Pocket Seeker

Binding Site Predictor

Theoretical Report

Marc Arreaza

Marta Pérez

A report submitted in partial fulfilment of the requirements of
the University of Pompeu Fabra for the master's degree of
Bioinformatics in Health Sciences in *Structural Bioinformatics and Introduction to Python courses*

April 22, 2025

Contents

Chapter 1

Theoretical background

The identification of protein binding sites is a cornerstone of the drug discovery field. In molecular biology, a ligand refers to any molecule that binds to a specific site on a target protein, typically known as the binding site. These sites—also referred to as binding pockets, active sites, or cavities—are often concave regions or holes on the protein surface that accommodate small molecules. Accurate prediction of these sites can significantly accelerate drug discovery by facilitating virtual screening and guiding molecular docking.

Traditional computational approaches to binding site prediction fall into three main categories: geometry-based methods, template-based methods, and energy-based methods. However, recent developments have seen the successful application of machine learning (ML) to this task, offering increased flexibility, scalability, and predictive power by learning complex patterns from molecular data.

As we know, the input features are very important for ML methods to achieve good prediction power. Generally, the existing methods mostly consider sequence, structure. and network features in their predictions. However, recent studies suggest that incorporating categories may significanly enhance model performance. These categories include residue interface features, concavity characteristics, and coevolutionary information. For the first one, it has been widely proven that the residues that are in the interface have more probabilities of being involved in a cavity related to binding processes (Ahmad and Sarai (2005)). In contrast concavity features consist in the geometry and flexibility of protein surfaces. These are biologically relevant because functional regions (e.g., binding sites, allosteric sites) often correspond to low-frequency collective motions, and the residues involved typically reside in hinge regions critical for conformational changes. Conversely, high-frequency motions are associated with geometric irregularities and residues important for structural stability. On the other hand, for sequence evolution, residue conservation has been used to infer functional sites in proteins. The residue coevolution provides information on residue–residue interactions, and generally the residues with strong coevolutionary levels are essential for maintaining 3D structure and biological function. Now, this information has been extensively applied in protein spatial structure prediction as well as binding site prediction.

Given the importance of these features, integrating residue interaction potentials, surface concavity, and coevolution signals may significantly improve the identification of binding sites.

Moreover, in scenarios where only limited labeled data is available, chosing an optimal feature subset becomes essential to prevent overfitting and enhance model generalizability.

Popular feature selection techniques include Minimum Redundancy Maximum Relevance, XGBoost, and Random Forest (RF)-based selection. In this work, we adopt RF for its balance between interpretability and performance.

The flowchart of PocketSeeker approach is illustrated in Figure 1.1.

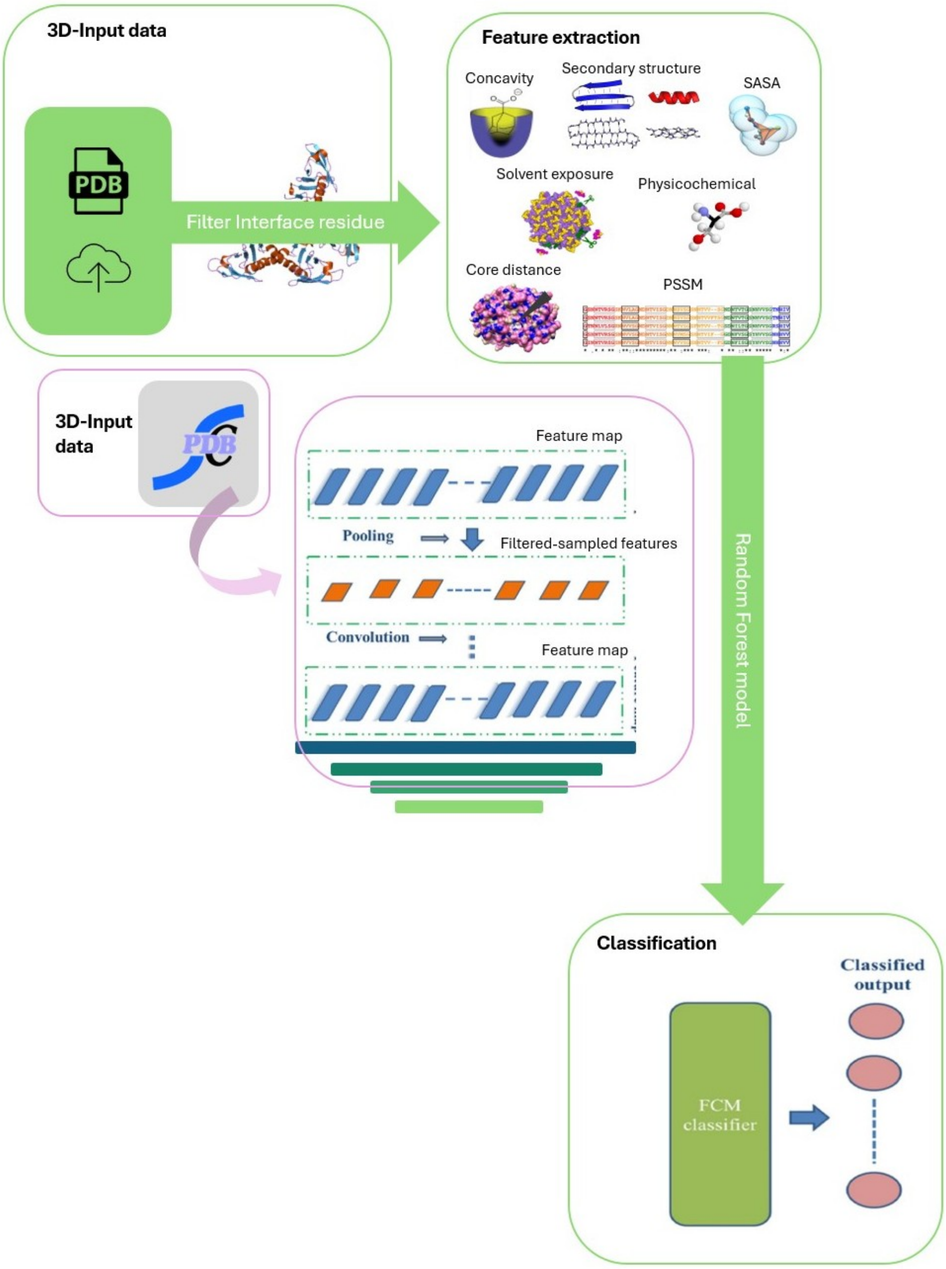

Figure 1.1: Flowchart of the PocketSeeker approach. Steps from feature extraction and selection to model training and binding site prediction are detailed.

Here, we developed PocketSeeker, a standalone Python-based tool designed to predict binding sites. The program uses an enhanced Random Forest (RF) model, improved through the integration of SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance and improve generalization.

A detailed description the model architecture, training process, and evaluation metrics is provided in the Model Performance report, while supporting information can be found in the Supplementary Information report.

Chapter 2

Explanation of the Features

PocketSeeker integrates various features that are important for the identification and characterization of ligand-binding sites in proteins, based on previous studies (Krivák and Hoksza (2018); Carbery et al. (2024); Liu et al. (2023); Tao et al. (2024)). PocketSeeker includes the following features:

- Coevolution
- Secondary Structure
- Accessible Surface Area (ASA)
- Physicochemical Properties
- Solvent Exposure
- PSSM (Position Specific Scoring Matrix)
- Concavity
- Distance to Core

Each of these features provides valuable information that helps PocketSeeker accurately predict the ligand-binding sites in proteins.

These features are computed via automated Python scripts that parse PDB structures and extract the different features. Only standard amino acids are considered, using a three-letter to one-letter mapping. The results are saved into a CSV file for downstream integration and analysis.

1 Coevolution

Coevolution refers to the correlated evolutionary changes of amino acid residues that are functionally or structurally linked within a protein (Anishchenko et al. (2017)). Residues that interact—either through physical contact or allosteric effects—often exhibit coevolution to maintain the protein's stability, structural integrity, or function. For example, if a mutation alters a residue at a binding interface, compensatory changes in interacting residues may occur to preserve the interaction (Ju et al. (2021)).

To analyze coevolutionary patterns, a Multiple Sequence Alignment (MSA) of homologous protein sequences is first constructed. From this MSA, statistical methods can be applied to quantify coevolution (Morcos et al. (2013)). Two widely used metrics are Mutual Information (MI) and Direct Information (DI).

Mutual Information (MI)

Mutual Information measures the dependency or correlation between two positions in the MSA. It is sensitive to both direct and indirect correlations. (Simonetti et al. (2013))

$$\mathrm{MI}(i,j) = \sum_{x,y} P_{ij}(x,y) \log\left(\frac{P_{ij}(x,y)}{P_i(x)P_j(y)}\right)$$

where:

- $P_{ij}(x,y)$ is the joint probability of observing amino acids x and y at positions i and j, respectively.
- $P_i(x)$ and $P_j(y)$ are the marginal probabilities of observing amino acids x at position i, and y at position j.

While MI captures correlated positions, it cannot distinguish between direct and indirect interactions (e.g., due to transitive effects in the MSA).

Direct Information (DI)

Direct Information is derived from global statistical models like Direct Coupling Analysis (DCA), which aim to disentangle direct from indirect coevolutionary couplings. DI uses a maximum entropy model to estimate the direct interaction strength between pairs of residues. (Morcos et al. (2011))

$$\mathrm{DI}(i,j) = \sum_{x,y} P_{ij}^{\mathrm{dir}}(x,y) \log\left(\frac{P_{ij}^{\mathrm{dir}}(x,y)}{P_i(x)P_j(y)}\right)$$

where:

- $P_{ij}^{\mathrm{dir}}(x,y)$ is the direct coupling probability between residues x and y at positions i and j, estimated from the DCA model.

Unlike MI, DI is better at identifying physically interacting residues, making it particularly useful for contact prediction in protein structure modeling and analysis.

2 Secondary Structure

The secondary structure of a protein refers to the local folded conformations adopted by segments of the polypeptide chain. PocketSeeker takes into account both categorical and geometric descriptors to characterize the structure around each residue.

DSSP (Define Secondary Structure of Proteins) is the standard method for PocketSeeker used to assign secondary structure annotations to a protein structure (Kabsch and Sander (1983)). DSSP utilizes the atomic coordinates of a structure to assign the secondary codes, which are:

- **H**: Alpha helix (α-helix)
- **B**: Isolated beta bridge
- **E**: Extended beta strand (part of a beta sheet)
- **G**: 3_{10} helix
- **I**: Pi helix
- **T**: Turn
- **S**: Bend (irregular hydrogen-bonded structure)
- -: No defined secondary structure (coil or loop)

These annotations describe the local folding pattern and hydrogen bonding tendencies of each residue in the protein.

Backbone Torsion Angles and Geometric Features

To further characterize the conformation of the protein backbone, PocketSeeker also includes geometric descriptors:

- **Phi angle (ϕ)**: The torsion angle between the atoms C(n-1)–N(n)–C$_\alpha$(n)–C(n). It describes the rotation around the N–C$_\alpha$ bond.
- **Psi angle (ψ)**: The torsion angle between the atoms N(n)–C$_\alpha$(n)–C(n)–N(n+1). It describes the rotation around the C$_\alpha$–C bond.
- **Theta (θ)**: A geometric angle defined between three consecutive C$_\alpha$ atoms: $C\alpha_{i-1} \rightarrow C\alpha_i \rightarrow C\alpha_{i+1}$. It provides insight into the local bend of the backbone.
- **Tau (τ)**: A dihedral angle involving four C$_\alpha$ atoms: $C\alpha_{i-2} \rightarrow C\alpha_{i-1} \rightarrow C\alpha_i \rightarrow C\alpha_{i+1}$. It describes the torsion of the backbone across four sequential residues.

Together, the secondary structure types and geometric parameters (ϕ, ψ, θ, and τ) provide a comprehensive view of the local conformation and folding behavior of each residue of the protein. These features can be important for identifying regions with structural characteristics favorable for ligand binding. (Zvelebil et al. (1987))

3 Accessible Surface Area (ASA)

The Accessible Surface Area (ASA)—also referred to as Solvent-Accessible Surface Area (SASA)—measures the surface area of a protein residue or atom that is accessible to a solvent molecule, typically water. SASA can give us information to identify residues that are exposed on the surface and potentially involved in molecular interactions, such as ligand binding.

In PocketSeeker, ASA is computed at both the residue-level and the atom-level, providing a detailed view of solvent accessibility.

Residue-Level ASA

Three methods are used to estimate the solvent exposure of each residue:

- **DSSP-based Method**: Uses the DSSP package to calculate residue exposure based on empirical definitions of secondary structure and surface area from hydrogen bonding patterns, following the method of Sander and Rost. (Rost and Sander (1993))
- **Shrake-Rupley Method**: This geometric method simulates the rolling of a spherical probe (typically 1.4 Å radius to represent water) over the protein surface, calculating the solvent-accessible area for each atom and summing it to obtain residue-level values. (Shrake and Rupley (1973)).
- **Total SASA**: The total solvent-accessible surface area of all atoms in a residue.

Atom-Level SASA Decomposition

For a more detailed analysis, the total solvent-accessible area of each residue is broken down into categories based on atomic properties:

- **Main Chain SASA**: The contribution from backbone atoms (N, C_α, C, and O), which form the protein's structural scaffold.
- **Side Chain SASA**: The contribution from the side chain atoms—those unique to each amino acid and often involved in specific interactions.
- **Polar SASA**: The portion of SASA due to polar atoms (e.g., O, N), typically capable of hydrogen bonding and interactions with the aqueous environment.
- **Non-Polar SASA**: The SASA contributed by non-polar atoms (mainly carbon), usually hydrophobic and often buried in the protein core.

By integrating both residue- and atom-level ASA features, PocketSeeker captures the spatial exposure patterns essential for detecting ligand-binding sites. Highly exposed, polar residues near concave regions are often strong indicators of potential interaction points.

4 Physicochemical Properties of Amino Acids

The physicochemical features of amino acid residues provide essential insight into their behavior and potential interactions within a protein structure (Nikam et al. (2023)). PocketSeeker integrates a rich set of physicochemical descriptors that have been shown to correlate well with protein interface characteristics, especially in the context of ligand binding.

Each residue is represented by the following ten physicochemical features:

- **Number of Atoms**: The total count of atoms in the residue. Larger residues often have more complex interaction patterns and may contribute more steric bulk.
- **Number of Electrostatic Charges**: Represents the total count of formal charges carried by a residue. Charged residues, such as Aspartic acid (Asp), Glutamic acid (Glu), Arginine (Arg), and Lysine (Lys), play key roles in salt bridge formation and electrostatic interactions.

- **Hydrophobicity (Kyte-Doolittle Scale)** (Kyte and Doolittle (1982)): A quantitative measure of a residue's hydrophobic or hydrophilic nature. Hydrophobic residues (e.g., Leu, Ile, Val) are usually buried, while hydrophilic residues (e.g., Ser, Thr, Asp) are surface-exposed and more likely to participate in ligand interactions.

- **Polarity (Grantham Scale)**(Grantham (1974)): Indicates the polarity of a residue based on its side chain's chemical properties. Polar residues tend to participate in hydrogen bonds and are often found at protein surfaces or in binding pockets.

- **Isoelectric Point (pI)**: The pH at which a residue has no net electrical charge. This feature informs about the charge state of residues under different environmental pH conditions, potentially affecting binding affinity and protein stability.

- **Molecular Mass**: The molecular weight of the residue, reflecting its size and impact on steric interactions. Mass also contributes to overall residue dynamics.

- **Propensity (Structural Preference)**: A measure of how likely a residue is to be found in certain structural environments (e.g., helices, sheets, turns). This information helps to understand residue function and structural compatibility.

- **Expected Number of Contacts (within 14 Å Sphere)**: An estimate of how many neighboring atoms are within a 14 Å radius of the residue. This captures residue compactness and provides information on the local packing density and potential interaction hotspots.

- **Center of Mass**: The spatial average of atomic positions weighted by atomic mass. This feature reflects how mass is distributed within the residue, providing spatial context to its position in the protein.

These physicochemical descriptors collectively help PocketSeeker identify and characterize ligand-binding sites by modeling residue behavior, interaction tendencies, and spatial properties.

5 Solvent Exposure

Solvent exposure features describe how buried or exposed each amino acid residue is within the 3D structure of a protein. Here, we focus on two key solvent exposure descriptors:

- **Residue Depth**: The average distance of all atoms in a residue to the closest point on the solvent-accessible surface. Residue depth helps to identify how deeply buried a residue is within the protein structure. Buried residues often participate in the protein core stabilization, whereas surface residues are more likely to be involved in interactions.

  - The calculation is performed using the `ResidueDepth` module from `Bio.PDB`, which internally calls Michel Sanner's MSMS algorithm (Mih et al. (2018)). This tool computes a triangulated solvent-accessible surface using a rolling probe method.

- **Contact Number (CN)**: The number of neighboring Cα atoms located within a defined radius (typically 15 Å) of a residue's Cα atom. CN reflects the local packing density around each residue.

  - A higher CN value suggests the residue is in a densely packed region, likely in the protein core.
  - A lower CN indicates a more exposed residue, which may be accessible to solvent or potential binding partners.
  - CN is computed using the `ExposureCN` class from `Bio.PDB`'s `HSExposure` module.

6 Position-Specific Scoring Matrix (PSSM)

The Position-Specific Scoring Matrix (PSSM) is a probabilistic representation of residue conservation in a protein sequence. It is widely used in structural bioinformatics and protein function prediction (Tao et al. (2024); Nikam et al. (2023)). PSSMs are generated using PSI-BLAST, which aligns homologous sequences to calculate amino acid substitution probabilities at each position.

- Rows represent specific positions in the protein sequence.
- Columns represent the 20 standard amino acids.
- Each cell contains a log-odds score that reflects the likelihood of observing a particular amino acid at that position in related sequences.
- Highly conserved residues received high scores, suggesting structural or functional importance.

This information is particularly valuable for identifying functional regions, such as active sites or ligand-binding pockets.

Example PSSM:

Position	A	R	N	D	C	...
1	2	-3	0	1	-1	...
2	-1	1	2	0	-2	...
...	...	...	...	...	...	...

7 Concavity

Concavity describes the local 3D geometric environment of a residue in the protein structure. It reflects how buried or recessed a residue is relative to its surroundings, which can be indicative of pocket or cavity regions.

- Concavity is computed based on the spatial coordinates of atoms in the protein.
- Residues in deeper pockets exhibit higher concavity values.
- Tools that calculate surface accessibility and surface curvature (such as CASTp or concavity-specific metrics) are used.

This feature is especially useful for identifying potential ligand-binding residues, as ligands typically bind in concave or pocket-shaped regions.

8 Distance to Core

This distance measures how far a residue is from the structural center of the protein. It provides information about whether a residue is on the surface or buried deep within the protein.

- The core is often defined as the geometric center (centroid) of all atoms in the protein or of all Cα atoms.
- The distance is calculated between the Cα atom of each residue and the protein centroid.
- Residues with smaller distances are more likely to be buried and involved in structural stability, while larger distances suggest surface exposure.

This measure helps assess the likelihood of a residue being solvent-accessible or participating in core structural interactions. **?**

References

Ahmad, S. and Sarai, A. (2005), 'Pssm-based prediction of dna binding sites in proteins', *BMC bioinformatics* **6**, 1–6.

Anishchenko, I., Ovchinnikov, S., Kamisetty, H. and Baker, D. (2017), 'Origins of coevolution between residues distant in protein 3d structures', *Proceedings of the National Academy of Sciences* **114**(34), 9122–9127.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.1702664114*

Carbery, A., Buttenschoen, M., Skyner, R. et al. (2024), 'Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures', *Journal of Cheminformatics* **16**, 32.
URL: *https://doi.org/10.1186/s13321-024-00821-4*

Grantham, R. (1974), 'Amino acid difference formula to help explain protein evolution', *Science* **185**(4154), 862–864.

Ju, F., Zhu, J., Shao, B., Kong, L., Liu, T.-Y., Zheng, W.-M. and Bu, D. (2021), 'Copulanet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction', *Nature Communications* **12**(1), 2535.
URL: *https://doi.org/10.1038/s41467-021-22869-8*

Kabsch, W. and Sander, C. (1983), 'Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features', *Biopolymers* **22**, 2577–2637.

Kriván, R. and Hoksza, D. (2018), 'P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure', *Journal of Cheminformatics* **10**, 39.
URL: *https://doi.org/10.1186/s13321-018-0285-8*

Kyte, J. and Doolittle, R. F. (1982), 'A simple method for displaying the hydropathic character of a protein', *Journal of Molecular Biology* **157**(1), 105–132.

Liu, Y., Li, P., Tu, S. and Xu, L. (2023), 'Refinepocket: An attention-enhanced and mask-guided deep learning approach for protein binding site prediction', *IEEE/ACM Transactions on Computational Biology and Bioinformatics* pp. 1–8.
URL: *https://doi.org/10.1109/tcbb.2023.3265640*

Mih, N., Brunk, E., Chen, K., Catoiu, E., Sastry, A., Kavvas, E., Monk, J. M., Zhang, Z. and Palsson, B. O. (2018), 'ssbio: a python framework for structural systems biology', *Bioinformatics* **34**(12), 2155–2157.
URL: *https://doi.org/10.1093/bioinformatics/bty077*

Morcos, F., Jana, B., Hwa, T. and Onuchic, J. N. (2013), 'Coevolutionary signals across protein lineages help capture multiple protein conformations', *Proceedings of the National Academy of Sciences* **110**(51), 20533–20538.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.1315625110*

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T. and Weigt, M. (2011), 'Direct-coupling analysis of residue coevolution captures native contacts across many protein families', *Proceedings of the National Academy of Sciences* **108**(49), E1293–E1301.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.1111471108*

Nikam, R., Yugandhar, K. and Gromiha, M. M. (2023), 'Deepbsrpred: deep learning-based binding site residue prediction for proteins', *Amino Acids* **55**(10), 1305–1316.

Rost, B. and Sander, C. (1993), 'Prediction of protein secondary structure at better than 70% accuracy', *Journal of Molecular Biology* **232**(2), 584–599.

Shrake, A. and Rupley, J. A. (1973), 'Environment and exposure to solvent of protein atoms. lysozyme and insulin', *Journal of Molecular Biology* **79**(2), 351–371.

Simonetti, F. L., Teppa, E., Chernomoretz, A., Nielsen, M. and Marino Buslje, C. (2013), 'Mistic: Mutual information server to infer coevolution', *Nucleic Acids Research* **41**(Web Server issue), W8–14. PMC3692073.
URL: *http://mistic.leloir.org.ar*

Tao, L., Zhou, T., Wu, Z., Hu, F., Yang, S., Kong, X. and Li, C. (2024), 'Espdhot: An effective machine learning-based approach for predicting protein–dna interaction hotspots', *Journal of Chemical Information and Modeling* **64**(8), 3548–3557. PMID: 38587997.
URL: *https://doi.org/10.1021/acs.jcim.3c02011*

Zvelebil, M. J., Barton, G. J., Taylor, W. R. and Sternberg, M. J. (1987), 'Prediction of protein secondary structure and active sites using the alignment of homologous sequences', *Journal of Molecular Biology* **195**(4), 957–961.
URL: *https://www.sciencedirect.com/science/article/pii/0022283687905018*