UNIVERSITAT ROVIRA i VIRGILI

PROTEIN-LIGAND INTERMOLECULAR INTERACTION ATLAS (PLII ATLAS): APPLICATION TO THE ANALYSIS OF THE INTERMOLECULAR INTERACTIONS BETWEEN NON-COVALENT LIGANDS AND THEIR CO-CRYSTALLIZED TARGETS

Ignacio Miguel Rodríguez

BIOTECHNOLOGY FINAL DEGREE THESIS

Tarragona, January 2025

Tutor: Gerard Pujadas Anguiano (gerard.pujadas@urv.cat)

Supervisor: Santiago Garcia Vallve (santi.garcia-vallve@urv.cat)

Jo, Ignacio Miguel Rodríguez, amb DNI 73412282P, soc coneixedor de la guia de prevenció de plagi a la URV P*revenció, detecció i tractament del plagi en la docència: guia per a estudiants* (aprovada el juliol 2017) [1], i afirmo que aquest TFG no constitueix cap de les conductes considerades com a plagi per la URV.

Tarragona, 4 de gener de 2025

(signatura)

MIGUEL RODRIGUEZ IGNACIO - 73412282P

Firmado digitalmente por MIGUEL RODRIGUEZ IGNACIO - 73412282P
Fecha: 2025.01.04 21:25:53 +01'00'

[1] http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formaciocompetencies-nuclears/plagi/

Content

Abstract and keywords

Proteins are crucial macromolecules in biological systems, taking part in diverse functions such as structural support, enzymatic activity, and signaling. To do so, proteins often interact with ligands. Intermolecular recognition is a key process to various biological functions, where interactions governed by non-covalent forces such as hydrogen bonds, Van der Waals forces, electrostatic interactions, and hydrophobic effects, are essential. The ligand's binding affinity, a measure of the strength of these interactions, can be described using parameters like K_d, K_i, and IC_{50}. These metrics along with interaction data are pivotal in drug design, offering valuable information about the potency and stability of protein-ligand complexes.

Recent works in computer-aided drug discovery (CADD) have heavily transformed the process of drug development thanks to the usage of computational techniques like structure-based drug design (SBDD) and ligand-based drug design (LBDD). While SBDD uses the 3D structure of target proteins to design and screen potential ligands, LBDD builds models based on known ligands to predict interactions with target proteins. Both methods benefit from databases that catalog protein-ligand interactions, enabling more efficient drug development. This work proposes the development of a database that facilitates the study of protein-ligand affinities and interactions, aiding the refinement of CADD methodologies and that could be used to improve the prediction of binding affinities for drug discovery.

***Keywords**: CADD, database, protein-ligand, interactions, bioinformatics, affinity, cheminformatics*

1. Introduction

Proteins are essential macromolecules in biological systems, performing a wide range of functions, including structural support, enzymatic activity, signaling, and more. To carry out these roles effectively, proteins often interact with other molecules known as ligands.

Ligands are diverse in size and nature, ranging from small to larger organic molecules such as cofactors, inhibitors, and substrates. These molecules bind to proteins through specific non-covalent interactions with varying degrees of affinity. A higher binding affinity indicates a stronger interaction, while lower affinity reflects weaker binding. This precise process, known as **intermolecular recognition** [1], is fundamental to numerous biological processes.

1.1. Interactions

The non-covalent interactions that allow protein-ligand binding include hydrogen bonds, Van der Waals interactions, electrostatic forces and hydrophobic interactions [2].

According to the IUPAC, “the **hydrogen bond** is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation.” A shorter distance in the hydrogen bond is thought to be related with a bigger strength in the interaction [3].

Oxygen and nitrogen are the most common acceptor atoms in hydrogen bonds, while the donor is typically another electronegative atom [4], as shown in Figure 1.

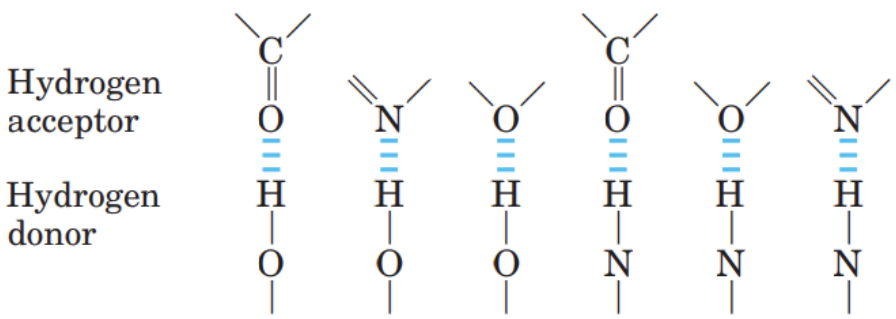


Figure 1. Most common hydrogen bonds in biological systems [4].

The bond between the donor and the acceptor is considered to be **directional** although **flexible**. In fact, the strength of the interactions is influenced by several factors including the **angle, distance and chemical properties** of the donor and acceptor. The length of a hydrogen bond is defined as the distance between the hydrogen atom of the donor and the acceptor atom. For instance, when oxygen or nitrogen act as the acceptor, the hydrogen bond length is approximately 2.8 Å, including the σ covalent bonds that are of around 1.0 Å [5], as shown in Figure 2a. These covalent bonds are 30 times stronger than hydrogen bonds, and this is also translated into smaller distances. The distribution of the interaction angle is shown in Figure 2b.

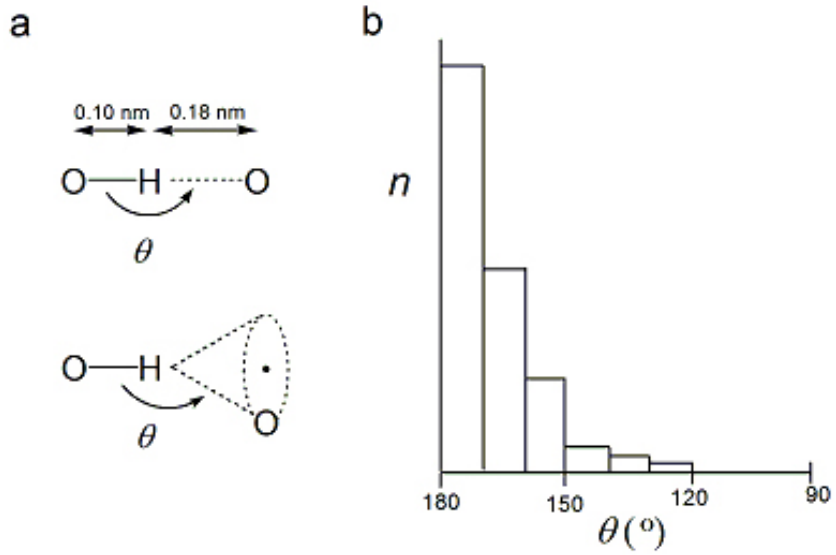


Figure 2. Geometry of hydrogen bonds [5].

The strength of hydrogen bonds is typically categorized into three levels: weak, normal and strong. Figure 3 illustrates the differences in potential energy between weak and strong hydrogen bonds. Strong hydrogen bonds are shorter and develop wider potential energy wells, meaning that there is a weak or even inexistent energy barrier.

Particularly, strong bonds showcase angles between 170 and 180º, while weak interactions establish angles lower than 120º [5].

A high-energy **barrier** exists in weak hydrogen bonds, preventing efficient hydrogen transfer. This is often due to differences in electronegativity or larger distances between the donor and acceptor atoms. On the other hand, strong hydrogen bonds significantly reduce or eliminate this barrier, allowing the effective sharing of the hydrogen between the atoms that make up the bond [6].

In conclusion, Figure 3 depicts that the strength of hydrogen bond is heavily influenced by distance and symmetry. Stronger interactions are originated by shorter distances, while symmetric bonding provides more stable energy profiles.

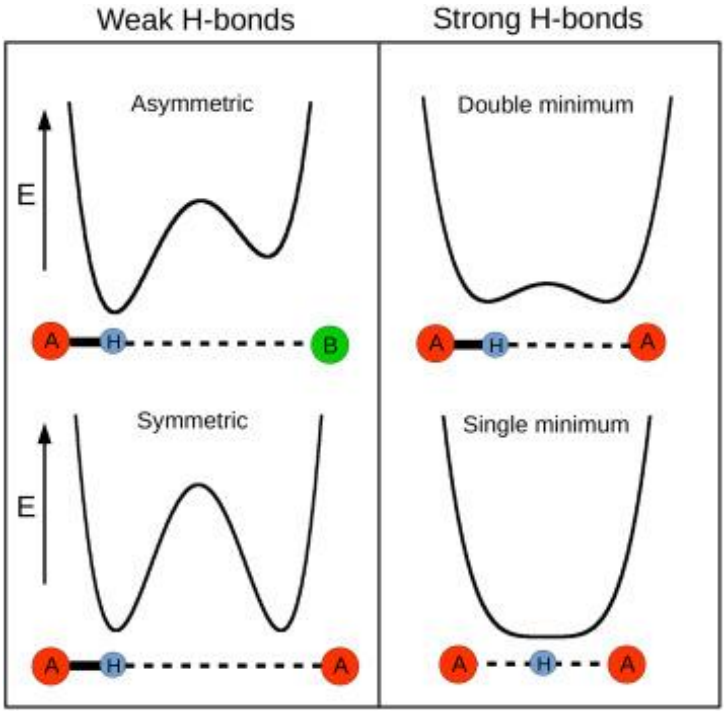


Figure 3. Potential energy comparison between weak and strong hydrogen bonds [6].

Van der Waals forces are a result of the addition of electrostatic forces that exist between neutral molecules at short distances. Among these electrically neutral molecules we can find permanent dipoles, induced dipoles, and nonpolar molecules and atoms. Based on the type of electrical interaction, van der Waals forces are typically classified into three main categories: (1) the interaction between permanent dipoles (referred to as the Keesom force); (2) the interaction between a permanent dipole and an induced dipole (namely, Debye force); and (3) the interaction between non-polar molecules or atoms (known as London force or dispersion force) [7].

Figure 4 represents how the strength of van der Waals forces varies with the distance between molecules. As one molecule gets closer to the other, the van der Waals force intensifies until reaching a maximum value. Beyond this point, the force decreases until it eventually turns repulsive [8].

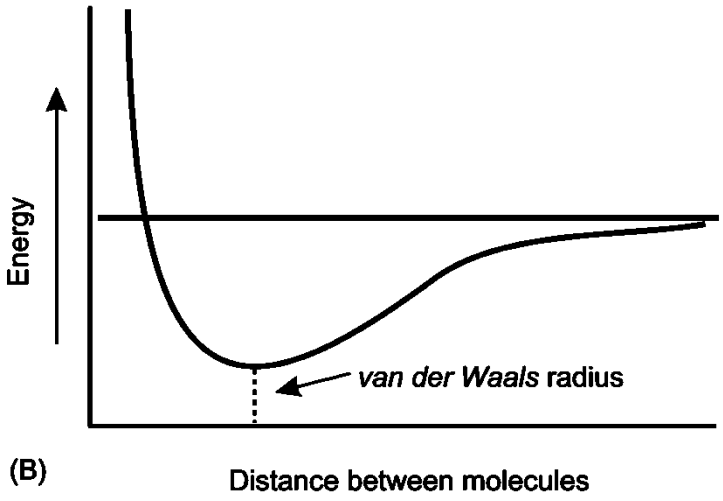


Figure 4. Van der Waals forces. Representation of energy vs. bond distance [8].

On the other hand, **hydrophobic interactions** are not true “forces”, but rather entropic effects driven by the behavior of water molecules [9]. When nonpolar molecules are introduced into water, they disrupt the hydrogen bonds established within water molecules. To maximize entropy, the hydrophobic (nonpolar) molecules aggregate, which minimizes their contact with water, as depicted in Figure 5.

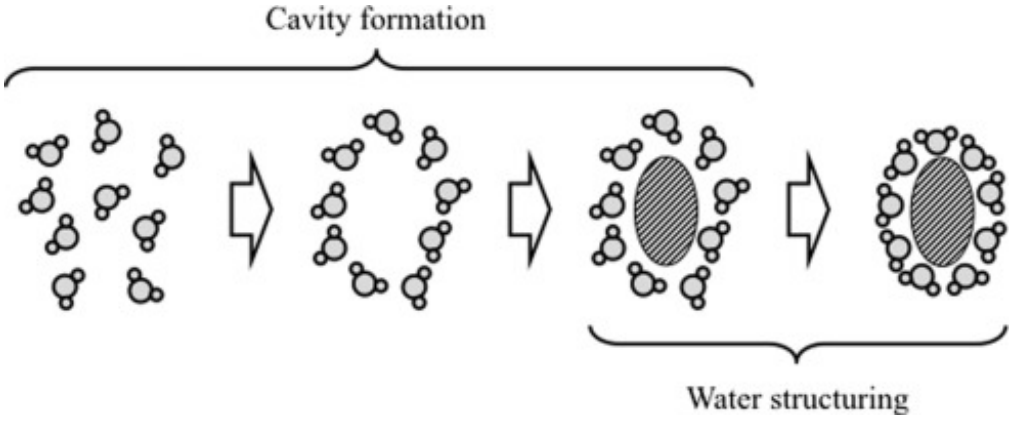


Figure 5. Structuring of water around a hydrophobic molecule [10].

Numerous theories have been postulated to explain the behavior of water when interacting with hydrophobic particles. One of the theories, the “iceberg” model, was first proposed in 1945 by Frank and Evans [11]. According to their work, water forms a structured “cage” around the solute in an ordered arrangement.

However, as stated in the review by Qiang Sun [9], several studies present contradictory conclusions, with some showing an increase in water ordering and others indicating a decrease in the structure order around hydrophobic molecules. This discrepancy raises questions about the validity of the “iceberg” theory.

Years later, in 1959, the idea of hydrophobic interactions was introduced by Kauzmann [12]. He built up his theory on the “iceberg” model, stating that when two “caged” hydrophobic molecules reach each other, the “structured” water between them could be dispersed into the surrounding water, increasing the entropy of the system. This entropy increase may be the reason why these particles are attracted to each other. As a result, hydrophobic interactions are traditionally considered to be driven by entropy, as previously stated.

In addition to these fundamental forces, other interactions such as pi-cation, metal acceptor and aromatic forces (including edge/edge and edge/face interactions) are also present in protein-ligand binding.

Pi-cation interactions are a type of electrostatic force where a cation with positive charge engages with the negatively charged electron cloud of π systems (Figure 6). This type of interaction is considered to be the strongest among noncovalent interactions [13].

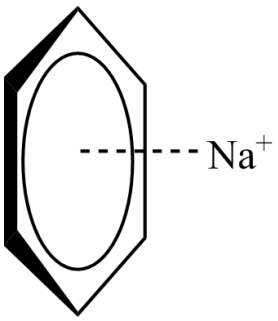


Figure 6. Pi-cation interaction between benzene and a sodium cation. Source: https://en.wikipedia.org/wiki/Cation%E2%80%93%CF%80_interaction

Eventually, let’s analyze aromatic forces. **Aromatic forces** involve interactions between aromatic molecules or parts of them as a consequence of their electronic structures. There exist two main types: **face-to-face** interactions, also known as π-π stacking and **edge-to-face** interactions, sometimes referred to as CH-π interactions.

Face-to-face interactions appear when two aromatic rings align in parallel, while edge-to-face interactions occur when the edge of one aromatic ring engages with the face of another aromatic ring. According to [14], edge-to-face interactions are more energetic than face-to-face forces among aromatic groups. Figure 7 shows the possible stacking of molecules in aromatic interactions. Offset-stacked is another type

of aromatic interaction; however, it will not be further discussed in this work as it is not utilized within the database.

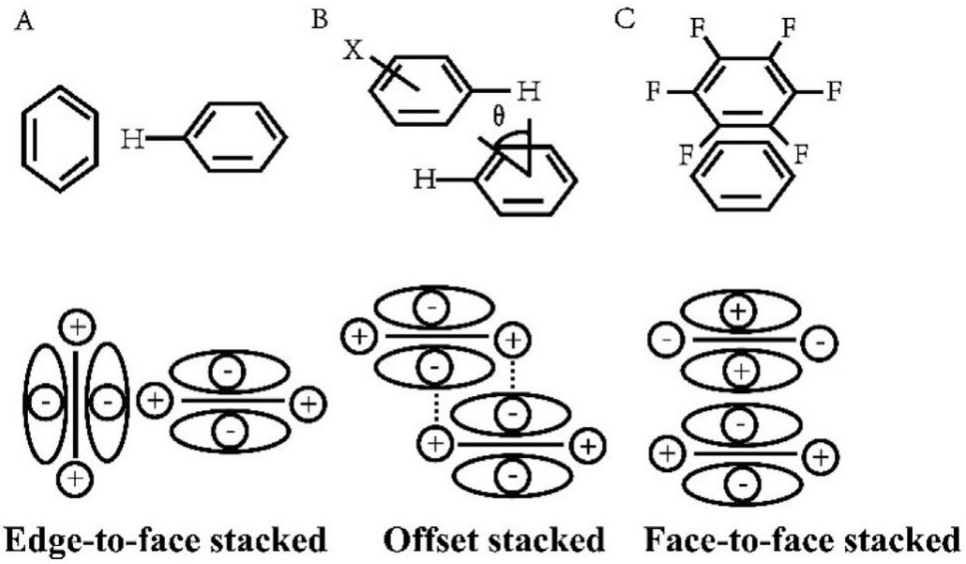


Figure 7. Drawing of the different geometric configurations of aromatic interactions [14].

1.1. Binding affinity

Understanding and studying the kinetics and parameters that build up these interactions in protein-ligand complexes is vital to new drug design. In the fields of biochemistry and pharmacology, several metrics are used to assess the strength of the interactions between proteins and their ligands/inhibitors. These include the **inhibition constant K_i**, the **dissociation constant K_d** and the **half-maximal inhibitory concentration IC_{50}**.

While both K_i and K_d describe the affinity of a ligand to its receptor (typically a protein), they provide different meanings. The inhibition constant K_i measures the potency of an inhibitor in preventing the binding of a ligand to its target. On the other hand, the dissociation constant K_d measures the equilibrium between the protein-ligand complex (referred to as PL) and its separate units ($P + L$ protein and ligand, respectively) [15].

$$\mathrm{PL} \underset{K_d^{-1}}{\overset{K_d}{\rightleftharpoons}} \mathrm{P} + \mathrm{L}$$

$$K_d = \frac{[\mathrm{P}][\mathrm{L}]}{[\mathrm{PL}]}$$

In the above equation, $[P]$ represents the concentration of free protein, $[L]$ the concentration of free ligand, and $[PL]$ refers for the concentration of the protein-ligand complex.

The inhibition constant involves a third party (the inhibitor, referred to as I). The equilibrium equation involving the inhibitor depends on the mechanism of inhibition (competitive, uncompetitive, non-competitive or mixed). Figure 8 shows the equilibria for different inhibition mechanisms.

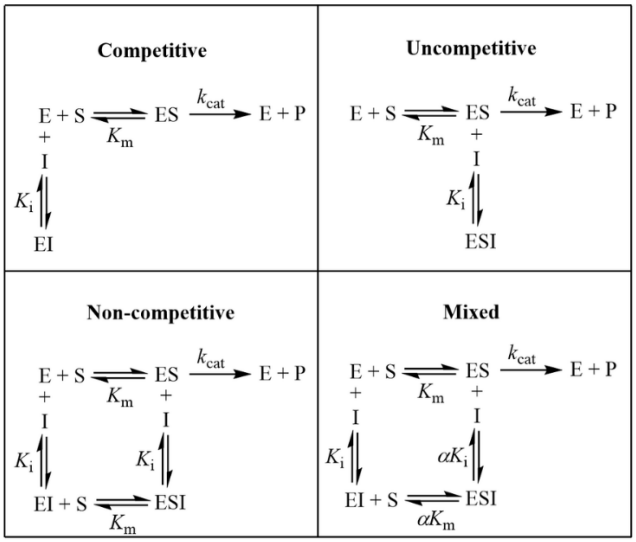


Figure 8. Equilibrium equations for different inhibition mechanisms. Source [16].

In these equations, E refers to the enzyme (the protein) and S refers to substrate (the ligand), which bind to create the enzyme-substrate ligand (ES), and that may dissociate into E and P (the product). In this system, the inhibitor I also plays a crucial role as a potential ligand to E. Comparing the mechanisms:

- In **competitive inhibition** the inhibitor can only bind to the free enzyme (E).
- In the **uncompetitive** approach, the inhibitor can only bind to the enzyme-substrate complex (ES).
- In **non-competitive** and **mixed inhibitions**, the inhibitor can bind to both the free enzyme (E) and the enzyme-substrate complex (ES). However, in the former (non-competitive inhibition) the binding of the substrate does not affect the binding of the inhibitor, in contrast to the general mechanism of mixed inhibition.

The inhibition constant is typically used when the binding is measured through inhibition kinetic methods. In contrast, when the measurement is done directly, the dissociation constant is used instead.

The concentration of inhibitor needed to reduce biological activity by half, known as IC_{50}, is not a direct measure of the affinity, as it does not correspond to any equilibrium constant. This makes it less precise than other measures like K_d or K_i. Additionally, IC_{50} values strongly depend on experimental conditions, such as the concentrations of protein, ligand and inhibitor during the measurement [17]. This variability complicates comparisons and highlights the need for dissociation constants that better represent binding affinity. Despite these limitations, IC_{50} values are still commonly used as an indirect measure of binding affinity.

1.2. Computer-aided drug discovery (CADD)

As previously discussed, knowing the values for these metrics is pivotal to finding and designing stronger drugs. Traditionally, releasing a new drug into the market was a long and costly process, both in terms of time and money. However, in recent years, the use of **computer-aided drug discovery (CADD)** has become crucial for streamlining the drug development process and reducing the associated costs, especially in the preliminary stages. Modern computational resources, such as supercomputers and parallel computing, have further accelerated this area in pharmaceutical research [18].

There are several methodologies in computational design of drugs. Usually, two categories are leveraged: **structured-based** and **ligand-based** methodologies. A schematic representation of the steps taken by both methods is shown in Figure 9.

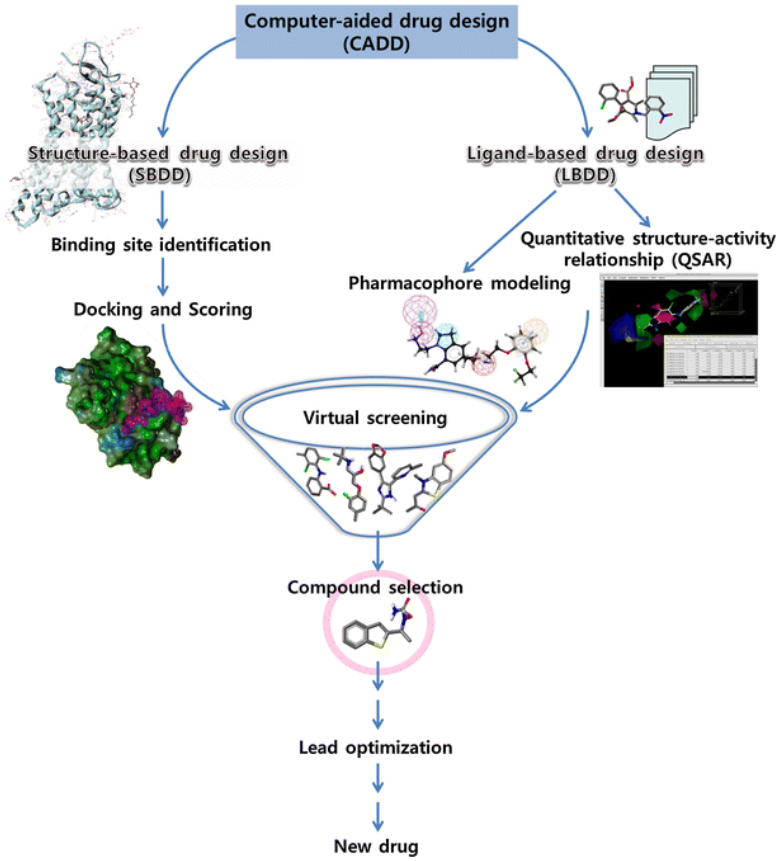


Figure 9. Computer-aided drug design (CADD) techniques [18].

1.2.1. Structured-based drug design (SBDD)

Structured-based drug design (SBDD) uses the three-dimensional structure of a target protein to obtain a set of potential ligands. These ligands are evaluated on their binding affinity based on the predicted interactions between the ligand and the protein's binding site.

The process begins with the identification of the **binding site** of the target protein. Two primary approaches can be utilized within SBDD: ***de novo* design** and **virtual screening (VS)**. The *de novo* approach focuses on designing new molecules from scratch, which allows for broader exploration of chemicals very efficiently [19].

On the other hand, virtual screening implies assessing pre-existing compounds from a database to determine their binding effectiveness with the target. This facilitates the identification of compounds that serve as substitutes for already known ligands in binding to target molecules or even for previously unknown targets, if their structural information is available. The **main drawback** of this technique is that it is not always possible to obtain powerful compounds with a high affinity with the target molecule.

Active or powerful compounds are said to be so if their IC_{50} or K_i values are of the order of μM or nM [20].

Docking is a widely used method in which constraints can be applied to limit solutions to desired features. If the key interactions needed for a ligand to bind to its target are known, this technique becomes particularly effective in identifying potential ligands. A series of assumptions are done throughout this process. First, the protein is considered to be rigid and receptive to binding, while the ligand is said to be flexible. The main goal is to find the position of the ligand within the binding site that minimizes the energy conformation [21].

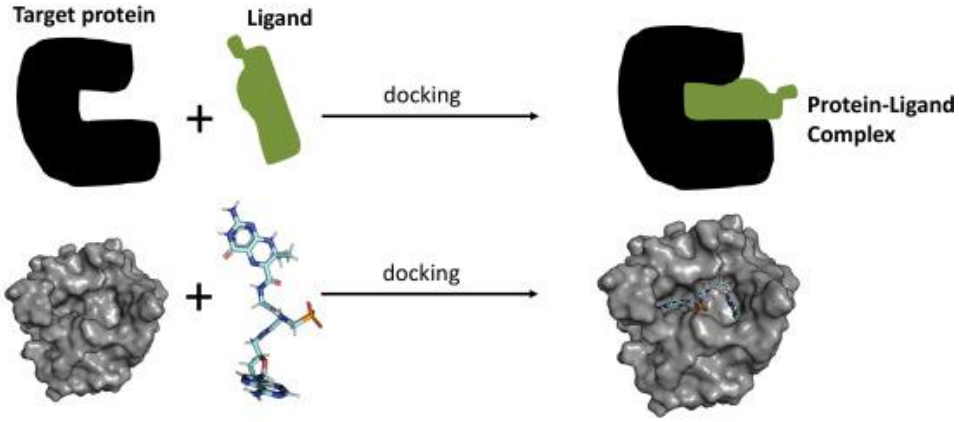


Figure 10. Protein-ligand molecular docking [21].

The process of protein-ligand docking begins with the selection of the target protein or receptor. The next step involves representing the binding pockets. This is a computationally-demanding process, as atomic radii and charges have to be described. Once the pockets are defined, focus is placed only on those for specific ligand-binding, rather than the whole receptor molecule. The search for a suitable ligand follows, using methods like deterministic, stochastic and simulative approaches to generate a diverse pool of potential ligands. This set of ligands is then assessed using scoring functions to rank the ligands based on their suitability for binding to the described pockets [21].

A database that provides detailed information on protein-ligand affinities and interactions can facilitate this process, as interactions can be better categorized based on binding strength. For that reason, this work proposes the development of such a database, aimed at learning from existing protein-ligand complexes that could be used for improving CADD processes, including constraint-based docking.

1.2.2. Ligand-based drug design (LBDD)

Unlike SBDD, the **ligand-based drug design (LBDD)** technique employs computational modeling techniques on a dataset of molecules with different structures and known potencies to create theoretical predictive models. This is especially useful when the 3D structure of the target protein is unavailable or unknown. By analyzing a set of known active ligands, it becomes possible to identify the molecular descriptors responsible for the observed affinity or activity [18].

Virtual screening based on ligands tends to be computationally less expensive compared to virtual screening in SBDD, as the size of the molecules is smaller [20].

One of the possible methods in LBDD involves using **pharmacophores**. A pharmacophore is composed of the steric and electronic features that provide better interactions with the target molecules and trigger a biological response [22]. It identifies the shared molecular interactions across compounds that bind to a specific target, such as hydrogen bonds, hydrophobic regions, and electrostatic interaction sites. As stated in [23], “the pharmacophore should be considered as the largest common denominator of the molecular interaction features shared by a set of active molecules”, emphasizing that a pharmacophore is something abstract and not an actual physical molecule. While 2D pharmacophores represent the minimal structural elements linking key binding groups, 3D pharmacophores define their spatial arrangement in the three-dimensional space.

Figure 11 represents a pharmacophore query, showing several interactions such as hydrophobic, aromatic, and hydrogen bonding interactions. The spheres illustrate the geometric constraint of these interactions. Furthermore, several of these features can be combined to perform the query, along with directional elements for those interactions where orientation is critical, like in hydrogen bonds.

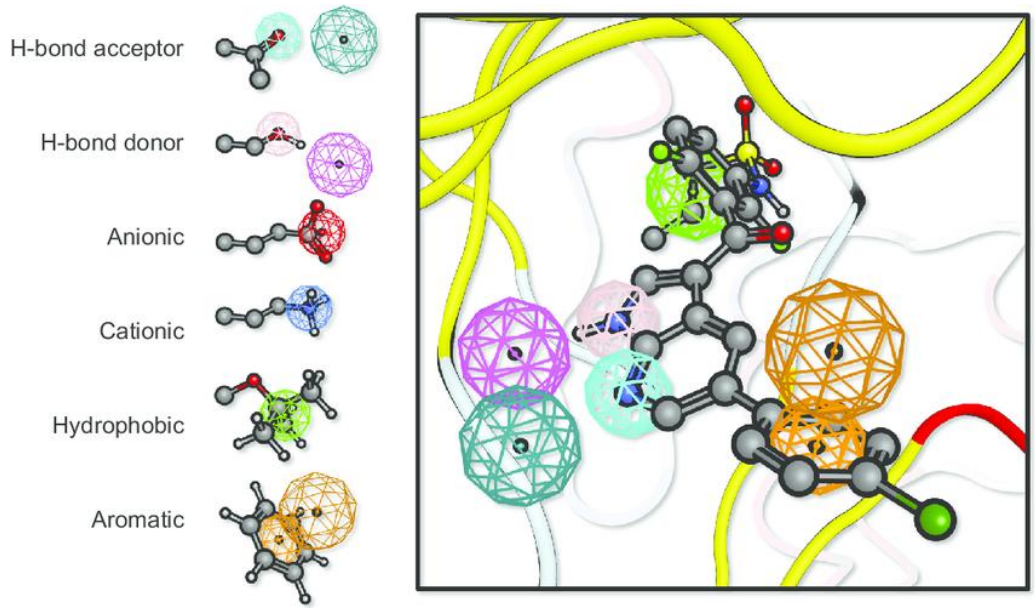


Figure 11. Pharmacophore query. Spheres represent the geometric constraints for the interactions [23].

By analyzing an active (or inactive) compound with known interactions, the key interaction sites can be identified. This enables the selection of the most suitable features to construct a pharmacophore. Our database can be utilized to pinpoint the most relevant binding sites, streamlining this process. Additionally, the database can be also valuable during the compound selection step illustrated in Figure 9.

2. Hypothesis and objectives

The main hypothesis of this work is that protein-ligand interaction data can be pivotal to obtain high bioactivity values in the design or search of new drugs.

The main goal of this project is **to develop a database**, the **P**rotein-**L**igand **I**ntermolecular **I**nteraction **Atlas** (PLII ATLAS), that includes interaction data about protein-ligand complexes from which their affinity is known. This database will enable users to perform searches based on protein or ligand information, as well as based on specific types of interactions. Additionally, it will also include plotting tools to represent data visually, supporting statistical and comparative studies and enhancing its utility for researchers and drug designers.

A secondary objective is to provide practical examples on the usage and functionality of the database.

3. Methodology

In this section, we explore the different methodologies utilized within this project, mainly focusing on data collection and preprocessing, and a summary of the used libraries and technologies.

3.1. Input data

The **PDBind database**, introduced by Wang et al. in 2003, compiled binding affinity data (K_d, K_i, and IC_{50}) for 1,359 complexes using information available in the Protein Data Bank (PDB) [24]. Their posterior work [25], provides a detailed explanation of the methodologies used to develop the database, including the algorithm for identifying and classifying protein-ligand complexes.

The database kept undergoing updates and new releases until the latest free version that was released in 2020. This release contains binding affinity data for a total of 23,496 biomolecular complexes from the PDB, which highly contrasts with the original number of complexes in the first version.

Table 1. Comparison of the number of complexes in PDBind for different versions [26].

Version	Entries in the PDB	All complexes with binding data	Protein-ligand complexes
2004	28,991	2,276	2,276
2020	157,974	23,496	19,443

As depicted in Table 1, the number of complexes in the PDBind has grown significantly from the 2004 release to the 2020 version. This wide data gathering simplifies drug discovery studies and avoids the tedious task of manually filtering and searching for protein-ligand complexes within the PDB database.

In this study, we have used the 2020 version of PDBind, applying a filter to keep only those complexes that establish non-covalent binds with their ligands.

3.1.1. IFP files

IFP (Interaction FingerPrint) files are output files generated with the iChem tool [27]. This tool allows to determine the molecular interactions between protein and ligands,

or between molecules in general, based on the information retrieved from the PDB database[2]. In this work, the data filtered from the PDBind serves as input to the iChem tool.

Inside an IFP file, each interaction is placed on a different line, and it consists of the information shown in Table 2 [28]. An example of IFP file is shown in Figure 12. As we can see, the angle of the interaction is only available for hydrogen bonds (HBond), while for the rest of interactions a "/" indicates a null value.

Table 2. Information of each row in IFP data files.

IFP file data								
Interaction type	Protein atom name	Protein atom number	Protein residue name and number	Ligand atom name	Ligand atom number	Ligand residue name and number	Interaction distance	Interaction angle (only for H-bonds)

```
HBond PROT          O    |172  |HOH 646-B   O17  |1    |ZOL   0-XX  3.03992 125.137
Metal Acceptor      MG   |167  | MG1303-B   O17  |1    |ZOL   0-XX  1.90938 /
Ionic PROT          CZ   |52   |ARG  89-B   O15  |4    |ZOL   0-XX  3.60955 /
Metal Acceptor      MG   |167  | MG1303-B   O10  |7    |ZOL   0-XX  2.07228 /
Ionic PROT          NZ   |166  |LYS 238-B   O10  |7    |ZOL   0-XX  3.86884 /
Metal Acceptor      MG   |168  | MG1304-B   O10  |7    |ZOL   0-XX  2.14341 /
Ionic PROT          NZ   |88   |LYS 174-B   O11  |8    |ZOL   0-XX  2.87419 /
Ionic PROT          NZ   |166  |LYS 238-B   O12  |9    |ZOL   0-XX  2.71442 /
Hydrophobic         CD2  |14   |LEU  77-B   C7   |11   |ZOL   0-XX  4.12181 /

|B    S76|B    L77|B    D80|B    D81|B    D84|B    A86|B    R89|B   L143|B   Q147|B
0000000001000000000000000000000000000000000000000000000001000000000000000000000
```

Figure 12. Example of IFP file (protein 2e91 and ligand ZOL).

3.1.2. Mol2 files

A **`.mol2` file** is a representation of a molecule, and it contains all the necessary information to fully reconstruct the molecule. It is written in ASCII characters [29]. The content of these files is crucial to construct the database, as it provides information

[2] **PDB database**: https://www.rcsb.org/

about the atoms taking part in the interactions (e.g., their coordinates, the name of the atom, etc.). A further explanation of all the information that can be collected from the .mol2 files is next discussed.

Mol2 files follow a specific structure. All the sections are defined by the `@<TRIPOS>` tag, followed by the name of the section (e.g., `@<TRIPOS>ATOM`). For this work, we will focus on the `@<TRIPOS>ATOM` (Figure 13) and `@<TRIPOS>BOND` (Figure 14) sections.

Table 3. Information provided in the @<TRIPOS>ATOM section of a .mol2 file.

@<TRIPOS>ATOM								
Atom ID	Atom name	X coordinate	Y coordinate	Z coordinate	SYBYL atom type	Substructure ID	Substructure name	Charge

```
#    Name: 2e91_1_ZOL_SITE
#    Modification time : Tue Apr 30 19:18:15 2024

@<TRIPOS>MOLECULE
2e91_1_ZOL_SITE
  209   183    33     0     1
PROTEIN
NO_CHARGES


@<TRIPOS>DICT
BIOPOLYMER macromol
@<TRIPOS>ATOM
      1 N          32.8540   42.9980   14.9440 N.am      1 SER76       0.0000
      2 CA         31.5490   43.3140   14.3790 C.3       1 SER76       0.0000
      3 C          31.3930   42.7230   12.9520 C.2       1 SER76       0.0000
      4 O          30.2740   42.3820   12.5590 O.2       1 SER76       0.0000
      5 CB         31.3280   44.8170   14.3260 C.3       1 SER76       0.0000
```

Figure 13. Mol2 file, atom section.

Table 3 provides a description of the information expected in the ATOM section, with Figure 13 illustrating an example of an actual .mol2 file.

The same applies to Table 4 and Figure 14 for the information written in the BOND section. Additionally, Table 5 shows the encoding of bond types adopted in. mol2 files,

where the 4th column represents the bond type. In Figure 14, various bond types are shown, including single, double and amide bonds.

Table 4. Information provided in the @<TRIPOS>BOND section of a. mol2 file.

@<TRIPOS>BOND			
Bond ID	Origin atom ID	Target atom ID	SYBYL bond type

Table 5. Codification of bond types used in the. mol2 files.

Bond type	
1	single
2	Double
3	triple
am	Amide
Ar	Aromatic
Du	Dummy
Un	unknown
Nc	Not connected

```
@<TRIPOS>BOND
     1    1    2 1
     2    2    3 1
     3    2    5 1
     4    3    4 2
     5    3    7 am
     6    5    6 1
     7    7    8 1
     8    8    9 1
     9    8   11 1
    10    9   10 2
```

Figure 14. Mol2 file, bond section.

3.2. Preprocessing

3.2.1. Interactions data

Based on the original files already presented in section 1, the goal is to construct a matrix (referred to as interactions table in the following sections) that agglutinates data from different files but about the same protein-ligand complex, so that the data is centralized in a single table. Particularly, from the previously introduced. mol2 and IFP files, we constructed the following matrix (for visualization purposes it is split into several submatrices):

Table 6. Ligand data included in the interactions table.

Ligand data						
Atomic number	Atom tag	Atom position in .mol2 file	X coordinate	Y coordinate	Z coordinate	SYBYL atom type

Table 7. Protein data included in the interactions table.

Protein data									
Atomic number	Atom tag	Residue	Residue position	Polypeptide chain	Atom position in .mol2 file	X coordinate	Y coordinate	Z coordinate	SYBYL atom type

Table 8. Interaction type and ligand-atom connectivity included in the interactions table.

Intermolecular interaction	**Ligand-atom connectivity**					
Interaction type	Atom position in .mol2 file	SYBYL bond type	X coordinate	Y coordinate	Z coordinate	SYBYL atom type

Table 9. Protein-atom connectivity included in the interactions table.

Protein-atom connectivity					
Atom position in .mol2 file	SYBYL bond type	X coordinate	Y coordinate	Z coordinate	SYBYL atom type

The content in Tables 6-9 is added up to compose a wide matrix used to show protein-ligand interaction data in the database. In fact, as stated before, we will refer to this matrix as “Interactions table” in the next sections.

3.2.2. Affinity data

As introduced in Section 1, the affinity data was retrieved from PDBind v2020 as a `.csv` file that included, amongst other information, the PDB ID, ligand ID and binding data. A logarithmic transformation was applied to the values of K_d, K_i or IC_{50} to standardize the binding data.

For instance, a value of $K_i = 190\ \mu M$ is converted as follows:

1) **Conversion to molar units**:

$$K_i = 190\mu M \cdot \frac{1\ M}{10^6\ \mu M} = 0.00019\ M$$

2) **Logarithmic normalization**:

$$pK_i = -log_{10}(0.00019)\ = 3.72$$

This normalization removes the unit dependency, making sure that all values are standardized and expressed in the same way, making it easier to compare different data.

3.3. Libraries and software used

The project has been developed entirely using **Python**[3], a versatile, easy and widely adopted programming language, particularly in the field of bioinformatics. In fact, as of

[3] **Python programming language**: https://www.python.org/

2024, Python remains the most used programming language, according to IEEE Spectrum ranking [30].

To ease the treatment of data and files, the *pandas*[4] library was used. This powerful Python library simplifies the tasks of reading, manipulating and writing files in tabular formats, which makes development more efficient and straightforward.

Additionally, the SQL database language has been employed to construct the database tables and handle user queries via the webpage. The integration of SQL into Python was done using SQLAlchemy[5], a tool that simplifies database handling and brings SQL to the Python language.

3.3.1. Server

The database is hosted on a server and made accessible from the Internet as a web page. The server has been developed using the *Flask*[6] framework for Python. This framework allows to easily create new endpoints and dynamically update HTML files in response to changes or triggered functions. For instance, it allows to make some computations in the back-end (server) and pass the results on to the front-end (browser) in order for the data to be shown.

4. Results and Discussion

In this work, we developed a protein-ligand interaction database comprising 6,125 protein-ligand complexes and 238,477 recorded interactions. These complexes contain information about the binding affinity between the proteins and the ligands. The affinity values range from 0.40 to 15.22 (having applied the normalization explained in Section 3.2.2).

As depicted in Figure 15, hydrophobic interactions are the most predominant, accounting for 94.2% of all interactions (a total of 224,628 hydrophobic interactions). On the contrary, the least common interaction is Aromatic Edge/Face, with only a 0.0159% of all interactions (only 38 Aromatic Edge/Face occurrences).

---

[4] **Pandas library:** https://pandas.pydata.org/

[5] **SQLAlchemy library**: https://www.sqlalchemy.org/

[6] **Flask framework:** https://flask.palletsprojects.com/en/stable/

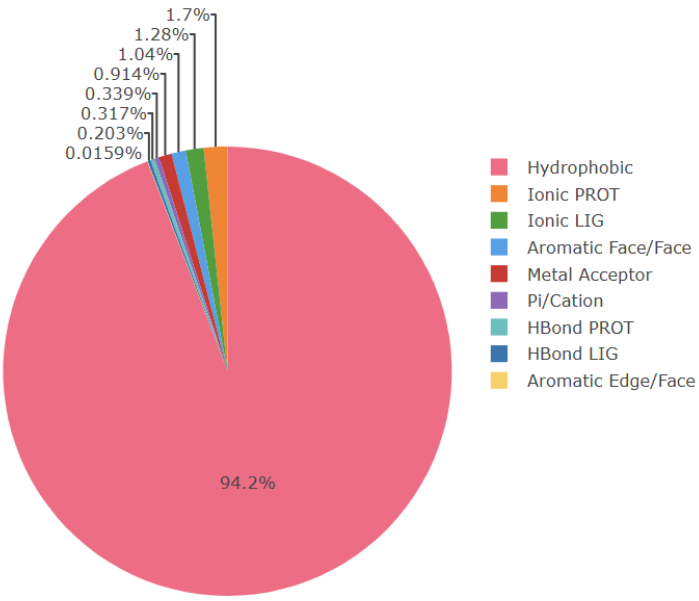


Figure 15. Pie chart indicating the frequency of each interaction type in the whole database.

When analyzing hydrogen bonds, we can observe the distribution and range of bond angles, as shown in Figure 16. The minimum angle is 120.99º and the maximum angle is 178.14º. The most common angles are concentrated around 150º.

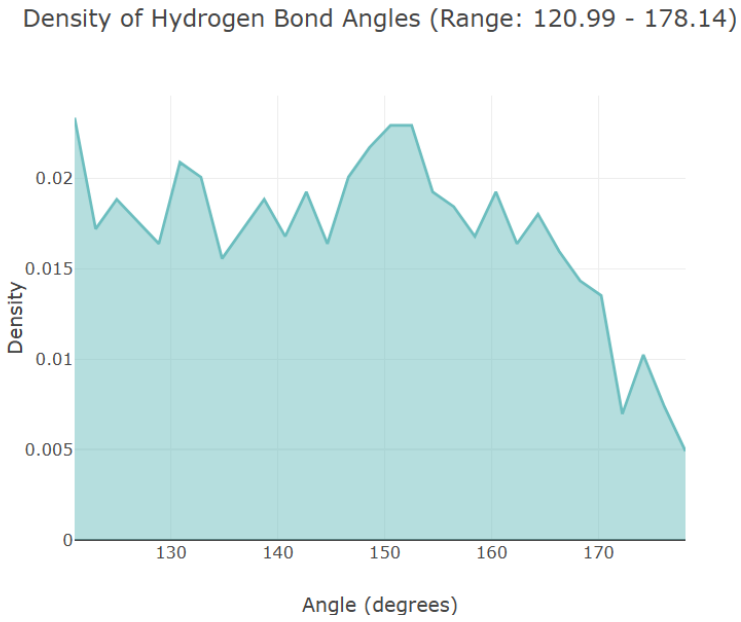


Figure 16. Density of hydrogen bond angles in the whole database. Angles are measured in degrees.

Another interesting aspect to study is the distance between the donor and acceptor in hydrogen bonds. As represented in Figure 17, these distances range from 2.52 Å to 3.48 Å, with the most frequent distances falling between 2.8 Å and 3.0 Å.

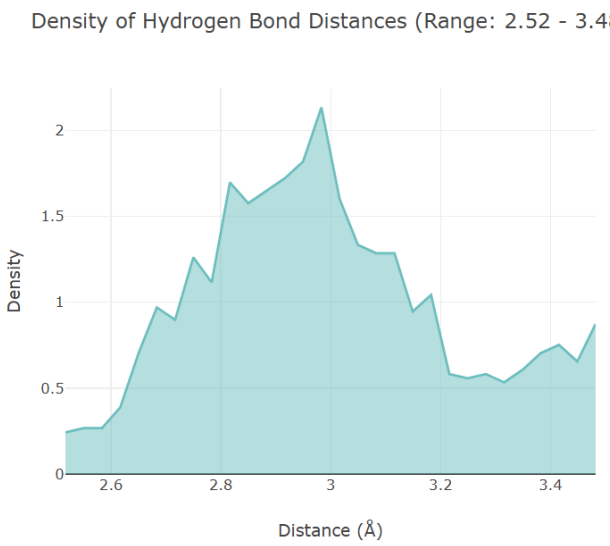


Figure 17. Density of hydrogen bond distances in the whole database. Distances are measured in angstroms.

Finally, Figure 18 illustrates the joint distribution of donor-acceptor distances and bond angles in hydrogen bonds. This plot can be useful to determine which distance-angle combinations are the most prevalent in the database. The regions with the highest density are painted in yellow, corresponding to a bond angle of approximately 150º and a donor-acceptor distance close to 3.0 Å, while lower-density regions are represented by darker colors. The side histograms complement this analysis by showing the distributions of both features separately, pinpointing their respective peaks.

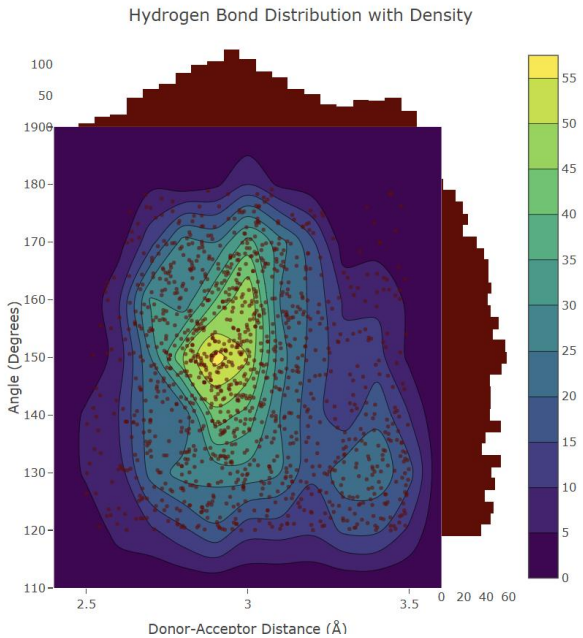


Figure 18. Joint distribution of donor-acceptor distances and bond angles in hydrogen bonds. Complementary histograms are represented to provide a better understanding of the figure.

4.1. Database organization

In order to store all the newly created data, it is necessary to decide how all the information will be related to each other based on the queries that will be performed.

Initially, the database was thought to be structured using two main tables:

1) **Affinity table** (`affinity_data`): table that stores the binding affinity between proteins and their ligands.
2) **Interactions table** (`interaction_data`): table that stores the information described in Section 3.2.

These two tables are supported by the following:

1) **Protein table** (`proteins`): table that stores all protein PDB IDs.
2) **Ligand table** (`ligands`): table that stores all ligand IDs.

Since several PDB IDs may be associated with a same UniProt ID, a complementary table was added, the **Protein Uniprot table** (`proteins_uniprot`). This table has information about the UniProt IDs and their corresponding names, organisms, etc. Finally, in order to link the initial information about the protein PDB IDs with those of the UniProt, the relationship `pdb_uniprot_association` was created. This relationship relates each PDB ID present in the database with the possible UniProt IDs.

Figure 19 illustrates the organization of the information within the database, including the data stored by each table and the relations between tables.

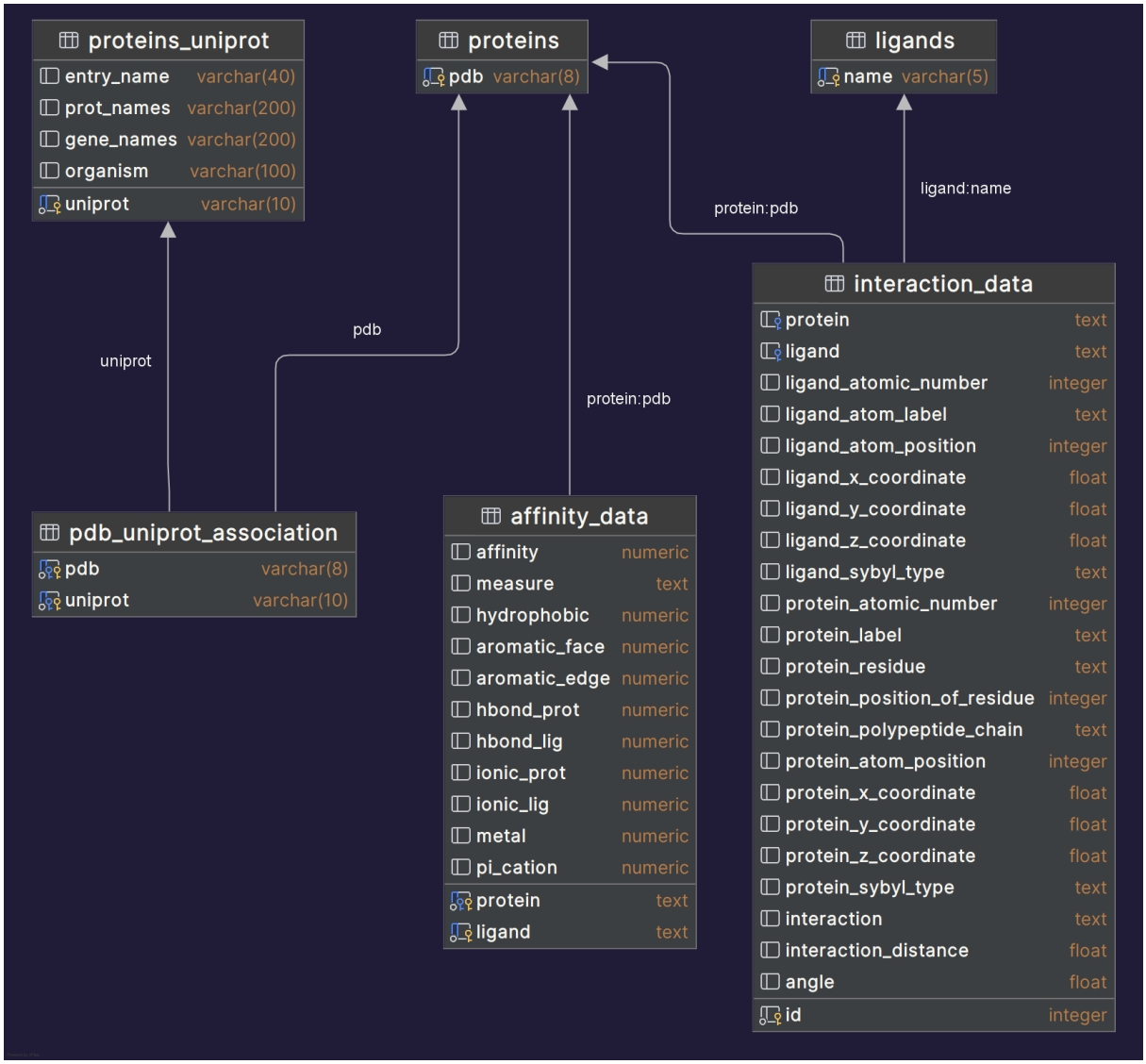


Figure 19. SQL database diagram. Tables and relationships.

4.2. Search web overview

This section provides a brief overview of the webpage that serves as the search engine for the PLII database. The home page presents various search parameters and displays global statistics for the entire database. In Figure 20, the search form is shown. Users can use constraints to filter the search by protein ID (including PDB and UniProt), ligand ID, or affinity value. Users can define a threshold for affinity values and choose between pK_d, pK_i or pIC_{50}.

Additionally, searches can be performed based on some PDB/UniProt IDs written into a text file. Finally, the type of interactions can also be specified as a filter (see Figure 21), also specifying the quantity of each of them (e.g., filter only those protein-ligand

complexes that contain at least 30 hydrophobic interactions and between 2 and 8 aromatic interactions). The global count of interactions and protein-ligand complexes are shown under the search form, with a total of 238,477 interactions and 6,125 protein-ligand complexes.

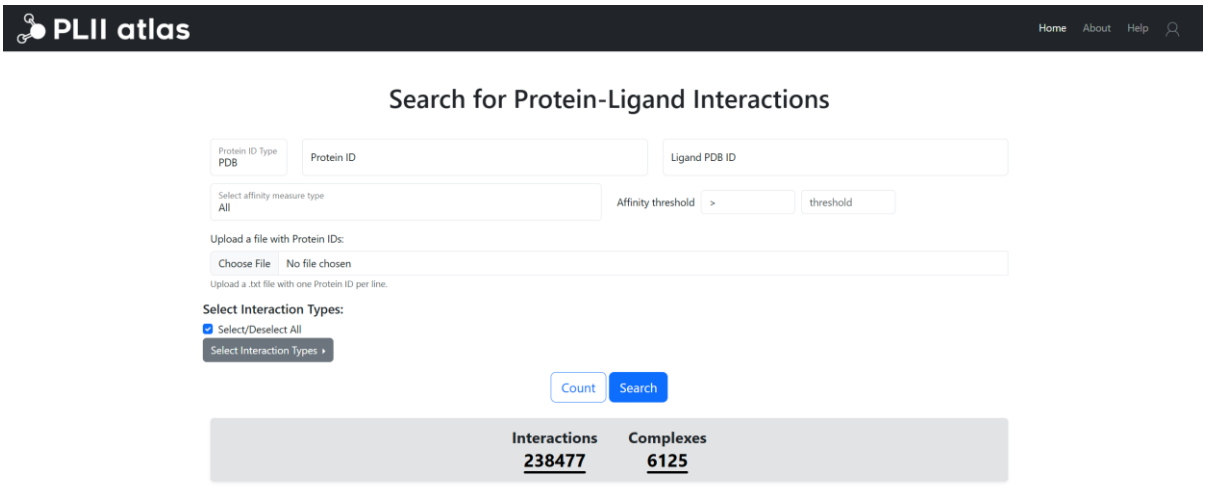


Figure 20. PLII search page (also referred to as home page).

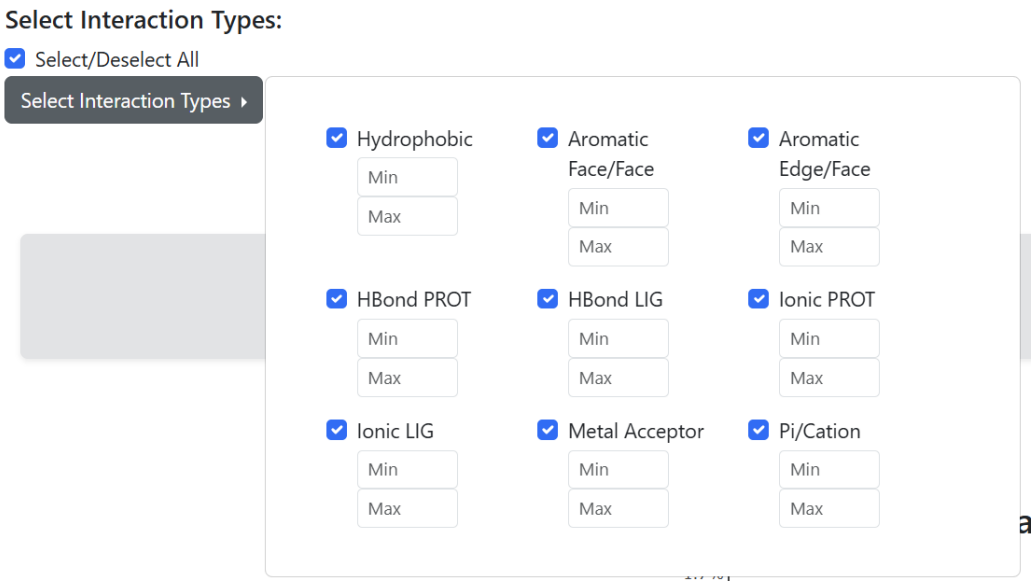


Figure 21. Interaction type filtering. Lower and upper bounds can be defined when filtering by interaction types.

To enhance the **usability** of the webpage, a **Help** section is included (Figure 22). This section includes a basic explanation on how to perform searches and includes practical examples with screenshots, to guide the user throughout the whole process.

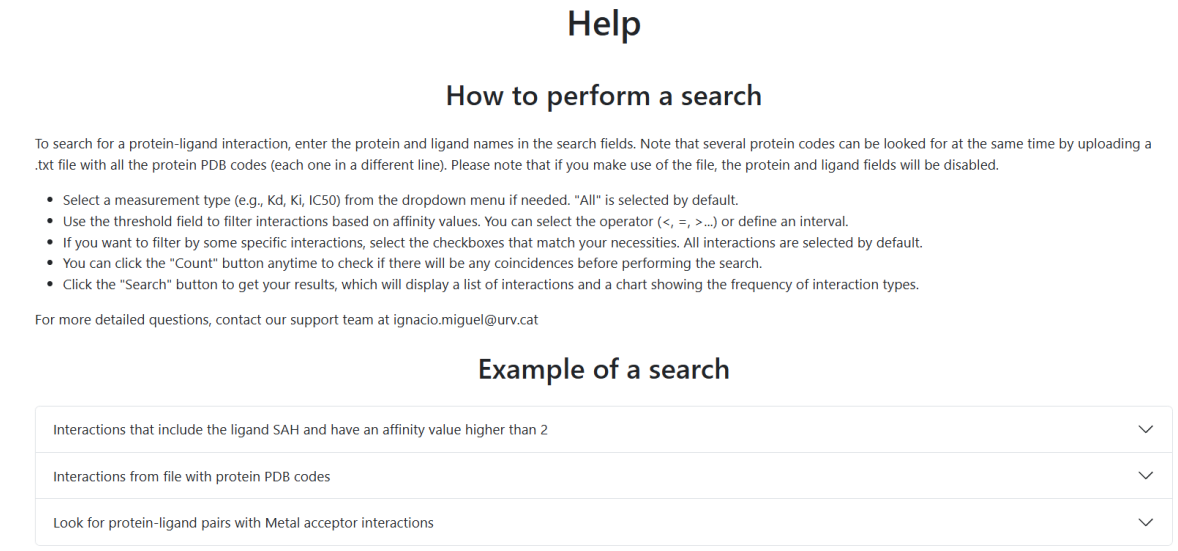


Figure 22. Help page in the PLII database.

For instance, the second example (Figure 23) shows how to perform searches using a `.txt` file. The examples are displayed as a collection of images and text explanations that provide a step-by-step guide.

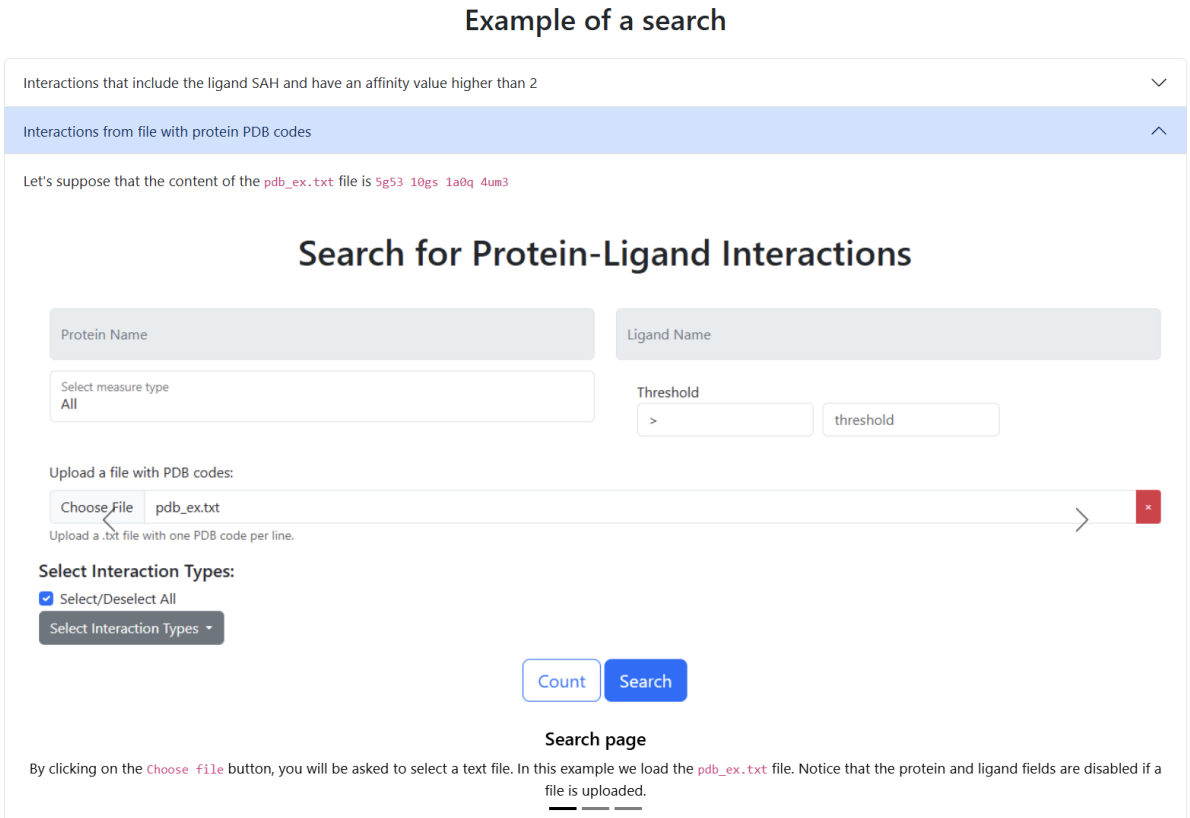


Figure 23. Particular example on how to perform a search in the PLII database.

4.3. Database functionality demonstration

In this section we will answer some questions that can be done to extract information about the protein-ligand complexes and arrive at some conclusions. By doing this, we will also demonstrate the usage of the database and the necessary steps to build up a query.

4.3.1. Analysis of P03367

As an example to illustrate how the database works, the protein with UniProt code **P03367**[7] has been chosen. This protein corresponds to the **Gag-Pol polyprotein** of HIV-1 (Human immunodeficiency virus type 1). Together with Gag polyprotein, Gag-Pol polyprotein mediates the crucial steps in virion assembly, including plasma membrane binding, establishing the necessary protein-protein interactions to create spherical particles, Env protein recruitment, and genomic RNA packaging [31]. The protein's function varies with the concentration: at low concentrations it fosters translation of genomic RNA, while at high concentrations it encapsidates genomic RNA and paralyzes translation. Figure 24 schematically illustrates the assembly, maturation and budding process of HIV-1, in which Gag-Pol polyprotein is highly relevant.

[7] **P03367 UniProt entry:** https://www.uniprot.org/uniprotkb/P03367/entry

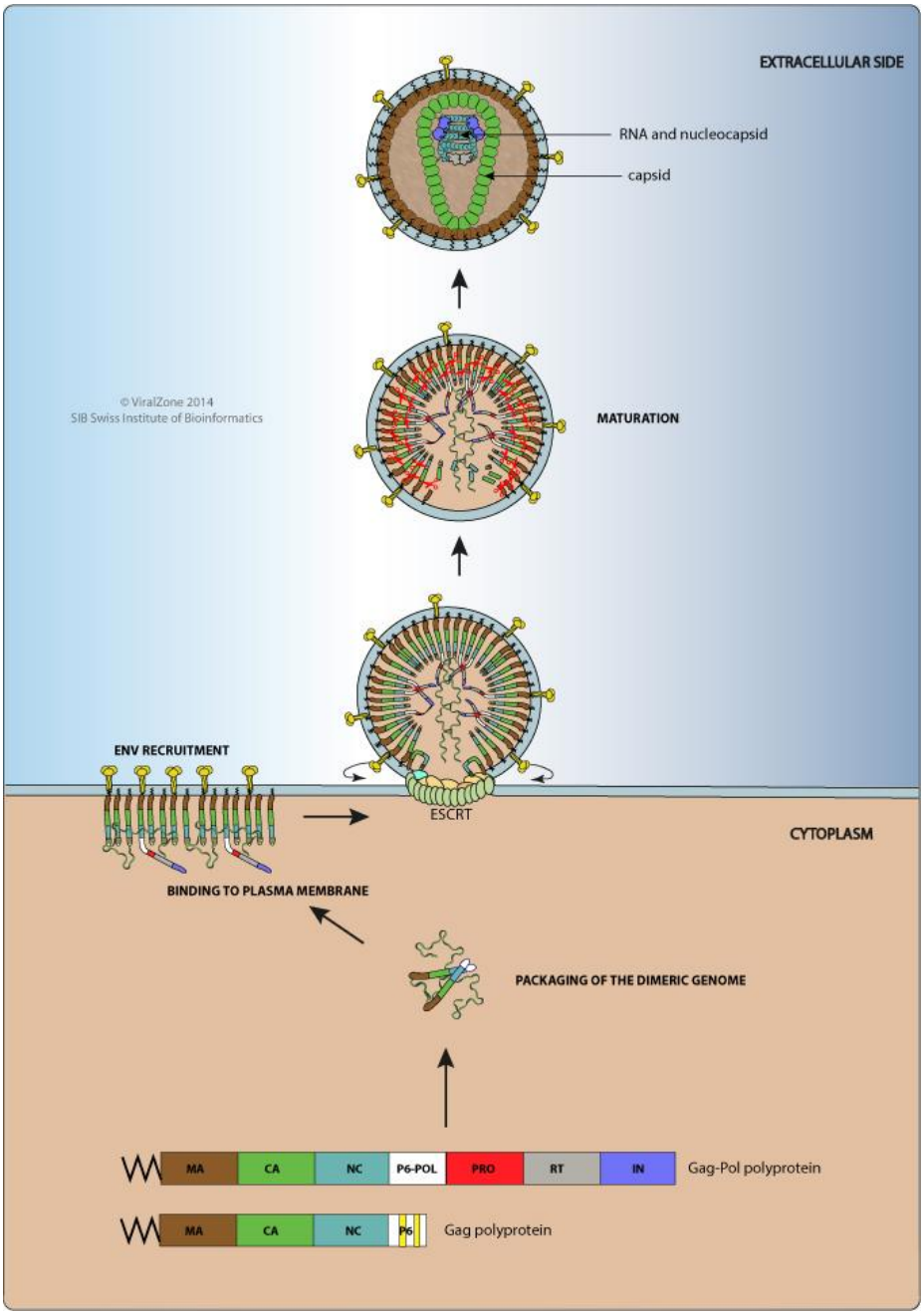


Figure 24. HIV-1 assembly, budding and maturation pathway.

Now, let's conduct a database search to demonstrate the webpage's functionality. On the search page, select **UniProt** as the Protein ID Type and write the UniProt code **P03367** in the Protein ID field. Optionally, we can click the **Count** button to display the number of occurrences for that specific search. This way, we avoid performing searches with no results. In this case, the number of results is 73, as shown in Figure 25 .

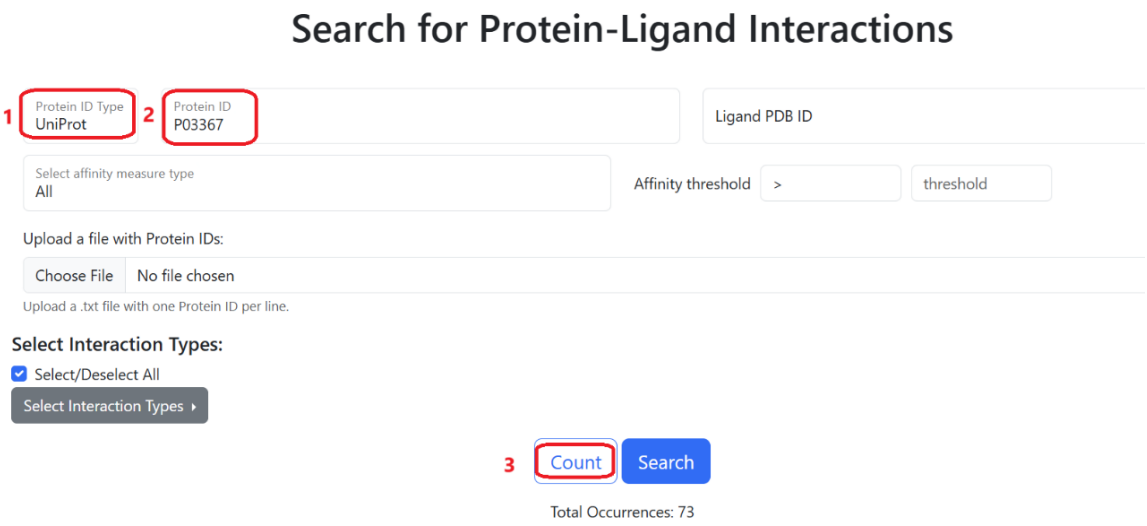


Figure 25. Search protein-ligand complexes by UniProt code (P03367).

To perform the search, click the **Search** button. This action sends a query to the server and retrieves the results accordingly. While the results are being processed, a loading dialog is displayed to provide better feedback to the user, as depicted in Figure 26.

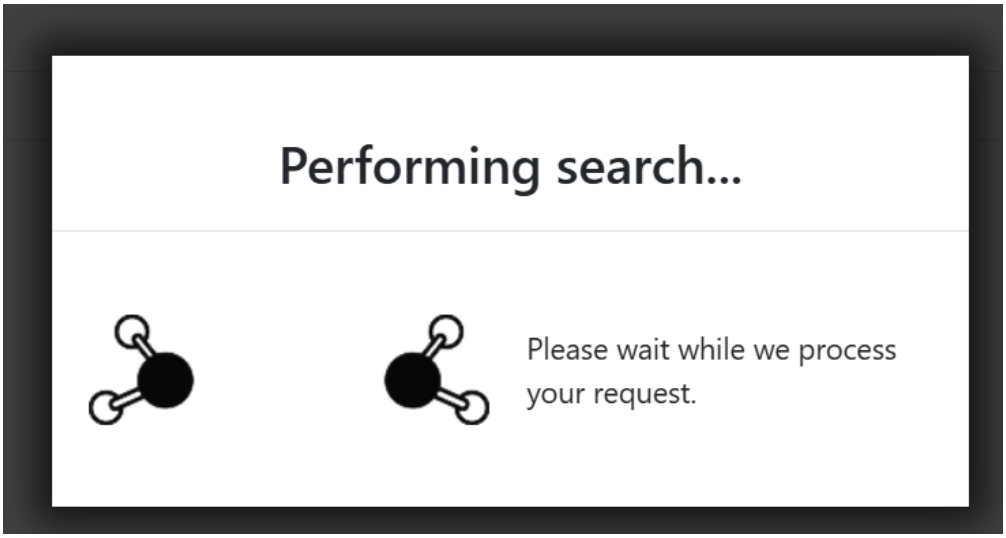


Figure 26. Loading dialog after performing the search shown while waiting for the results.

Upon receiving the information from the server, the browser displays the results in a tabular format. This table includes key information about the protein-ligand complexes matching our query, such as binding affinity and the count of each type of interaction, as can be seen in Figure 27. Additionally, the number of entries in the table corresponds to the number of occurrences obtained with the **Count** button.

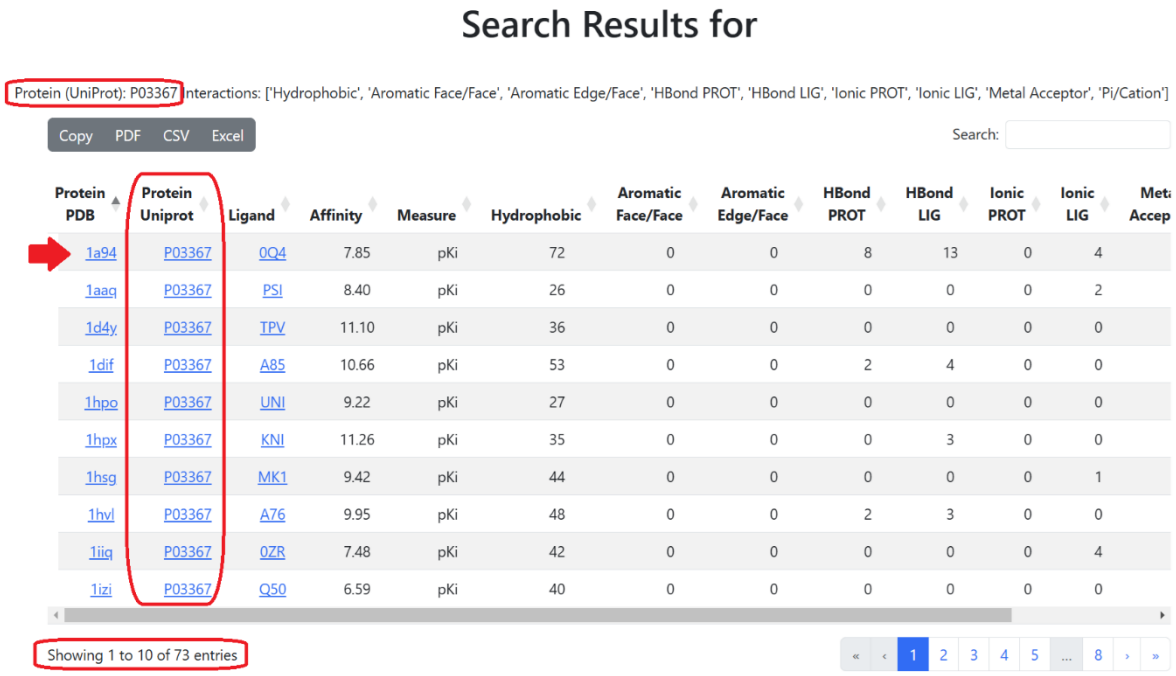


Figure 27. Results for search by UniProt code (P03367).

Now, we can access both protein and ligand information by clicking on the **Protein PDB** or **Ligand** links. For instance, let's visualize the details for the PDB code **1a94**. When opening a protein detailed page (Figure 28) we can observe:

1. **UniProt information (number 1)**: this table lists all UniProt codes associated with the selected PDB code, along with other valuable information like the protein names, gene names or organism.
2. **Protein-ligand interaction details (number 2)**: this section provides information about the ligands interacting with the protein. For each ligand, details such as the ligand name and its binding affinity are displayed. Below this, the main table shows detailed information about the protein-ligand interactions, corresponding to the matrix presented in Section 3.2.1. The table can be exported into several formats (such as PDF or CSV), a feature available for all tables in the database.
3. **Three-dimensional representation (number 3)**: structural representation of the protein and its ligands. The structure is directly downloaded from the PDB and rendered on the webpage.

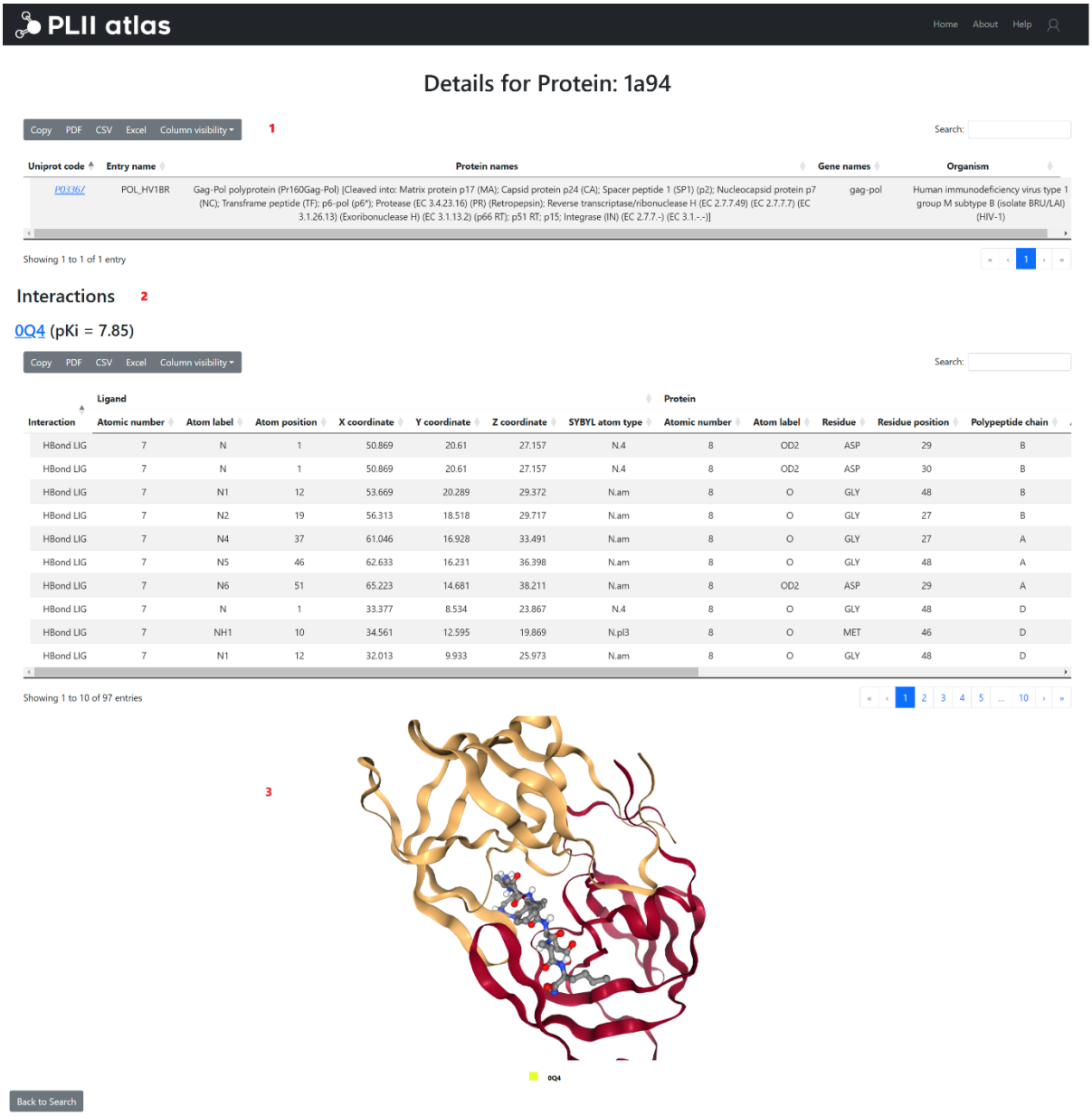


Figure 28. Detailed information page for protein with PDB code 1a94.

By clicking on the ligand code, we can discover if the database tracks other complexes involving the selected ligand (**0Q4**). In Figure 29, number 1 points at the table that lists all recorded binding affinities for this ligand. From this table, we can identify the PDB codes associated with this ligand and the corresponding binding affinities. Number 2 highlights the section where the 3D structure of the ligand is displayed.

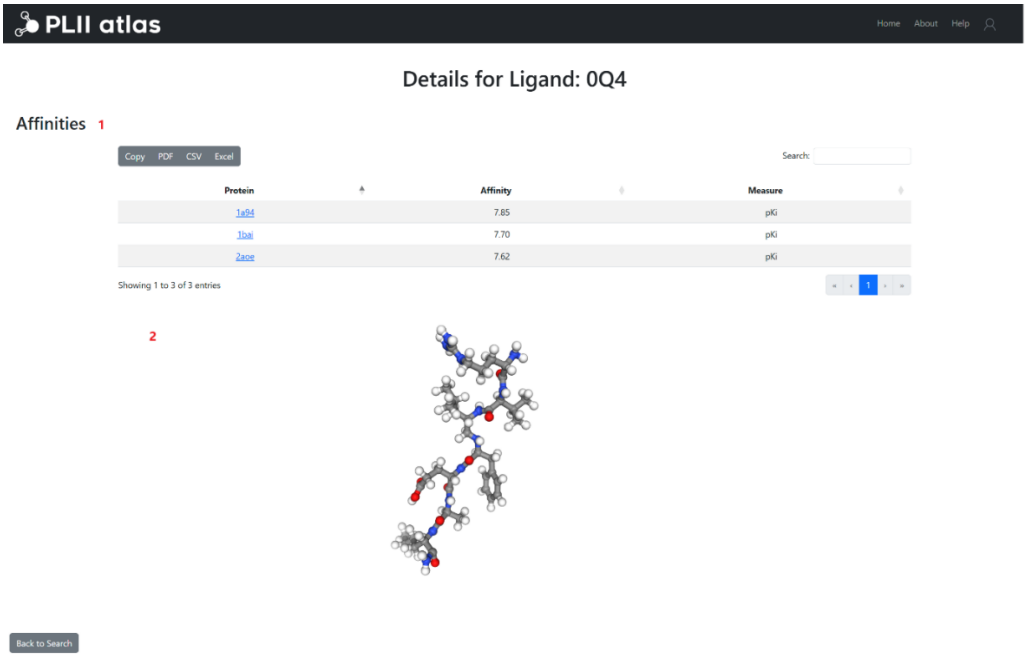


Figure 29. Detailed information page for ligand 0Q4.

On the **results page** (Figure 27) we can find a statistical section below the protein-ligand table under the name of "**Global stats**". This section offers a comprehensive comparison between the whole database and the current search results. This comparative analysis allows users to arrive at valuable conclusions about the target study structures, such as identifying trends.

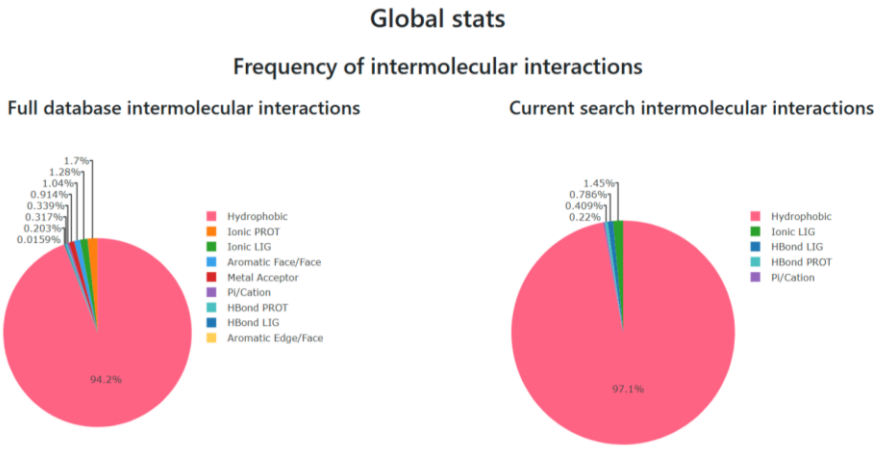


Figure 30. Comparison of intermolecular interaction frequencies between the entire database and the current search results for UniProt code P03367.

Figure 30 compares the frequency of each type of interaction between the whole database and the current search. We can deduce that **hydrophobic interactions** are the most prevalent interactions in both datasets, with over 90% of the total interaction count.

However, significant differences are evident in other interaction types. **Aromatic interactions** (e.g., Face/Face or Edge/Face) and **metal acceptor interactions** are completely absent in the current search results. Apart from that, **ionic interactions** and **hydrogen bonds** are also represented in the search results, foretelling that most of the resulting protein-ligand complexes results are likely to exhibit high affinities, as these interactions are not among the weakest.

Figure 31 shows the comparison of the frequency of atom hybridizations between the whole database and the current search. For proteins, the aromatic carbon type (**C.ar**) appears significantly less compared to the full database. On the other hand, the carboxyl oxygen (**O.co2**) gains more importance in the current results.

Furthermore, the current results illustrate an absence of metal ions, such as magnesium (**Mg2+**) or zinc (**Zn2+**), which are slightly present in the full database. This lack of metal atoms suggests that the proteins in the current search may not include metalloproteins or structures bound to metal ions.

Halogens like chlorine (**Cl**) and fluorine (**F**) have less presence in the current search dataset for ligands, which points out that the ligands in this query are less likely to contain halogenated functional groups.

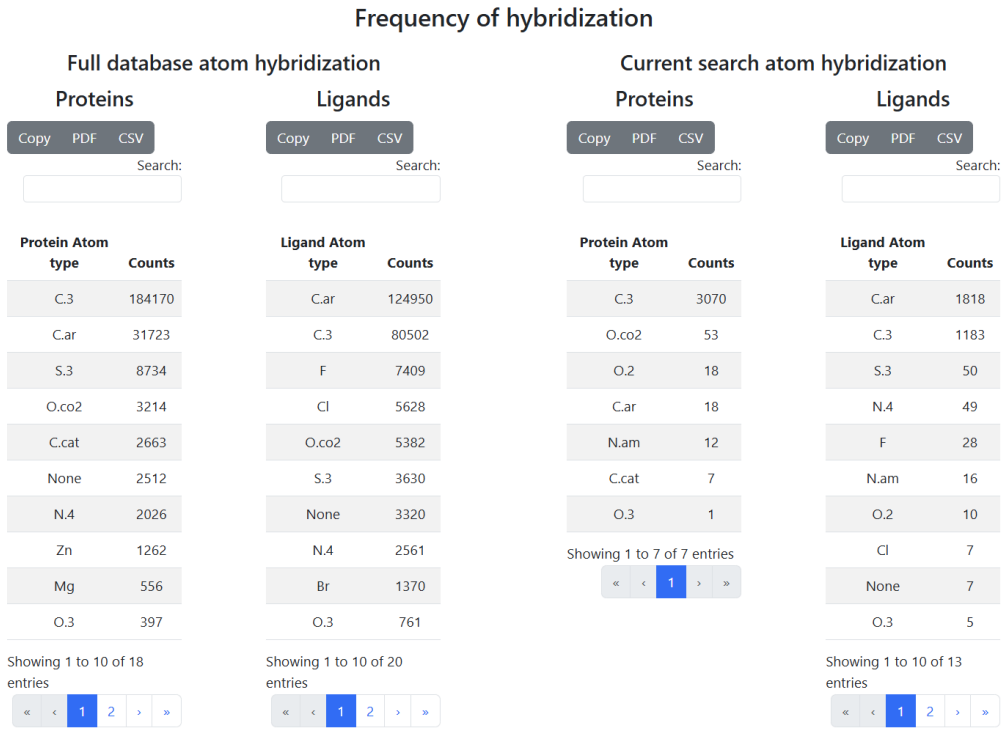


Figure 31. Comparison of atom hybridization frequencies between the entire database and the current search results.

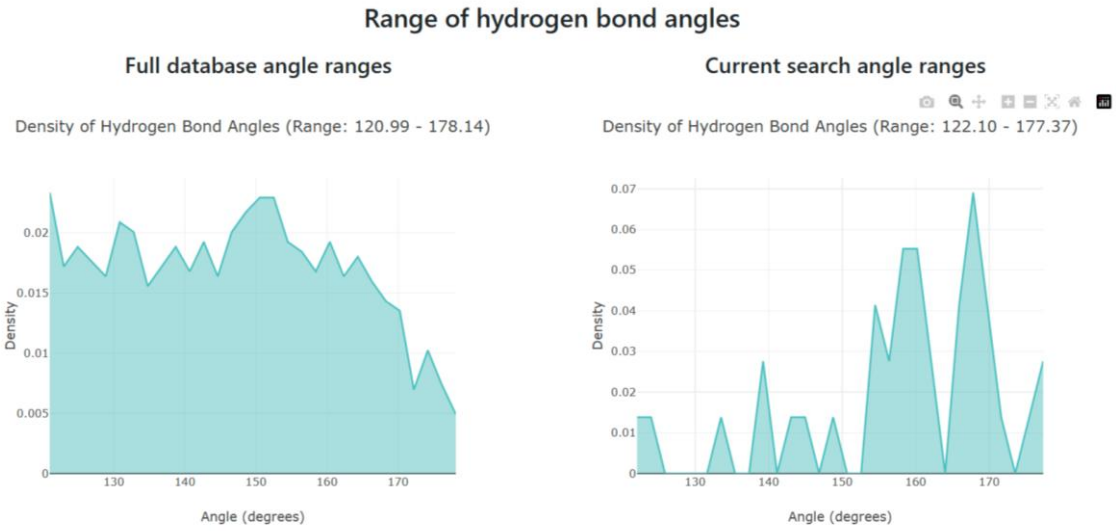


Figure 32. Comparison of hydrogen bond angle ranges: full database vs. current search.

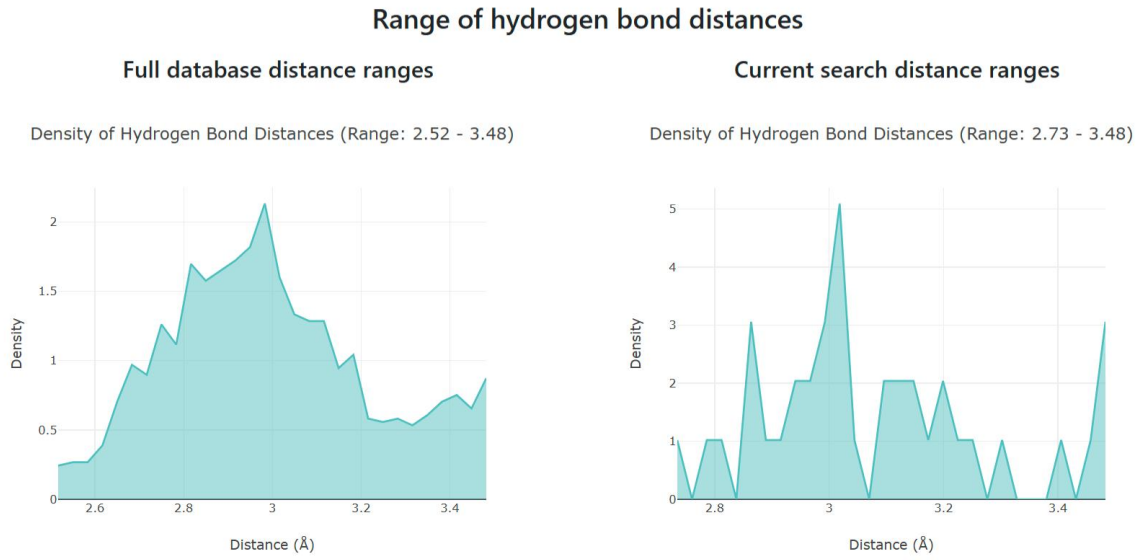


Figure 33. Comparison of hydrogen bond angle distances: full database vs. current search.

Analyzing Figure 32 and Figure 33 we get almost no new insights, since the ranges of both hydrogen bond angles and distances closely matches those of the full database, with only a slight increase in the minimum bound.

Nevertheless, a remarkable difference is observed in the density distribution of hydrogen bond angles. The current results **show peaks around 170º**, which differs significantly from the peaks found in the full database plot and that clarifies that the searched protein establishes strong hydrogen bonds with its ligands.

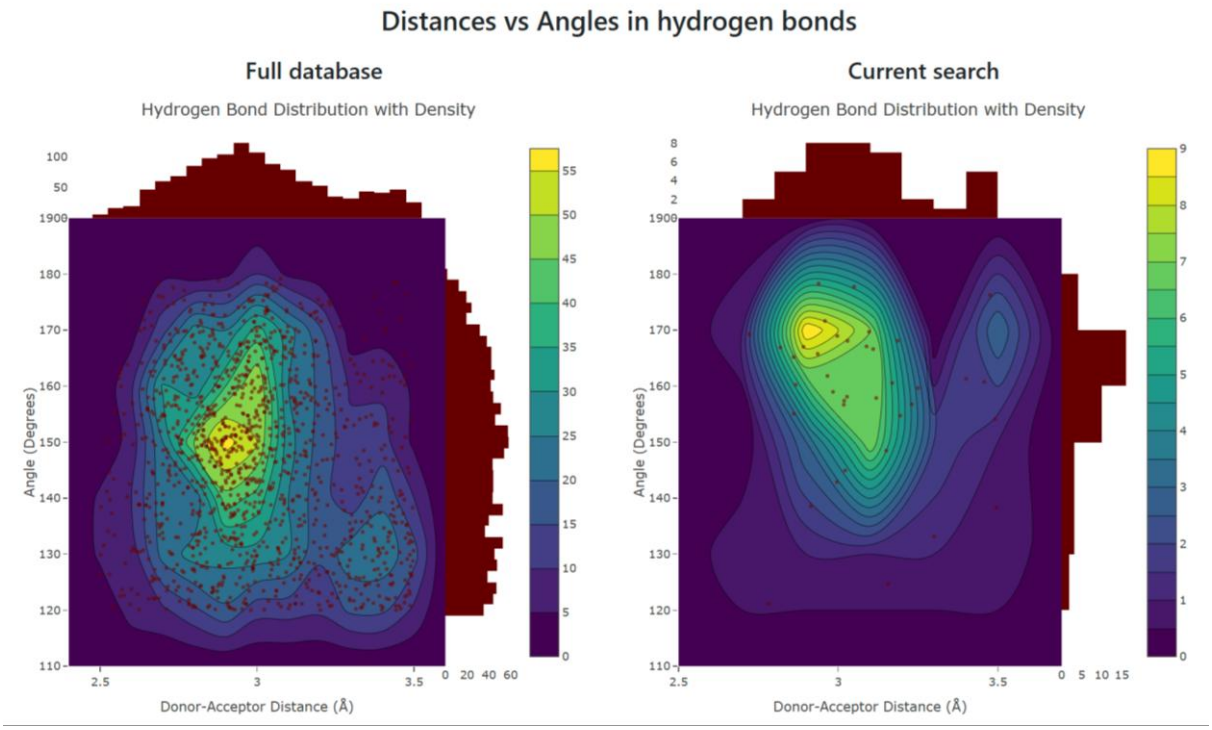


Figure 34. Comparison of hydrogen bond distance and angle distributions: full database vs. current search with density contour visualization.

Figure 34 helps us clearly identify that the hydrogen bond distance-angle comparison does not follow exactly the same fashion as the whole database, but is rather **concentrated around 170º and 2.9 Å**, with a smaller cluster near 170º and 3.5 Å. Notably, the latter represents a low-density region in the global database statistics. Interestingly, the higher density regions in yellow and green coincide with the angles present in stronger hydrogen bonds, as stated in the introduction of this work.

According to [32], "high-affinity binding was defined K_d, K_i, or $IC_{50} \leq 250$ nM" which translates to a normalized logarithmic scale value of 6.6. Based on this value we can refine the search to include only results with high affinities. To do so, just select the appropriate threshold in the search form (Figure 35). Out of the 73 total results obtained previously, **60** meet the condition of having a high affinity.

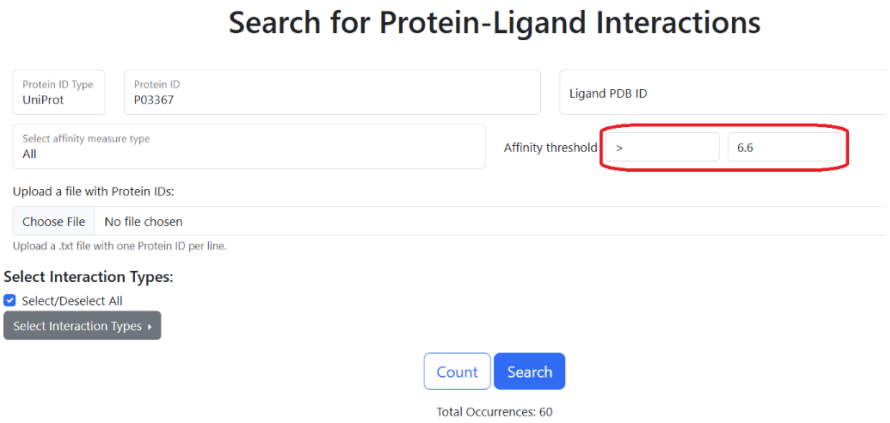
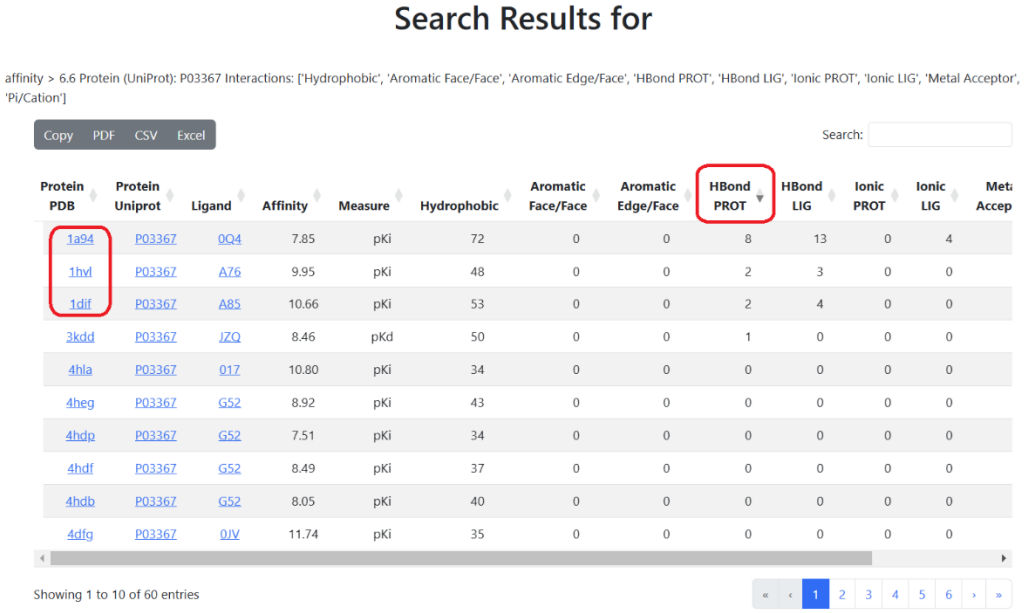


Figure 35. Search refinement including a threshold for the binding affinity to be greater than 6.6.

After filtering only those protein-ligand complexes with high affinities, let's try to define a **pharmacophore** for the **Gag-Pol polyprotein (P03367)**. Using the filtered search results, we analyze the interactions within the identified complexes. Since the number of hydrophobic interactions is very high, a manual analysis would be infeasible. Instead, we focus on those interactions that are likely critical and occur with low enough frequency to allow for detailed manual examination.

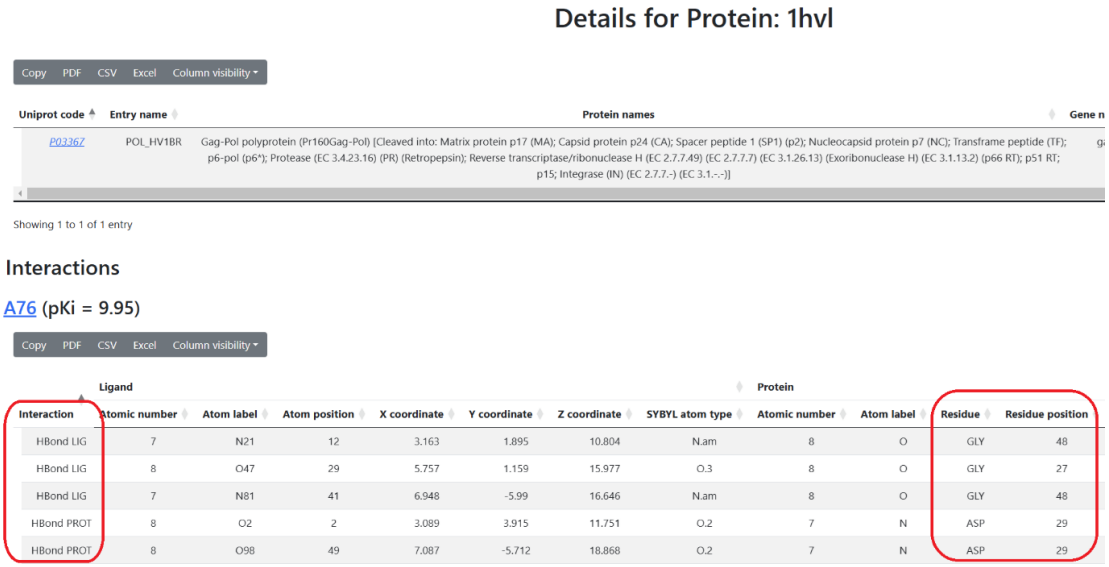
The results table can be sorted in a descending order based on interaction types (Figure 36). This approach helps to easily identify the complexes with the highest counts of specific interaction types. By opening multiple windows for the most representative complexes, we can perform a comparative analysis to identify which **interactions are common** to all complexes.



Protein PDB	Protein Uniprot	Ligand	Affinity	Measure	Hydrophobic	Aromatic Face/Face	Aromatic Edge/Face	HBond PROT	HBond LIG	Ionic PROT	Ionic LIG	Meta Accep
1a94	P03367	0Q4	7.85	pKi	72	0	0	8	13	0	4	
1hvl	P03367	A76	9.95	pKi	48	0	0	2	3	0	0	
1dif	P03367	A85	10.66	pKi	53	0	0	2	4	0	0	
3kdd	P03367	JZQ	8.46	pKd	50	0	0	1	0	0	0	
4hla	P03367	017	10.80	pKi	34	0	0	0	0	0	0	
4heg	P03367	G52	8.92	pKi	43	0	0	0	0	0	0	
4hdp	P03367	G52	7.51	pKi	34	0	0	0	0	0	0	
4hdf	P03367	G52	8.49	pKi	37	0	0	0	0	0	0	
4hdb	P03367	G52	8.05	pKi	40	0	0	0	0	0	0	
4dfg	P03367	0JV	11.74	pKi	35	0	0	0	0	0	0	

Figure 36. Filtered results with affinity threshold greater than 6.6, for protein P03367.

Complexes **1a94-0Q4**, **1hvl-A76**, **1dif-A85**, **1hpx-KNI**, **2qnn-QN1** and **1m0b-0ZQ** were analyzed.



Details for Protein: 1hvl

Copy PDF CSV Excel Column visibility

Uniprot code	Entry name	Protein names	Gene n
P03367	POL_HV1BR	Gag-Pol polyprotein (Pr160Gag-Pol) [Cleaved into: Matrix protein p17 (MA); Capsid protein p24 (CA); Spacer peptide 1 (SP1) (p2); Nucleocapsid protein p7 (NC); Transframe peptide (TF); p6-pol (p6*); Protease (EC 3.4.23.16) (PR) (Retropepsin); Reverse transcriptase/ribonuclease H (EC 2.7.7.49) (EC 2.7.7.7) (EC 3.1.26.13) (Exoribonuclease H) (EC 3.1.13.2) (p66 RT); p51 RT; p15; Integrase (IN) (EC 2.7.7.-) (EC 3.1.-.-)]	g

Showing 1 to 1 of 1 entry

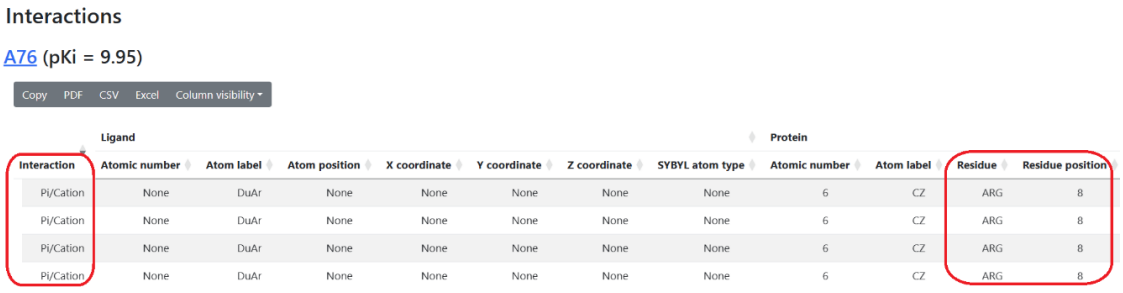
Interactions

A76 (pKi = 9.95)

Copy PDF CSV Excel Column visibility

	Ligand							Protein			
Interaction	Atomic number	Atom label	Atom position	X coordinate	Y coordinate	Z coordinate	SYBYL atom type	Atomic number	Atom label	Residue	Residue position
HBond LIG	7	N21	12	3.163	1.895	10.804	N.am	8	O	GLY	48
HBond LIG	8	O47	29	5.757	1.159	15.977	O.3	8	O	GLY	27
HBond LIG	7	N81	41	6.948	-5.99	16.646	N.am	8	O	GLY	48
HBond PROT	8	O2	2	3.089	3.915	11.751	O.2	7	N	ASP	29
HBond PROT	8	O98	49	7.087	-5.712	18.868	O.2	7	N	ASP	29

Figure 37. Hydrogen bond interactions of complex 1hvl-A76.



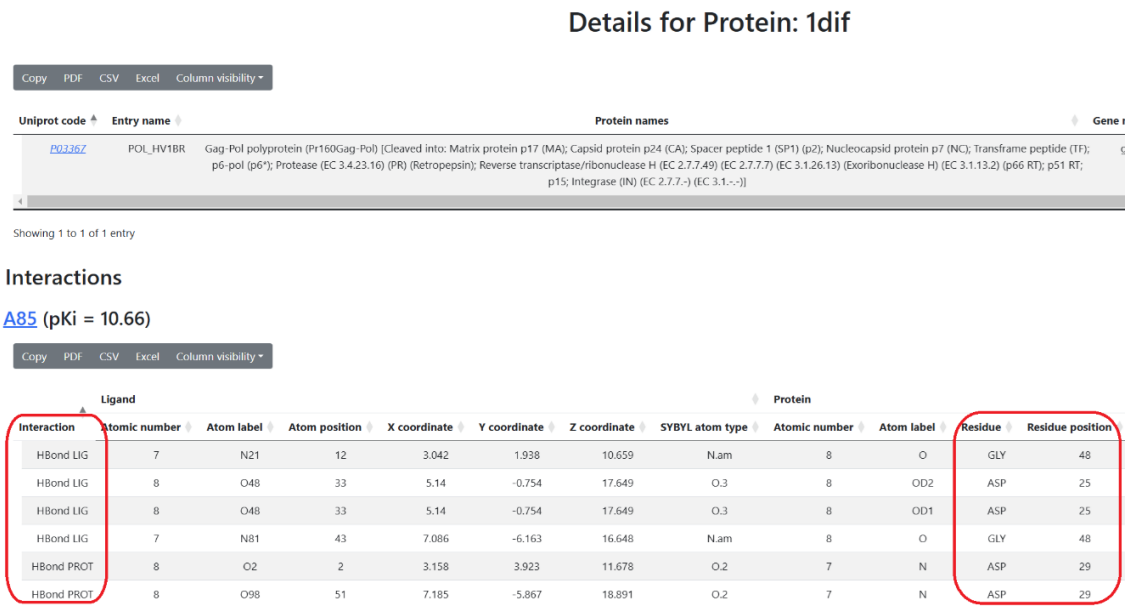
Interactions

A76 (pKi = 9.95)

Copy PDF CSV Excel Column visibility

	Ligand							Protein			
Interaction	Atomic number	Atom label	Atom position	X coordinate	Y coordinate	Z coordinate	SYBYL atom type	Atomic number	Atom label	Residue	Residue position
Pi/Cation	None	DuAr	None	None	None	None	None	6	CZ	ARG	8
Pi/Cation	None	DuAr	None	None	None	None	None	6	CZ	ARG	8
Pi/Cation	None	DuAr	None	None	None	None	None	6	CZ	ARG	8
Pi/Cation	None	DuAr	None	None	None	None	None	6	CZ	ARG	8

Figure 38. Pi/Cation interactions of complex 1hvl-A76



Details for Protein: 1dif

Copy PDF CSV Excel Column visibility

Uniprot code	Entry name	Protein names	Gene
P03367	POL_HV1BR	Gag-Pol polyprotein (Pr160Gag-Pol) [Cleaved into: Matrix protein p17 (MA); Capsid protein p24 (CA); Spacer peptide 1 (SP1) (p2); Nucleocapsid protein p7 (NC); Transframe peptide (TF); p6-pol (p6*); Protease (EC 3.4.23.16) (PR) (Retropepsin); Reverse transcriptase/ribonuclease H (EC 2.7.7.49) (EC 2.7.7.7) (EC 3.1.26.13) (Exoribonuclease H) (EC 3.1.13.2) (p66 RT); p51 RT; p15; Integrase (IN) (EC 2.7.7.-) (EC 3.1.-.-)]	g

Showing 1 to 1 of 1 entry

Interactions

A85 (pKi = 10.66)

Copy PDF CSV Excel Column visibility

	Ligand							Protein			
Interaction	Atomic number	Atom label	Atom position	X coordinate	Y coordinate	Z coordinate	SYBYL atom type	Atomic number	Atom label	Residue	Residue position
HBond LIG	7	N21	12	3.042	1.938	10.659	N.am	8	O	GLY	48
HBond LIG	8	O48	33	5.14	-0.754	17.649	O.3	8	OD2	ASP	25
HBond LIG	8	O48	33	5.14	-0.754	17.649	O.3	8	OD1	ASP	25
HBond LIG	7	N81	43	7.086	-6.163	16.648	N.am	8	O	GLY	48
HBond PROT	8	O2	2	3.158	3.923	11.678	O.2	7	N	ASP	29
HBond PROT	8	O98	51	7.185	-5.867	18.891	O.2	7	N	ASP	29

Figure 39. Hydrogen bond interactions of complex 1dif-A85.

Interactions

A85 (pKi = 10.66)

Copy PDF CSV Excel Column visibility

	Ligand							Protein			
Interaction	Atomic number	Atom label	Atom position	X coordinate	Y coordinate	Z coordinate	SYBYL atom type	Atomic number	Atom label	Residue	Residue position
Pi/Cation	None	DuAr	None	None	None	None	None	6	CZ	ARG	8
Pi/Cation	None	DuAr	None	None	None	None	None	6	CZ	ARG	8
Pi/Cation	None	DuAr	None	None	None	None	None	6	CZ	ARG	8

Figure 40. Pi/Cation interactions of complex 1dif-A85.

Details for Protein: 1hpx

Copy PDF CSV Excel Column visibility

Uniprot code	Entry name	Protein names	Gene
P03367	POL_HV1BR	Gag-Pol polyprotein (Pr160Gag-Pol) [Cleaved into: Matrix protein p17 (MA); Capsid protein p24 (CA); Spacer peptide 1 (SP1) (p2); Nucleocapsid protein p7 (NC); Transframe peptide (TF); p6-pol (p6*); Protease (EC 3.4.23.16) (PR) (Retropepsin); Reverse transcriptase/ribonuclease H (EC 2.7.7.49) (EC 2.7.7.7) (EC 3.1.26.13) (Exoribonuclease H) (EC 3.1.13.2) (p66 RT); p51 RT; p15; Integrase (IN) (EC 2.7.7.-) (EC 3.1.-.-)]	

Showing 1 to 1 of 1 entry

Interactions

KNI (pKi = 11.26)

Copy PDF CSV Excel Column visibility

	Ligand							Protein			
Interaction	Atomic number	Atom label	Atom position	X coordinate	Y coordinate	Z coordinate	SYBYL atom type	Atomic number	Atom label	Residue	Residue position
HBond LIG	7	N2	15	3.955	2.384	10.893	N.am	8	O	GLY	48
HBond LIG	7	N3	23	4.496	0.982	13.492	N.am	8	O	GLY	27
HBond LIG	8	O2	30	4.949	0.934	16.427	O.3	8	OD2	ASP	25

Figure 41. Hydrogen bond interactions of complex 1hpx-KNI.

Details for Protein: 2qnn

Copy PDF CSV Excel Column visibility

Uniprot code	Entry name	Protein names	Gene r
P03367	POL_HV1BR	Gag-Pol polyprotein (Pr160Gag-Pol) [Cleaved into: Matrix protein p17 (MA); Capsid protein p24 (CA); Spacer peptide 1 (SP1) (p2); Nucleocapsid protein p7 (NC); Transframe peptide (TF); p6-pol (p6*); Protease (EC 3.4.23.16) (PR) (Retropepsin); Reverse transcriptase/ribonuclease H (EC 2.7.7.49) (EC 2.7.7.7) (EC 3.1.26.13) (Exoribonuclease H) (EC 3.1.13.2) (p66 RT); p51 RT; p15; Integrase (IN) (EC 2.7.7.-) (EC 3.1.-.-)]	

Showing 1 to 1 of 1 entry

Interactions

QN1 (pKi = 7.15)

Copy PDF CSV Excel Column visibility

	Ligand							Protein			
Interaction	Atomic number	Atom label	Atom position	X coordinate	Y coordinate	Z coordinate	SYBYL atom type	Atomic number	Atom label	Residue	Residue position
Ionic LIG	7	N22	26	14.13	-18.167	-18.2	N.4	8	OD1	ASP	25
Ionic LIG	7	N22	26	14.13	-18.167	-18.2	N.4	8	OD2	ASP	25
Ionic LIG	7	N22	26	14.13	-18.167	-18.2	N.4	8	OD1	ASP	25
Ionic LIG	7	N22	26	14.13	-18.167	-18.2	N.4	8	OD2	ASP	25

Figure 42. Ionic interactions of complex 2qnn-QN1.

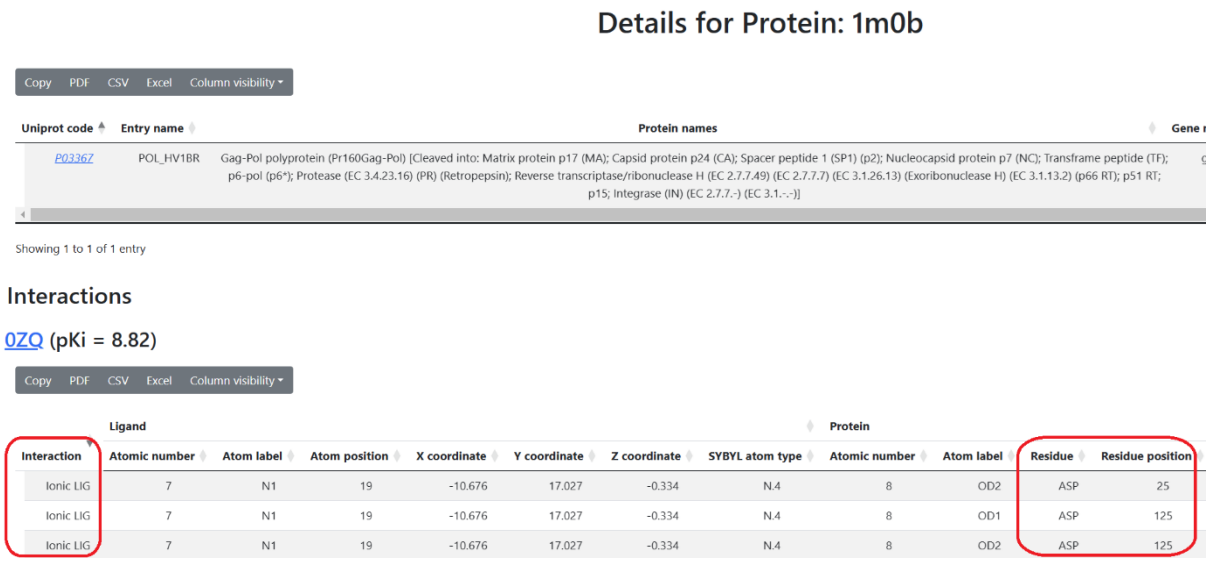


Figure 43. Ionic interactions of complex 1m0b-0ZQ.

After performing the manual comparison based on the information shown in Figures 28, 37-43, I could identify the following key interactions that could be used to define a pharmacophore:

Table 10. Key interactions found to define a pharmacophore for protein P03367.

Interaction	Protein residue
Hydrogen bond	ASP29, GLY27, GLY48
Pi/cation	ARG8
Ionic interaction	ASP25

We can use **interaction representation tools like PDBe**[8]. Interactions are computed with the **Arpeggio**[9] software, which may differ slightly from the ones computed by iChem. For instance, Figure 44 illustrates the interactions between the **1a94** PDB code protein and the **0Q4** ligand. The residues specified in Table 10 are highlighted as red squares for easier visualization.

[8] **Protein Data Bank in Europe (PDBe):** https://www.ebi.ac.uk/pdbe/

[9] **Arpeggio:** https://github.com/PDBeurope/arpeggio

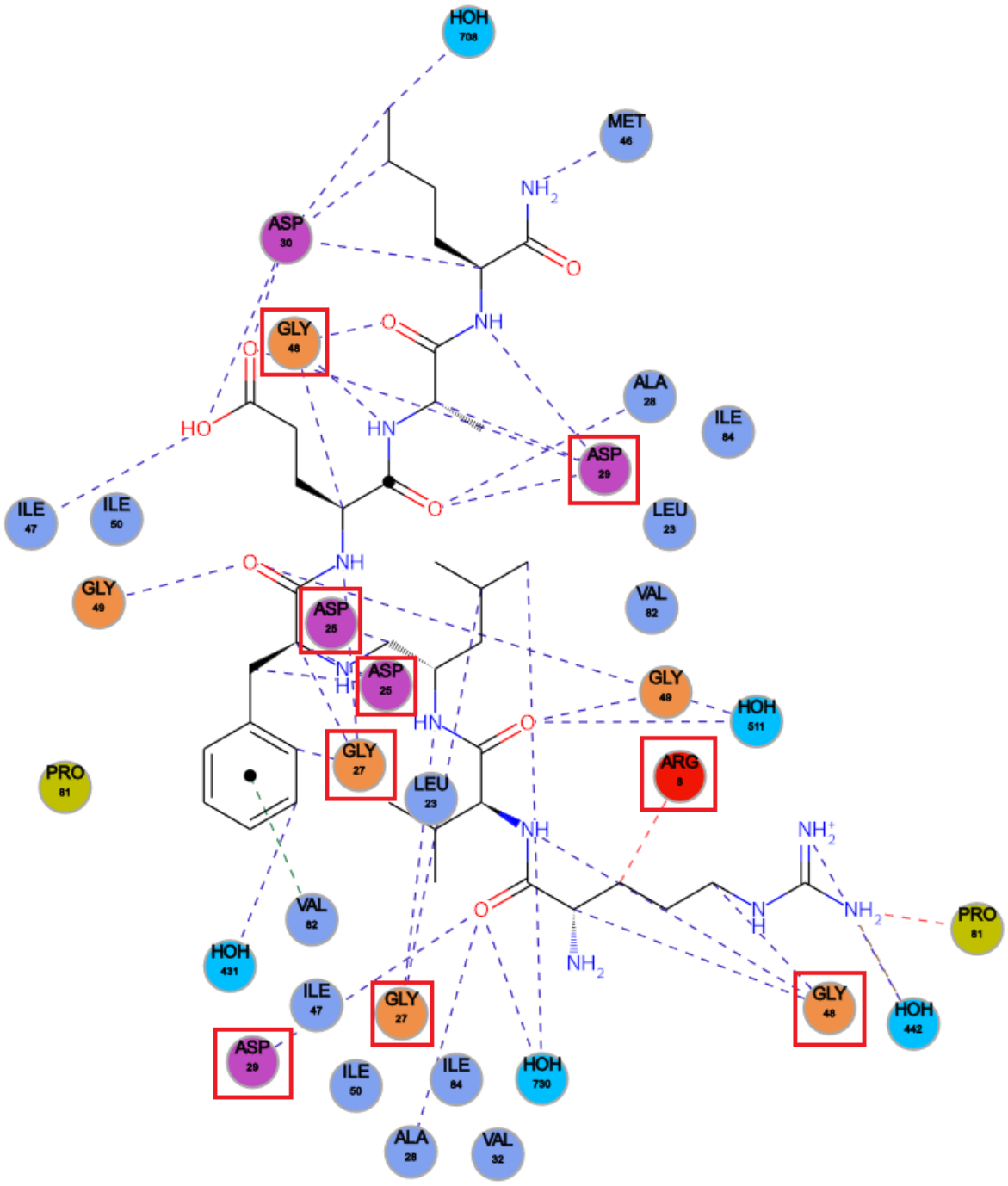


Figure 44. Interactions between protein 1a94 and ligand 0Q4. Source: https://www.ebi.ac.uk/pdbe/entry/pdb/1a94/bound/0Q4

5. Conclusions

Protein-ligand interaction and affinity data is pivotal to the development of new drugs in CADD. Agglutinating all this information into a centralized database has proven to be really useful in the extraction of insights into hybridization types, the most prevalent interactions, and the nature of hydrogen bonds present in protein-ligand complexes.

The database, developed using Python, is designed to be deployed into a server hosted by the Cheminformatics and Nutrition research group at URV for public use. The webpage from which this database will be accessible allows for personalized searches. Users can apply filters based on the protein, ligand, interaction types, a binding affinity threshold, or even upload files with protein codes, among other options. These filters enhance efficiency and allow for more specific and targeted queries.

To demonstrate the database's functionality, the Gag-Pol polyprotein (UniProt code P03367) was analyzed. The results revealed that its interactions were mostly hydrophobic and ionic interactions as well as hydrogen bonds, while lacked others such as metal acceptors or aromatic interactions. A closer examination of the hydrogen bonds showed that the angles were distributed mostly around 170º, which indicates high affinities for those protein-ligand complexes. Moreover, the manual exploration of the filtered results allowed for the definition of a pharmacophore, which can be used in CADD for the development of potential therapeutic compounds.

5.1. Future work

In future iterations of this work, we propose to include support for other emerging interaction-computing tools like PDBe Arpeggio. Providing support for various formats will help our application to be more versatile and interoperative. Moreover, the database could provide a tool to find the most common interactions among a specific search to easily identify pharmacophores. This is particularly helpful and time-saving considering the huge amount of data that should be analyzed by hand. Finally, the deployment of the server must be done in the headquarters of the Cheminformatics and Nutrition research group at URV.

6. Self-assessment

First, I would like to thank Dr. Santiago Garcia and Dr. Gerard Pujadas for welcoming me into the Cheminformatics and Nutrition research group and for offering me the opportunity to take part in a project with them. We communicated periodically to share the updates and discuss about which steps should be taken next.

The data preprocessing process took longer than initially expected due to the lack of uniform format in the input files and the discovery of additional cases not covered in the user guide of IChem. This made us take decisions on how to process data appropriately to avoid loss of information and improve the overall quality of the database.

The development of this protein-ligand interaction database has helped me learn about concepts that I wasn't taught about during the degree. Since this work is about bioinformatics, I have been able to use my knowledge in both the computer science and the biotechnology fields, arriving at a synergy that has played a key role in the success of the project.

The expertise and guidance of my academic tutors has also helped me gain insights into concepts like Computer-Aided Drug Design (CADD) and the importance of knowing protein-ligands interactions in order to better understand and predict new drugs.

Finally, this final degree thesis has taught me how to conduct research in cutting-edge technologies, and it has provided me with the opportunity to discover a fascinating and exciting field that I wouldn't have been aware of otherwise.

Bibliography

[1] X. Du *et al.*, “Insights into protein–ligand interactions: Mechanisms, models, and methods,” Jan. 26, 2016, *MDPI AG*. doi: 10.3390/ijms17020144.

[2] T. I. Chandel *et al.*, “A mechanistic insight into protein-ligand interaction, folding, misfolding, aggregation and inhibition of protein aggregates: An overview,” Jan. 01, 2018, *Elsevier B.V.* doi: 10.1016/j.ijbiomac.2017.07.185.

[3] D. Herschlag and M. M. Pinney, “Hydrogen Bonds: Simple after All?,” Jun. 19, 2018, *American Chemical Society*. doi: 10.1021/acs.biochem.8b00217.

[4] J. Boyle, “Lehninger principles of biochemistry (4th ed.): Nelson, D., and Cox, M.,” *Biochemistry and molecular biology education.*, vol. 33, no. 1, pp. 74–75, doi: 10.1007/978-3-662-08289-8.

[5] D. Fahrney and J. Hansen, “Hydrogen Bond Lengths and Angles,” 2019, *Colorado State University*. [Online]. Available: https://bc401.bmb.colostate.edu/appendix/h-bonds.php

[6] L. M. Molina, “Theoretical Description and Modeling of Hydrogen Bonds at Solid Surfaces,” in *Encyclopedia of Interfacial Chemistry*, K. Wandelt, Ed., Oxford: Elsevier, 2018, pp. 175–180. doi: https://doi.org/10.1016/B978-0-12-409547-2.13030-6.

[7] X.-J. Zhang, “Van der Waals Forces,” in *Encyclopedia of Tribology*, Q. J. Wang and Y.-W. Chung, Eds., Boston, MA: Springer US, 2013, pp. 3945–3947. doi: 10.1007/978-0-387-92897-5_457.

[8] C. P. Gerba, I. L. Pepper, and D. T. Newby, “Microbial Transport in the Subsurface,” in *Environmental Microbiology: Third Edition*, Elsevier Inc., 2015, pp. 319–337. doi: 10.1016/B978-0-12-394626-3.00015-6.

[9] Q. Sun, “The Hydrophobic Effects: Our Current Understanding,” Oct. 01, 2022, *MDPI*. doi: 10.3390/molecules27207009.

[10] B. Kronberg, “The hydrophobic effect,” Apr. 01, 2016, *Elsevier Ltd*. doi: 10.1016/j.cocis.2016.02.001.

[11] H. S. Frank and M. W. Evans, “Free volume and entropy in condensed systems III. Entropy in binary liquid mixtures; Partial molal entropy in dilute solutions;

Structure and thermodynamics in aqueous electrolytes," *J Chem Phys*, vol. 13, no. 11, pp. 507–532, 1945, doi: 10.1063/1.1723985.

[12] W. Kauzmann, "Some Factors in the Interpretation of Protein Denaturation," in *Advances in Protein Chemistry*, vol. 14, C. B. Anfinsen, M. L. Anson, K. Bailey, and J. T. Edsall, Eds., Academic Press, 1959, pp. 1–63. doi: https://doi.org/10.1016/S0065-3233(08)60608-7.

[13] A. S. Mahadevi and G. N. Sastry, "Cation-π interaction: Its role and relevance in chemistry, biology, and material science," Mar. 13, 2013. doi: 10.1021/cr300222d.

[14] W. R. Zhuang *et al.*, "Applications of π-π stacking interactions in the design of drug-delivery systems," Jan. 28, 2019, *Elsevier B.V.* doi: 10.1016/j.jconrel.2018.12.014.

[15] D. S. Spassov, "Binding Affinity Determination in Drug Design: Insights from Lock and Key, Induced Fit, Conformational Selection, and Inhibitor Trapping Models," Jul. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ijms25137124.

[16] The Science Snail, "The difference between Ki, Kd, IC50, and EC50 values." Accessed: Dec. 26, 2024. [Online]. Available: https://www.sciencesnail.com/science/the-difference-between-ki-kd-ic50-and-ec50-values

[17] J. Barbet and S. Huclier-Markai, "Equilibrium, affinity, dissociation constants, IC5O: Facts and fantasies," *Pharm Stat*, vol. 18, no. 5, pp. 513–525, Oct. 2019, doi: 10.1002/pst.1943.

[18] S. J. Y. Macalino, V. Gosu, S. Hong, and S. Choi, "Role of computer-aided drug design in modern drug discovery," Sep. 22, 2015, *Pharmaceutical Society of Korea*. doi: 10.1007/s12272-015-0640-5.

[19] Y. Tang, R. Moretti, and J. Meiler, "Recent Advances in Automated Structure-Based De Novo Drug Design," Mar. 25, 2024, *American Chemical Society*. doi: 10.1021/acs.jcim.4c00247.

[20] A. Gimeno *et al.*, "The light and dark sides of virtual screening: What is there to know?," Mar. 02, 2019, *MDPI AG*. doi: 10.3390/ijms20061375.

[21] A. Sharma and R. M. Yennamalli, “Chapter 16 - Docking strategies,” in *Basic Biotechniques for Bioprocess and Bioentrepreneurship*, A. K. Bhatt, R. K. Bhatia, and T. C. Bhalla, Eds., Academic Press, 2023, pp. 243–258. doi: https://doi.org/10.1016/B978-0-12-816109-8.00016-7.

[22] A. Tiwari and S. Singh, “Chapter 13 - Computational approaches in drug designing,” in *Bioinformatics*, D. B. Singh and R. K. Pathak, Eds., Academic Press, 2022, pp. 207–217. doi: https://doi.org/10.1016/B978-0-323-89775-4.00010-9.

[23] X. Qing *et al.*, “Pharmacophore modeling: Advances, Limitations, And current utility in drug discovery,” Nov. 11, 2014, *Dove Medical Press Ltd*. doi: 10.2147/JRLCR.S46843.

[24] R. Wang, X. Fang, Y. Lu, and S. Wang, “The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures,” *J Med Chem*, vol. 47, no. 12, pp. 2977–2980, Jun. 2004, doi: 10.1021/jm030580l.

[25] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, and S. Wang, “The PDBbind Database: Methodologies and Updates,” *J Med Chem*, vol. 48, no. 12, pp. 4111–4119, Jun. 2005, doi: 10.1021/jm048957q.

[26] “Beginner’s Guide to the PDBbind Database (v.2020).” Accessed: Dec. 26, 2024. [Online]. Available: http://www.pdbbind.org.cn/download/pdbbind_2020_intro.pdf

[27] F. Da Silva, J. Desaphy, and D. Rognan, “IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions,” *ChemMedChem*, vol. 13, no. 6, pp. 507–510, Mar. 2018, doi: 10.1002/cmdc.201700505.

[28] J. Desaphy, D. A. Silva, G. Bret, and D. Rognan, “IChem User guide IChem: A Toolkit for detecting, comparing and predicting protein-ligand interactions,” 2023.

[29] Tripos Developers, “Tripos Mol2 File Format.” Accessed: Dec. 10, 2024. [Online]. Available: https://zhanggroup.org/DockRMSD/mol2.pdf

[30] Stephen Cass, “The Top Programming Languages 2024 > Typescript and Rust are among the rising stars,” *IEEESpectrum*, Aug. 2024, Accessed: Dec. 13, 2024. [Online]. Available: https://spectrum.ieee.org/top-programming-languages-2024

[31] B. M. T, C. Andrea, C. John, C. S. P, and L. Jeremy, “Human Immunodeficiency Virus Type 1 Gag Polyprotein Multimerization Requires the Nucleocapsid Domain and RNA and Is Promoted by the Capsid-Dimer Interface and the Basic Region of Matrix Protein,” *J Virol*, vol. 73, no. 10, pp. 8527–8540, Oct. 1999, doi: 10.1128/jvi.73.10.8527-8540.1999.

[32] H. A. Carlson, R. D. Smith, N. A. Khazanov, P. D. Kirchhoff, J. B. Dunbar, and M. L. Benson, “Differences between high- and low-affinity complexes of enzymes and nonenzymes,” *J Med Chem*, vol. 51, no. 20, pp. 6432–6441, Oct. 2008, doi: 10.1021/jm8006504.