites

To predict the residues conforming binding sites in a given protein, the following command needs to be written:

```
python3     main.py     [-h,--help]     -model     {path_to/model_file}     -pdb
{path_to/pdb_file_or_folder} -out {output_file}
```

Options:

-h, -- help: displays usage information for the script.
-model: specifying the path to the desired model. The package already contains a model, which performs an acceptable accuracy for any given problem protein.
-pdb: the path to the pdb file or folder with pdb files of the protein(s) of interest, whose binding sites will be predicted.
-out: path to the output folder.

The output will consist on three main files:

- **pdbID_predictions.csv**: contains the list of residues and the prediction 0 (non binding site) and 1 (binding site) and the probability of the prediction of the given protein(s).
- **pdbID_binding.csv**: contains the list of residues predicted to be in the binding site of the given protein.
- **pdbID_binding.pdb**: a pdb file which has the corresponding coordinates of the residues from the predicted binding site, which can be visualized in a molecular graphics software (Chimera, PyMOL).

4.3. Create a new Random Forest Classification model

To create a new model with a different dataset, it is possible to use the model_generator.py script with the following command-line options:

```
python3   model_generator.py   [-h,--help]   -pdb   {path_to/pdb_dir}   -labels
{path_to/labels_dir} -out {output_file} -test {test_file}
```

Options:

-h, – help: displays usage information for the script.
-pdb: Specifies the path to the directory containing the protein structure files (*.pdb) you want to include in your training/test dataset.
-labels: Specifies the path to the directory containing the label files (*.csv) you want to include in your training/test dataset.The CSV files should be structured as follows, with each column separated by commas labeled accordingly:

- ○ Column A (res_name): residue unique identifier written as PDBid_chain_resname_resindex.
- ○ Column B (prediction): could be empty, it is just a matter of format of the label files of the dataset used during model generation.
- ○ Column C (class): binary identifier for the presence or absence of the residual in the binding site.

-out: Specifies the output file where the generated model will be saved.
-test: Specifies the test file with the features extracted for evaluating the model.

4.4. Model evaluation

The evaluation of our Random Forest model involves assessing its performance on a test dataset and potentially displaying metrics including:

- **Accuracy** indicates the proportion of correctly classified instances out of the total in the test dataset. It provides a general overview of the model's predictive power. HIgher values indicate higher predictive power of the model.
- **AUC (Area Under the Curve):** The AUC represents the area under the ROC curve.
- **Precision** measures the proportion of correctly predicted positive cases out of all cases predicted as positive. High precision values imply that the model has a low false positive rate, while low precision values suggest a high false positive rate.
- **Recall** (sensitivity), calculates the proportion of true positive cases correctly identified by the model. High values indicate that the model can capture a large portion of positive cases, while low values suggest that some positive cases are being missed.
- **F1-score** is the harmonic mean of precision and recall, providing a balance between them, which is useful when there is an uneven class distribution. Higher F1-score values indicate better balance between precision and recall, signifying a more reliable model performance.

In addition to these metrics, it is possible to display:

- **ROC Curve:** The ROC curve visually represents the trade-off between the True Positive Rate (sensitivity) and the False Positive Rate (1 - specificity) for different threshold values.
- **Confusion Matrix:** The confusion matrix is a tabular representation of actual versus predicted class labels. It provides a more detailed breakdown of the model's performance by showing the counts of true positive, false positive, true negative, and false negative predictions.

These additional visualizations, such as the ROC curve and confusion matrix, complement the numerical metrics and provide deeper insights into the model's performance and behavior.

To evaluate the model, these are the provided command-line options:

```
python3 accuracy.py [-h,--help] -model {path_to/model} -test {path_to/test.csv}
-metrics -roc {roc_file.png} -pred {predictions_file.csv}
```

Options:

-h, --help: displays usage information for the script.
-model: file path to the trained machine learning model.
-test: file path to the test dataset.
-metrics: flag to display accuracy, AUC, precision, recall, F1-score, and confusion matrix.
-roc: file path to save the ROC curve plot.
-pred: file path to save the predictions.

By running the evaluation script with these options, an analysis of the model performance and assessment of its effectiveness in predicting ligand binding sites can be obtained.

5. Examples

5.1. Crystal Structure of the Mouse-Muscle Adenylosuccinate Synthetase bound to GTP (1LOO)

This pdb file presents the crystal structure of the mouse-muscle adenylosuccinate synthetase (basic isozyme) bound to GTP [14]. The complex reveals crucial insights into the ligand recognition and conformational changes associated with catalysis in this enzyme.

The mouse-muscle adenylosuccinate synthetase, a vital enzyme in purine nucleotide metabolism, exhibits distinct properties compared to its bacterial counterpart. Unlike the Escherichia coli enzyme, the mouse basic isozyme demonstrates a unique active site conformation and ligand recognition pattern.

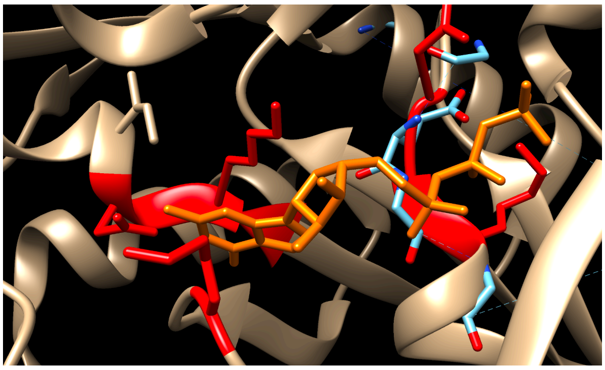
As described in literature, this protein binding site contains a GTP loop (residues 448–452), a guanine recognition element (residues 362−365) and a P-loop (residues 38−46). Moreover, residue K448 plays a role in the binding to GTP [14]. In figure 3 it is possible to visualize a comparison between the predicted residues with our method and the actual residues.

Predicted binding sites with our method are the following:

- GLY at position 42.
- GLU at position 44.
- GLY at position 45.
- GLY at position 47.
- HIS at position 105.
- GLY at position 168.
- VAL at position 277.
- MET at position 345.

The correct predicted residues are the ones belonging to the P-loop.

Figure 3. Image showing binding site of interest. The GTP molecule is colored in orange. Main residues implicated in the binding site are colored in red. Residues predicted with our method are shown in blue, most of them belonging to the P-loop.



Note that, when looking for a binding site in a protein, predictions of each residue with its corresponding probability will be obtained. The default threshold for filtering these predicted residues is 0.8 and therefore, only those residues above this threshold will be included in the PDBid_binding.pdb file for visualization requirements.

5.2. X-ray structure of the E58A mutant of Ribonuclease T1 complexed with 3'-guanosine monophosphate (1LOV)

This PDB file presents the X-ray crystal structure of the E58A mutant of Ribonuclease T1 (RNase T1) in complex with 3'-guanosine monophosphate (3'-GMP) [15]. RNases are enzymes responsible for catalyzing the cleavage of phosphodiester bonds in RNA.

High-resolution structures of RNases with 3'-mononucleotide substrates reveal the accommodation of RNA's nucleophilic 2'-OH group in a binding pocket containing the catalytic base and a charged hydrogen bond donor.

As described in the literature, the binding sites for the protein 1LOV include residues 38, 40-46, 77, 92, 98, and 100. Among these, residues 40-42 are located in an antisense parallel beta sheet, and residues 40 and 92 are reported as active sites [15].

The predicted binding sites obtained from our method are as follows:

- TYR at position 38.
- TYR at position 42.
- ASN at position 43.
- ASN at position 44.
- TYR at position 45.
- GLU at position 46.
- ARG at position 77.
- ASN at position 98.
- PHE at position 100.

As evident from the results, nearly all predicted binding sites align accurately with the known binding residues, indicating a high degree of prediction accuracy. Specifically, residues 40, 41, and 92 were identified with a probability of approximately 0.7, which, although slightly below the predetermined threshold of 0.8, still demonstrates a relatively high probability. No instances of incorrect predictions were observed within the final results. (Figure 4) allows for a comparison between the actual binding sites and the predicted binding sites.

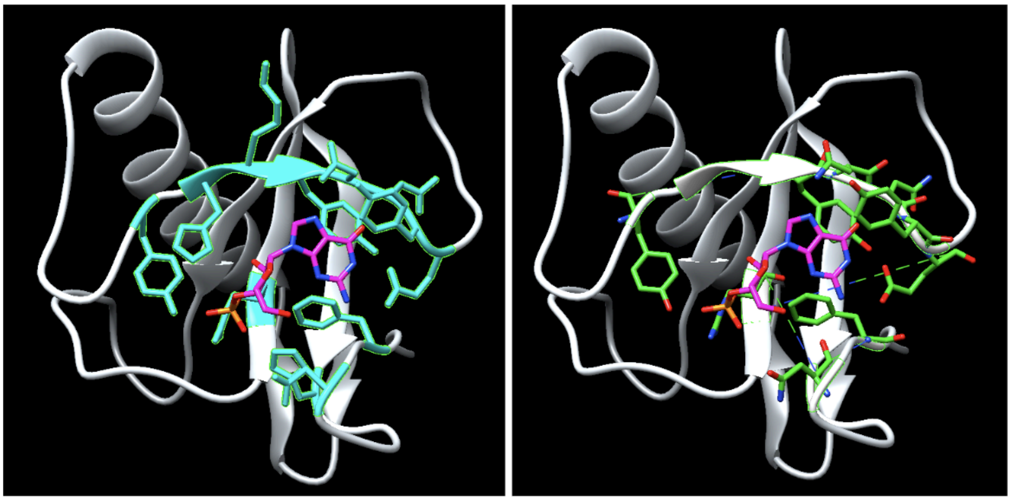


Figure 4. Chimera visualization. The 1LOV PDB structure is highlighted in white, while the 3'-GMP ligand is colored pink. In the left image, the actual binding sites are marked in cyan. In the right image, the predicted binding sites are highlighted in green.

Bibliography

1. Zhao J, Cao Y, Zhang L. Exploring the computational methods for protein-ligand binding site prediction. Comput Struct Biotechnol J. 2020 Feb 17;18:417-426. doi: 10.1016/j.csbj.2020.02.008. PMID: 32140203; PMCID: PMC7049599.
2. Kandel J, Tayara H, Chong KT. PUResNet: prediction of protein-ligand binding sites using deep residual neural network. J Cheminform. 2021 Sep 8;13(1):65. doi: 10.1186/s13321-021-00547-7. PMID: 34496970; PMCID: PMC8424938.
3. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Improving detection of protein-ligand binding sites with 3D segmentation. Sci Rep. 2020 Mar 19;10(1):5035. doi: 10.1038/s41598-020-61860-z. PMID: 32193447; PMCID: PMC7081267.
4. Konc J, Janežič D. Protein binding sites for drug design. Biophys Rev. 2022 Dec 9;14(6):1413-1421. doi: 10.1007/s12551-022-01028-3. PMID: 36532870; PMCID: PMC9734416.
5. Sotriffer C., Klebe G. Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. *Il Farmaco.* 2002;57:243–251.
6. Zhu H., Pisabarro M.T. MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets. *Bioinformatics.* 2010;27:351–358.
7. Glaser F. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics.* 2003;19:163–164.
8. Brylinski M., Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci.* 2008;105:129–134.
9. Jayatilake SMDAC, Ganegoda GU. Involvement of Machine Learning Tools in Healthcare Decision Making. J Healthc Eng. 2021 Jan 27;2021:6679512. doi: 10.1155/2021/6679512. PMID: 33575021; PMCID: PMC7857908.
10. Alipanahi B., Delong A., Weirauch M.T., Frey B.J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat Biotechnol.* 2015;33:831.
11. Santana CA, Silveira SA, Moraes JPA, Izidoro SC, de Melo-Minardi RC, Ribeiro AJM, Tyzack JD, Borkakoti N, Thornton JM. GRaSP: a graph-based residue neighborhood strategy to predict binding sites. Bioinformatics. 2020 Dec 30;36(Suppl_2):i726-i734. doi: 10.1093/bioinformatics/btaa805. PMID: 33381849.
12. Fassio AV, Martins PM, Guimarães SDS, Junior SSA, Ribeiro VS, de Melo-Minardi RC, Silveira SA. Vermont: a multi-perspective visual interactive platform for mutational analysis. BMC Bioinformatics. 2017 Sep 13;18(Suppl 10):403. doi: 10.1186/s12859-017-1789-3. PMID: 28929973; PMCID: PMC5606220.
13. Fassio AV, Santos LH, Silveira SA, Ferreira RS, de Melo-Minardi RC. nAPOLI: A Graph-Based Strategy to Detect and Visualize Conserved Protein-Ligand Interactions in Large-Scale. IEEE/ACM Trans Comput Biol Bioinform. 2020 Jul-Aug;17(4):1317-1328. doi: 10.1109/TCBB.2019.2892099. Epub 2019 Jan 10. PMID: 30629512.

14. Iancu CV, Borza T, Fromm HJ, Honzatko RB. IMP, GTP, and 6-phosphoryl-IMP complexes of recombinant mouse muscle adenylosuccinate synthetase. J Biol Chem. 2002 Jul 26;277(30):26779-87. doi: 10.1074/jbc.M203730200. Epub 2002 May 9. PMID: 12004071.
15. Mignon P, Steyaert J, Loris R, Geerlings P, Loverix S. A nucleophile activation dyad in ribonucleases. A combined X-ray crystallographic/ab initio quantum chemical study. J Biol Chem. 2002 Sep 27;277(39):36770-4. doi: 10.1074/jbc.M206461200. Epub 2002 Jul 16. PMID: 12122018.