Pocket Seeker

Binding Site Predictor

Supplementary Information 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

Chapter 1

Dataset Description and Benchmarking

PocketSeeker is a machine learning model based on a random forest algorithm, trained on a curated dataset of 153 protein-ligand complex structures derived from the sc-PDB database. The dataset used is listed in Table 1.1. The PDB IDs of the protein structures in the dataset were obtained from the Protein Data Bank (PDB).

Table 1.1: Additional protein–DNA complexes with their PDB chain identifiers. These structures are used in PocketSeeker's analysis pipeline.

1a2b_1	1ffu_2	1fkg_1	1fm7_2	1fp6_1	1fw6_2	1g16_3	1g4k_2	1g6o_2	1hmp_1
1a2n_1	1ffy_1	1fkh_1	1fm8_1	1fpq_1	1fwk_1	1g17_2	1g4o_1	1g7u_1	1hn2_1
1a4i_1	1fg6_1	1fki_2	1fm9_1	1fpy_6	1fwm_2	1g1a_4	1g4p_1	1g7v_1	1hna_1
1a4l_2	1fg8_1	1fkj_1	1fm9_2	1fq1_1	1fwy_1	1g1d_1	1g4s_2	1g8j_3	1hnc_2
1a4r_1	1fgc_1	1fkl_1	1fmb_1	1fr8_2	1fx1_1	1g1l_1	1g4t_1	1g8o_1	1hnd_1
1a7k_4	1fgi_1	1fko_1	1fmj_1	1frb_1	1fx9_1	1g1y_1	1g5q_2	1g93_1	1i2l_1
1a7x_1	1fgx_1	1fkp_1	1fml_2	1fro_3	1fxf_1	1g27_3	1g5s_1	1g9a_1	1i3k_2
1a8g_1	1fi7_1	1fkw_1	1fmw_1	1ftl_2	1fxo_6	1g28_2	1g63_7	1g9b_1	
1a8k_1	1fi9_1	1fkx_1	1fnb_1	1fue_1	1fxs_1	1g2a_2	1g67_1	1g9c_1	
1a8p_1	1fin_2	1fl2_1	1fnc_1	1fuy_1	1fxu_1	1g2k_1	1g67_2	1g9d_1	
1fej_1	1fjx_1	1flm_2	1fnd_1	1fv0_1	1fxv_1	1g2v_6	1g69_1	1g9r_1	
1fel_1	1fk8_1	1fls_1	1fnd_2	1fvo_2	1fy7_1	1g35_1	1g69_2	1ga8_1	
1fem_1	1fk9_1	1fm1_1	1fnn_2	1fvp_1	1fyf_1	1g38_1	1g6c_1	1gad_2	
1ff0_1	1fkb_1	1fm4_2	1foa_1	1fvt_1	1g05_1	1g3l_5	1g6c_2	1hku_1	
1fff_1	1fkd_1	1fm6_2	1fp1_1	1fvv_1	1g0n_1	1g3m_1	1g6k_3	1hld_2	
1ffi_1	1fkf_1	1fm6_3	1fp2_1	1fw0_1	1g0r_10	1g4j_1	1g6n_1	1hlk_1	
1ffo_1	1flm_1	1fb9_1							

Chapter 2

Features Description

To effectively characterize the structural and physicochemical properties of ligand-binding sites, PocketSeeker utilizes a comprehensive set of 48 features derived from each residue. These features are selected to capture relevant biological and geometric information necessary for accurate prediction. They are grouped into eight categories based on their biological role and computational derivation: Coevolution, Secondary Structure, Accessible Surface Area (ASA), Physicochemical Properties, Solvent Exposure, PSSM (Position Specific Scoring Matrix), Concavity, and Distance to Core.

Table 2.1: The total 131 features used in this study, grouped by category.

Feature Symbol	Description	Category
Na	Number of atoms	Physicochemical
Nec	Number of electrostatic charges	Physicochemical
pI	Isoelectric point	Physicochemical
Mass	Molecular mass	Physicochemical
Enc	Expected Number of Contacts	Physicochemical
Hidrofobicity	Hydrophobicity	Physicochemical
Polarity	Polarity	Physicochemical
Center_of_mass_x	Center of mass (x-axis)	Physicochemical
Center_of_mass_y	Center of mass (y-axis)	Physicochemical
Center_of_mass_z	Center of mass (z-axis)	Physicochemical
A	Probability of Alanine	PSSM
R	Probability of Arginine	PSSM
N	Probability of Asparagine	PSSM
D	Probability of Aspartic acid	PSSM
C	Probability of Cysteine	PSSM
Q	Probability of Glutamine	PSSM
E	Probability of Glutamic acid	PSSM
G	Probability of Glycine	PSSM
H	Probability of Histidine	PSSM
I	Probability of Isoleucine	PSSM
L	Probability of Leucine	PSSM
K	Probability of Lysine	PSSM
M	Probability of Methionine	PSSM
F	Probability of Phenylalanine	PSSM
P	Probability of Proline	PSSM
S	Probability of Serine	PSSM
T	Probability of Threonine	PSSM
W	Probability of Tryptophan	PSSM
Y	Probability of Tyrosine	PSSM
V	Probability of Valine	PSSM
MI	Mutual information score	Coevolution
DI	Direct information score	Coevolution
SASA	Solvent-accessible surface area (Shrake-Rupley method)	ASA
Total_SASA	Total SASA of all atoms in the residue	ASA
Main-chain_SASA	SASA contributed by backbone atoms (N, C_α, C, O)	ASA
Side-chain_SASA	SASA from side chain atoms unique to each amino acid	ASA
Polar_SASA	SASA from polar atoms (e.g., O, N)	ASA
ASA	Relative solvent accessibility (DSSP-based method)	ASA
Non-polar_SASA	SASA from non-polar atoms (mainly carbon)	ASA
SS	Secondary structure (e.g., helix, sheet, coil)	Secondary Structure
Phi	Phi torsion angle	Secondary Structure
Psi	Psi torsion angle	Secondary Structure
Theta(i-1⇒i+1)	Local backbone angle	Secondary Structure
Tau(i-2⇒i+2)	Extended backbone angle	Secondary Structure
Residue_Depth	Residue depth from surface	Solvent exposure
CN	Contact number	Solvent exposure
Distance_to_Core	Distance to Core	Distance to Core
Concavity	Concavity	Concavity