**PDB data analysis**

<https://colab.research.google.com/drive/1_utRJwHa9YKO_mBq9UoMeMFwFoSJQNRj>

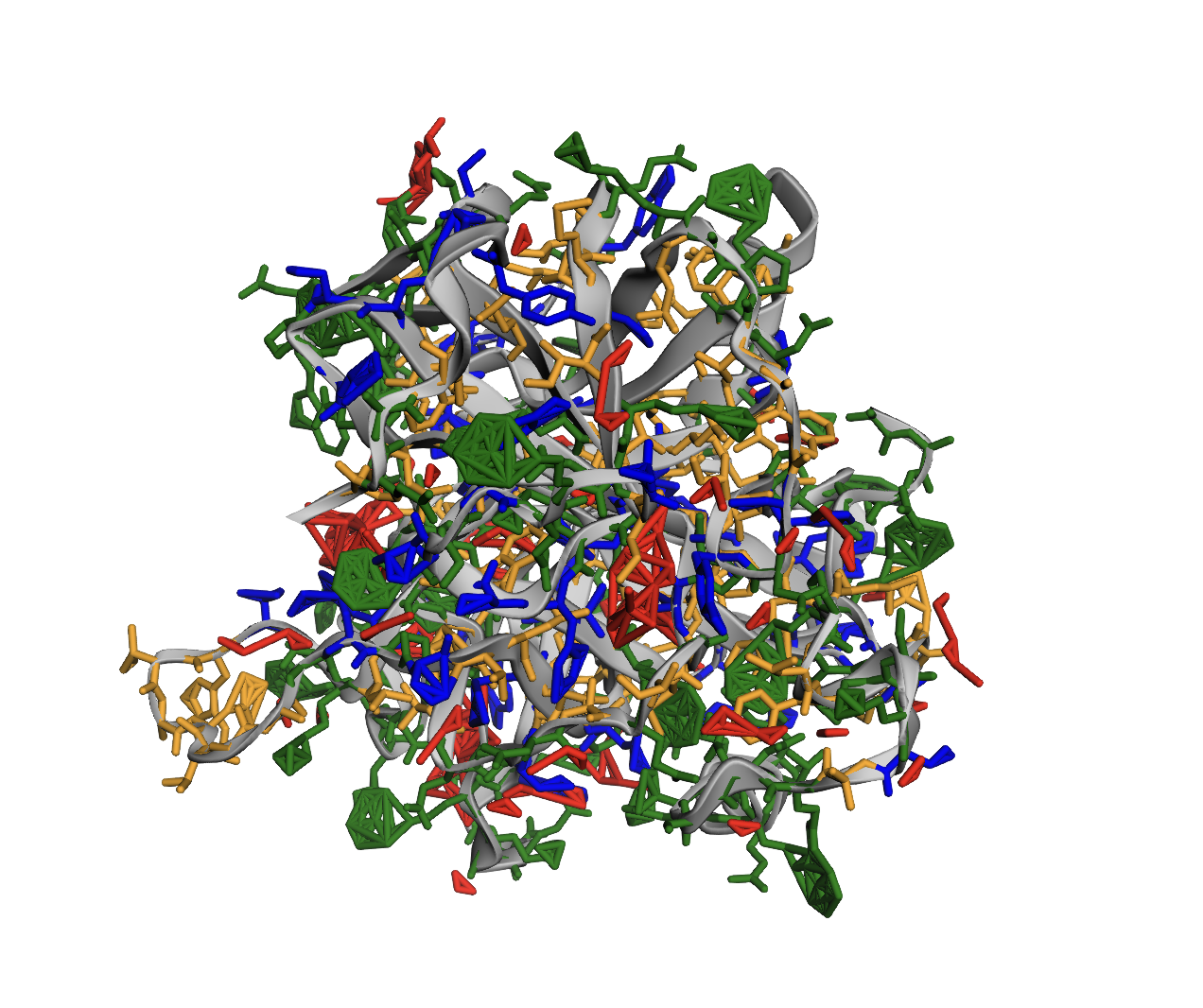
We performed a comprehensive analysis of GPCR-ligand complexes by downloading selected PDB files from the RCSB Protein Data Bank and parsing them using Biopython. It identifies protein and ligand residues within each structure by distinguishing standard amino acids from hetero-residues, excluding water molecules. Ligands are extracted along with their chain information and visualized using py3Dmol. The protein structure is rendered as a cartoon, while ligands are highlighted as red sticks.

To identify the binding pocket, the code uses a 5-angstrom distance cutoff to search for protein residues near ligand atoms with the help of Biopython’s NeighborSearch. These pocket residues are further classified based on their properties: hydrophobic, polar, or others. The visualization highlights these residues using color-coded sticks (orange for hydrophobic, blue for polar, green for others).

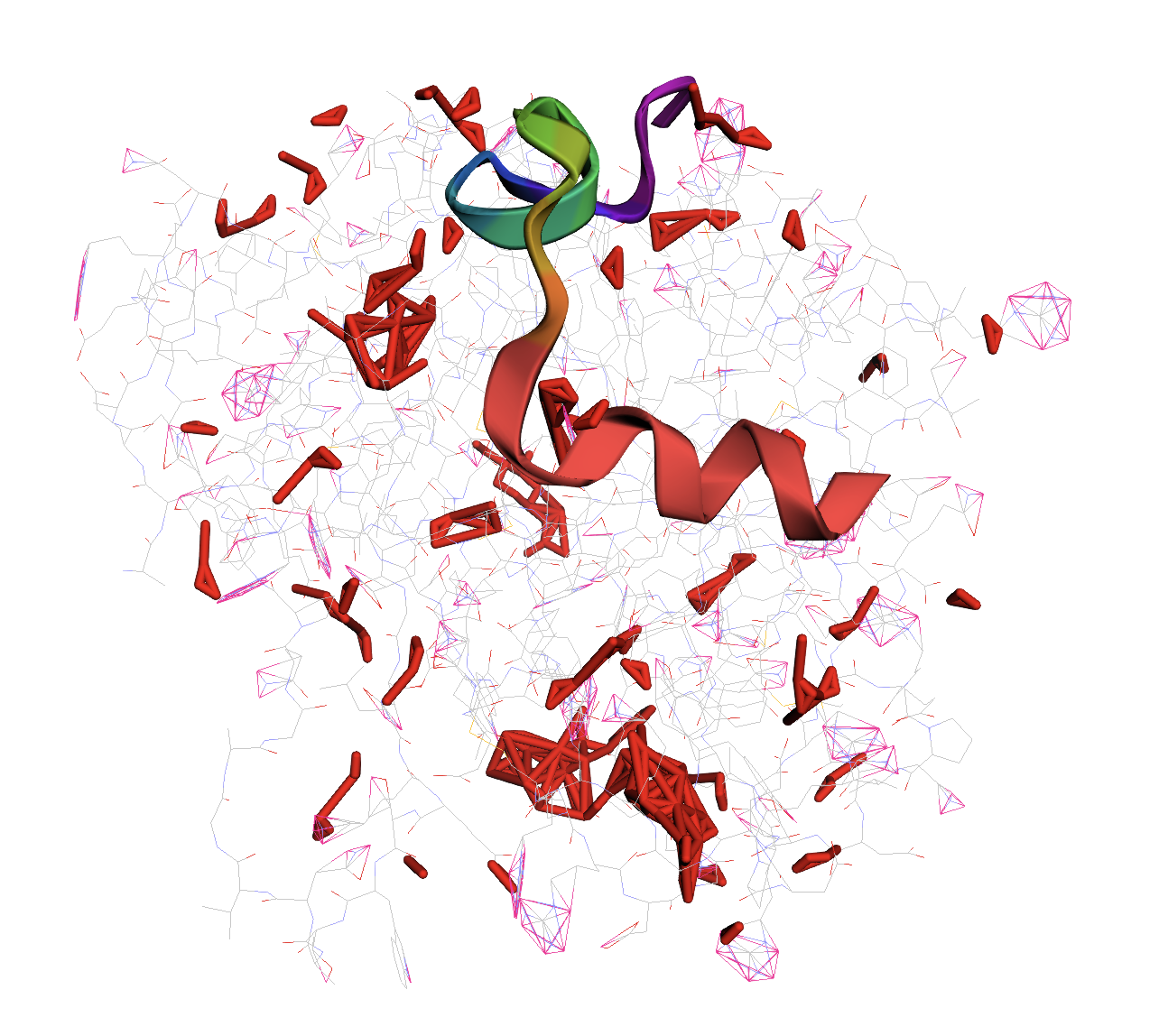
In addition, the pipeline downloads the ideal SDF file for each ligand from the RCSB ligand repository and uses RDKit to calculate key molecular descriptors such as molecular weight, LogP, number of hydrogen bond donors, and acceptors. The amino acid composition of the binding pocket is analyzed, and a pie chart is generated to visualize the frequency of each residue type. Overall, this workflow provides a structured approach for identifying and characterizing ligand binding sites in GPCR structures.

**Plots:**

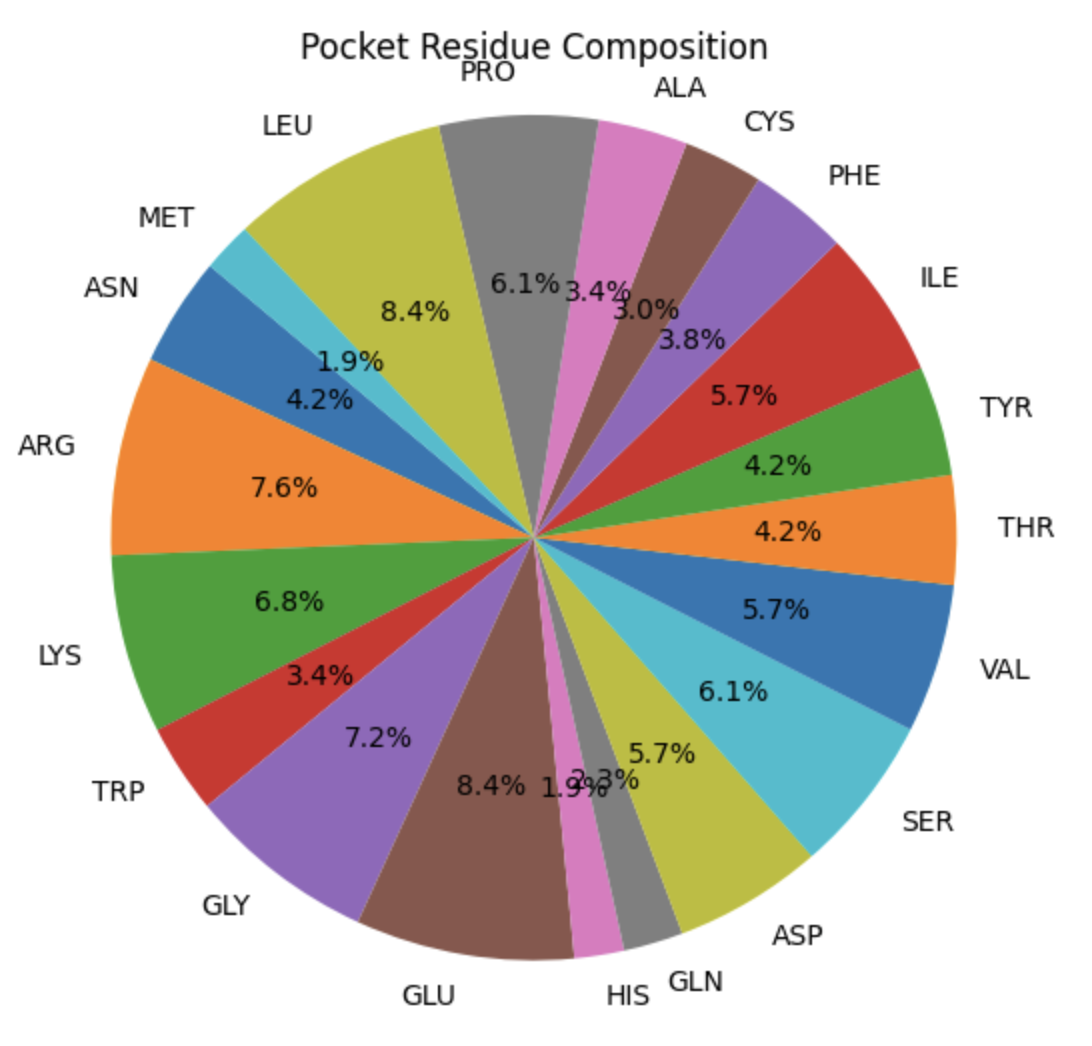
1. Multi-color ligand plot: 🔴 Ligand 🟧 Hydrophobic 🔵 Polar 🟩 Other residues



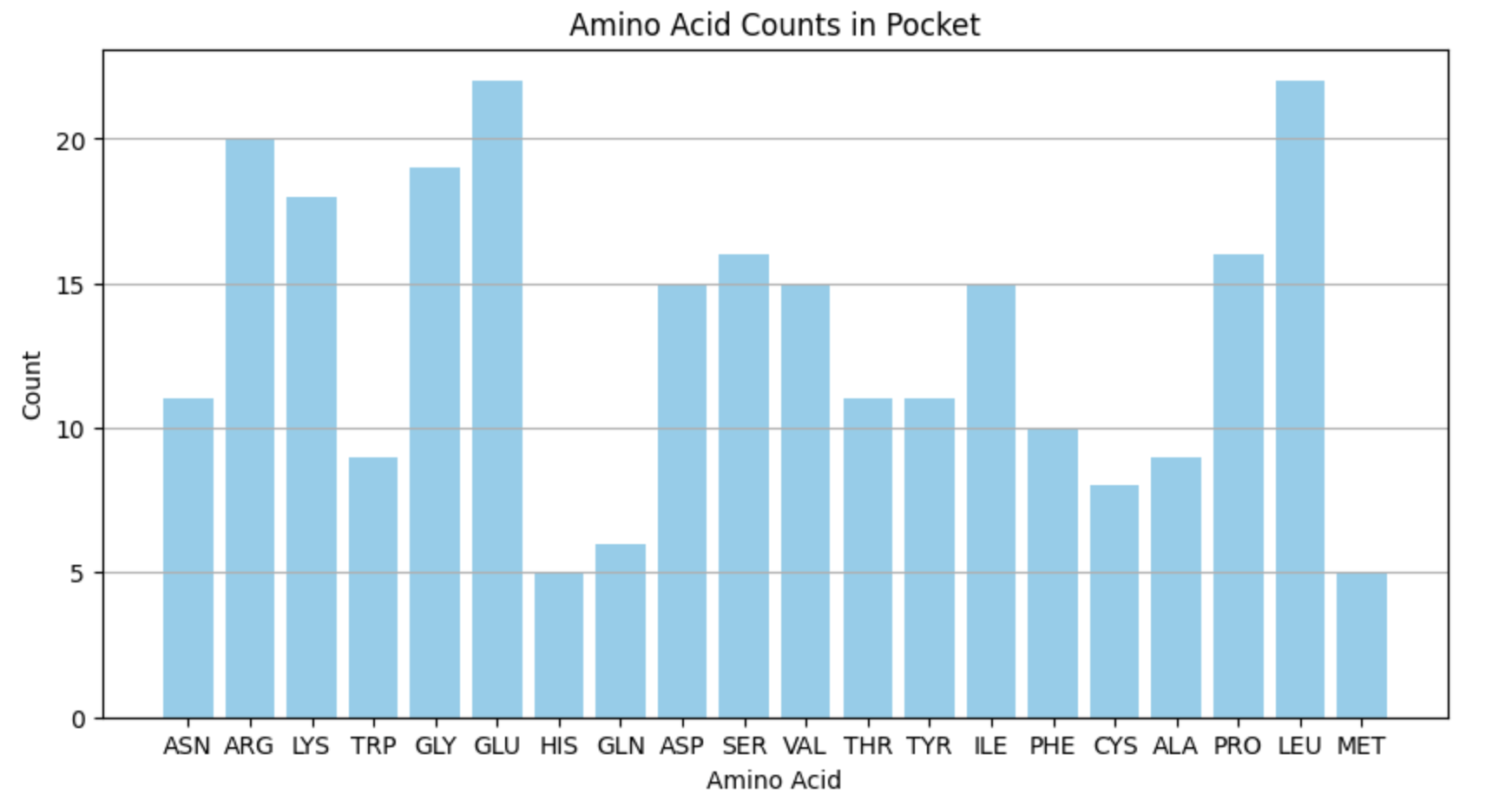
1. Chains A and L: multicolor (spectrum); ligands: red



1. Pie chart of pocket residue composition



1. Bar chart: amino acid counts

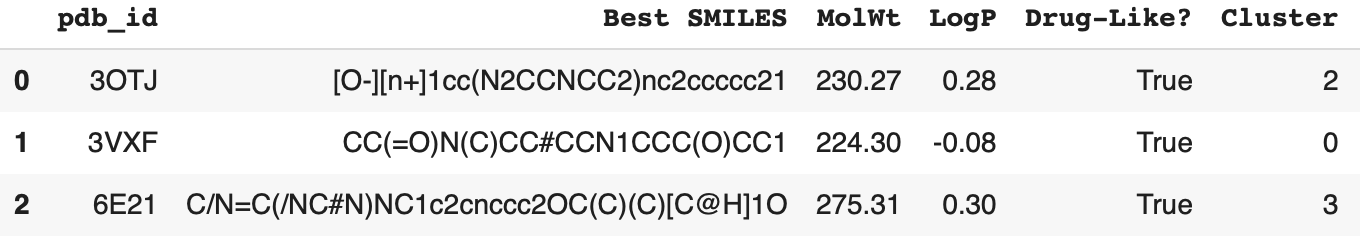


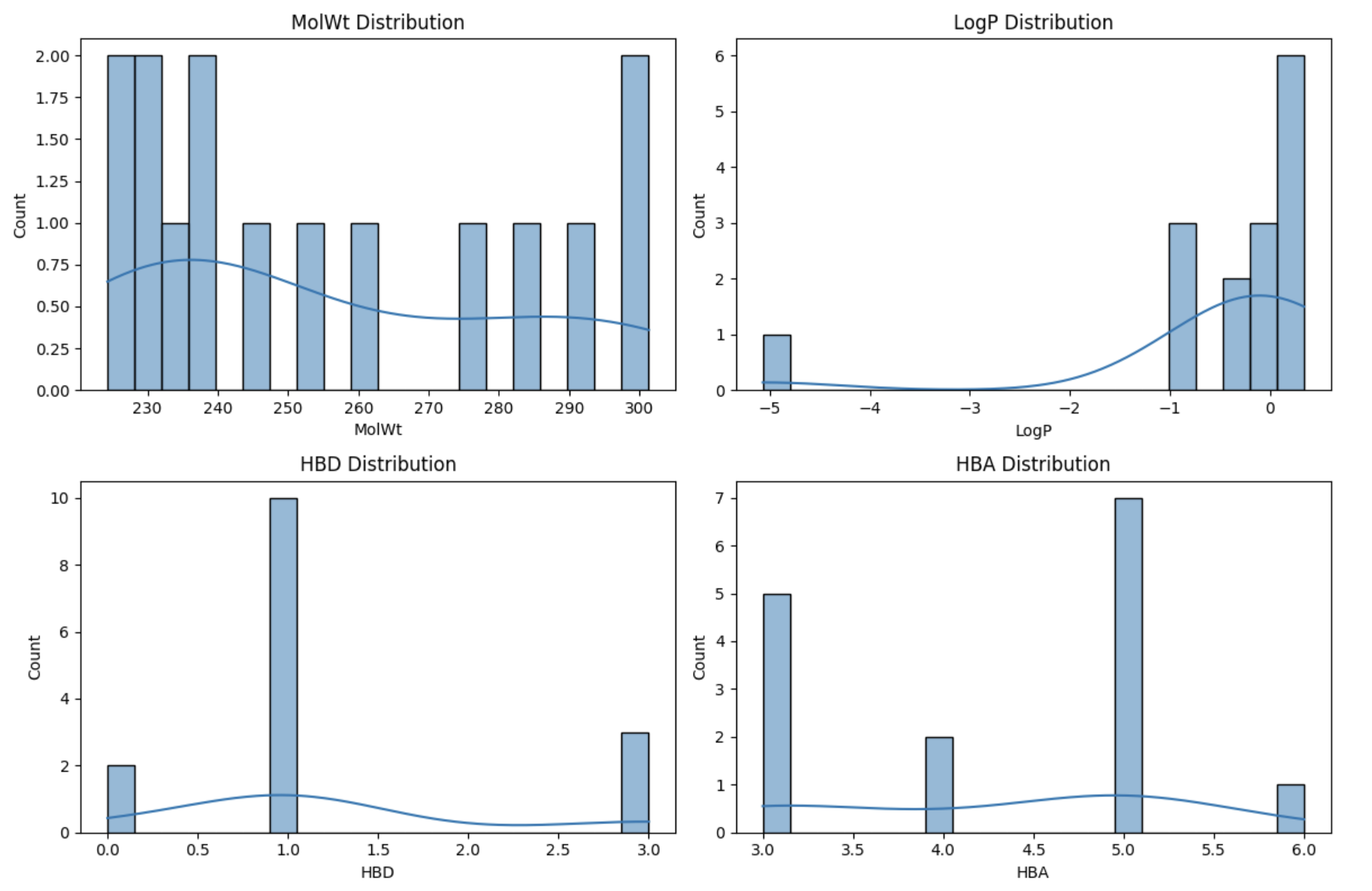
**Structure-Guided Ligand Design and Clustering for GPCR Binding Pockets**

<https://colab.research.google.com/drive/16m9KgdA18I2DLjxcUBU2ehXcPAKp6EXb#scrollTo=kfPRlmQ0V4qE>

A pipeline for designing, selecting, and analyzing small molecule ligands for GPCR (G-Protein Coupled Receptor) binding pockets using structural data from the Protein Data Bank and chemical information from the ChEMBL database.

A summary of how it works: First, it downloads a set of PDB structures of GPCRs, parses them using Biopython, and analyzes the composition of the binding pocket in each structure—counting how many residues are present, especially hydrophobic ones. Based on this information, it creates a prompt to guide ligand generation, specifying desired molecular weight, LogP, hydrogen bond acceptors/donors, and rotatable bonds. Using these constraints, it queries ChEMBL to fetch real compounds that match the desired properties. The retrieved SMILES are validated using RDKit to compute molecular descriptors like molecular weight, LogP, and hydrogen bonding features. All validated ligands are compiled into a dataframe. Molecular fingerprints are then calculated and used for PCA to reduce dimensionality, followed by KMeans clustering to group similar ligands. A scatter plot is generated to visualize how these ligands are distributed in chemical space. Finally, for each PDB structure, the ligand with the lowest molecular weight is selected as the best candidate. These top ligands are marked with placeholder "Docking Score" values and exported as a summary CSV file for further analysis.





**TRIAL 1 -** 07.07.2025

<https://colab.research.google.com/drive/14NnWS1t2Fea0fUyiypr_JRx2vtNN3PSJ>

*To evaluate compound performance, the system retrieves experimental records (ELN entries) and assay results (CDD data) for a given compound—e.g., Compound 9831—using a mock API. These datasets are then formatted into structured text and fed into a custom prompt. This prompt is processed by a BioGPT language model via a HuggingFace pipeline integrated with LangChain. The model generates a concise, human-readable summary that synthesizes the compound’s efficacy, safety, and observed effects based on the input data.*

**Output**   
=== ELN Data ===  
- Transfection with Compound 9831: HEK293T cells treated with compound 9831 at 10 µM. 85% inhibition. Moderate cytotoxicity.  
- Dose Response Study for Compound 9831: Compound 9831 tested at 1, 10, 50 µM. IC50 = 8.2 µM. Minimal off-target effects.  
- Compound 9831 on SH-SY5Y Cells: No inhibition on SH-SY5Y cells. Likely cell-line specific effect.  
  
=== CDD Data ===  
- Compound 9831 assays:  
 \* HEK293T Inhibition = 85 %  
 \* IC50 HEK293T = 8.2 µM  
 \* CYP450 Off-target = Low qualitative  
  
=== LLM Summary ===  
{'eln': 'Transfection with Compound 9831: HEK293T cells treated with compound 9831 at 10 µM. 85% inhibition. Moderate cytotoxicity.\nDose Response Study for Compound 9831: Compound 9831 tested at 1, 10, 50 µM. IC50 = 8.2 µM. Minimal off-target effects.\nCompound 9831 on SH-SY5Y Cells: No inhibition on SH-SY5Y cells. Likely cell-line specific effect.', 'cdd': 'Compound 9831:\n- HEK293T Inhibition: 85 %\n- IC50 HEK293T: 8.2 µM\n- CYP450 Off-target: Low qualitative', 'text': ' In summary, compound 9831 is a safe and potent small molecule for in vitro use.'}

***Explaining the output*:** The summary is returned as a dictionary containing: **`eln`**: The raw concatenated ELN text input. `**cdd`**: The raw concatenated CDD assay text input. **`text`**: The **generated summary by the language model**, which says: "In summary, compound 9831 is a safe and potent small molecule for in vitro use."