Prediction and visualization of cryptic binding regions

Lukáš Polák

Faculty of Mathematics and Physics, Charles University

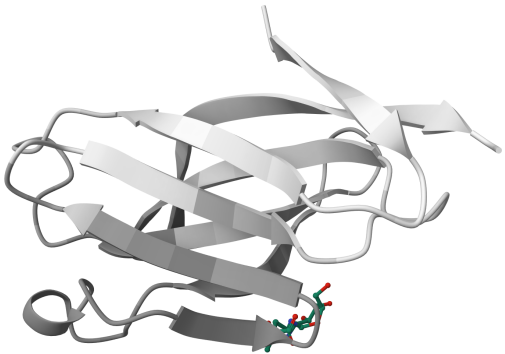
Prague, September 9, 2025

Outline

Proteins, Amino Acids, and Protein Structures

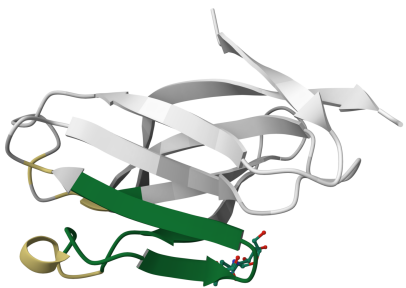
...TVFQGVAGQSLQVSCPYDSMKHW ...

- **Proteins**: essential biomolecules, diverse functions
- Composed of **amino acid (residue)** chains
- **Primary structure**: amino acid sequence
- **Tertiary structure**: 3D folding, functional form



Binding Sites

- **Binding sites**: regions on proteins where ligands (e.g., substrates, inhibitors) bind
- **Cryptic binding sites**: hidden or transient sites, not apparent in static structures
- **Apo/Holo conformations**: different states of a protein (apo: unbound, holo: bound)
- Crucial for **protein function**, regulation, and interaction with other molecules
- Essential for **drug design** and discovery



Cryptic Binding Residues Prediction: CryptoBench

- A benchmark dataset developed at FMP CUNI for evaluating cryptic binding site prediction methods
- Provides a diverse set of protein structures
- Includes a model for predicting individual cryptic binding **residues** (not sites)

Outline

Objectives

Clustering

Cluster **residue-level** cryptic predictions (CryptoBench) into contiguous **cryptic binding sites**.

Software (CryptoShow)

Build a client–server web app to submit structures, run residue predictions, cluster them into sites, and visualize results. Additionally, provide animations using AHoJ[1].

[1] AHoJ is a tool developed at FMP CUNI for finding apo-holo protein pair analogs.

Data Processing Pipeline

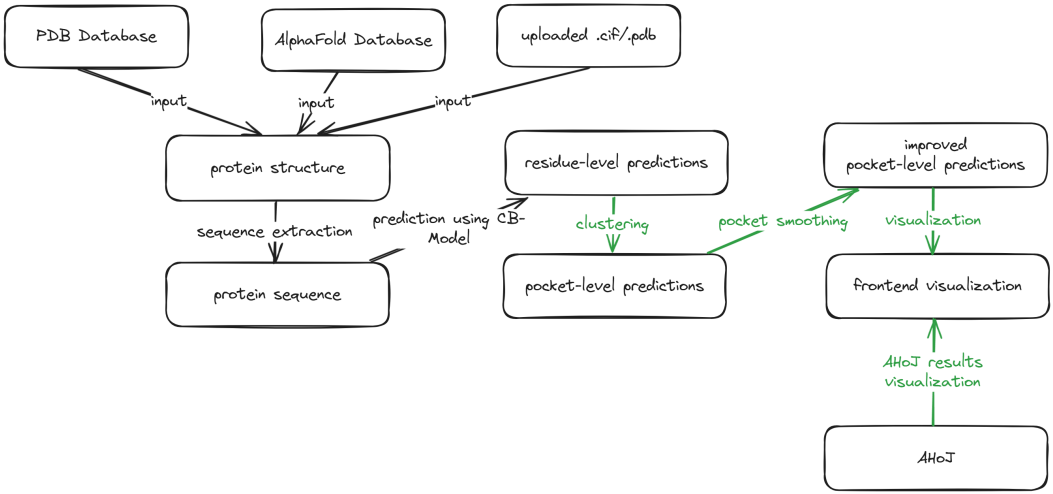


Figure: Data processing pipeline. Green arrows indicate newly introduced techniques.

Clustering

Approach 1

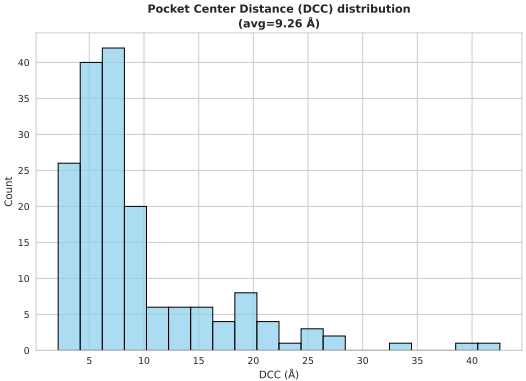
- DBSCAN algorithm
- Cluster residues within 5 Å
- Based only on residues exceeding a probability threshold

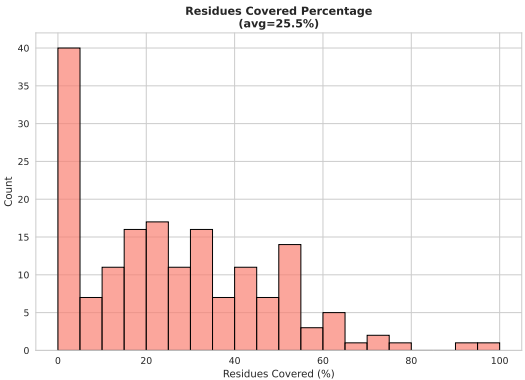
Approach 2

- DBSCAN algorithm + smoothing model
- Improve clustering by adding residues with lower probabilities

Evaluation on the CryptoBench dataset: Approach 1

"Simple" clustering is not enough...





Approach 2: Smoothing Model Training

Let's train another model!

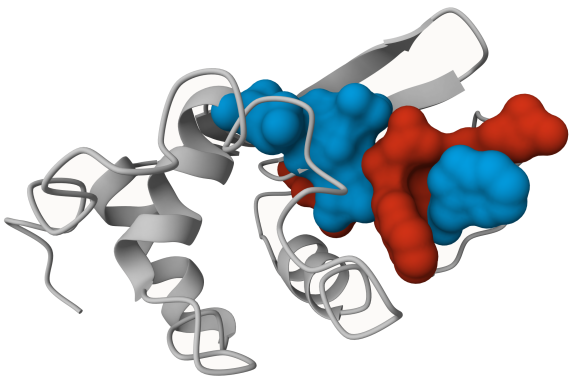
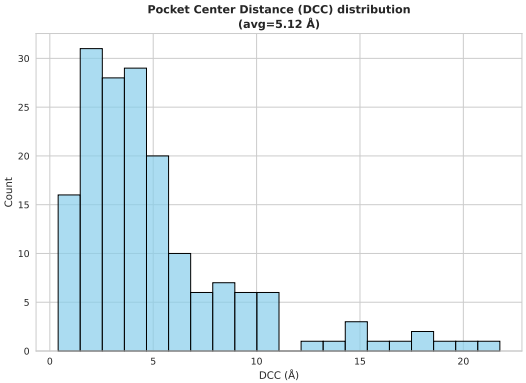
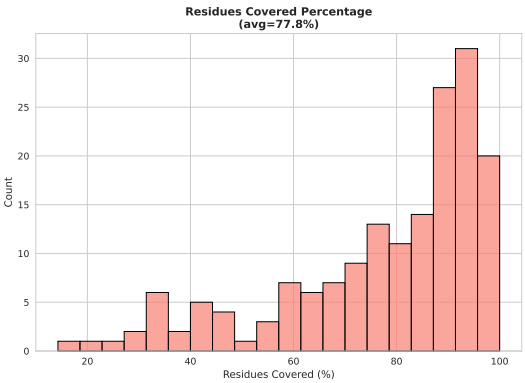


Figure: Results. Blue residues: CryptoBench, red residues: smoothing model.

Evaluation on the CryptoBench dataset: Approach 2

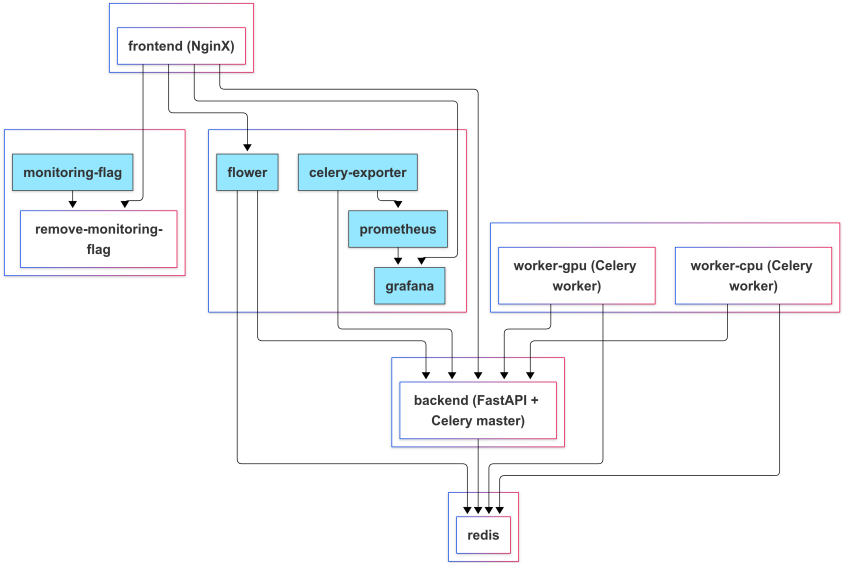
. . . and it seems like the new model works!





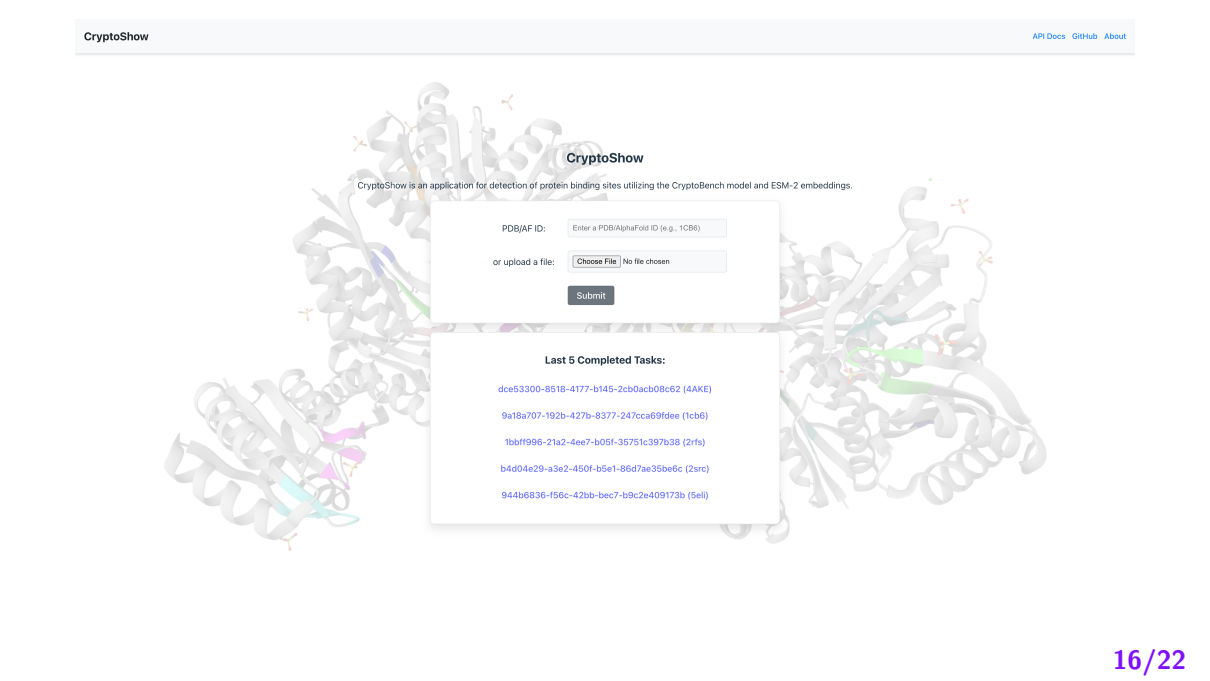
Outline

Architecture



Technology Stack

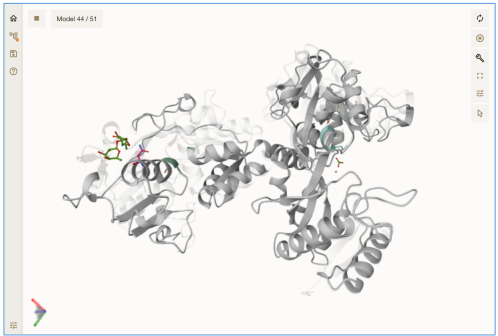
- **Backend**: Python, notable libraries and tools: Celery (asynchronous tasks in Python), FastAPI, Flower, PyTorch, BioPython, Biotite, Scikit-Learn, MDAnalysis, Gemmi, Redis, uv
- **Frontend**: TypeScript, notable libraries, tools, and frameworks: React, NginX, Mol*, Bun, Vite
- **DevOps**: Docker (Docker Compose), GitHub Actions


CryptoShow
API Docs GitHub About
CryptoShow
CryptoShow is an application for detection of protein binding sites utilizing the CryptoBench model and ESM-2 embeddings.
PDB/AF ID:
Enter a PDB/AlphaFold ID (e.g., 1CB6)
or upload a file:
Choose File No file chosen
Submit
Last 5 Completed Tasks:
dce53300-8518-4177-b145-2cb0acb08c62 (4AKE)
9a18a707-192b-427b-8377-247cca69fdee (1cb6)
1bbff996-21a2-4ee7-b05f-35751c397b38 (2rfs)
b4d04e29-a3e2-450f-b5e1-86d7ae35be6c (2src)
944b6836-f56c-42bb-bec7-b9c2e409173b (5eli)

CryptoShow

API Docs GitHub About

3D Structure Viewer: 1cb6



Polymer Representation: Cartoon

Pocket Representation: Cartoon

Reset camera | Remove superposition

To control the animation, use the button next to the "Model" label in the 3D viewer.

The animation shows the transformation of the AHoJ structure (shown in white) to the query structure (shown with transparency). Both the pockets and the ligands are kept for clarity.

Pocket 1

Score: 0.860

Predictions: 0.785, 0.807, 0.927, 0.925, 0.858

Residue IDs: A_1395, A_1396, A_1465, A_1466, A_1467

PyMOL visualization: `select s, 1cb6 and ( (chain A and resi 1395+1396+1465+1466+1467) )`

Job Completed — AHoJ Job ID: AWW86 (note that the results table is scrollable)

Structure	OG Chains	Type	RMSD (Å) ↑	SASA (Å²)	Chains	Ligands	Animation
1lfh	A	HOLO	0.35	45.09	A	CL	Play
1lfi	A	HOLO	3.42	36.07	A	CO3, CU	Loaded
1bka	A	HOLO	3.44	72.43	A	FE, OXL	Play
1fck	A	HOLO	3.46	42.06	A	CO3, CE	Play
P02788	A	AlphaFold	3.48	47.11	A	N/A	Play

Pocket 2

Score: 0.718

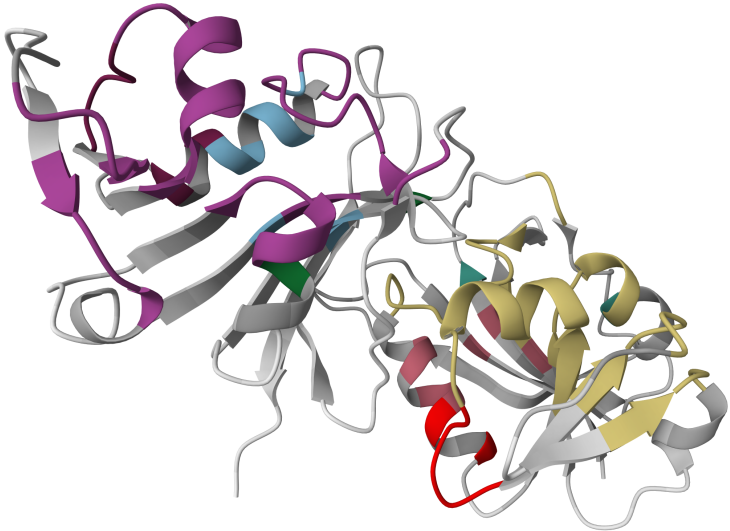
Predictions: 0.787, 0.604, 0.512, 0.731, 0.893, 0.914, 0.848, 0.594, 0.578

Residue IDs: A_1061, A_1117, A_1118, A_1120, A_1121, A_1122, A_1123, A_1124, A_1210

PyMOL visualization: `select s, 1cb6 and ( (chain A and resi 1061+1117+1118+1120+1121+1122+1123+1124+1210) )`

Run AHoJ

Download Results



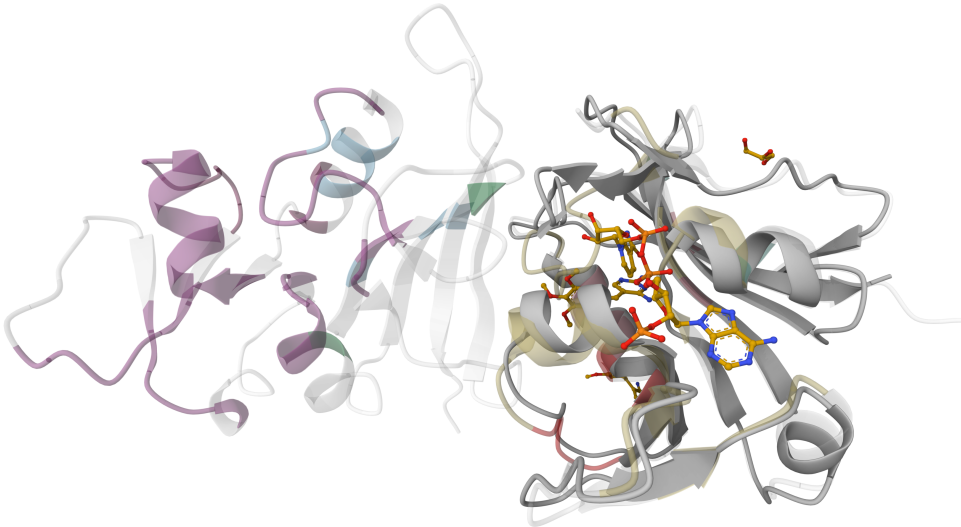


Figure: Animation of the 2W9T protein structure. Video link

Questions?

Response to Reviewer Feedback

Text

- AFAIK no other data sources available
- Minor text issues acknowledged
- Residue distance and clustering algorithm selection slightly discussed in Section 2.4

Software

- UX improvements
- Backend with docstrings vs. Frontend without them
- MVC in React components fixed
- GitHub PR

The End