Universitat Pompeu Fabra Barcelona

Final Report Project SBI and Introduction to Python

Regina Rodriguez Durant Reyes
Liam McBride

April 22, 2025

INDEX

Introduction

In this project, a Python-based pipeline was developed to visualize potential ligand binding sites on protein structures using a geometry-driven approach. The final visualization was rendered in PyMOL, providing an intuitive representation of the predicted pockets on the protein surface.

Objective

The core objective was to create a structure-only method to identify and rank potential ligand binding pockets from protein PDB files, leveraging purely geometric features such as surface depth and curvature.

Methods

The project was organized into a modular pipeline under the folder “ligand-site-predictor-geo”, with each Python script responsible for a specific step:

- main.py – The main entry point. It coordinates all the scripts to run the full pipeline from start to finish:

- PDBparser.py : Acts as the entry point of the pipeline, ingesting the input .pdb file and extracting the atomic coordinates necessary for subsequent geometric analysis

- surface_analysis.py – Perform surface geometry analysis on the protein structure. Using convex hulls, it calculates point-wise geometric features such as surface depth, curvature, and spatial position. These descriptors form the foundation for identifying potential ligand-binding pockets.

- pocket_detection.py – Detects and clusters deep surface points using the DBSSCAN algorithm, this density-based clustering enables the identification of concave surface regions with limited accessibility, geometric traits indicating ligand-binding pockets.

- scoring.py – Scores detected pocket based on a combination of heuristic criteria: mean and maximum depth, pocket compactness, enclosure, and cluster size. This scoring step allows the pipeline to rank pockets by their structural plausibility as ligand-binding sites.

- visualization.py – Generates a PyMOL script to visualize the top pockets.Pockets are visualized as color-coded spheres mapped onto the protein surface, enabling intuitive spatial inspection and comparison of predicted sites within a 3D structural context.

The program operated through the following sequential steps:

1. **PDB Parsing**
    - A parser ingests the .pdb file and extracts relevant atomic coordinates for surface analysis.
    - Output: parsed_json
2. **Surface Analysis**
    - Calculates point-wise geometric properties: depth from convex hull, curvature, and surface coordinates.
    - **Output:** The computed features are stored in *surface_json and surface_plot*
3. **Pocket Detection**
    - Filters surface points by depth and clusters them using **DBSCAN**.
    - **Output:** *pockets_json* contains detailed cluster data, including coordinates and geometric statistics.
4. **Scoring**
    - Applies heuristics such as mean/max depth, cluster size, and compactness.
    - **Output:** *scored_json* with ranked pockets.
5. **Visualization**
    - Converts the top-ranked pockets into **PyMOL**-compatible scripts using color-coded spheres.
    - **Output:** *pockets_visualization.pml*
6. **Pipeline**
    - Integrates all components, manages user input/output, and runs the full process in sequence.

Key Features of the Geometry-Based Method

- Geometry-Only Approach: Focused on surface depth and density-based clustering.
- Educational Focus: Lightweight dependencies and a clear, modular design.
- Visualization-Ready: Output easily rendered in PyMOL for 3D exploration.

Discussion and Results

The use of PyMOL for visualizing geometric features on protein surfaces significantly enhances the interpretability of binding site prediction results. In this project, geometric descriptors such as surface curvature and depth were central to detecting and ranking potential ligand binding pockets. The final visual output rendered as color-coded spheres in PyMOL offered an intuitive 3D representation of pocket locations and spatial properties, which is essential for evaluating the physical plausibility of these sites [1].

LIGSITE, developed by Hendlich *et al.,* 1997, presents an early and influential approach to automatic binding site detection on protein surfaces. Like this pipeline, it uses a grid-based method to identify potential ligand-binding pockets by scanning the protein's 3D structure. However, while LIGSITE relies on a fixed cubic grid to locate concave regions suggestive of pockets, our pipeline employs a more data-driven approach involving Delaunay triangulation, t-SNE dimensionality reduction, and DBSCAN clustering to capture complex surface depressions and rank them based on geometric features [2].

The approach implemented in this pipeline shares key conceptual elements with PocketPicker (Weisel *et al*., 2007), a method designed to identify and describe protein binding pockets using shape-based descriptors and buriedness calculations. PocketPicker's use of a grid-based buriedness metric to detect surface concavities aligns with our use of accessibility and geometric clustering (DBSCAN) to identify potential ligand-binding pockets. Both approaches emphasize the role of shape and depth in determining pocket relevance.

While PocketPicker generates a correlation vector for comparing pocket geometries, this script computes heuristic scores using volume, depth, enclosure, and curvature to rank predicted pockets. This reflects a simplified, yet complementary, strategy to PocketPicker's descriptor-based comparisons. Furthermore, our integration of PCA/t-SNE for visualization and PyMOL scripts for structural inspection adds an interactive layer to the analysis, supporting exploratory and comparative binding site assessment [3].

The method used in this project, also parallels aspects of Q-SiteFinder (Laurie & Jackson, 2005), an energy-based approach that predicts ligand-binding sites by mapping energetically favorable positions for a van der Waals probe. Q-SiteFinder clusters these low-energy probe positions and ranks clusters by total interaction energy, effectively highlighting physicochemically attractive pockets for ligands. In contrast, this pipeline relies on geometry-driven heuristics clustering of dense surface point regions with limited accessibility

and subsequently refines pocket detection via t-SNE-based dimensionality reduction and scoring heuristics.[4].

The script used in this project, offers a flexible framework for ligand-binding site prediction by integrating geometric analysis, unsupervised learning, and interactive visualization. Unlike traditional grid-based methods such as LIGSITE and PocketPicker, our approach leverages Delaunay triangulation, t-SNE dimensionality reduction, and DBSCAN clustering to capture intricate surface depressions and buried regions with high spatial resolution. By combining depth, curvature, volume, and enclosure into a scoring heuristic, we provide an interpretable and scalable method to rank predicted pockets. Furthermore, the seamless integration with PyMOL visualization rendering pockets as color-coded spheres, enhances the intuitive evaluation of spatial and geometric plausibility, making our tool not only accurate but also highly accessible for exploratory structural analysis. This blend of computational rigor and visual interactivity makes the program a powerful platform for binding site discovery and drug design applications.

Conclusion

This project successfully showcased a modular, geometry-based pipeline for predicting and visualizing ligand binding sites on protein structures. By relying solely on structural features such as surface depth, curvature, and enclosure without the need for ligand-bound complexes, this method offers a fast, interpretable, and visually rich approach for pocket detection. The integration with PyMOL enables intuitive 3D inspection, making the pipeline not only a valuable educational tool but also a practical framework for early-stage drug discovery, comparative structural analysis, and future method development in structural bioinformatics.

References

[1] Seeliger, D., & De Groot, B. L. (2010). Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *Journal Of Computer-Aided Molecular Design*, *24*(5), 417-422. https://doi.org/10.1007/s10822-010-9352-6

[2] Hendlich, M., Rippmann, F., & Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal Of Molecular Graphics And Modelling*, *15*(6), 359-363. https://doi.org/10.1016/s1093-3263(98)00002-3

[3] Weisel, M., Proschak, E. & Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central Journal* 1, 7 (2007). https://doi.org/10.1186/1752-153X-1-7

[4] *Bioinformatics*, Volume 21, Issue 9, May 2005, Pages 1908–1916, https://doi.org/10.1093/bioinformatics/bti315