UNIVERSITAT ROVIRA i VIRGILI

STUDY OF INTERMOLECULAR INTERACTIONS IN DIVERSE SARS-COV-2 MPRO INHIBITORS: WHAT INSIGHTS CAN WE GAIN?

Said Trujillo de León

BIOTECHNOLOGY FINAL DEGREE THESIS

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

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

Ariadna Llop Peiró (ariadna.llop@urv.cat)

Aleix Gimeno Vives (aleix.gimeno@urv.cat)

June 2025

Jo, Said Trujillo de León, amb DNI 79354371R, 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, 5 de juny de 2025

(signatura)

TRUJILLO DE LEON SAID - 79354371R

Firmado digitalmente por TRUJILLO DE LEON SAID - 79354371R
Fecha: 2025.06.05 12:12:12 +02'00'

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

Content

Abstract and keywords

The COVID-19 pandemic has led to an unprecedented accumulation of structural and computational data on potential therapeutic targets. Among these, the SARS-CoV-2 main protease (M_{PRO}) has become a key focus due to its role in viral replication and the absence of human homologs, making it an attractive target for antiviral drug development. Hundreds of experimentally determined protein-ligand complexes, particularly noncovalent inhibitors bound to M_{PRO}, are now available in public databases. These structures offer an opportunity to deepen our understanding of computational drug discovery through the design and validation of virtual screening protocols, docking approaches, and extended QSAR analyses. However, the volume and complexity of the data make manual analysis impractical.

To support QSAR (Quantitative Structure-Activity Relationships) analysis and streamline the identification of potential pharmacological candidates, an open-source Python library called CORAL-PIC (COrellation of Residues and Activities for Ligands with pIC_{50}) has been developed. This tool aims to detect and rapid summarize non-covalent ligand-protein interactions. By systematically extracting and comparing key interaction features across large datasets, CORAL-PIC facilitates the identification of recurrent binding patterns linked to strong inhibitory activity. In a proof-of-concept study involving 378 M_{PRO}–inhibitor complexes, CORAL-PIC condensed more than 300 MB of raw interaction data into a lightweight, 100 KB summary file, enabling rapid frequency analysis of interaction types. Results revealed that hydrophobic and weak polar contacts dominate overall binding, while directional interactions, such as hydrogen bonds and π–type contacts with residues like CYS145 and GLY143, correlate most strongly with high pIC_{50} values. Additionally, key anchoring residues within the S1 subsite (GLU166, MET165, HIS163, MET49, ASN142) were identified as conserved hotspots across potent inhibitors. These findings demonstrate CORAL-PIC's ability to pinpoint both frequent and rare interaction motifs, offering a streamlined workflow for QSAR model development and the rational design of new M_{PRO} inhibitors.

Keywords: *COVID-19, SARS-CoV-2 main protease (M_{PRO}), antiviral drug development, protein-ligand interactions, QSAR, binding patterns, non-covalent inhibitors*

1. Introduction

This section describes the structural features of M_{PRO}, including the nature of its catalytic centre and the organisation of its binding subsites, with the aim of understanding how it interacts with different non-covalent inhibitors. It will also review the main types of molecular interactions involved and address experimental parameters such as IC_{50} and its conversion to pIC_{50}, which are relevant for the evaluation of the inhibitory activity of the complexes under study.

1.1. SARS-CoV-2 Main Protease

SARS-CoV-2 requires its viral replication machinery to spread efficiently within the host. An important part of this machinery is the main protease of the virus, known as M_{PRO} or 3-chymotrypsin-like protease (3CL-Pro). This enzyme is vital in the post-translational maturation of the viral polypeptide. The viral replicase gene encodes two large overlapping polyproteins, pp1a and pp1ab, which must be proteolytically cleaved to release the non-structural proteins (nsps) required for viral RNA replication and transcription [1].

M_{PRO} is enclosed within these polyproteins and is released through an autolytic cleavage process as its first enzymatic step [2]. This protease acts functionally as a homodimer, each monomer being a 306-residue chain that folds into three defined domains: domains I and II form antiparallel beta-barrel-like structures, while domain III consists of a set of five alpha helices that regulate dimerization [3].

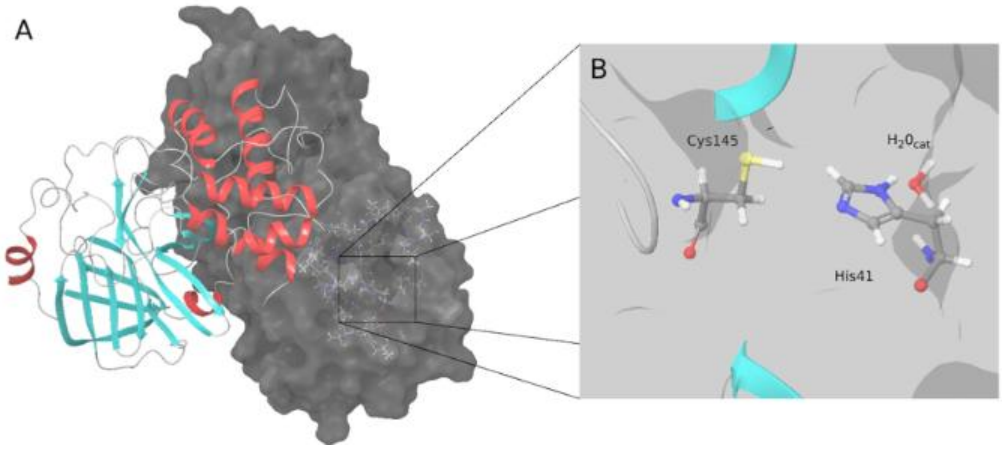

Figure 1. (A) Dimeric assembly with one protomer shown as a cartoon (coloured by secondary structure) and the other as a molecular surface. (B) Catalytic site detail showing the key residues (Cys145, His41) and catalytic water molecule. [1]

The active site of M_{PRO} is located at the interface between domains I and II and has unusual features compared to other chymotrypsin-like proteases. Instead of the classical catalytic triad (Ser/Cys-His-Asp/Glu), M_{PRO} has a catalytic dyad composed of His41 and Cys145. As illustrated in Figure 1, the dyad's activity is stabilized by a structurally conserved water molecule, which forms hydrogen bonds with His41 and

may interact with adjacent residues like His164 and Asp187. This structure suggests a compensatory role analogous to that of the third residue of the traditional triad [4].

The substrate-binding region of M_{PRO} is organized into four subsites (S4, S2, S1, and S1′), which contribute to its catalytic specificity [5]. Their spatial arrangement around the catalytic dyad is shown in Figure 2, while the detailed residue composition for each subsite, including side chain and main chain contributions, is represented in Table 1.

Table 1. Residue composition of M_{PRO} substrate-binding subsites. Residues are classified by their structural contribution:main chain or side chain.

	S4	**S2**	**S1**	**S1'**
Side Chain	Met165 Leu167 Pro168 Ala191 Gln192	His41 Met49 Tyr54 Asp187	Phe140 Asn142 Ser144 Cys145 His163 Glu166 His172	Thr25 His41 Val42 Asn119 Cys145
Main Chain	Glu166 Arg188 Thr190	Arg188	Leu141 Gly143 His164 Met165	Thr26 Gly143

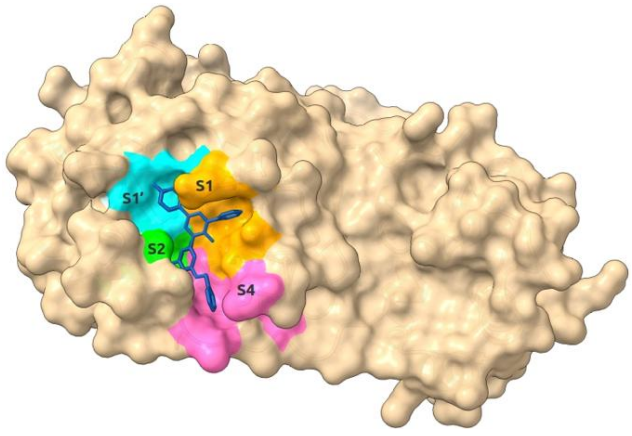

Figure 2. Spatial location of M_{PRO} subsites [6].

The substrate specificity of M_{PRO} is also peculiar, as it differs from human proteases, showing a strong preference for glutamine at the P1 position of the cleaved sequence. This uniqueness, together with its absence in human cells and its high conservation

among coronaviruses, makes M_{PRO} an attractive target for the development of antivirals [7].

Although allosteric binding sites on M_{PRO} have been proposed, most of the non-covalent inhibitors crystallised to date bind to the active site of the enzyme [8].

To study these interactions, numerous studies have used molecular docking techniques to predict the binding affinity and preferred orientation of candidate compounds within the active site. However, the absence of well-validated reference inhibitors makes it difficult to assess the accuracy of these computational approaches. In addition, docking has intrinsic limitations, such as the treatment of the protein as a statically rigid body, and the uncertainty associated with scoring functions, which often do not accurately estimate the binding free energy [9].

Given these constraints, it is necessary to complement docking with other computational approaches, such as molecular dynamics simulations, which allow the structural flexibility of the protein to be incorporated, and QSAR models, which help to identify relationships between the chemical structure of compounds and their biological activity. The integration of multiple strategies improves the robustness of predictions and facilitates the identification of new inhibitor candidates with therapeutic potential [10].

1.2. Interactions

The interactions between an enzyme and its inhibitor can vary significantly in nature and strength. Some may occur rarely but contribute strongly to binding, while others appear frequently but with weaker effects. Understanding these differences is key, as the effectiveness of inhibition depends not only on binding affinity but also on the desired pharmacological behaviour. In some cases, tight binding is preferred for sustained inhibition, while weaker, more dynamic interactions may be advantageous when modulation of inhibition (e.g., concentration-dependent effects), is required.

Understanding protein-ligand molecular recognition requires detailed characterization of the physicochemical interactions that stabilize the complex. These interactions can be quantified using structural, energetic, and statistical descriptors that set out how atoms and functional groups behave in three-dimensional space.

The spatial positioning of atoms involved in intermolecular interactions determine the strength and specificity of the association. One key parameter is the interatomic distance, typically measured in angstroms (Å), which controls how two atoms or groups can interact. As separation increases, interaction strength generally decreases. Equally important is the angle formed between atoms participating in directional interactions such as hydrogen or halogen bonds. The efficiency of these interactions depends not only on proximity but also on proper geometric alignment. A donor-hydrogen-acceptor angle close to 180º maximizes hydrogen bond strength, while deviations from linearity reduce it [11].

Beyond spatial parameters, potential energy quantifies the stability of intermolecular interactions in thermodynamic terms (e.g., kcal/mol). It reflects the relative positions and orientations of molecules, such as ligands and their targets, and varies as a function of conformation and distance. For example, conformational changes such as bond rotation, as illustrated in Figure 3, alter potential energy by modulating steric hindrance and electronic interactions: energy is maximized in eclipsed conformations and minimized in staggered ones.

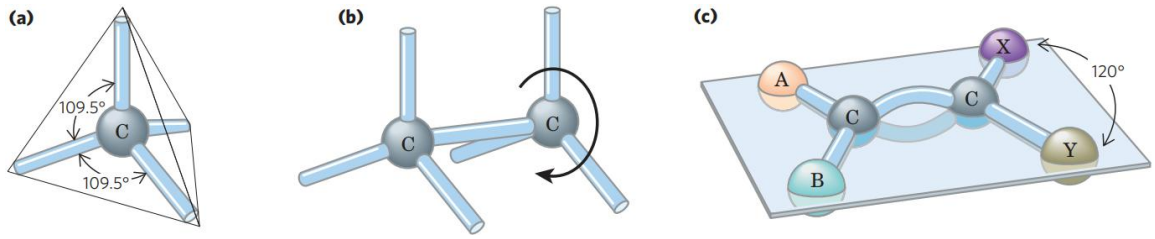

Figure 3. Bonding geometry of carbon atoms. Carbon atoms exhibit a tetrahedral arrangement around single bonds. Single C–C bonds allow free rotation, as exemplified by ethane, while double bonds are shorter and restrict rotational freedom, maintaining atoms in a fixed planar arrangement [12].

This energy landscape dictates the likelihood of favourable binding geometries. From a thermodynamic perspective, potential energy can be understood as the capacity of a molecular system to perform work as it approaches a more stable, lower-energy state [12]. In docking and simulations, this energy supports scoring functions to predict affinity and prioritize ligand candidates.

Entropy reflects the thermodynamic trade-offs associated with the loss of molecular freedom upon ligand binding. When an inhibitor binds to its target, it experiences a significant reduction in conformational flexibility as rotatable bonds become constrained, generating an unfavourable entropic penalty. This penalty must be compensated by favourable enthalpic interactions such as hydrogen bonds or hydrophobic effects for binding to occur. Concurrently, the displacement of ordered water molecules from the binding site provides a favourable entropic contribution due to the increase in solvent disorder [13]. The net entropy change therefore reflects the balance between ligand rigidification and solvent reorganization.

1.2.1. Atom-Atom

The definitions and conceptual framework presented herein are based primarily on the textbook *Inorganic Chemistry by Miessler, Fischer, and Tarr (5th edition, Pearson, 2014)* [11], which provides a comprehensive treatment of chemical bonding, molecular geometry, and intermolecular forces. Where relevant, additional biochemical context is included to bridge the principles of inorganic chemistry with the structure and function of proteins.

Covalent interactions arise when two atoms share one or more electron pairs. It represents the strongest type of atomic bonding within molecules. In proteins, disulfide

bond (S–S) formed between two cysteine residues are covalent crosslinks that stabilize tertiary and quaternary structures.

Hydrogen bond (X–H···B) occurs when a hydrogen atom covalently bound to a highly electronegative atom X (commonly N, O or F) interacts with a lone pair donor. The strength and geometry of hydrogen bonds depend on the participating atoms and their alignment, with stronger interactions tending toward linearity and involving more electronegative donors and acceptors. In proteins, key hydrogen bond donors and acceptors include:

- Strong donors: hydroxyl groups of serine, threonine, and tyrosine; amide N–H groups of asparagine and glutamine.
- Strong acceptors: backbone carbonyls and side-chain C=O groups of asparagine and glutamine.
- Weak donors: thiol groups of cysteine and aromatic C–H groups.

Analogous to hydrogen bonding, halogen bonds involve an electropositive region (σ-hole) on a halogen atom (typically Cl, Br or I) interacting with a lone-pair donor (O, N or S). These interactions are generally directional and nearly to linear geometries.

Ionic interactions result from the electrostatic attraction between oppositely charged ions. These interactions are stronger in hydrophobic environments but become weaker in aqueous solutions due to dielectric screening by water. Positively charged residues such as lysine and arginine interact with negatively charged residues like aspartate and glutamate to form salt bridges or ion pairs.

Metal complexes in biological systems consist of a central metal ion (e.g., Zn^{2+}, Mg^{2+}, $Fe^{2+/3+}$) coordinated by ligands, often side chains of amino acids or small molecules. These ligands donate electron pairs to the metal centre, forming covalent bonds.

In aqueous environments, nonpolar residues and ligands associate to minimize disruption of the hydrogen-bonding network of water. This entropically driven process, known as hydrophobic effect, is augmented by weak but additive London dispersion forces between nonpolar groups. Though individually weak, these interactions are collectively significant and contribute to protein folding, stability, and ligand binding. Relevant residues include aliphatic side chains (leucine, isoleucine, valine, alanine, methionine) and aromatic residues (phenylalanine, tryptophan).

Carbonyl groups (C=O) participate in hydrogen bonding, metal coordination, and polar-π interactions. Their reactivity arises from the polarized C–O bond, where the oxygen lone pairs act as hydrogen-bond acceptors, and the π* orbital can engage in back-donation with metal centres.

In proteins, the carbonyl groups collaborate to their structural integrity and functional versatility, forming hydrogen bonds that stabilize secondary structures like α-helices and β-sheets while also participating in interactions such as metal ion coordination through acidic and amide side chains.

Polar interactions arise from dipole–dipole attractions between molecules with permanent dipole moments. These interactions occur when bonded atoms differ in electronegativity, creating a molecular dipole that aligns to maximize attraction between opposite partial charges. Highly polar amino acids such as serine, threonine, asparagine, glutamine, and histidine help stabilize active sites through their dipolar character and hydrogen-bonding capacity. Other residues with weaker dipoles, including glycine, alanine, and proline, also influence local folding and ligand orientation, often subtly and in combination with more dominant electrostatic effects.

1.2.2. Atom-Plane

Atom–Plane interactions represent a class of noncovalent contacts in which a polarized atom or bond approaches the π-electron cloud of an aromatic ring. Despite their modest individual strengths, these interactions share common geometric and physicochemical features that render them biologically significant, particularly in the context of protein-ligand interfaces.

A defining feature of atom–π interactions is their directional geometry. The interacting atom or bond is generally oriented toward the centroid of the aromatic ring, frequently adopting an edge-to-face (T-shaped) or perpendicular configuration relative to the ring plane. The interactions are ruled by a combination of electrostatic forces, dispersion effects, and, in certain cases, orbital overlap between the electron donor and the π-system [14], [15], [16].

These contacts can be highly cooperative, particularly within hydrophobic environments of enzyme active sites, where multiple weak interactions collectively enhance binding affinity.

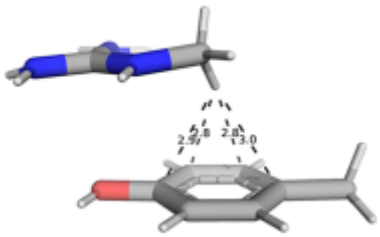

Figure 4. Tyr-Arg carbon-π interaction [14]

Among the main subclasses, carbon–π (CH–π) interactions involve polarized C–H bonds (commonly from aliphatic or aromatic residues) acting as a weak proton donor to the π-electron cloud of the aromatic rings. Figure 4 shows an example of this type of interaction between an arginine residue and the aromatic ring of a tyrosine side chain. Cation–π interactions result from the attraction between a positively charged species, such as a protonated lysine or arginine side chains, and the electron-rich face of an aromatic ring. Donor–π interactions involve lone-pair-bearing atoms like oxygen, nitrogen, or sulphur approaching the π-system and partially donating electron density

into the aromatic cloud; these are often enhanced when the ring is electron-deficient [16].

Halogen–π interactions occur when a covalently bound halogen (Cl, Br, or I) with a positive σ-hole interacts directionally with the π-electrons of an aromatic system [15]. Lastly, sulphur–π interactions involve the divalent sulphur atom of methionine approaching the face of an aromatic ring such as phenylalanine, tyrosine or tryptophan [17].

1.2.3. Plane-Plane

Non-covalent π–π interactions, are defined as attractive forces between aromatic systems that contain delocalized π-electron clouds, represent one of the main classes of directional interactions that contribute to the structural and functional integrity of proteins, as well as to ligand binding specificity. These interactions, often described as aromatic–aromatic interactions, occur between planar components such as the side chains of aromatic residues [18].

In protein structures, π–π interactions primarily adopt two geometries: the offset stacked (parallel-displaced) and edge-to-face (T-shaped) arrangements. These orientations are not random but reflect stereospecific constraints imposed by the three-dimensional fold of the protein [19]. Figure 5 illustrates representative examples of these configurations.

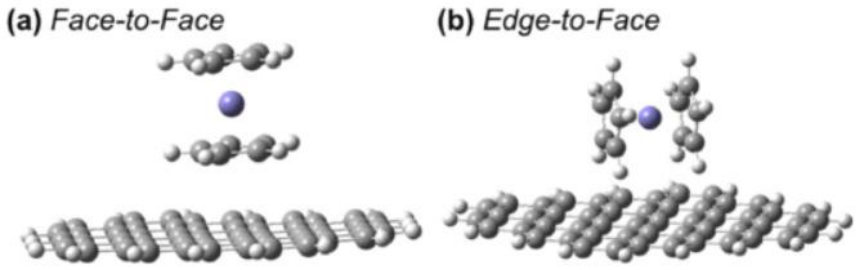

Figure 5. Orientation scheme for π–π interactions between aromatic side chains in proteins. (a) Offset-stacked (parallel-displaced) configuration; (b) Edge-to-face (T-shaped) configuration [20].

1.2.4. Group-Group/Plane

Amide–π interactions arise between the electron-rich π-system of aromatic rings and a polar amide group, which may be located either in the protein backbone or in side chains such as those of asparagine and glutamine. Depending on the geometry, the interaction may involve the amide hydrogen (NH/π), the carbonyl lone pair (O/π), or a planar alignment of the amide and aromatic systems resembling π–π stacking [21].

Among these, NH/π interactions in a face-on orientation, where the amide hydrogen points directly toward the centre of the aromatic ring, are energetically most favourable. Slightly less optimal geometries, such as those in which the hydrogen points toward a specific atom or bond of the ring, still retain a substantial stabilizing character. Parallel stacking between the amide plane and the aromatic ring can also

be favourable, due to electron correlation effects. In contrast, O/π interactions tend to be repulsive in isolation but may become stabilizing in the complex electrostatic environment of a protein active site, particularly when coupled with other favourable interactions such as hydrogen bonding or electrostatic complementarity [21].

1.3. Binding Affinity Metrics

The half-maximal inhibitory concentration (IC_{50}) is widely employed to describe the potency of an inhibitor in reducing the activity of a target protein. Although frequently used, IC_{50} is not a direct measure of binding affinity, as it does not correspond to an equilibrium constant. Instead, it reflects the inhibitor concentration required to achieve 50% inhibition under specific experimental conditions, which makes it highly sensitive to factors such as enzyme, substrate, and inhibitor concentrations.

To address the limitations associated with the direct use of IC_{50} values, it is common practice to convert them into their negative base-10 logarithm, known as pIC_{50}. This transformation not only facilitates the comparison of results across different studies but also provides a more interpretable scale for analysing inhibitor potency [22]. The conversion requires that IC_{50} values be expressed in molar units (mol/L), and follows the formula:

$$pIC_{50} = -log_{10}(IC_{50}\ [mol/L])$$

A higher pIC_{50} value indicate lower concentrations needed to achieve 50% inhibition, which corresponds to greater inhibitor potency.

2. Hypothesis and objectives

This study hypothesizes that a detailed understanding of the interactions between the SARS-CoV-2 M^{PRO} and a set of known inhibitors may be useful for guiding the identification and development of new therapeutic agents targeting the virus.

The primary objective of this project is to **analyse a set of molecular complexes between M^{PRO} and several inhibitors in order to identify interaction patterns** that could serve as potential leads in the development of drugs against SARS-CoV-2.

As a secondary objective, this work aims to **develop a Python-based library** for the automated analysis of protein-ligand interaction files. This tool is intended to simplify the extraction, visualization, and filtering of relevant interaction data, thereby facilitating the screening process in future drug discovery workflows.

3. Methodology

This section outlines the comprehensive workflow employed in the study, from data collection to output generation. First, a curated dataset of SARS-CoV-2 M^{PRO} crystallographic complexes was assembled from the Protein Data Bank, followed by the integration of experimental bioactivity data. The input files, comprising atomic

interaction profiles, activity measurements, and subpocket definitions, were structured to enable precise filtering and analysis. A custom-built library was then developed to automate the extraction, organization, and filtering of interaction data from JSON files generated by the Arpeggio software. The library supports a variety of configurable parameters, enabling tailored analyses based on interaction types, subunits, activity thresholds, and more. Additionally, it offers built-in visualization tools and exports results in a structured *.xlsx* format, facilitating both interpretability and post-processing data handling.

3.1. Data Collection

To retrieve all crystallographic structures of the SARS-CoV-2 M_{PRO}, a search was performed in the Protein Data Bank (PDB) in February 2025. This search yielded 1590 entries. These structures were then filtered using the PDB-CAT tool [23] according to the following criteria:

- Presence of a non-covalently bound ligand.
- No sequence deviations relative to the M_{PRO} reference sequence (January 2020).

Subsequently, the literature was surveyed to compile experimental IC_{50} values for each ligand. To facilitate numerical handling and enable direct comparison between the complexes, all IC_{50} values were converted into pIC_{50}. With this transformation, higher pIC_{50} values correspond to lower IC_{50} values, and therefore indicate greater inhibitory potency. For instance, the complex 7gb0 has a pIC_{50} of 4.00, corresponding to an IC_{50} of approximately 100 µM, while 8w1u has a pIC_{50} of 8.05, equivalent to an IC_{50} of around 9 nM, thus reflecting substantially higher potency. After applying these selection steps, a final set of 378 non-covalent inhibitors, each co-crystallized with M_{PRO} and associated with a measured bioactivity (full list in the Appendix), was established. Interaction profiles between each inhibitor and M_{PRO} were then computed using the Arpeggio software [24]. The overall workflow for structure retrieval, filtering, bioactivity annotation, and interaction profiling is summarized in Figure 6.

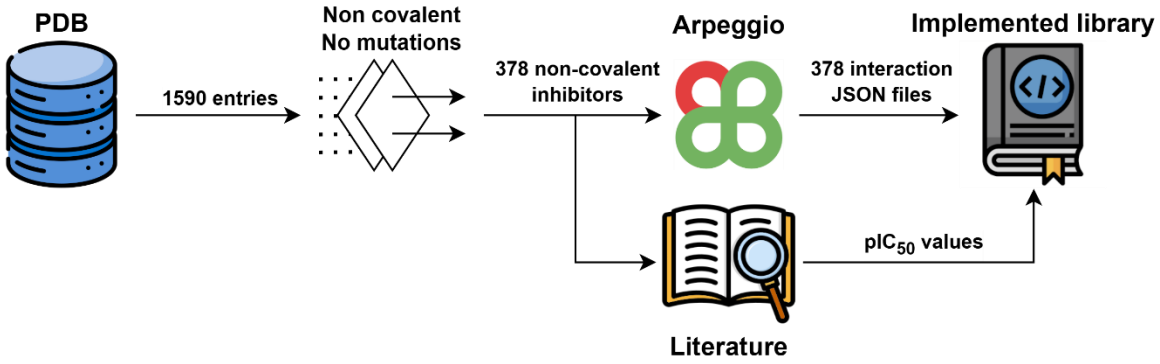

Figure 6. Data acquisition and preprocessing workflow for the study.

3.2. Input Data

This section describes the different types of input files used in the analysis, including how they were generated, structured, and filtered to extract relevant interaction and activity data for the study.

3.2.1. Arpeggio Files

For each of the 378 inhibitor-M_{PRO} complexes, Arpeggio generates a JSON file containing a detailed list of atomic interactions. To ensure proper identification of `INTER` interactions the inhibitor and M_{PRO}, the PDB code of the ligand must be specified using the `-s` option when running Arpeggio. Figure 7 shows an example of a JSON input file generated by Arpeggio, illustrating the structure and key fields used in the data extraction process. In this example, a hydrophobic interaction can be observed between a carbon atom of the ligand and the MET49 residue of the protein. The library development focused on extracting and structuring this information as follows:


```
"bgn": {
    "auth_asym_id": "A_2",
    "auth_atom_id": "C12",
    "auth_seq_id": 1001,
    "label_comp_id": "UGG",
    "label_comp_type": "B",
    "pdbx_PDB_ins_code": " "
},
"contact": [
    "proximal",
    "hydrophobic"
],
"distance": 3.74,
"end": {
    "auth_asym_id": "A_2",
    "auth_atom_id": "CE",
    "auth_seq_id": 49,
    "label_comp_id": "MET",
    "label_comp_type": "P",
    "pdbx_PDB_ins_code": " "
},
"interacting_entities": "INTER",
"type": "atom-atom"
```


Figure 7. Example of an input JSON file generated by Arpeggio.

- `interacting_entities`: Only entries labelled `INTER` were retained.
- Atom assignment (`bgn` and `end`): Each interacting atom is annotated with a `label_comp_type` (`P` for protein, other for ligand). If both are labelled `P`, the `label_comp_id` (three-letter residue code) is used to distinguish natural amino acids (protein) from non-natural residues (ligand).

- Residue and atom identifiers:
  - `auth_asym_id`: Chain identifier within the PDB entry
  - `auth_atom_id`: Atom name
  - `auth_seq_id`: Residue number (relevant for protein only)
  - `label_comp_id`: Three-letter residue code (relevant for protein)
- Interaction descriptors:
  - `contact`: Enumerates specific atomic contacts
  - `type`: Characterizes the nature of the interaction

  Table 2 summarizes the interaction types detected by Arpeggio. Those marked with a green tick were included in the present study, while the rest were excluded from the analysis.
- `distance`: Although available, distance metrics were not retained for this study.

Table 2. List of interaction types detected by Arpeggio. A green tick (✓) indicates interaction types that were considered in this study, while a red cross (✗) marks those that were excluded from the analysis.

type	contact	Interaction Name	Selected
atom-atom	clash	Clash	✗
	covalent	Covalent	✓
	vdw_clash	VdW Clash	✗
	vdw	VdW	✗
	proximal	Proximal	✗
	hbond	Hydrogen Bond	✓
	weak_hbond	Weak Hydrogen Bond	✓
	xbond	Halogen Bond	✓
	ionic	Ionic	✓
	metal	Metal Complex	✓
	aromatic	Aromatic	✓
	hydrophobic	Hydrophobic	✓
	carbonyl	Carbonyl	✓
	polar	Polar	✓
	weak_polar	Weak Polar	✓
atom-plane	CARBONPI	Carbon-PI	✓
	CATIONPI	Cation-PI	✓
	DONORPI	Donor-PI	✓

	HALOGENPI	Halogen-PI	✓
	METSULPHURPI	Sulphur-PI	✓
plane-plane	-	-	✓
group-group/plane	AMIDEAMIDE	amide - amide	✓
	AMIDERING	amide - ring	✓

3.2.2. Template File

To ensure reusability across diverse cases, a flexible filtering system was implemented. Filters are defined in a configuration file and applied sequentially to each JSON.

Figure 8 shows two specific filters defined for this study. The first filter retained only those interactions where the field `interacting_entities` is equal to `INTER` and where the `contact` field includes at least one interaction type from a predefined list of interest. The second filter also required `interacting_entities` to be `INTER`, but instead of considering the `contact` field, it selected interactions where the `type` field is equal to `plane-plane`.

```
{
    "contact": [
        "covalent", "hbond", "aromatic", "hydrophobic", "polar", "ionic", "xbond", "metal",
        "carbonyl", "CARBONPI", "CATIONPI", "DONORPI", "HALOGENPI", "METSULPHURPI",
        "AMIDEAMIDE", "AMIDERING", "weak_hbond", "weak_polar"
    ],
    "interacting_entities": "INTER",
    "type": null
},
{
    "contact": null,
    "interacting_entities": "INTER",
    "type": "plane-plane"
}
```

Figure 8. JSON configuration file for filters applied in the study.

This modular approach allows users to tailor which interaction categories and molecular contexts are extracted without modifying the library.

3.2.3. Activity File

To enable filtering based on the inhibitory activity of the compounds, the pIC_{50} values corresponding to each inhibitor were compiled into a *.csv* file. This file maps the bioactivity data directly to the PDB identifier of each ligand-M_{PRO} complex, facilitating cross-referencing during the analysis. Although optional within the general use of the library, this file was included in the present study to enable activity-based filtering and

ranking of ligands. Table 3 provides example entries from the activity file used in this study.

Table 3. Input CSV file example: PDB IDs with corresponding pIC_{50} activity values.

pdb_id	pIC_{50}
7gnf	6.66976282750156
7gne	5.42908880221149
7gga	6.54725354142573
7gg2	5.76611887119722

3.2.4. Regions File

To explore the spatial distribution of ligand interactions across the binding site, a subpocket classification scheme was implemented using a second *.csv* file. This file allows users to define custom residue sets (tagged with `<main>` (backbone) or `<side>` (R-group)) to filter interactions by structural role and was applied here to formalize the four subpockets (S4, S2, S1, S1′) listed in Table 1. Though optional for library functionality, this file associates each zone with a specific set of amino acid residues, enabling region-specific analyses of interaction frequency and relevance. This approach aims to identify whether certain subpockets consistently attract more interactions, which may indicate regions of greater pharmacological relevance within the binding site.

3.3. Library

The implemented library, CORAL-PIC, enables the analysis of large datasets in a time-efficient manner, significantly reducing the processing time compared to manual analysis. For instance, the JSON files analysed in this study may contain over 45,000 lines of data, rendering manual curation not only time-consuming but also prone to human error. In contrast, the developed tool performs information parsing and filtering within seconds, while also offering additional functionalities. The library is openly available at https://github.com/31ldts/CORAL-PIC.

During the initial analysis phase, the user can configure specific parameters to tailor the output, such as considering biological activity or displaying/suppressing particular types of information. Once the initial parsing is complete, the tool allows for various filtering operations, including the selection of specific interactions, subunits, or chains. Furthermore, data can be sorted based on thresholds of activity or interaction count.

The tool also includes capabilities for data visualization, supporting the generation of bar charts, pie charts, and heatmaps to facilitate data interpretation.

Finally, all processed data can be exported to an *.xlsx* file. This file contains two sheets: one with detailed information on interactions, residues, and inhibitors; and

another serving as a legend, explaining the assignment of IDs and colour codes used in the visualizations.

Figure 9 illustrates the overall workflow of the library, highlighting the main functionalities and the specific steps in which each input file is required.

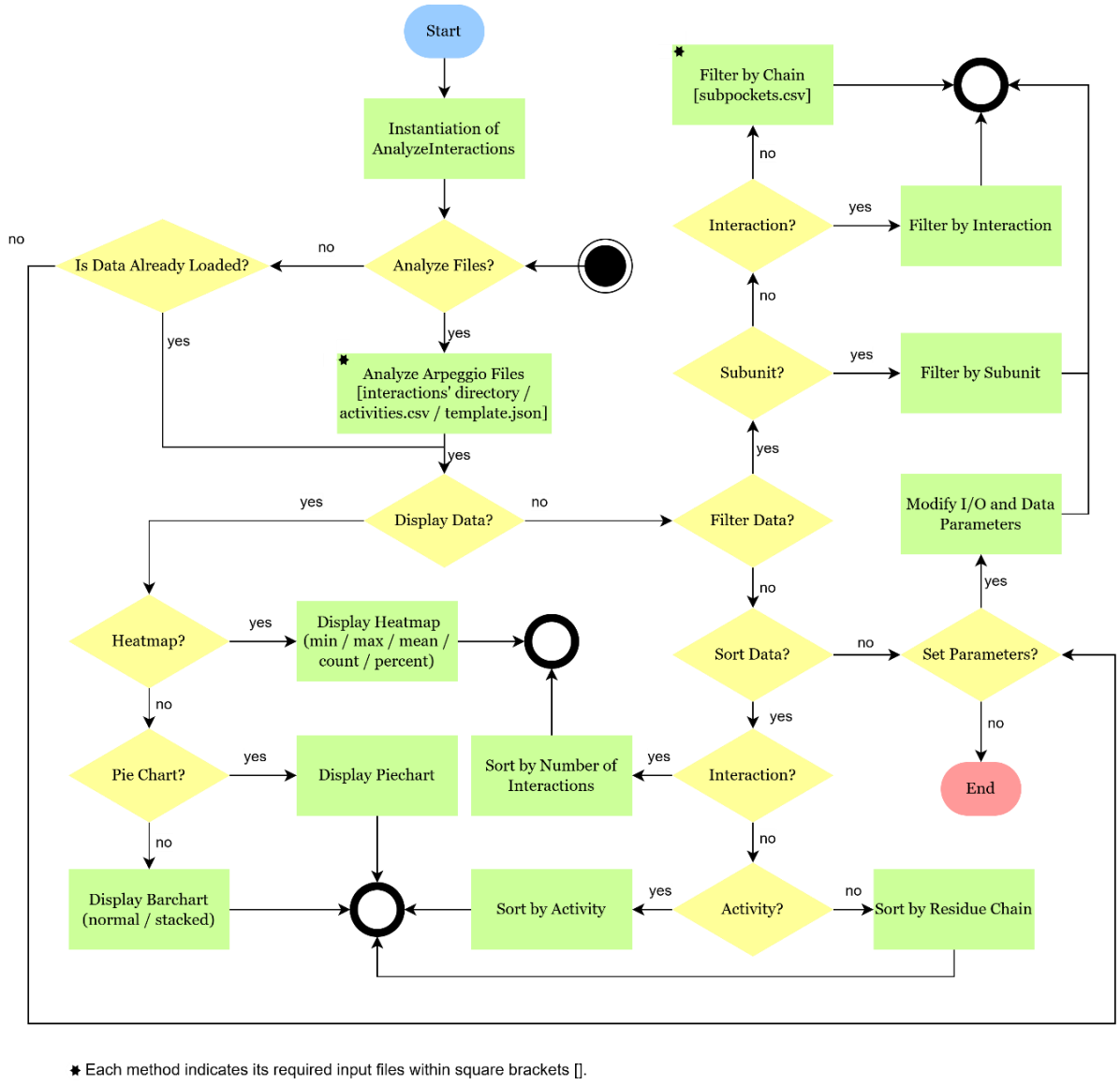

Figure 9. Workflow of the proposed library. To improve visual clarity and avoid overlapping arrows, white circles have been introduced to act as entry points for transitions between methods, while a black circle serves as a common exit point. These elements do not represent computational stages, but rather function as a schematic simplification to organize return paths within the diagram.

3.4. Output Data

As previously mentioned, the library permits the storage of both raw and filtered data derived from the file analysis. This information is exported into an *.xlsx* file composed of two distinct sheets.

The first sheet, titled *Matrix*, presents the interaction data in a structured matrix format. One axis lists the inhibitors, optionally annotated with their activity values if provided in the corresponding activity file. The other axis represents the protein residues, which can appear either as residue-position pairs (e.g., MET 49) or accompanied by the subunit identifier (e.g., MET 49-A), depending on the parameters specified during the analysis phase.

If an interaction is detected between a specific inhibitor and residue, the corresponding cell will display detailed information as illustrated in Figure 10. This includes the interaction ID, the interacting atoms of the protein and the ligand, and optionally the subunit, again depending on the chosen configuration. If the same interaction type occurs multiple times between a given residue and ligand, these instances are grouped within vertical bars ("| |"). In the case of multiple different interactions, entries are separated by semicolons (";").

Figure 10 shows an example where two interaction types, 17 and 18, are observed. The specific names of these interactions can be found in the *Attributes* sheet, which will be discussed later. Each interaction type appears twice, indicating two occurrences of interaction 17 and two of interaction 18. In all four cases, the interacting protein atom is an oxygen, while the ligand atom is carbon 3, although atom identities may vary in general. Additionally, the interactions are distributed across two different subunits, A and A_2.

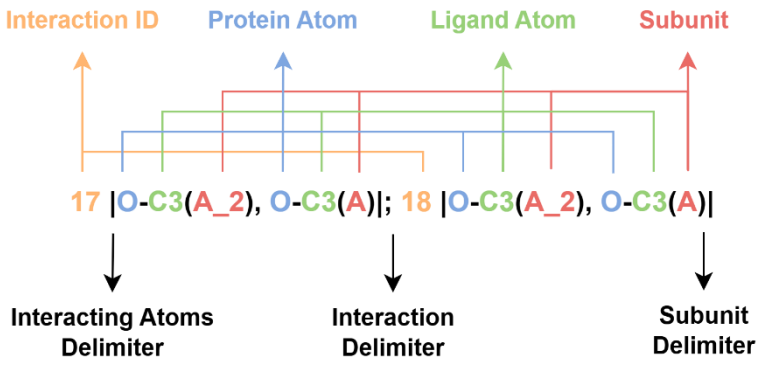

Figure 10. Text-based interaction representation schema based on the Matrix worksheet

The second sheet, named *Attributes*, serves as a legend to aid the interpretation of both the *Matrix* sheet and the generated visualizations. Figure 11 provides an example corresponding to the dataset used in this study. The sheet contains several columns:

- `id`: A unique identifier for each type of interaction.
- `interactions`: Name of the interaction type.
- `colors`: The assigned colour code used in graphical outputs.
- `ligand` and `protein`: Boolean values reflecting user preferences during Arpeggio file analysis. A value of `True` indicates that atoms from the ligand or

protein should be represented in the matrix and/or plots; `False` indicates omission.

- `subunit`: Also a Boolean value; when `True`, subunits are included in residue labels (e.g., distinguishing MET 49-A from MET 49-B). Otherwise, subunit information is incorporated only in the interaction details as shown in Figure 10. It also lists all protein subunits identified during analysis, enabling targeted filtering by the user.
- `mode`: Indicates the analysis mode used. For this study, the mode is consistently set to `arpeggio`, although the library supports data from other structural interaction analysis tools, such as IChem [25].

id	interactions	colors	ligand	mode	protein	subunit
1	AMIDEAMIDE	#e6194B	True	arpeggio	True	False (A, AAA, AAA_2, A_2, B, BBB, B_2, C)
2	AMIDERING	#3cb44b				
3	aromatic	#ffe119				
4	CARBONPI	#4363d8				
5	carbonyl	#f58231				
6	CATIONPI	#911eb4				
7	covalent	#46f0f0				
8	DONORPI	#f032e6				
9	HALOGENPI	#bcf60c				
10	hbond	#fabebe				
11	hydrophobic	#008080				
12	ionic	#e6beff				
13	metal	#9a6324				
14	METSULPHURPI	#fffac8				
15	plane-plane	#800000				
16	polar	#aaffc3				
17	weak_hbond	#808000				
18	weak_polar	#ffd8b1				
19	xbond	#000075				

Figure 11. Screenshot of the Attributes worksheet from the analysis XLSX file

Beyond data storage and matrix representation, CORAL-PIC generates customizable plots to facilitate interaction analysis. Bar plots (stacked or unstacked) quantify interactions per residue or ligand, revealing frequency distributions. Pie charts provide an overview of global interaction-type prevalence. Heatmaps correlates residues against interaction types, highlighting metrics such as maximum/minimum/mean activity, occurrence counts, and percentages. Examples of these visualizations are included in the Results and Discussion section.

4. Results and Discussion

Due to the absence of a dedicated tool for efficiently analysing output files generated by Arpeggio, a Python library was developed to streamline this process. This tool enables users to extract, filter, and visualize interaction data in a simple and timely manner.

As an illustration of the efficiency of the library, the original Arpeggio files used in this study amount to a total size of 365 MB. In contrast, the processed output provided by the library (summarized in XLSX format) occupies only 109 KB. This represents a reduction of approximately 3420% in the amount of information that must be manually reviewed, facilitating downstream analysis.

To evaluate the performance of the library and simultaneously characterise the interaction patterns between SARS-CoV-2 M_{PRO} and a variety of non-covalent inhibitors, a quantitative analysis was conducted on a dataset of 378 M_{PRO} non-covalent inhibitor complexes, using the filtering and classification criteria described in the Methodology section. All structural interaction analyses and figures presented in this section were generated using the CORAL-PIC library.

4.1. Frequency and nature of molecular interactions

Figure 12 summarises the relative distribution of the types of interactions observed. The most frequent interactions correspond to hydrophobic contacts (17.93%), followed by weak polar interactions (17.86%), weak hydrogen bonds (16.54%), polar interactions (8.67%), carbon–π interactions (7.69%), hydrogen bonds (7.36%) and, in minor measure, plane–plane interactions (5.23%). This order of predominance suggests that the stabilisation of the ligand-protein complex is mainly dominated by interactions of a non-directional nature, especially those of a hydrophobic and weak polar type, which probably contribute to a basal affinity of the ligand for the active pocket.

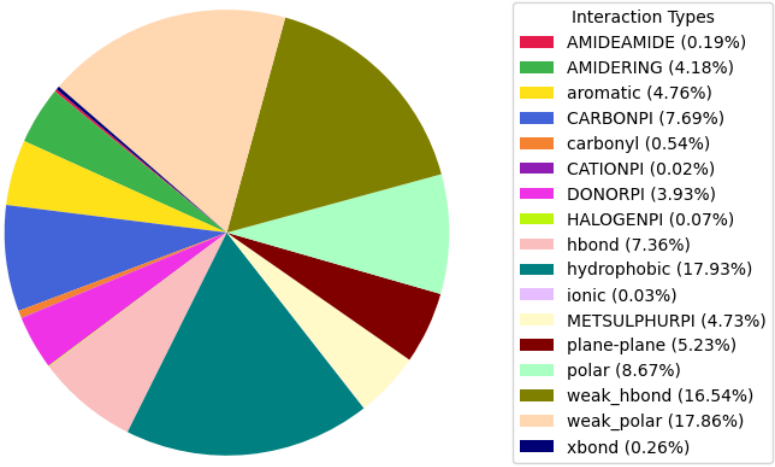

Figure 12. Relative frequency distribution of interaction types between M_{PRO} and 378 inhibitors.

To identify the main residues involved in these interactions, a bar chart was generated (Figure 13) quantifying the total number of interactions per type for each M_{PRO} residue. In order to facilitate the interpretation of the results, a filtering threshold was applied by excluding those residues with less than 200 cumulative interactions.

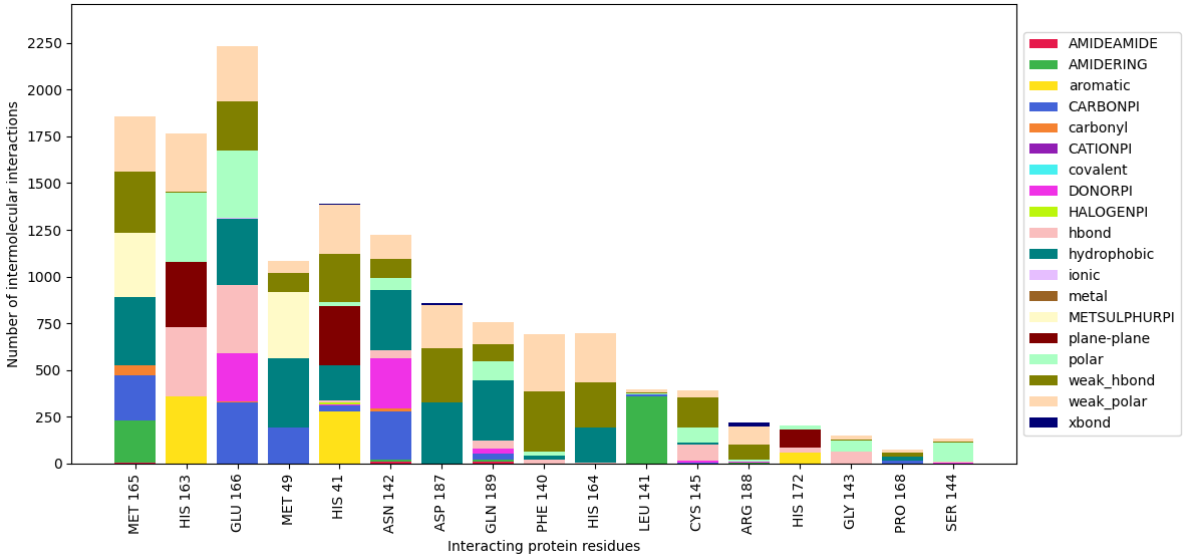

Figure 13. Total interaction counts per residue for M_{PRO} (threshold: ≥200 interactions), stratified by interaction type.

Table 4 lists the residues with the highest number of interactions, together with the M_{PRO} subsite to which they belong. It is notable that HIS41 and CYS145, two of the most interactive residues identified, constitute the catalytic site, which is consistent with the functional relevance of these areas in ligand recognition. However, not all of the most interacting residues are located in the active centre, indicating the possible contribution of adjacent regions to the stabilisation of the complex.

Table 4. Main M_{PRO} residues with the highest number of interactions with inhibitors, their associated subsites, and predominant interaction types.

Residue	Subsite(s)	Predominant Interaction Types
MET 165	S1, S4	Hydrophobic, Met-sulfur–π, Weak hydrogen bond
HIS 163	S1	Hydrogen bond, Polar, Aromatic
GLU 166	S1, S4	Hydrogen bond, Polar, Hydrophobic
MET 49	S2	Hydrophobic, Met-sulfur–π, Carbon–π
HIS 41	S1’, S2	π–π, Aromatic, Weak polar, Weak hydrogen bond
ASN 142	S1	Hydrophobic, Carbon–π, Donor–π
ASP 187	S2	Hydrophobic, Weak hydrogen bond, Weak polar
GLN 189	-	Hydrophobic, Weak polar, Weak hydrogen bond
PHE 140	S1	Weak hydrogen bond, Weak polar
HIS 164	S1	Weak polar, Weak hydrogen bond, Hydrophobic
LEU 141	S1	Amide–π
CYS 145	S1, S1’	Weak hydrogen bond, Hydrogen bond, Polar

The subsite distribution shows a predominant involvement of the S1 subsite, which includes several of the residues with the highest number of interactions, such as HIS163, GLU166, PHE140, LEU141 and CYS145. This suggests that S1 plays a central role in ligand recognition and binding, especially through polar interactions and hydrogen bonding.

4.2. Correlation between interaction count and inhibitory activity

The relationship between the number of molecular interactions and the inhibitory activity of the compounds was also examined. Both the total number of interactions per ligand and the distribution by interaction type were considered, with the aim of identifying potential correlations between interaction density and potency.

As shown in Figure 12 and Figure 15, among the most frequently represented interaction types, carbon–π and plane–plane interactions exhibit comparatively low activity values and involve a limited set of residues. These interactions appear to contribute less significantly to high-affinity binding, likely due to their weaker energetic profile or geometric constraints within the binding pocket.

Furthermore, the analysis of average activity values associated with different residue–interaction pairs reveals a clear trend: the observed activity tends to depend more strongly on the identity of the interacting residue than on the interaction type itself. This suggests that, for a given interaction class, the residue involved plays a more decisive role in determining ligand potency. In other words, even frequent interaction types may lead to different activity outcomes depending on the specific residue involved.

4.3. Interaction patterns in highly active compounds

To explore the relationship between specific interactions and ligand potency, a series of heatmaps was generated, focusing on the most active compounds. The first heatmap (Figure 14) illustrates the maximum pIC_{50} value observed for each residue-interaction pair, with activity values ranging approximately from 4.0 to 8.0. When compared to the heatmap of mean activities (Figure 15) for the same interactions, it becomes evident that extreme values tend to correspond to isolated cases. This suggests that some interactions are rare and not consistently present across multiple compounds. This interpretation is further supported by a third heatmap (Figure 16), which reports the frequency of each interaction across the compound dataset.

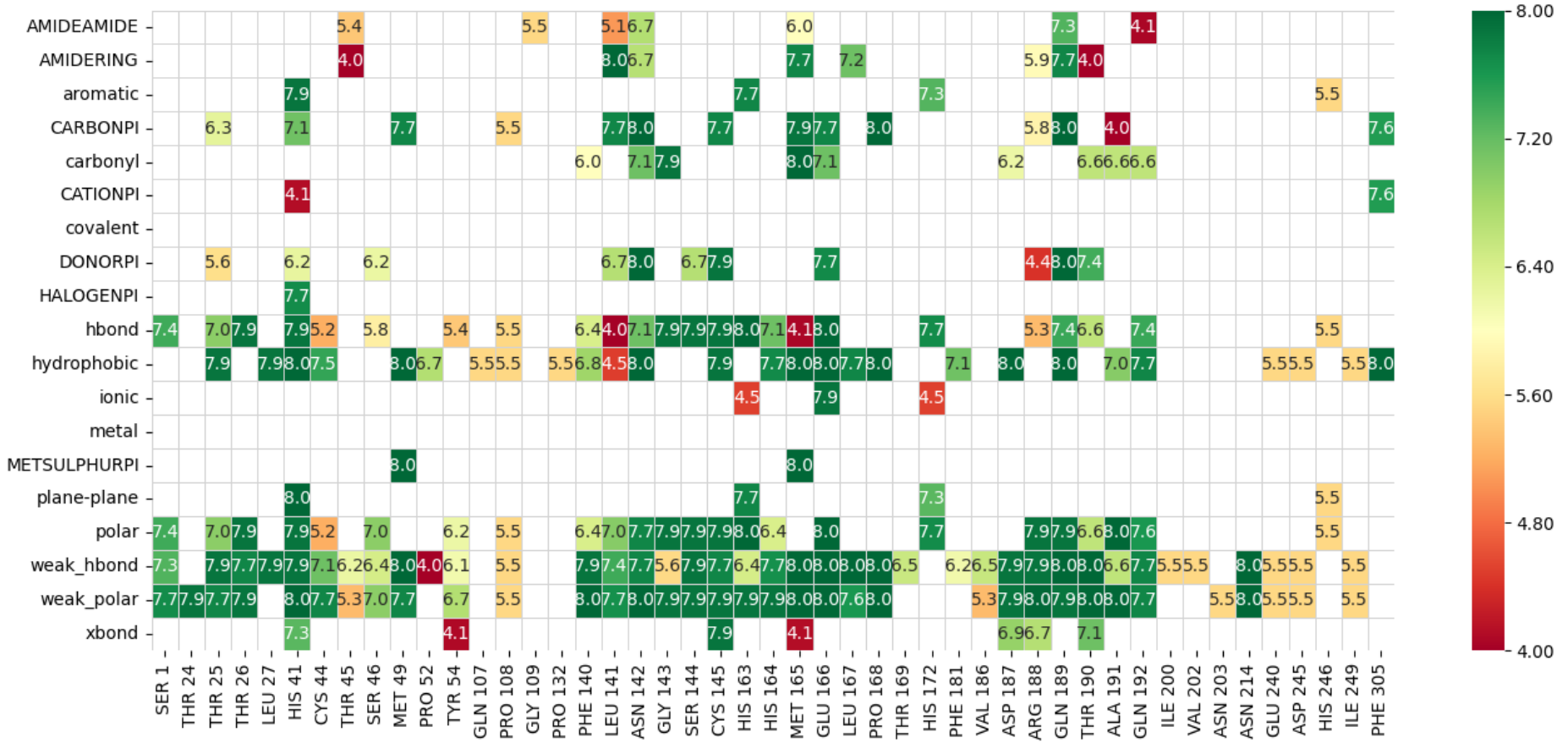

Figure 14. Heatmap of maximum pIC_{50} values for residue-interaction pairs.

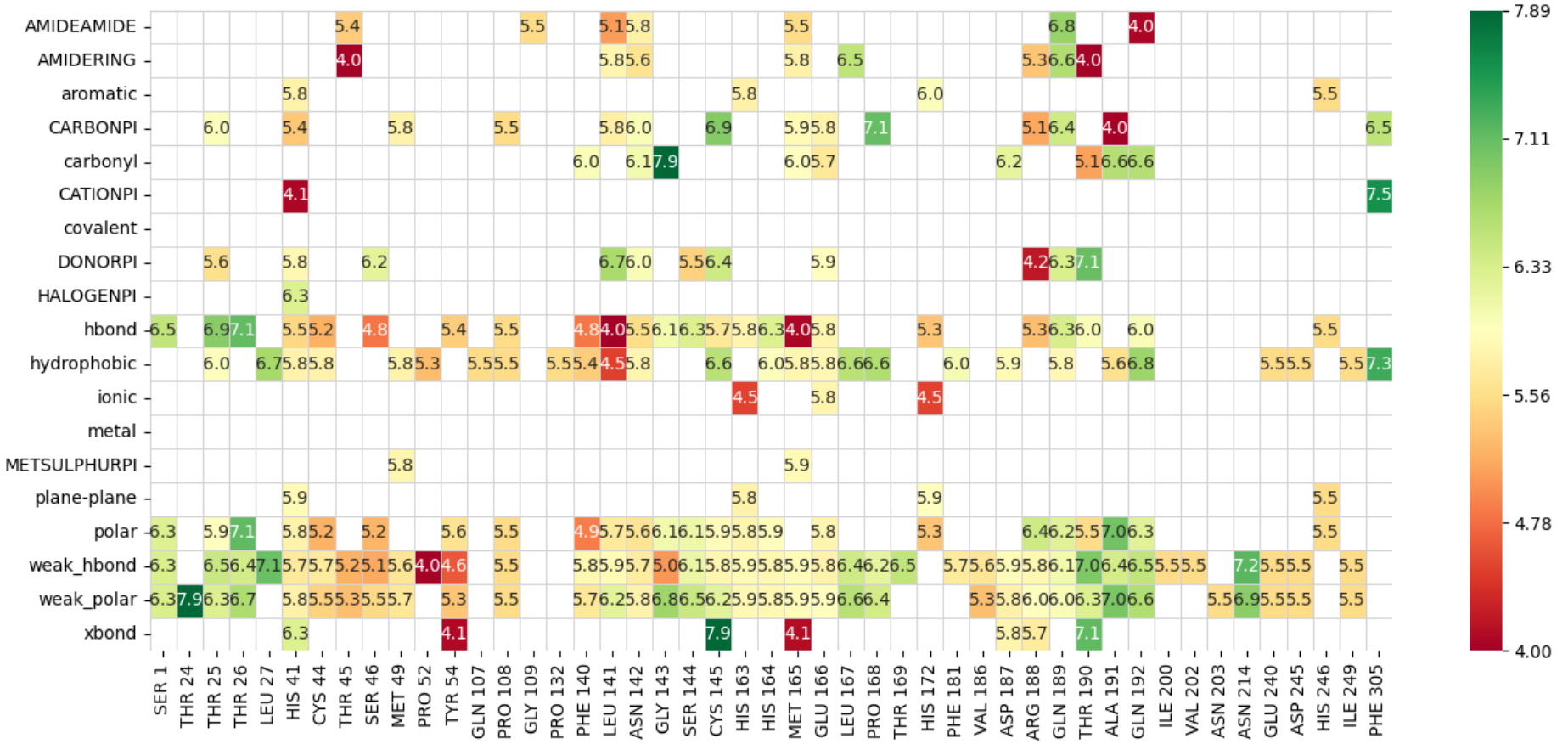

Figure 15. Heatmap of mean pIC_{50} values for residue-interaction pairs.

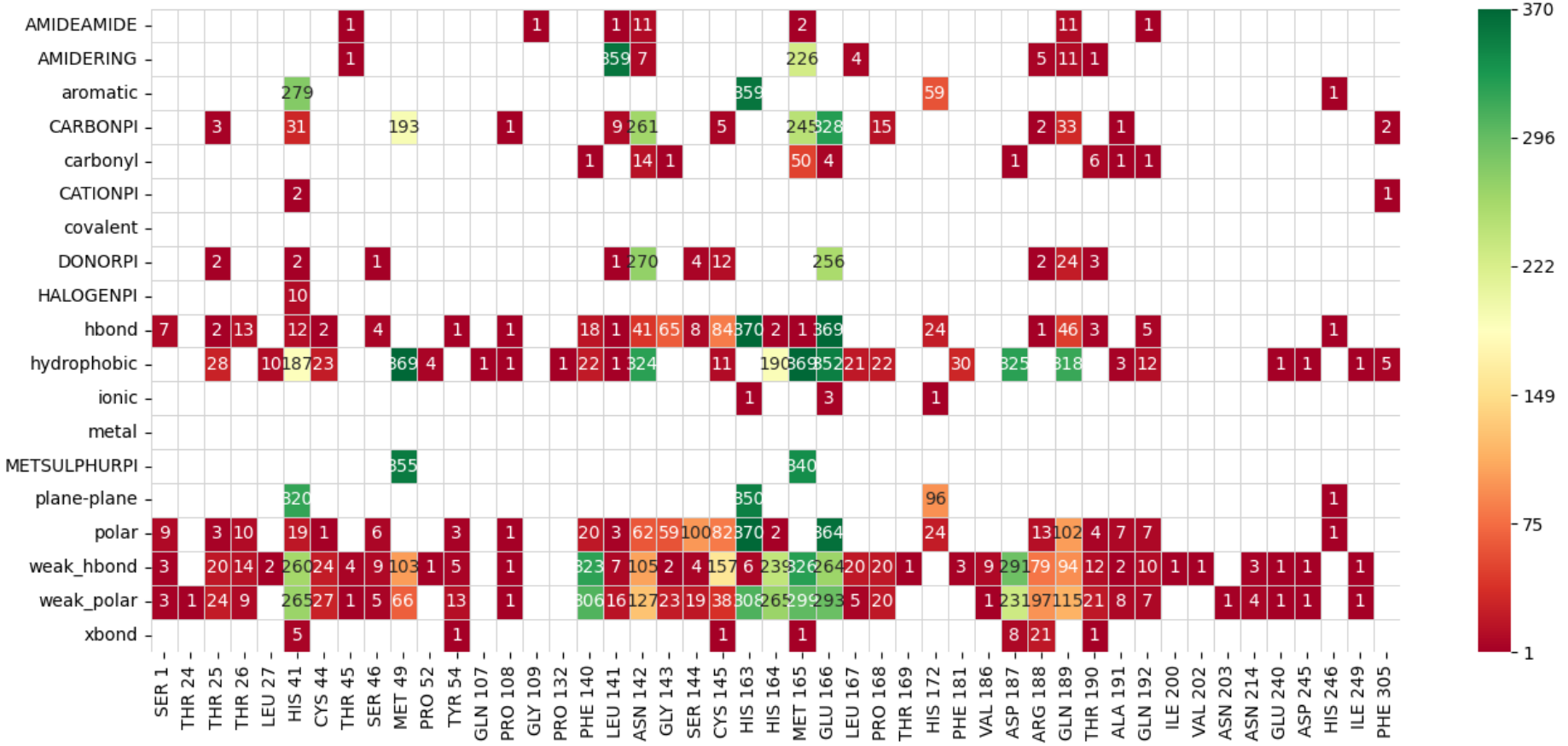

Figure 16. Heatmap of interaction frequencies for residue-interaction pairs.

To refine the analysis, a filtering step was applied, retaining only those compounds with a pIC_{50} above 7.4. This selection yielded a subset of 16 ligands classified as highly active. The heatmap of interaction counts (Figure 17) for this subset does not reveal novel residue-interaction pairs that were not already present in the full dataset. In other words, while the distribution of interactions changes in frequency, no new specific contact emerges as unique to high-activity compounds. This indicates that there is no exclusive interaction pattern or consensus motif that uniformly underlies the high potency of these ligands.

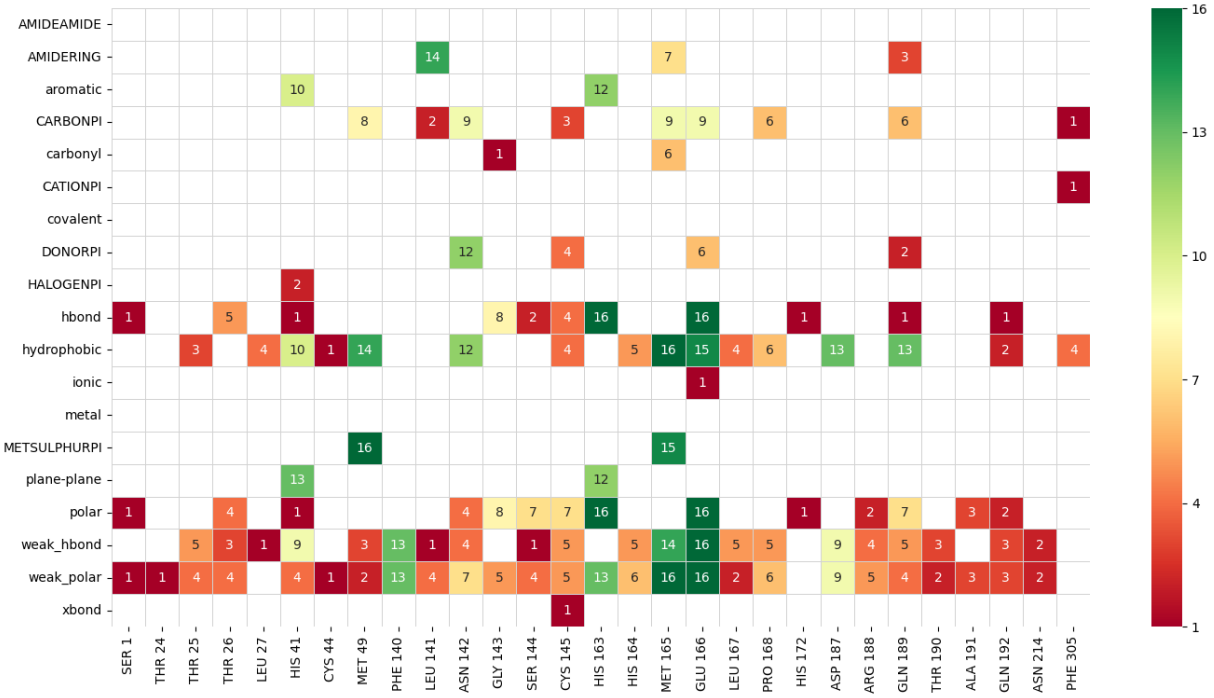

Figure 17. Heatmap of interaction frequencies for residue-interaction pairs ($pIC_{50} \geq 7.4$).

To further characterise this subset and isolate the interactions most likely involved in the molecular recognition (as opposed to those contributing to differential activity), an additional filtering step was applied. Only residue–interaction pairs shared across multiple high-activity compounds were retained, and those lacking consensus were excluded. This approach produced a simplified and more interpretable heatmap (Figure 18), highlighting a consistent interaction framework. Although these interactions do not explain variations in activity, they likely represent the fundamental set of contacts required for correct ligand orientation and anchoring within the M_{PRO} binding pocket.

Notable among these are interactions with GLU166, MET165, HIS163, MET49, and ASN142, which appear in multiple interaction categories (hydrophobic, polar, π-based and hydrogen-bonding) underscoring their role in maintaining the structure and stability of the complex. These residues, predominantly located in the S1 subsite, likely act as anchoring points that enable the ligand to adopt a binding mode, thus constituting a structural consensus for molecular recognition.

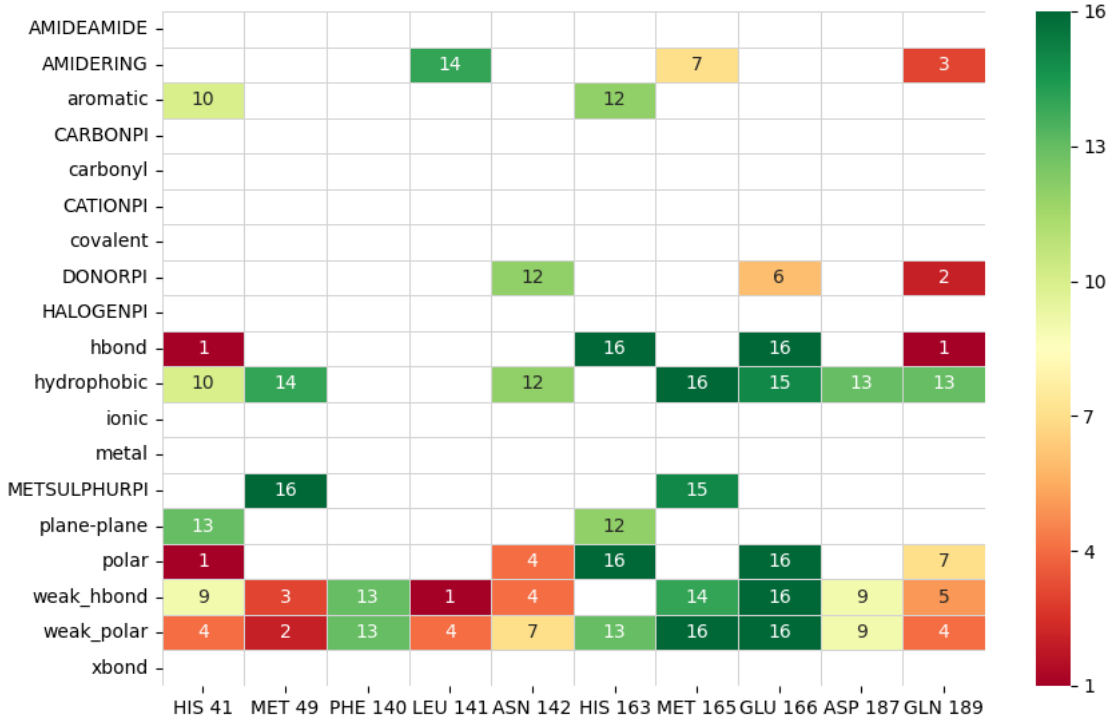

Figure 18.Heatmap of interaction frequencies for residue-interaction pairs ($pIC_{50} \geq 7.4$). Only interactions shared across multiple ligands were retained, while non-consensual contacts were excluded to facilitate the visual identification of core interactions.

4.4. Subsite-specific interaction analysis

In order to further investigate the spatial distribution of ligand-residue interactions and their possible influence on inhibitory activity, the observed interactions were characterised along the S1, S1', S2 and S4 subsites of M_{PRO}.

According to Figure 19, S1' subsite present the lowest cumulative number of interactions across the dataset. Within this site, HIS41 was the most frequently involved residue, forming recurrent interactions with a large number of ligands. While HIS41 shows a high maximum pIC_{50}, its average pIC_{50} is not particularly elevated, suggesting that its contribution is necessary for structural recognition but not sufficient to confer high inhibitory potency on its own. In contrast, other residues such as THR26, GLY143 and CYS145, despite being involved in fewer interactions, are associated with average pIC_{50} values above 7. These residues participate in specific interaction types such as hydrogen bonding (THR26), carbonyl interactions (GLY143) and halogen bonding (CYS145), which appear to be more influential in driving potent inhibition.

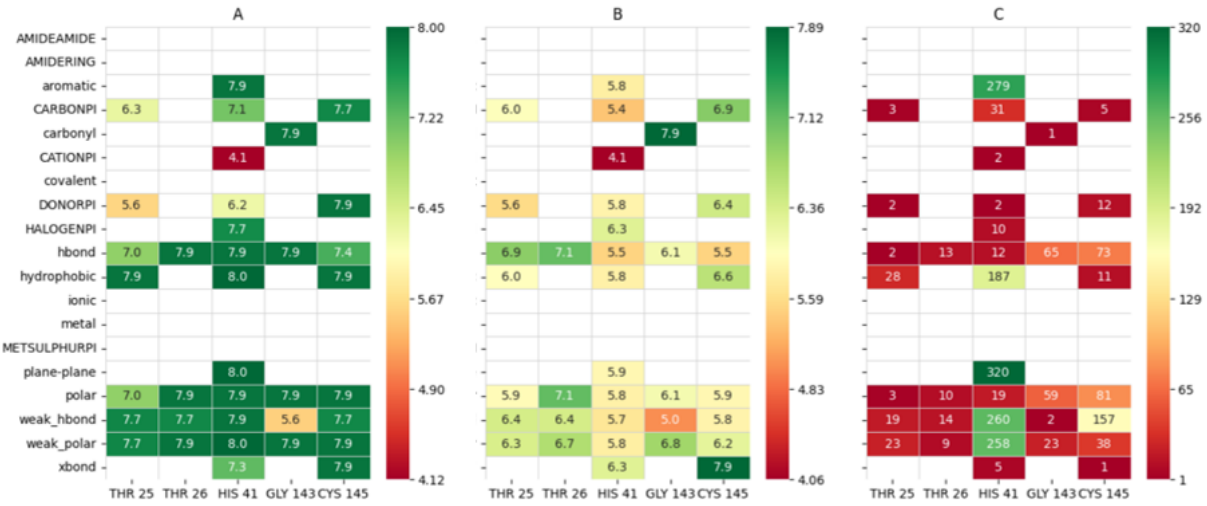

Figure 19. Interactions in the S1' subsite of M_{PRO}. A) Maximum pIC_{50} value between the residues of S1' subsite and each type of interaction. B) Average pIC_{50} value between the residues of S1' subsite and each type of interaction. C) Count of complexes where each residue of S1' establishes a specific interaction.

The S2 subsite displayed a moderate level of consensus (Figure 20), with several ligands forming interactions within this region, but the average pIC_{50} values remained below 6.5. This suggests that, although S2 may participate in the initial anchoring or orientation of the ligand, it is not a critical determinant of binding strength.

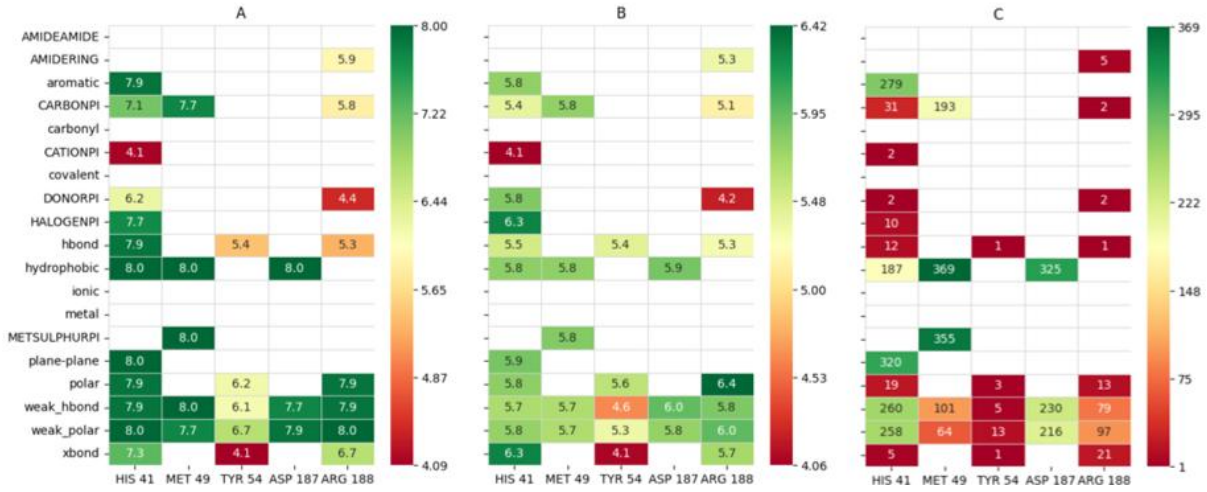

Figure 20. Interactions in the S2 subsite of M_{PRO}. A) Maximum pIC_{50} value between the residues of S2 subsite and each type of interaction. B) Average pIC_{50} value between the residues of S2 subsite and each type of interaction. C) Count of complexes where each residue of S2 establishes a specific interaction

S1 was identified as the subsite comprising the largest number of interactive residues (Figure 21), including many of those with high total interaction counts, such as GLU166, PHE140, LEU141 and CYS145. However, similar to HIS41, several of these residues, despite their high frequency, are associated with only moderate average activity, reinforcing the idea that frequent interaction does not necessarily imply strong binding affinity. Notably, GLY143 and CYS145, also part of S1', again stand out as residues with high activities. This convergence highlights their relevance as dual anchoring points.

The S4 subsite revealed a more heterogeneous pattern (Figure 22). MET165 and GLU166 appear across many complexes, yet their interaction averages do not correlate with high pIC_{50} values, indicating that these residues may contribute more to recognition than to potency. Conversely, residues such as PRO168 and THR190, involved in less frequent but structurally specific interactions (e.g., carbon–π with PRO168, donor–π and halogen bonds with THR190), were found in complexes with higher pIC_{50} values.

Taken together, the results indicate that interaction frequency alone does not predict inhibitory potency. Instead, a subset of specific interactions, less frequent but associated with higher activity, stand out as more informative indicators for structure–activity relationships. In particular, residues such as CYS145 and GLY143 appear across various subsites and correlate with higher inhibitory activity, reinforcing their relevance as important pharmacophoric hotspots in the M_{PRO} binding site.

Figure 21. Interactions in the S1 subsite of M_{PRO}. A) Maximum pIC_{50} value between the residues of S1 subsite and each type of interaction. B) Average pIC_{50} value between the residues of S1 subsite and each type of interaction. C) Count of complexes where each residue of S1 establishes a specific interaction

Figure 22. Interactions in the S4 subsite of M_{PRO}. A) Maximum pIC_{50} value between the residues of S4 subsite and each type of interaction. B) Average pIC_{50} value between the residues of S4 subsite and each type of interaction. C) Count of complexes where each residue of S4 establishes a specific interaction

5. Conclusions

This study hypothesized that a detailed understanding of the interactions between SARS-CoV-2 M_{PRO} and a set of known non-covalent inhibitors would be useful for guiding the identification and development of new therapeutic agents targeting the virus. The primary objective was to analyze a set of molecular complexes between M_{PRO} and various non-covalent inhibitors in order to identify interaction patterns that could serve as leads in antiviral drug design. As a secondary objective, a Python-based library named CORAL-PIC was developed for automated analysis of protein–ligand interaction files, simplifying extraction, visualization, and filtering of relevant data to facilitate future drug discovery workflows. The source code and documentation are openly available at https://github.com/31ldts/CORAL-PIC.

The analysis pipeline reduced the original 365 MB of Arpeggio output to a 109 KB XLSX summary, enabling rapid inspection of interaction frequencies and types across 378 M_{PRO}–inhibitor complexes. Frequency distributions showed that non-directional contacts, particularly hydrophobic and weak polar interactions, dominate ligand stabilization within the active site, while directional contacts (hydrogen bonds, π-type interactions) occur less frequently but often correspond to higher pIC_{50} values when engaging specific residues. Subsite mapping confirmed that S1 is the principal recognition region, with residues GLU166, MET165, HIS163, MET49, and ASN142 acting as core anchoring points via mixed hydrophobic, polar, and π-based contacts. Although HIS41 also ranked among the most interactive residues, its high interaction counts did not uniformly translate into elevated activity. Instead, structurally specific, low-frequency contacts, such as halogen bonds with CYS145 or carbonyl interactions with GLY143, correlated most strongly with $pIC_{50} > 7$, underlining that interaction quality (residue identity and contact nature) outweighs interaction count number in determining inhibitory activity.

Within the subset of highly active ligands ($pIC_{50} \geq 7.4$), no novel residue–interaction pair was observed. The simplified consensus heatmap for high-activity compounds highlights a conserved framework, primarily involving GLU166, MET165, HIS163, MET49, and ASN142, necessary for correct ligand orientation and binding. Subsite-specific analyses further revealed that rare but targeted interactions in S1′ (e.g., halogen bonding with CYS145) or S4 (e.g., carbon–π with PRO168, donor–π with THR190) are more indicative of high potency than interactions in more crowded subsites.

Taken together, these results validate the primary objective by identifying both frequent and rare interaction motifs that inform structure–activity relationships for M_{PRO} inhibition. The secondary objective was also achieved: the Python library proved capable of automating large-scale interaction profiling, reducing manual review time and producing publication-quality visualizations.

6. Future work

The results of this study provide a foundation for multiple future research directions focused on the design of M_{PRO} inhibitors. First, it would be of interest to expand the analysis to covalent and mixed (covalent/non-covalent) compounds in order to contrast the relative contributions of each interaction type to ligand affinity and specificity. This comparison could reveal general principles of molecular recognition applicable to other viral enzymatic systems.

From a computational perspective, future work will aim to extend the functionality of the CORAL-PIC library by incorporating interaction pattern clustering, support for additional file formats derived from other docking or molecular dynamics platforms, and new filtering options to enable more complex and customizable analyses. These enhancements would extend its applicability in large-scale virtual screening campaigns.

7. Self-assessment

This project has allowed me to deepen my understanding of protein–ligand interactions, particularly in the context of enzyme inhibition. By analysing interaction files for various SARS-CoV-2 M_{PRO}–inhibitor complexes, I have gained insights into the types of non-covalent interactions that specific amino acid residues can or cannot establish. This knowledge helped me recognize that some residues are structurally or chemically constrained, limiting the interactions they can participate in.

Additionally, I learned how enzymes such as M_{PRO} are organized into functional subsites, which influence ligand binding. For example, some subsites show a high density of interactions, while others exhibit structural limitations that make certain interaction types unlikely. Understanding this spatial organization has improved my ability to interpret interaction data in a biologically meaningful way.

From a technical perspective, I worked extensively with Arpeggio output files and developed strategies to extract relevant data efficiently. This involved designing filters, as described in earlier sections, to preserve meaningful information while avoiding unintended data modifications. Building these filters forced me to think carefully about how to structure and automate the analysis pipeline.

Moreover, this experience has helped me appreciate the dual nature of ligand–protein interactions. While some interactions clearly enhance inhibitory activity, others are essential for molecular recognition and proper binding orientation. These interactions, although not always correlated with high potency, should not be overlooked as they are fundamental to establishing the initial protein–ligand complex.

Overall, this project has enhanced my knowledge of structural bioinformatics and enzyme–ligand recognition, while also strengthening my ability to develop practical

tools for large-scale interaction analysis. It has shown me how to approach scientific questions with both a molecular and computational perspective.

8. Bibliography

[1] G. Macip, P. Garcia-segura, J. Mestres-truyol, B. Saldivar-espinoza, G. Pujadas, and S. Garcia-Vallvé, “A review of the current landscape of SARS-CoV-2 main protease inhibitors: Have we hit the bullseye yet?,” Jan. 01, 2022, *MDPI*. doi: 10.3390/ijms23010259.

[2] C. P. Chuck, H. F. Chow, D. C. C. Wan, and K. B. Wong, “Profiling of substrate specificities of 3C-like proteases from group 1, 2a, 2b, and 3 coronaviruses,” *PLoS One*, vol. 6, no. 11, Nov. 2011, doi: 10.1371/journal.pone.0027228.

[3] Z. Jin *et al.*, “Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors,” *Nature*, vol. 582, no. 7811, pp. 289–293, Jun. 2020, doi: 10.1038/s41586-020-2223-y.

[4] G. La Monica, A. Bono, A. Lauria, and A. Martorana, “Targeting SARS-CoV-2 Main Protease for Treatment of COVID-19: Covalent Inhibitors Structure-Activity Relationship Insights and Evolution Perspectives,” Oct. 13, 2022, *American Chemical Society*. doi: 10.1021/acs.jmedchem.2c01005.

[5] S. Zev, K. Raz, R. Schwartz, R. Tarabeh, P. K. Gupta, and D. T. Major, “Benchmarking the Ability of Common Docking Programs to Correctly Reproduce and Score Binding Modes in SARS-CoV-2 Protease Mpro,” *J Chem Inf Model*, vol. 61, no. 6, pp. 2957–2966, Jun. 2021, doi: 10.1021/acs.jcim.1c00263.

[6] A. Llop-Peiró, G. Macip, S. Garcia-Vallvé, and G. Pujadas, “Are protein–ligand docking programs good enough to predict experimental poses of noncovalent ligands bound to the SARS-CoV-2 main protease?,” Oct. 01, 2024, *Elsevier Ltd*. doi: 10.1016/j.drudis.2024.104137.

[7] A. Gimeno *et al.*, “Prediction of novel inhibitors of the main protease (M-pro) of SARS-CoV-2 through consensus docking and drug reposition,” *Int J Mol Sci*, vol. 21, no. 11, Jun. 2020, doi: 10.3390/ijms21113793.

[8] D. D. Nguyen, K. Gao, J. Chen, R. Wang, and G. W. Wei, “Unveiling the molecular mechanism of SARS-CoV-2 main protease inhibition from 137 crystal structures using algebraic topology and deep learning,” *Chem Sci*, vol. 11, no. 44, pp. 12036–12046, Nov. 2020, doi: 10.1039/d0sc04641h.

[9] G. L. Warren *et al.*, “A critical assessment of docking programs and scoring functions,” *J Med Chem*, vol. 49, no. 20, pp. 5912–5931, Oct. 2006, doi: 10.1021/jm050362n.

[10] L. qin Gao, J. Xu, and S. dong Chen, “In Silico Screening of Potential Chinese Herbal Medicine Against COVID-19 by Targeting SARS-CoV-2 3CLpro and Angiotensin Converting Enzyme II Using Molecular Docking,” *Chin J Integr Med*, vol. 26, no. 7, pp. 527–532, Jul. 2020, doi: 10.1007/s11655-020-3476-x.

[11] G. L. Miessler, P. J. Fischer, and D. A. Tarr, *Inorganic Chemistry*, 5th ed. Boston, MA: Pearson, 2013.

[12] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 6th ed. New York, NY: W. H. Freeman and Co, 2012.

[13] G. Zhu, Z. Xu, and L. T. Yan, “Entropy at Bio-Nano Interfaces,” Aug. 12, 2020, *American Chemical Society*. doi: 10.1021/acs.nanolett.0c02635.

[14] R. Calinsky and Y. Levy, “Aromatic Residues in Proteins: Re-Evaluating the Geometry and Energetics of π-π, Cation−π, and CH−π Interactions,” *Journal of Physical Chemistry B*, Sep. 2024, doi: 10.1021/acs.jpcb.4c04774.

[15] P. Auffinger, F. A. Hays, E. Westhof, and P. Shing Ho, “Halogen bonds in biological molecules,” 2004. [Online]. Available: www.pnas.orgcgidoi10.1073pnas.0407607101

[16] A. K. Tewari and R. Dubey, “Emerging trends in molecular recognition: Utility of weak aromatic interactions,” Jan. 01, 2008. doi: 10.1016/j.bmc.2007.09.023.

[17] C. C. Valley *et al.*, “The methionine-aromatic motif plays a unique role in stabilizing protein structure,” *Journal of Biological Chemistry*, vol. 287, no. 42, pp. 34979–34991, Oct. 2012, doi: 10.1074/jbc.M112.374504.

[18] P. Chakrabarti and R. Bhattacharyya, “Geometry of nonbonded interactions involving planar groups in proteins,” Sep. 2007. doi: 10.1016/j.pbiomolbio.2007.03.016.

[19] U. Ahmed, D. Sundholm, and M. P. Johansson, “The effect of hydrogen bonding on the π depletion and the π-π stacking interaction,” *Physical Chemistry Chemical Physics*, vol. 26, no. 43, pp. 27431–27438, Oct. 2024, doi: 10.1039/d4cp02889a.

[20] L. Zhang *et al.*, “Ferrocene-decorated graphene nanosheets built by edge-to-face π-π interaction for room temperature ppb-level NO sensing,” *Talanta*, vol. 285, Apr. 2025, doi: 10.1016/j.talanta.2024.127365.

[21] Y. N. Imai, Y. Inoue, I. Nakanishi, and K. K. Aura, “Amide-π interactions between formamide and benzene,” *J Comput Chem*, vol. 30, no. 14, pp. 2267–2276, Nov. 2009, doi: 10.1002/jcc.21212.

[22] A. Thakur, A. Kumar, V. Sharma, and V. Mehta, “PIC50: An open source tool for interconversion of PIC_{50} values and IC_{50} for efficient data representation and analysis,” Oct. 18, 2022. doi: 10.1101/2022.10.15.512366.

[23] A. Llop-Peiró, G. Pujadas, A. Gimeno, and S. Garcia-Vallvé, “PDB-CAT: A User-Friendly Tool to Classify and Analyze PDB Protein-Ligand Complexes,” Aug. 12, 2024. doi: 10.26434/chemrxiv-2024-54073.

[24] H. C. Jubb, A. P. Higueruelo, B. Ochoa-Montaño, W. R. Pitt, D. B. Ascher, and T. L. Blundell, “Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures,” *J Mol Biol*, vol. 429, no. 3, pp. 365–371, Feb. 2017, doi: 10.1016/j.jmb.2016.12.004.

[25] 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.

Appendix

Table 5. PDB codes of protein complexes, ligands, and reported activity (pIC_{50}) of the 378 complexes used in this study.

5RGV, UGG (4.24)	7GH5, Q2G (4.19)	7GLU, RD5 (6.425)
5RGX, UGP (4.212)	7GH6, Q2U (5.104)	7GLV, R76 (7.705)
5RH2, UH7 (4.609)	7GH7, Q36 (4.26)	7GLW, RDK (6.581)
5RH3, UHA (4.582)	7GHB, Q4R (4.004)	7GLX, RDQ (7.177)
5RHD, US7 (4.002)	7GHE, Q5R (4.56)	7GLY, RDX (6.922)
6M2N, 3WL (6.027)	7GHM, Q99 (5.957)	7GLZ, REU (6.908)
7B2J, SQ2 (4.699)	7GHN, QBR (6.999)	7GM0, RFF (5.597)
7CA8, FNO (6.401)	7GI4, QCC (5.44)	7GM1, RFR (6.54)
7DDC, H3F (4.498)	7GI5, QCO (6.14)	7GM2, RG3 (7.129)
7EN8, J7R (7.143)	7GI6, QD4 (6.001)	7GM3, RG9 (5.537)
7EN9, J7O (7.149)	7GI7, QD9 (6.542)	7GM4, RGQ (6.59)
7GAV, KFU (6.117)	7GI8, QDF (6.35)	7GM5, RGX (6.952)
7GAW, KG9 (7.434)	7GI9, QDU (5.766)	7GM6, RHI (6.711)
7GB0, KLR (4.002)	7GIA, QE3 (4.912)	7GM7, RI1 (6.476)
7GB5, KOI (4.27)	7GIB, QER (5.732)	7GM8, RI6 (6.468)
7GB6, KP0 (5.515)	7GIC, QEX (5.318)	7GM9, RIJ (6.654)
7GBD, KT9 (4.464)	7GID, QF5 (4.854)	7GMB, RIU (6.049)
7GBE, KU6 (6.218)	7GIE, QF9 (6.366)	7GMC, RIY (6.828)
7GBI, KVX (4.373)	7GIF, QFL (5.304)	7GME, RJ3 (5.419)
7GBJ, KW9 (4.002)	7GIG, QFU (5.738)	7GMF, RJF (6.443)
7GBL, KWR (4.053)	7GIH, QG3 (6.245)	7GMG, RJO (6.922)
7GBN, KXR (4.603)	7GII, QGC (5.853)	7GMH, RJX (5.787)
7GBO, KY0 (4.002)	7GIJ, QGO (5.534)	7GMI, RK6 (5.772)
7GBR, KZC (4.002)	7GIK, QGX (5.471)	7GMJ, RKC (6.176)
7GBS, KZX (4.018)	7GIL, QH6 (6.426)	7GMK, RKR (6.137)
7GBV, L3I (4.013)	7GIM, QHI (6.576)	7GML, RL0 (5.521)
7GBW, L6R (4.761)	7GIN, QHU (6.686)	7GMM, RL8 (6.046)
7GBX, L6D (5.081)	7GIO, QI4 (6.12)	7GMN, RLH (5.667)

7GBZ, L4U (4.004)	7GIP, QI7 (6.108)	7GMO, RLR (5.908)
7GCK, LF3 (4.222)	7GIQ, QIM (6.016)	7GMP, RM3 (5.169)
7GCQ, LRC (4.333)	7GIR, QIB (6.005)	7GMQ, RMI (5.548)
7GCV, LUC (5.44)	7GIS, QIQ (6.001)	7GMR, RN0 (5.573)
7GCZ, LWO (5.364)	7GIT, QIT (5.323)	7GMS, RNI (7.402)
7GD3, Y6J (5.398)	7GIV, QJ6 (5.967)	7GMT, ROZ (5.958)
7GD4, R30 (5.788)	7GIW, QJF (6.644)	7GMU, RPK (6.4)
7GD5, M0X (5.746)	7GIX, P4R (6.16)	7GMW, RQ6 (6.814)
7GDC, M93 (4.864)	7GIZ, PUU (6.45)	7GMX, RQO (5.475)
7GDD, 860 (6.138)	7GJ0, QJL (5.301)	7GMY, RQF (6.495)
7GDI, M5I (4.525)	7GJ1, QJR (5.734)	7GMZ, RR0 (4.116)
7GDZ, N3I (4.434)	7GJ2, QK3 (5.541)	7GN0, RRD (6.118)
7GE1, N43 (5.693)	7GJ3, PJ6 (7.076)	7GN1, RRU (6.919)
7GE2, N4L (4.974)	7GJ4, QKB (6.492)	7GN2, RS6 (6.091)
7GE3, N5L (5.374)	7GJ5, PJX (6.609)	7GN3, RSL (5.795)
7GE4, N6X (4.794)	7GJ6, QKI (5.696)	7GN4, RT4 (6.948)
7GE8, NB0 (4.62)	7GJ8, QKR (5.827)	7GN5, RT9 (6.676)
7GE9, NB6 (4.686)	7GJ9, QL3 (6.175)	7GN6, RTS (6.711)
7GEB, NDI (6.195)	7GJA, QLC (6.06)	7GN7, RV0 (7.057)
7GEC, NEL (4.002)	7GJB, QLO (5.477)	7GN8, RPZ (7.117)
7GEF, NJE (4.002)	7GJD, QM9 (6.325)	7GN9, RVL (7.257)
7GEG, NJU (4.002)	7GJF, QMX (6.095)	7GNB, RW0 (7.301)
7GEH, NKU (4.002)	7GJG, QN9 (5.001)	7GNC, RW9 (6.972)
7GEI, NM0 (5.095)	7GJI, QO0 (6.512)	7GND, RVR (6.688)
7GEJ, NO0 (5.028)	7GJJ, QOO (5.777)	7GNE, RWO (5.429)
7GEK, NOI (4.027)	7GJK, QOU (6.249)	7GNF, RWT (6.67)
7GEL, NQ3 (4.088)	7GJL, QOC (5.84)	7GNG, RXU (6.588)
7GEM, NQO (4.002)	7GJM, QP0 (6.805)	7GNH, RYB (6.794)
7GEN, NRC (4.625)	7GJO, QIZ (6.92)	7GNI, RZF (6.844)
7GEO, NRX (4.002)	7GJP, QPQ (6.382)	7GNK, S0X (7.084)
7GER, NU0 (5.369)	7GJQ, QQ6 (6.672)	7GNL, S1U (7.553)

7GES, NUR (4.464)	7GJR, QQF (5.03)	7GNQ, S5L (7.469)
7GET, NV9 (4.388)	7GJS, QQO (5.214)	7GNR, RZU (7.301)
7GEU, NVO (4.665)	7GJT, QQU (6.281)	7GNS, S6K (6.972)
7GEV, NW0 (4.002)	7GJU, QR5 (5.12)	7GNT, S7C (7.128)
7GEW, NWI (5.154)	7GJV, QP6 (6.691)	7KX5, X7V (6.699)
7GEX, NX9 (4.002)	7GJW, QR9 (5.422)	7L0D, 0EN (5.602)
7GEY, NYR (4.263)	7GJX, QNU (6.54)	7L10, XEY (5.396)
7GEZ, NZK (4.122)	7GJY, QRS (5.543)	7L11, XF1 (6.854)
7GF0, O0C (5.1)	7GJZ, QRF (6.142)	7L12, XF4 (6.893)
7GF1, O0R (5.144)	7GK0, QS3 (6.25)	7L13, XF7 (7.745)
7GF2, O0X (4.173)	7GK1, QSF (6.215)	7L14, XFD (6.77)
7GF3, O1I (4.002)	7GK2, QSX (5.194)	7LMD, Y6A (6.963)
7GF5, O2R (4.606)	7GK3, QT3 (5.522)	7LMF, Y6G (6.83)
7GF6, O3I (4.496)	7GK4, QTC (5.745)	7LTJ, YD1 (6.167)
7GFA, Z26 (5.92)	7GK5, QC3 (5.529)	7M8M, YSG (6.921)
7GFB, NSR (6.679)	7GK6, QTL (5.948)	7M8N, YSP (7.0)
7GFC, O87 (5.42)	7GK7, QU9 (6.208)	7M8O, YSM (7.432)
7GFD, O8L (4.113)	7GK8, QUQ (6.417)	7M8P, YSJ (7.699)
7GFE, O9O (5.422)	7GK9, QV0 (6.311)	7M8X, YTJ (6.328)
7GFG, OAO (4.621)	7GKA, QV9 (4.404)	7M8Y, YTM (6.959)
7GFH, OBO (4.877)	7GKB, QVG (6.503)	7M8Z, YTV (6.602)
7GFI, OCI (4.949)	7GKC, QVJ (5.739)	7M90, YTS (6.602)
7GFJ, OD7 (4.411)	7GKD, QVU (5.984)	7M91, YU4 (7.602)
7GFK, ODX (4.705)	7GKE, QW1 (6.602)	7N44, 06I (7.377)
7GFL, OE6 (5.236)	7GKF, QWL (6.602)	7NBT, U7W (5.678)
7GFM, OEO (6.546)	7GKG, QWU (5.831)	7NEO, U9H (5.161)
7GFN, OFX (4.798)	7GKH, QX3 (6.01)	7O46, V18 (6.481)
7GFO, OGF (5.848)	7GKI, QX9 (6.27)	7P2G, 4N0 (6.0)
7GFP, OHC (4.384)	7GKJ, QXI (5.598)	7QBB, V1B (6.409)
7GFQ, OI4 (5.306)	7GKK, QXR (4.64)	7RLS, 5YN (6.538)
7GFR, OIE (5.469)	7GKL, QXX (4.741)	7RM2, 5YJ (6.538)

7GFS, OIK (5.264)	7GKM, QY6 (5.889)	7RMB, 5Z7 (6.215)
7GFT, OIX (5.489)	7GKN, QYI (6.786)	7RME, 5Z3 (5.854)
7GFU, OJ9 (5.825)	7GKO, QYN (6.663)	7RMT, 5ZN (5.208)
7GFX, OKW (5.287)	7GKP, QYR (6.508)	7RMZ, 5ZJ (5.194)
7GFZ, ONU (4.002)	7GKQ, QZ0 (6.56)	7RN4, H69 (5.056)
7GG0, OGV (4.002)	7GKR, QZC (6.854)	7S4B, 87H (5.678)
7GG1, OGO (4.002)	7GKS, QZL (7.288)	7UR9, O5F (7.741)
7GG2, OO6 (5.766)	7GKT, QZU (6.06)	7URB, O5O (7.455)
7GG4, OPU (4.002)	7GKU, R08 (6.808)	7US4, O69 (7.659)
7GG5, OQF (5.314)	7GKV, R0F (7.103)	7VIC, ODN (5.666)
7GG7, OQX (5.944)	7GKW, R0Q (6.242)	7VTH, 7XB (5.066)
7GG8, ORR (4.475)	7GKX, R1I (5.645)	8ACD, LQ6 (6.398)
7GGA, OSI (6.547)	7GKY, R1U (5.825)	8ACL, LQL (6.398)
7GGB, OT6 (4.437)	7GKZ, R2L (4.49)	8CYU, P5X (6.155)
7GGC, OTV (4.597)	7GL0, R2X (6.336)	8CYZ, P6I (6.046)
7GGE, OV4 (5.296)	7GL2, R43 (7.192)	8CZ4, P6R (5.638)
7GGF, OVF (6.304)	7GL3, R4X (6.801)	8CZ7, P7L (5.301)
7GGG, OVX (5.037)	7GL4, R5H (7.155)	8DII, U2I (4.26)
7GGH, OWC (5.943)	7GL5, R5O (6.837)	8DZ0, 7YY (7.886)
7GGI, OWX (5.138)	7GL6, R66 (4.043)	8I4S, OU3 (7.0)
7GGJ, OYF (5.509)	7GL7, R6L (6.115)	8Q71, KKO (6.678)
7GGK, OYX (5.006)	7GL9, QM3 (6.609)	8R11, XI0 (5.252)
7GGL, OZC (4.107)	7GLA, R7F (6.452)	8R12, XH9 (4.616)
7GGM, OZX (4.391)	7GLC, R87 (6.386)	8R14, XHW (5.886)
7GGN, P0X (4.801)	7GLD, R8I (7.079)	8R16, XJ9 (5.638)
7GGQ, OQL (6.638)	7GLE, R8O (6.92)	8SXR, WZK (5.347)
7GGR, P6O (4.683)	7GLF, R8X (6.981)	8UDO, WBK (8.0)
7GGS, P7R (4.685)	7GLG, R95 (6.934)	8UDP, WBO (8.0)
7GGT, P9O (5.21)	7GLH, R9E (5.944)	8UDQ, WC0 (7.155)
7GGW, PKW (4.885)	7GLI, R9I (6.938)	8UDW, WDK (7.699)
7GGX, PQ6 (5.299)	7GLJ, R9R (7.246)	8UE0, WDF (5.523)

7GGZ, PVR (4.002)	7GLK, R9Z (6.996)	8UEF, WEK (6.523)
7GH0, PWR (4.51)	7GLL, RAQ (6.819)	8UEG, WEO (7.0)
7GH1, PZ6 (4.085)	7GLM, RBM (7.071)	8UEH, WEX (6.301)
7GH3, Q1C (4.308)	7GLN, RBX (7.043)	8UR9, XEK (7.357)
7GH4, Q1U (6.047)	7GLT, RC9 (7.222)	8YKJ, X77 (5.553)